The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English: A Corpus- and Web-Based Study of Pronoun Case Variation

Dissertation zur Erlangung des Grades des Doktors der Philosophie an der Fakultät Geisteswissenschaften der Universität Hamburg im Promotionsfach Englische Sprache, Literatur und Kultur

vorgelegt von

Georg Maier

aus Freiburg im Breisgau

Hamburg, November 2012
Betreuer/Gutachter: Prof. Dr. Peter Siemund
weiterer Gutachter oder Betreuer/Gutachter: Prof. Dr. Bernd Kortmann
Datum der Disputation: Hamburg, den 08.05.2013
ACKNOWLEDGEMENTS

First and foremost, I want to thank my thesis supervisor, Peter Siemund, to whom I am deeply indebted for his support and faith throughout the last years. Without him, this book would not have been possible.

I also want to express my gratitude to the people who awakened my interest in linguistics in the first place. Among them, Manfred Krug and Bernd Kortmann deserve special mention. Moreover, I would like to thank all the people who have contributed to my education and share my passion for linguistics. In particular, I want to thank Thomas Berg, Christian Mair, Lukas Pietsch, Günter Radden, Anatol Stefanowitsch and Donald Winford.

I would also like to sincerely thank my colleagues at the University of Hamburg, especially Micah Corum, Dennis Büscher-Ulbrich, Oliver Lehrbaß, Bärbel Rieckmann and Rebekka Rohleder. Kalynda Beal deserves a special word of gratitude for being such a fantastic last-minute proof-reader.

Moreover, I am grateful to the people who introduced me to statistics and who helped me to get a grip on it, particularly to Arne Lohmann, Hagen Peukert and Martin Schweinberger.

Speaking of Martin Schweinberger, he and my other office mate Florian Dolberg deserve special thanks. I have always enjoyed not only the fruitful and stimulating discussions of linguistics, statistics and science in general, but also the rather controversial debates over basically all other issues. Thank you for such a great time.

Finally, I would like to sincerely thank my family and friends. I am particularly indebted to my parents Barbara and Heinrich and to Christina, Christoph, Martin, and Laura for their endless faith, support and love. Thank you.
# TABLE OF CONTENTS

1 INTRODUCTION ........................................................................................................... 1  
   1.1 THE PROBLEM: VARIABILITY IN THE USE OF PRONOUN CASE FORMS IN PRESENT-DAY ENGLISH ................................................. 1  
   1.2 THE SCOPE: THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS ............................. 3  
   1.3 THE MAIN OBJECTIVES OF THIS STUDY ............................................................... 6  
   1.4 THE METHODOLOGICAL AND THEORETICAL ORIENTATION OF THIS STUDY ................................................................. 7  
   1.5 THE ORGANISATION OF THIS STUDY .................................................................. 8  

2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN ENGLISH ................................. 10  
   2.1 PRONOUN CASE IN STANDARD ENGLISH .......................................................... 10  
   2.2 THE DISTRIBUTION OF PRONOUN CASE FORMS ACROSS VARIETIES OF ENGLISH .............................................................. 16  
      2.2.1 Pronoun Case across Varieties: Pronominal Paradigms between Retention, Extension and Idiosyncratic Variation .................. 16  
      2.2.2 Pronoun Case in Pidgins and Creoles: Emerging Pronominal Paradigms between Simplification and Innovation .................. 23  
      2.2.3 Pronoun Case across Varieties: The Reanalysis and Re-Functionlalisation of Pronominal Case Forms and Paradigms ............ 30  
   2.3 CONTEXTS PERMITTING VARIABILITY IN THE USE OF PRONOUN CASE FORMS IN STANDARD ENGLISH AND ACROSS VARIETIES OF ENGLISH ......................................................................... 36  
      2.3.1 Variability in the Use of Pronoun Case Forms: The Environments ...................... 36  
      2.3.2 Variable Usage of Pronoun Case Forms across Varieties of English .................. 44  
   2.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN ENGLISH: INTERIM SUMMARY ................................................................. 55  

3 PREVIOUS APPROACHES TO THE DISTRIBUTION OF PRONOUN CASE FORMS IN ENGLISH ...... 57  
   3.1 FORMAL APPROACHES TO THE DISTRIBUTION OF PRONOUN CASE FORMS ................................................................. 58  
   3.2 ‘POSITIONAL’ APPROACHES TO THE DISTRIBUTION OF PRONOUN CASE FORMS .............................................................. 64  
   3.3 THE ‘WEAK VS. STRONG’ PRONOUN DISTINCTION ........................................... 68  
   3.4 FUNCTIONAL APPROACHES TO THE DISTRIBUTION OF PRONOUN CASE FORMS .............................................................. 74  
   3.5 THE DISTRIBUTION OF PRONOUN CASE FORMS IN LINGUISTIC THEORY: INTERIM SUMMARY .................................................. 78  

4 A FOCUS-ORIENTED APPROACH TO PRONOUN CASE DISTRIBUTION IN SUBJECT PREDICATIVE COMPLEMENTS ........................................................................... 80  
   4.1 THE NOTION OF FOCUS AS USED IN THIS STUDY .................................................. 82  
   4.2 FROM CASE TO FOCUS MARKERS: THE REANALYSIS OF SUBJECT PRONOUN FORMS AS FOCUS MARKERS IN SUBJECT PREDICATIVE COMPLEMENTS .............................................................. 86  
   4.3 FROM SUBJECT PRONOUNS TO FOCUS MARKERS: AN EXPLANATION FOR CHOOSING AND USING THE SUBJECT FORMS .................................................................................. 91  
   4.4 A FOCUS-ORIENTED APPROACH TO PRONOUN CASE DISTRIBUTION IN SUBJECT PREDICATIVE COMPLEMENTS: INTERIM SUMMARY .................................................. 95  

5 AIMS, RESEARCH QUESTIONS AND CENTRAL HYPOTHESES OF THIS STUDY ....................... 97  
   5.1 RESEARCH QUESTIONS ......................................................................................... 97  
   5.2 CENTRAL HYPOTHESES ..................................................................................... 102  

6 VARIABLES AND STATISTICAL MODELLING ...................................................................... 106  
   6.1 THE VARIABLES: DESCRIPTION AND MOTIVATION .................................................................................................................. 106  
      6.1.1 The Dependent Variable ................................................................................. 107  
      6.1.2 The Independent Variables ............................................................................ 109  
         6.1.2.1 Construction .......................................................................................... 109  
         6.1.2.1.1 Inter-Constructional Variation: It BE Sentences versus it-Clefts .................. 109  
         6.1.2.1.2 Intra-Constructional Variation: The Influence of the Independent Variables in the IT BE Sentences and it-Clefts .................. 111  
         6.1.2.2 Person ...................................................................................................... 113  
         6.1.2.3 Number .................................................................................................... 116
6.1.2.4 Mode of Discourse .................................................................117
6.1.2.5 Focus ..............................................................................122
6.1.2.6 Regional Variety ..............................................................134
6.1.2.7 Summary of the Independent Variables ................................136

6.1.3 The Independent Variables: The Difficulty of their Classification and Implications for their Interpretation .................................................................138

6.2 THE STATISTICAL DESIGN OF THE STUDY ..................................................139

6.3 VARIABLES AND STATISTICAL MODELLING: INTERIM SUMMARY .............147

7 DATA USED FOR THIS STUDY .......................................................................149

7.1 CORPUS DATA ...........................................................................149

7.1.1 The International Corpus of English ....................................................150

7.1.2 The British National Corpus ............................................................151

7.1.3 The Corpus of Contemporary American English ..................................152

7.2 INTERNET DATA ........................................................................155

7.2.1 The World Wide Web as a Corpus Linguistic Resource: An Introduction ....156

7.2.2 The World Wide Web as a Corpus Linguistic Resource: Problems and Pitfalls 160

7.2.3 The World Wide Web as a Corpus Linguistic Resource: Prospects and Opportunities ...165

7.2.4 The Internet Data Collection for this Study ............................................169

7.2.4.1 The Collection of the Raw Data .....................................................170

7.2.4.2 Refining the Raw Data: The Cleaning and Verification of the Data ..........174

7.2.4.3 Refining the Raw Data: The Classification of the Data .........................176

7.2.4.3.1 The Classification of the Data: The Spoken Data .........................179

7.2.4.3.2 The Classification of the Data: The Written Data .........................180

7.2.4.3.3 The Classification of the Data: The Computer-Mediated Communication Data 184

7.3 DATA POINTS EXCLUDED FROM THE STATISTICAL ANALYSIS ..................188

7.4 THE DATABASE USED FOR THE STATISTICAL ANALYSES .......................194

7.5 THE NOTIONS OF REPRESENTATIVENESS, BALANCEDNESS AND COMPARABILITY OF THE DATA AND THEIR REPERCUSSIONS FOR THE PRESENT STUDY .................................................196

7.6 DATA USED FOR THIS STUDY: INTERIM SUMMARY ................................203

8 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE BRITISH NATIONAL CORPUS ..............205

8.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE BRITISH NATIONAL CORPUS .................................................................205

8.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE BRITISH NATIONAL CORPUS .................................................................211

8.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE BRITISH NATIONAL CORPUS ..................................................................................215

8.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE BRITISH NATIONAL CORPUS: INTERIM SUMMARY .................................................................220

9 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH ........................................................................................................222

9.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH ........................................................................................................222

9.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH ........................................................................................................229

9.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH ........................................................................................................233

9.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH: INTERIM SUMMARY .................................................................239

10 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE BRITISH (.UK) INTERNET DATA ........241

10.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE BRITISH (.UK) INTERNET DATA ........................................................................................................241

10.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE BRITISH (.UK) INTERNET DATA ........................................................................................................249

10.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE BRITISH (.UK) INTERNET DATA ........................................................................................................254

10.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE BRITISH (.UK) INTERNET DATA: INTERIM SUMMARY .................................................................260
11 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE AUSTRALIAN (.AU) INTERNET DATA 262
11.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE AUSTRALIAN (.AU) INTERNET DATA ................................................................. 262
11.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE AUSTRALIAN (.AU) INTERNET DATA ................................................................. 268
11.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE AUSTRALIAN (.AU) INTERNET DATA ................................................................. 271
11.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE AUSTRALIAN (.AU) INTERNET DATA: INTERIM SUMMARY ........................................................................................................ 277

12 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE IRISH (.IE) INTERNET DATA........ 279
12.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE IRISH (.IE) INTERNET DATA .................................................................................................................. 279
12.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE IRISH (.IE) INTERNET DATA ................................................................. 285
12.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE IRISH (.IE) INTERNET DATA ................................................................. 288
12.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE IRISH (.IE) INTERNET DATA: INTERIM SUMMARY ........................................................................................................ 292

13 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE SOUTH AFRICAN (.ZA) INTERNET DATA ........................................................................................................ 295
13.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE SOUTH AFRICAN (.ZA) INTERNET DATA .................................................................................................................. 295
13.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE SOUTH AFRICAN (.ZA) INTERNET DATA ................................................................. 302
13.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE SOUTH AFRICAN (.ZA) INTERNET DATA ................................................................. 305
13.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE SOUTH AFRICAN (.ZA) INTERNET DATA: INTERIM SUMMARY ........................................................................................................ 311

14 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE INDIAN (.IN) INTERNET DATA........ 314
14.1 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN THE INDIAN (.IN) INTERNET DATA .................................................................................................................. 314
14.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT BE SENTENCES IN THE INDIAN (.IN) INTERNET DATA ................................................................. 319
14.3 THE DISTRIBUTION OF PRONOUN CASE FORMS IN IT-CLEFTS IN THE INDIAN (.IN) INTERNET DATA ................................................................. 321
14.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN THE INDIAN (.IN) INTERNET DATA: INTERIM SUMMARY ........................................................................................................ 325

15 GENERAL DISCUSSION ........................................................................................................ 328
15.1 POSITION AND PRONOUN CLASS AS DETERMINANTS OF THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN VARIETIES OF ENGLISH .................................................................................................................. 328
15.2 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS IN VARIETIES OF ENGLISH: THE DETERMINANTS AND THEIR IMPACTS .................................................................................................................. 332
15.3 THE REANALYSIS OF SUBJECT PRONOUN CASE FORMS AS FOCUS MARKERS IN SUBJECT PREDICATIVE COMPLEMENTS .................................................................................................................. 337
15.4 THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS ACROSS VARIETIES OF ENGLISH .................................................................................................................. 342
15.5 THE POTENTIAL OF GOOGLE DATA FOR QUANTITATIVE STUDIES OF GRAMMATICAL VARIATION .................................................................................................................. 350
15.6 GENERAL DISCUSSION: INTERIM SUMMARY .................................................................................................................. 354

16 CONCLUSION AND OUTLOOK .......................................................................................... 356
16.1 CONCLUSION .................................................................................................................. 356
16.2 OUTLOOK .................................................................................................................... 359

APPENDIX A: MULTICOLLINEARITY STATISTICS .................................................................. 364
APPENDIX B: GOODNESS OF FIT STATISTICS ...................................................................... 373
REFERENCES .................................................................................................................... 380
PRIMARY SOURCES ........................................................................................................... 380
SECONDARY SOURCES ............................................................................................................. 380
WEB SOURCES .......................................................................................................................... 406
LISTE DER AUS DIESER DISSERTATION HERVORGEGANGENEN VORVERÖFFENTLICHUNGEN ......... 408
SUMMARY OF THE MAJOR FINDINGS/ZUSAMMENFASSUNG DER ERGEBNISSE DER DISSERTATION ........................................................................................................................................ 409
A.) ENGLISH SUMMARY: THE DISTRIBUTION OF PRONOUN CASE FORMS IN SUBJECT PREDICATIVE COMPLEMENTS: A CORPUS- AND WEB-BASED STUDY OF PRONOUN CASE VARIATION ............. 409
B.) ZUSAMMENFASSUNG DER ERGEBNISSE DER DISSERTATION: DIE VERTeilung PRONOMINALER KASUSFORMEN IN SUBJEKTKOMPONENTEN IN VARIETÄTEN DES ENGLISCHEN: EINE KORPUS- UND WEBBASIerte STUDIE ZUR VARIATION PRONOMINALER KASUSFORMEN ................................................... 419
EIDESSTATTLICHE ERKLÄRUNG ................................................................................................ 431
LIST OF TABLES

Table 1: Variable Contexts of Pronoun Case Distribution in Present-Day English ................................................................. 1
Table 2: The Pronominal Paradigm of Present-Day English (Biber et al. 1999: 328) ............................................................... 11
Table 3: The Case-Sensitive Pronoun Forms of Present-Day English (Biber et al. 1999: 328) ..................................................... 14
Table 4: Personal Pronouns in Samoan Plantation Pidgin English (Mühlhäusler 1997: 148) ................................................... 26
Table 5: Subject Pronouns in Ghanaian Pidgin English (Huber 1999: 197) ................................................................... 28
Table 6: Object Pronouns in Ghanaian Pidgin English (Huber 1999: 199) ................................................................. 28
Table 7: The Token Frequencies of Case-Sensitive Personal Pronoun Forms in the British National Corpus (Kilgariff 2006b) .................................................................................................................. 92
Table 8: The Dependent Variable of the Following Analyses .............................................................................................. 107
Table 9: The Independent Variable Construction of the Following Analyses ................................................................. 111
Table 10: The it-Cleft-Specific Independent Variable Co-reference of the Following Analyses .............................................. 112
Table 11: The Independent Variable Person of the Following Analyses ........................................................................ 115
Table 12: The Independent Variable Number of the Following Analyses .............................................................................. 117
Table 13: The Independent Variable Mode of Discourse of the Following Analyses ............................................................ 121
Table 14: The Independent Variable Focus of the Following Analyses ........................................................................... 134
Table 15: Overview of the Independent Variables Tested in the Multivariate Analyses ..................................................... 137
Table 16: Token Numbers Obtained from the Closed Corpora for the Statistical Analyses ................................................... 194
Table 17: Token Numbers Obtained from the Web-Derived Datasets for the Statistical Analyses ........................................ 194
Table 18: Subject Predicative Complements in the BNC: The Token Numbers ................................................................. 206
Table 19: Subject Predicative Complements in the BNC: Logistic Regression Results ....................................................... 207
Table 20: it BE Sentences in the BNC: The Token Numbers .......................................................................................... 211
Table 21: it BE Sentences in the BNC: Logistic Regression Results .................................................................................. 212
Table 22: it-Clefts in the BNC: The Token Numbers ........................................................................................................... 215
Table 23: it-Clefts in the BNC: Logistic Regression Results ........................................................................................... 216
Table 24: Subject Predicative Complements in the COCA: The Token Numbers ............................................................. 222
Table 25: Subject Predicative Complements in the COCA: Logistic Regression Results ..................................................... 224
Table 26: it BE Sentences in the COCA: The Token Numbers ................................................................................................. 230
Table 27: it BE Sentences in the COCA: Logistic Regression Results .................................................................................. 230
Table 28: it-Clefts in the COCA: The Token Numbers .......................................................................................................... 233
Table 29: it-Clefts in the COCA: Logistic Regression Results .......................................................................................... 235
Table 30: Subject Predicative Complements in the.uk Data: The Token Numbers ............................................................ 242
Table 31: Subject Predicative Complements in the.uk Data: Logistic Regression Results .......................................................... 244
Table 32: it BE Sentences in the.uk Data: The Token Numbers .......................................................................................... 250
Table 33: it BE Sentences in the.uk Data: Logistic Regression Results .................................................................................. 250
Table 34: it-Clefts in the.uk Data: The Token Numbers ........................................................................................................ 255
Table 35: it-Clefts in the.uk Data: Logistic Regression Results .......................................................................................... 256
Table 36: Subject Predicative Complements in the.au Data: The Token Numbers ............................................................. 262
Table 37: Subject Predicative Complements in the.au Data: Logistic Regression Results ...................................................... 264
Table 38: it BE Sentences in the.au Data: The Token Numbers .......................................................................................... 268
Table 39: it BE Sentences in the.au Data: Logistic Regression Results .................................................................................. 269
Table 40: it-Clefts in the.au Data: The Token Numbers ........................................................................................................ 272
Table 41: it-Clefts in the.au Data: Logistic Regression Results .......................................................................................... 273
Table 42: Subject Predicative Complements in the.je Data: The Token Numbers .............................................................. 280
Table 43: Subject Predicative Complements in the.je Data: Logistic Regression Results .......................................................... 282
Table 44: it BE Sentences in the.je Data: The Token Numbers .......................................................................................... 285
Table 45: it BE Sentences in the.je Data: Logistic Regression Results .................................................................................. 286
Table 46: it-Clefts in the.je Data: The Token Numbers ........................................................................................................ 289
Table 47: it-Clefts in the.je Data: Logistic Regression Results .......................................................................................... 290
Table 48: Subject Predicative Complements in the.za Data: The Token Numbers .............................................................. 295
Table 49: Subject Predicative Complements in the.za Data: Logistic Regression Results .......................................................... 296
Table 50:it BE Sentences in the.za Data: The Token Numbers .......................................................................................... 302
Table 51: it BE Sentences in the.za Data: Logistic Regression Results .................................................................................. 303
Table 52: it-Clefts in the.za Data: The Token Numbers ........................................................................................................ 306
Table 53: it-Clefts in the.za Data: Logistic Regression Results .......................................................................................... 306
TABLE 54: SUBJECT PREDICATIVE COMPLEMENTS IN THE .IN DATA: THE TOKEN NUMBERS.................................................. 314
TABLE 55: SUBJECT PREDICATIVE COMPLEMENTS IN THE .IN DATA: LOGISTIC REGRESSION RESULTS........................................ 316
TABLE 56: IT BE SENTENCES IN THE .IN DATA: THE TOKEN NUMBERS ........................................................................... 320
TABLE 57: IT-CLEFTS IN THE .IN DATA: THE TOKEN NUMBERS .................................................................................. 322
TABLE 58: IT-CLEFTS IN THE .IN DATA: LOGISTIC REGRESSION RESULTS .............................................................. 323
LIST OF FIGURES

Figure 1: The Distribution of Pronoun Case Forms in Government and Binding Theory (adapted from Haegeman and Guéron 1999: 129) ................................................................. 59
Figure 2: The Distribution of Pronoun Case Forms in Minimalist Models (adapted from Radford 2009: 122) ........ 62
Figure 3: Quinn’s (2005b: 15) Model of Pronoun Case Distribution in English ....................................................... 72
Figure 4: Usage Restrictions of Yahoo .................................................................................................................. 173
1 Introduction

1.1 The Problem: Variability in the Use of Pronoun Case Forms in Present-Day English

As early as the 1930s, Jespersen (1933) noted that the “right use” of pronominal case forms “is one of the knottiest points in English grammar” (Jespersen 1933: 132). The reason for this knottiness is the fact that the distribution of subject and object pronoun case forms in English is not as “straightforward” as is often assumed (Biber et al. 1999: 335). Although subject pronoun forms are mainly used as subjects of finite clauses and object pronoun forms largely as objects of transitive verbs and as complements of prepositions (cf. Biber et al. 1999: 335), Present-Day English also exhibits a considerable number of contexts which permit a choice between the uses of subject or object pronoun forms. As summarised in Table 1, prominent examples of these contexts that allow for variability in the use of pronoun case forms include the focal position of *it*-clefts, predicative complements following *it* and a form of *BE*, i.e. *it BE* sentences (cf. Quinn 2005a), coordinated noun phrases in subject and object position, pronoun–noun phrase constructions and *than*-comparatives, to mention only a few (cf. Sections 2.3.1 and 2.3.2):

<table>
<thead>
<tr>
<th>Contexts Exhibiting Variability in the Use of Pronoun Case Forms</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>It</em>-clefts</td>
<td>It is they/them who are to blame</td>
</tr>
<tr>
<td><em>It BE</em> sentences</td>
<td>[Who is it? – ] It is I/me</td>
</tr>
<tr>
<td>Coordinated NPs in Subject Position</td>
<td>Rita and I/me will get divorced</td>
</tr>
<tr>
<td>Coordinated NPs in Object Position</td>
<td>This issue has to stay between you and I/me</td>
</tr>
<tr>
<td>Pronoun–Noun Phrase Constructions</td>
<td>We/Us Irish are a happy people</td>
</tr>
<tr>
<td><em>Than</em>-Comparatives</td>
<td>Susan is younger than he/him</td>
</tr>
</tbody>
</table>

Table 1: Variable Contexts of Pronoun Case Distribution in Present-Day English

---

1 This study uses the terms *subject pronoun* and *object pronoun* instead of the also common labels *subjective pronoun* and *objective pronoun* or *nominative and accusative pronoun* because they are the most neutral ones referring simply to the prototypical function of these pronouns in a clause (cf. Section 2.1).
Consequently, variable use of pronoun case forms causes some uncertainty among speakers and writers, which can be easily observed in data obtained from closed corpora (cf. example (1)) and from the World Wide Web (cf. example (2)) (cf. Chapter 7):

(1) Uncertainty about the Use of Pronoun Case Forms as Observed in Closed Corpora

When I tell my community college students to say, "It is I," instead of, "It is me," they look skeptical, perhaps believing this is the jargon of English teachers. I delay the full explanation until we study nominative and objective pronouns. (COCA/ACAD/Education)

(2) Uncertainty about the Use of Pronoun Case Forms as Observed on the World Wide Web

Unfortunately, someone has to say it and this time it is me (or I): INTERNATIONAL GRAND SLAM TENNIS IS THE MOST BORING SPORT EVER DEVISED BY THE WIT OF MAN, WOMAN OR DOG." (.au/it is me/14.07.2008)

Moreover, those contexts which allow for variability in the use of pronoun case forms have also attracted a considerable amount of scholarly interest in the distribution of pronoun case forms in Present-Day English and thus have triggered a lively linguistic debate (e.g. Angermeyer and Singler 2003; Erdmann 1978; Hudson 1995; Kjellmer 1986; Quinn 2005a, 2005b, 2009; Parker, Riley and Meyer 1988; Shorrocks 1992; Sobin 1997; Wales 1996).

Although much scholarly attention has been devoted to those contexts allowing for variability, so far only a few studies analyse these phenomena quantitatively (e.g. Angermeyer and Singler 2003; Biber et al. 1999; Erdmann 1978; Maier 2013; Quinn 2009). Furthermore, this scarcity of quantitative studies is particularly true for the first two contexts which allow for variability in the use of subject pronoun case forms listed in Table 1, i.e. it-clefts and it BE sentences. These are subsumed under the heading of subject predicative

---

2 This tag indicates that this example is taken from the corpora used for this study (cf. Chapter 7). The first abbreviation given in the brackets, i.e. COCA, indicates the corpus from which this example is taken. In this study, COCA stands for Corpus of Contemporary American English and BNC for British National Corpus. The information after the first forward slash specifies the genre of the data. The information after the second forward slash in the parentheses indicates the exact subgenre of the data point. Thus, if one wishes to obtain further information on this data point, the interested reader is referred to Appendix C, i.e. the accompanying CD-ROM, which contains a copy of all the data used for this study.

3 This tag indicates that this example is taken from the Web-based data collection specifically conducted for this study (cf. Chapter 7). The first abbreviation given in the brackets, i.e. .au, indicates the top-level domain from which this example is taken. As will be outlined in Section 7.2.4.1, .uk stands for the United Kingdom, .au for Australia, .ie for Ireland, .za for South Africa and .in for India. The information after the first forward slash specifies the exact search string of the query, which in the present case is it is me. Thus, if one wishes to obtain further information on this data point, such as the text type, mode of discourse or the Website from which it originates, the interested reader is referred to Appendix C, i.e. the accompanying CD-ROM, which contains a copy of all the data used for this study. Finally, the information after the second forward slash in the parentheses indicates the collection date of the data. Hence, this data point was obtained from a query for it is me on the Australian top-level domain on July 14, 2008.
complements (cf. Biber et al. 1999: 335-336; Maier 2013) and are the main focus of this study.

1.2 The Scope: The Distribution of Pronoun Case Forms in Subject Predicative Complements

Indeed, there are not many quantitative studies to date – let alone multivariate ones – that examine and try to account for the distribution of pronoun case forms in it BE sentences and it-clefs subsumed under the general category of subject predicative complements. However, there are several good reasons why a quantitative analysis of these contexts in particular can provide very valuable insights, not only regarding the contexts under consideration, but also for other contexts allowing for variability, as well as for the distribution of pronoun case forms in Present-Day English in general. However, before this section outlines why the analysis of it-clefs and it BE sentences may be a particularly rewarding field of study to account for the distribution of pronoun case forms in English, some general remarks on the notions of subject predicative complements, it BE sentences and it-clefs are in order to clearly outline what is referred to by these terms in this study.

As indicated by the respective examples in Table 1, it serves as the subject of the clause in both it-clefs and it BE sentences, which is then followed by a finite form of the copula BE. Furthermore, in both contexts, the personal pronoun form following it and the form of BE functions as the subject predicative complement of the clause. In the first sentence in Table 1, however, the pronoun additionally serves as the focal pronoun of a cleft construction (cf. Sections 6.1.2.1.1 and 6.1.2.5). Thus, we distinguish between these two types of subjective predicatives by adopting Quinn’s (2005a) terminology in referring to simple subject predicatives as it BE sentences (cf. (3)) and to those also serving as focal pronouns of an it-cleft sentence as it-clefs (cf. (4)).

(3) Subject Predicative Complement: it BE Sentence
Who wrote this? – It was I/me.

(4) Subject Predicative Complement: it-Cleft
It was he/him who wrote it.
At first sight, it might appear strange to attribute subject status to it in both it BE sentences and it-clefts. However, in each of these sentence types, it fulfils seven of the eight criteria for subjecthood proposed by Huddleston and Pullum (2002: 235-244). The reason the eighth criterion, i.e. case, is not met is that it is simply not applicable. Furthermore, comprehensive reference grammars of English clearly identify these constructions as (subject) predicative complements (Biber et al. 1999: 335-336; Huddleston and Pullum 2002: 251-252). In light of these considerations, the application of the term subject predicative complement to cover both it BE sentences and it-clefts is justified.

Having introduced some of the most important terminological concepts of this study, let us turn to the motivation for this study’s particular focus on the distribution of pronoun case forms in subject predicative complements, i.e. it BE sentences and it-clefts. To begin with, the distribution of pronoun case forms in subject predicative complements has been the subject of linguistic analysis for more than a century (e.g. Sweet 1875: 495). However, even though the variability in the use of pronoun case forms in both focal pronoun position of it-clefts and it BE sentences is still addressed and discussed today in most comprehensive grammar books of English (e.g. Biber et al. 1999; Greenbaum 1996a; Jespersen 1933; Quirk et al. 1985; Huddleston and Pullum 2002), there are hardly any quantitative studies examining the distribution of pronoun case forms in these two contexts – neither within nor across varieties of English (cf. Maier 2013; Quinn 2009). Thus, by analysing two mega-corpora as well as five Web-derived datasets, this study intends to shed more light on the use and distribution of pronoun case forms in it-clefts and it BE sentences in the analysed varieties, i.e. British, American, Australian, Irish, South African and Indian English.

Moreover, much of the current linguistic theory discussing variable contexts in the use of pronoun case forms assumes that the distribution of pronoun case forms in it-clefts and it BE sentences should be the same (cf. Sections 3.1–3.3). By analysing these two constructions, this study will be able to test and evaluate the predictions and validity of some well-established theoretical approaches addressing the distribution of pronoun case forms in Present-Day English.

The third reason why the analysis of the pronoun case distribution in it BE sentences and it-clefts is very promising is a methodological one. When looking for it BE sentences in a corpus, one also automatically obtains the corresponding it-clefts (cf. Sections 7.1 and 7.2.4.1). Thus, one receives results for two contexts allowing for variability in the use of
pronoun case forms at the same time. This advantage is not to be underestimated when taking into consideration that most contexts which allow for variability in the use of pronoun case forms are extremely rare in corpus data, which is one of the major reasons why there are so few quantitative studies (cf. Maier 2013; Quinn 2009).

Finally, the focus of this study on the distribution of pronoun case forms in subject predicative complements also results from this study’s central assumption, i.e. that subject pronoun case forms have been reanalysed as Focus\(^4\) markers in subject predicative complements. Indeed, this study argues that the more focussed a subject predicative complement context is, the more likely it is to observe a subject pronoun case form in a subject predicative complement context. Admittedly, this assumption may initially appear rather unorthodox given the expectations and approaches of much of the current linguistic theory (cf. Sections 3.1–3.3) as well as of accounts associating subject pronouns with preverbal, unstressed topical occurrences (e.g. Pietsch 2007: 167, 2009: 146). However, based on a comprehensive survey of the distribution and use of pronoun case forms in varieties of English and a thorough survey of previous research, this study is able to show that functional and particularly pragmatic factors seem to be very important to the distribution of pronoun case forms both in varieties of English and in other contexts allowing for variability (cf. Chapters 2 and 3). As a consequence, this study proposes and subsequently empirically tests a Focus-oriented approach to the distribution of pronoun case forms in subject predicative complements that can account for such a re-functionalisation process. Moreover, this Focus-oriented approach helps to explain the considerable extent of variation that can be observed in the distribution of pronoun case forms in *it BE* sentences and *it*-clefts in Present-Day English (cf. Chapter 4; Maier 2013).

\(^4\) Whenever this study uses the concept of Focus as defined in the framework of Functional Grammar, it also adopts the convention of spelling Focus with a capital letter. In Functional Grammar, Focus is defined as the relatively most important, salient or highlighted information in a clause (cf. Chapter 4 for a detailed discussion; Dik 1978: 130; Siewierska 2004: 159; Weinert and Miller 1996: 179).
1.3 The Main Objectives of this Study

On a very general level, the main objectives of this study are to analyse the distribution of pronoun case forms in subject predicative complements in British, American, Australian, Irish, South African and Indian English and shed further light thereby on the factors and variables influencing and determining the distribution of pronoun case forms in English in general.

Moreover, this study sets out to demonstrate that functional and also pragmatic factors – particularly Focus – play a prominent role in the distribution of pronoun case forms in Present-Day English (cf. Chapter 4). In order to do this, this study uses multivariate statistical models with the help of which two mega-corpora of British and American English, i.e. the *British National Corpus* (Davies 2004–) and the *Corpus of Contemporary American English* (Davies 2008–), as well as five prospectively compiled, Web-derived datasets of British, Australian, Irish, South African and Indian English will be analysed. By providing a thorough quantitative analysis of the distributions of pronoun case forms in six regional varieties of English, this study will not only be able to assess which factors may influence the distribution of pronoun case forms within varieties of English and to what extent, but can also find out whether there are any cross-varietal differences or trends observable in the distribution of pronoun case forms in subject predicative complements.

The precise research questions of this study as well as the hypotheses this study intends to test will be introduced after the theoretical foundations have been introduced on which these research questions and hypotheses actually rest (cf. Chapter 5). Until then, the following two questions will provide a rough guide for this study’s main objectives:

Q: How are pronoun case forms distributed in subject predicative complements in and across varieties of English?

Q: Which factors influence the distribution of pronoun case forms in these contexts?

Although they are only preliminary, the research questions outlined above may serve as an orientation for the reader to obtain a better understanding of the rationales underlying the surveys discussing the distribution of pronoun case forms in English in Chapter 2 as well as
the different approaches accounting for the distribution of pronoun case forms in English presented in Chapter 3.

1.4 The Methodological and Theoretical Orientation of this Study

As outlined in the previous section (cf. Section 1.3), the main objective of this study is to account for the distribution of pronoun case forms in subject predicative complements in six varieties of English. Hence, this study is embedded in the field of variationist research.

As far as this field of research is concerned, variationist research or research on morphosyntactic variation can be very broadly subdivided into two major branches. On the one hand, there are studies that rather focus on morphosyntactic variation within varieties of English (e.g. Gries 2003; Hilpert 2008; Lohmann 2011). On the other hand, there are also studies analysing morphosyntactic variation across varieties of English (e.g. Kortmann and Szmrecsanyi 2011; Schweinberger 2011).

The present study is located at the intersection of these two branches. Although this study examines whether there are cross-varietal differences and similarities in the distribution of pronoun case forms in subject predicative complements, this can only be reasonably done once the distribution of pronoun case forms as well as the factors determining this distribution is known for at least one variety. This variety can then serve as yardstick for subsequent cross-varietal comparisons. However, to the best of the author’s knowledge, no multifactorial analysis of the distribution of pronoun case forms in subject predicative complements is so far available which could function as such a yardstick for comparing different varieties. Thus, this study starts with the analysis of the British National Corpus, which is often considered the best corpus available (cf. Section 7.1.2), and then examines and discusses the distributions and factors influencing these distributions for each of the remaining corpora and datasets. By doing so, we will not only obtain an idea of how the distribution of pronoun case forms in subject predicative complements looks within the seven datasets examined for this study, but we will also be able to eventually filter out cross-varietal differences and similarities on the basis of these analyses (cf. Section 6.1.2.6).

As noted above, this study uses multivariate statistical methods to conduct the analyses of the different corpora and datasets. This is due to the fact that it has become
increasingly accepted in variationist research that a certain variation phenomenon is often subject to several factors that influence its shape or – to use the terminology of statistics – that affect the outcome of the dependent variable (e.g. Gries 2003; Lohmann 2011; Schweinberger 2011; Szmrecsanyi 2006). Moreover, as the discussion of the results in the respective chapters will show (cf. Chapters 8–14), the application of multifactorial statistical models also allows the determination of the exact magnitude and direction of impact for each significant factor (cf. Section 6.2). As far as its methodology is concerned, this study is therefore a corpus- and Web-based, quantitative account of morphosyntactic variation in and across varieties of English.

However, observing differences and similarities is one thing, accounting for them another. Thus, in order to explain and interpret the results obtained from the multifactorial analyses, this study mainly draws on concepts and insights of traditional Functional Grammar (e.g. Dik 1989; 1997; Siewierska 1991), Functional Typology (e.g. Bisang 2004; Kortmann 2004b; Siemund 2008) and also functional morphological theory (Haspelmath 2006; Mayerthaler 1981). In principle, however, the analyses, results and assumptions of this study should be compatible with a wide range of current functional approaches to the study of language.

1.5 The Organisation of this Study

As far as the structure of this thesis is concerned, the present study is organised as follows: Chapter 2 provides a comprehensive survey of the use and distribution of pronoun case forms in Present-Day English. This includes not only the use and distribution as postulated for Standard English, but also the distribution and use of pronoun case forms as observed in a wide range of English varieties. Moreover, this chapter offers a detailed survey of the contexts that allow for variability in the use of pronoun case forms in Present-Day English as well as the treatment and discussion of these contexts in the variationist literature.

Chapter 3 offers a survey of the most important approaches attempting to account for the distribution of pronoun case forms in English and the variation observed in many contexts. This survey includes and discusses the most important formal, positional and
functional accounts as well as approaches that try to explain the distribution of pronoun case forms in Present-Day English by means of different pronoun classes.

After having reviewed the current state of research, Chapter 4 introduces and motivates the Focus-oriented approach proposed and adopted in this study. In a nutshell, it is assumed that subject pronoun case forms have been re-functionalised as Focus markers in subject predicative complements and that the more focussed a certain subject predicative complement context is, the more likely it is to observe a subject form in a subject predicative complement context. Moreover, why and how such a reanalysis may have come about and what evidence corroborates such an assumption is also discussed in this chapter.

Based on the theoretical foundations laid down in Chapters 2–4, Chapter 5 presents the major research questions this study wants to answer and converts the most important of them into the central hypotheses that are tested in this study.

In order to test these hypotheses, Chapter 6 introduces and motivates the dependent and independent variables employed in this study as well as their operationalisation. In addition, this chapter discusses the statistical method used to analyse the data.

Speaking of data, Chapter 7 presents the databases used for this study. First of all, this chapter introduces the two closed corpora used for this study. Moreover, Chapter 7 also provides a thorough discussion of the problems and opportunities of using Web data and a meticulous description of the compilation, classification and refining of the data contained in the Web-derived datasets specifically constructed for this study.

Chapters 8–14 constitute the empirical backbone of this study. In particular, these chapters offer a statistical analysis of the data obtained from the closed corpora and Web-derived datasets examined in this study as well as a discussion of the distributions of pronoun case forms as observed in these varietal datasets.

By adopting a bird’s eye perspective, Chapter 15 provides a general discussion of the results obtained from the statistical analyses of the data particularly with regard to their repercussions for the central hypotheses presented in Chapter 5.

Finally, Chapter 16 concludes this study with a brief summary and an outlook for future research.
2 The Distribution of Pronoun Case Forms in English

In order to obtain an enhanced understanding of the mechanisms and factors influencing the distribution of pronoun case forms in subject predicative complements, the following chapter provides a survey of the distribution of pronoun case forms in Present-Day English. In particular, this survey helps to identify trends as well as factors that seem to influence the distribution of pronoun case forms in and across varieties of English. First of all, Section 2.1 describes the distribution of pronoun case forms in Standard English and addresses some important terminological notions as well as the range of phenomena to which the concept of case can or should be applied. In Section 2.2, a comprehensive survey of the distribution of pronoun case forms and the pronominal paradigms across varieties of English is provided, followed by Section 2.3, which focuses on those contexts that permit a choice between the uses of either subject or objects pronoun case forms in Present-Day English, including subject predicative complements. In particular, Section 2.3.1 introduces the most important contexts allowing for variation according to the body of literature, while Section 2.3.2 depicts how and to what extent these contexts are discussed in the framework of variationist linguistics. Finally, Section 2.4 provides a summary of the major insights obtained from this chapter.

2.1 Pronoun Case in Standard English

Case marking in Present-Day English is severely restricted, since English lost virtually all case affixes due to the dramatic changes it underwent during the transition from Old to Middle English (e.g. Denison 1993: 20-23; Lass 2006: 51-52, 69-75). Thus, all former functions of case, like the identification of grammatical roles and syntactic functions (e.g. Butt 2006: 3-6), have been taken over by other means such as word order or prepositions, and the syntactic function of a constituent now relies chiefly on its position within the clause (e.g. Denison 1993: 20-23; Lass 2006: 51-52, 69-75). As a consequence, different case forms in English are, for the most part, confined to the domain of pronouns, where the former case system or exponents thereof are partly retained, though to a rather modest extent, as Table 2 indicates. Indeed the former case system is most distinctively preserved in the distinction
between the relative pronouns *who* and *whom* and even more systematically in the domain of personal pronouns, particularly in the different sets of subject and object pronouns given in Table 2 (cf. Biber et al 1999: 328; Huddleston and Pullum 2002: 458, Quirk et al. 1985 335).

<table>
<thead>
<tr>
<th>PERSON</th>
<th>PERSONAL PRONOUN</th>
<th>POSSESSIVE</th>
<th>REFLEXIVE PRONOUN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUBJECT</td>
<td>OBJECT</td>
<td>DETERMINER</td>
</tr>
<tr>
<td>1st SINGULAR</td>
<td>I</td>
<td>me</td>
<td>my</td>
</tr>
<tr>
<td>1st PLURAL</td>
<td>we</td>
<td>us</td>
<td>our</td>
</tr>
<tr>
<td>2nd SINGULAR</td>
<td>you</td>
<td>you</td>
<td>your</td>
</tr>
<tr>
<td>2nd PLURAL</td>
<td>you</td>
<td>you</td>
<td>your</td>
</tr>
<tr>
<td>3rd SINGULAR</td>
<td>he</td>
<td>him</td>
<td>his</td>
</tr>
<tr>
<td></td>
<td>she</td>
<td>her</td>
<td>her</td>
</tr>
<tr>
<td></td>
<td>it</td>
<td>it</td>
<td>its</td>
</tr>
<tr>
<td>3rd PLURAL</td>
<td>they</td>
<td>them</td>
<td>their</td>
</tr>
</tbody>
</table>

Table 2: The Pronominal Paradigm of Present-Day English (Biber et al. 1999: 328)

Although Present-Day English exhibits only a few remnants of the formerly elaborate morphological system of case marking, there is an ongoing controversy in linguistic theory as to how many cases or case forms there are in Modern English. One extreme position was adopted by some 19th century grammarians who tried to transfer the Latin case system onto the English language. Sweet, for example, distinguished seven different cases for English, namely nominative, vocative, accusative, dative, genitive, instrumental and locative (1892: 49-52).

However, probably the most common classification of English with regard to its case system is a binary distinction between common and possessive case for nouns (cf. (5) and (6)) and a tripartite distinction between subject, object and possessive forms for pronouns (cf. Table 2).

(5) Common Case in Nouns in Present-Day English
a. king, queen, prince
b. kings, queens, princes

(6) Possessive/Genitive Case in Nouns in Present-Day English
a. king’s, queen’s, prince’s
b. kings’, queens’, princes’
This classification has a long-standing tradition and has been accepted by many scholars of different eras and linguistic frameworks (e.g. Mason 1885: 31-36; Sapir 1921: 176; Jespersen and Haislun 1949: 220; Quirk et al. 1985: 336; Haegeman 1993: 142-144; Denison 1993: 20-21; Hudson 1995: 375-378; Huddleston and Pullum 2002: 456; van Gelderen 2010: 13, 83-84).

Although this classification scheme has found many proponents, there has also been much discontent with this view, especially with regard to the status of the possessive or genitive case. Palmer, as early as the 1920s, only distinguished between “nominative” and “oblique”, and employed this distinction only for pronouns (1929: 42). According to this account, nouns do not inflect for case (1929: 33), since “genitive [z], [s] or [iz]” is not a real inflection because “[o]n logical grounds it is an independent word, for it may be separated from the noun and attached to the last member of the noun group” (Palmer 1929: 36). Possessive pronouns are therefore classified simply as determiners and not as genitive pronouns in this account (Palmer 1929: 45).

(7) The former president of France’s wife is a singer-songwriter.

The example in (7) illustrates what Palmer (1929: 36) criticises with regard to the classification of possessive ’s as genitive marker. The possessive marker attaches to the end of the whole noun phrase the former president of France. If it were, however, a real genitive marker, possessive ’s would attach to the head of the noun phrase, i.e. president, and not to the noun phrase as a whole:

(8) Die Ehefrau de-s ehemalige-n Präsident-en von Frankreich ist eine Chansonsängerin.

In (8), the German translation of (7), it can be observed that the head of the genitive noun phrase as well as the dependent article and the adjective modifying the nominal head are all marked for genitive case. The translation of the postmodifying noun phrase of France is, however, unmarked. Thus, there is a recognisable difference between the ‘genuine’ genitive marking in a language with a rich case morphology, such as German, and the system found in English. In fact, these observed differences have given risen to an ongoing and controversial debate whether or not possessive ’s and the possessive pronouns can indeed be considered as instances of case marking (cf. Hudson 1995).

Following up Palmer’s (1929) lines of argumentation, some accounts even oppose the idea of English having case distinctions altogether (Hudson 1995). After having discussed the
use and distribution of personal pronoun forms, possessive pronouns and determiners and possessive ‘s, Hudson concludes that there is no convincing evidence that there are any actual morphological case distinctions left in Present-Day English (cf. Hudson 1995: 390). Taking into consideration that the former functions of morphological case marking are now fulfilled by other strategies such as word order and prepositions, this perspective is far from absurd and indeed very interesting, since it also leaves room for reanalysis of the former case forms to reassign new functions to them. Such a reanalysis is not only part of one of the central assumptions of this study (cf. Sections 4.2 and 5.2), but can also be attested for many regional varieties of English (cf. Sections 2.2.2 and 2.2.3).

After having briefly touched upon the controversy of what can and cannot be considered an exponent of a case system, we will now focus on those pronoun forms that are in the centre of interest for this study. This study aims first and foremost to analyse and account for the distribution of pronoun case forms in subject predicative complements, where a particular alternation between two of the pronoun classes in Table 2 is observed. Hence, the study’s subsequent focus will be on the groups of pronouns which have been labelled as subject and object pronouns in Table 2. While other labels for these sets of pronouns are available and also quite common, such as nominative or accusative (e.g. Huddleston and Pullum 2002: 458) and subjective or objective (e.g. Quirk et al. 1985: 336), this study uses the labels subject form or subject case form to refer to those forms that are prototypically used as subjects of finite sentences. Moreover, this study uses the terms object pronoun or object case form to designate those pronouns that are prototypically used to refer to objects of transitive verbs and prepositions. The reason the terms subject and object pronoun are preferred to labels such as nominative or accusative pronoun is that the former are more theoretically neutral than the latter. Similarly, if this study uses the terms case, case-sensitive or case forms, this merely implies an acknowledgement that English once possessed a productive morphological case system, the remnants of which can still be observed in parts of the pronominal paradigm. The use of these terms does not imply, however, that this study shares the views of accounts assuming that English still has a productive case system or assumptions of abstract case as putatively universal language property (e.g. Haegeman 1993: 141-144). Instead, the functional outlook of this study is hereby once again affirmed. In addition, the focus of this study on the subject and object pronoun forms in Table 2 also allows us to evade the ongoing controversy as to whether
possessive ’s and possessive pronouns can or should be considered as case forms, although this study is, in principle, sympathetic to Palmer’s (1929) view on this matter.

To be even more precise, this study will focus primarily on those subject and object pronouns that are case-sensitive, i.e. those which have retained different forms for subject and object functions in Present-Day English. This restriction is of course due to the fact that variation between these subject and object pronoun forms in subject predicative complements is the main focus of this study (cf. Table 3):

<table>
<thead>
<tr>
<th>PERSON</th>
<th>PERSONAL PRONOUN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUBJECT</td>
</tr>
<tr>
<td>1st SINGULAR</td>
<td>I</td>
</tr>
<tr>
<td>1st PLURAL</td>
<td>we</td>
</tr>
<tr>
<td>2nd SINGULAR</td>
<td>he</td>
</tr>
<tr>
<td>2nd PLURAL</td>
<td>she</td>
</tr>
<tr>
<td>3rd SINGULAR</td>
<td>they</td>
</tr>
<tr>
<td>3rd PLURAL</td>
<td></td>
</tr>
</tbody>
</table>

*Table 3: The Case-Sensitive Pronoun Forms of Present-Day English (Biber et al. 1999: 328)*

Since the pronouns for the second person singular and plural as well as for the third person neuter pronoun do not have distinct subject or object forms (cf. Table 2), they will not play a prominent role in either the cross-varietal survey in Section 2.2 or in the multivariate analyses in Chapters 8–14.

So far, we have briefly addressed the controversy as to whether Present-Day English possesses any case distinctions and if so, how many. In addition, we have also already introduced the nomenclature to be used as well as the pronoun forms of which the distribution will be analysed in the subsequent chapters. However, we still have to address how the different subject and object forms are distributed – at least prototypically – in contemporary Standard English. With regard to the distribution of subject and object pronoun case forms, Biber et al. state that “[t]he distribution of the forms is generally straightforward: the nominative is used in subject position, while the object is used in object position and as the complement of prepositions” (1999: 335).
Thus, contexts in which we are supposed to observe only and obligatorily subject pronouns are non-inverted subjects (9a), inverted subjects in direct questions (9b), inverted subjects in questions tags (9c) and subjects of finite subordinate clauses (9d) (Huddleston and Pullum 2002: 459):

(9) Obligatory Subject Pronoun Contexts in Standard English
a. He loves ice cream
b. Does she also love ice cream?
c. Children love ice cream, don’t they?
d. Jane is sure that we love ice cream.

In a similar vein, the contexts in which we are supposed to observe only and obligatorily object pronoun forms are direct objects of verbs (10a), indirect objects of verbs (10b), objects of prepositions (10c) and subjects of non-finite infinitival clauses introduced by for (10d) (Huddleston and Pullum 2002: 461):

(10) Obligatory Object Pronoun Contexts in Standard English
a. Henry loves her.
b. Henry’s wife gave him a kiss.
c. The Johnsons used to live next to me.
d. For him to play for Celtic Glasgow is all he ever wanted.

Although at first glance this distribution seems to be “straightforward” (Biber et al. 1999: 335), reality is much more complicated. On the one hand, there are many contexts which permit a choice between the use of either a subject or an object pronoun form (cf. Sections 1.1 and 2.3). These contexts include not only *it BE* sentences and *it*-clefts, but also many other contexts such as coordinate noun phrases, comparatives, left and right dislocations and many more, which have been briefly touched upon in Section 1.1 and will be thoroughly discussed in Sections 2.3.1 and 2.3.2. On the other hand, regional varieties exhibit very different patterns of pronoun case distribution as compared to the Standard varieties, since many of them possess different pronominal paradigms or have refunctionalised the existing ones. Thus, before we discuss the distribution of pronoun case forms in those contexts that permit a choice between subject and object pronoun forms even in the Standard varieties (cf. Section 2.3.1), this study will provide a comprehensive summary of the major trends of pronoun case distribution across varieties of English. This survey will also help us to recognise the cross-varietal importance of factors which have thus far been rather neglected in the discussion of the distribution of pronoun case forms. More
specifically, this survey will demonstrate that pragmatic factors play a tremendous role in
the distribution of pronoun forms in varieties of English and that it is therefore likely that
this is also the case in subject predicative complements and in other contexts allowing for
variability. In particular, since this study assumes that the distribution of pronoun case forms
in subject predicative complements may be influenced by the pragmatic factor Focus (cf.
Chapter 4), this survey will lay the necessary foundation for the motivation of this
assumption (cf. Section 4.2).

2.2 The Distribution of Pronoun Case Forms across Varieties of
English

Compared to the standard paradigm outlined in the previous Section 2.1, a huge scope of
variation can be observed across varieties of English as far as the distribution of pronoun
case forms is concerned (e.g. Kortmann and Szmrecsanyi 2004; Kortmann 2004a; Milroy
and Milroy 1993; Quinn 2009; Wales 1996). Though this scope of variation is very diversified, the
following sections attempt to identify the major trends of pronoun case variation across
varieties of English, including phenomena such as retention and simplification, extension and
innovation as well as reanalysis and re-functionalisation.

2.2.1 Pronoun Case across Varieties: Pronominal Paradigms between
Retention, Extension and Idiosyncratic Variation

As stated above (cf. Section 2.1), the case marking system of English – including the
pronominal paradigm – has been subject to massive changes from Old English to the present
day (e.g. Lass 2006: 51-52, 69-75). However, not all varieties of English have undergone
these changes at the same pace. Hence, there are varieties that have retained an earlier
stage of pronominal case marking when compared to the Standard English system. The most
prominent retention phenomenon in terms of pronoun case forms, which can be observed
particularly in traditional northern and western varieties of British English, is the continued
use of distinct subject and object forms for the second person. In fact, the preservation of
forms such as thou/thee, thy, thine and thyself is attested for several varieties and has been

Shetland English is one of the traditional northern dialects in which a subject–object distinction for the second person singular has been preserved. As the examples in (11) indicate, *du* is used as second person singular subject pronoun, whereas *dee* is used as second person singular object pronoun. With regard to the second person plural, Shetland English conforms to Standard English in using *you* as both subject and object plural form (Melchers 2004: 38, 42-45).

(11) Shetland English (Melchers 2004: 44-45)
   a. Sees du yon, boy?
      ‘Boy, do you see that?’
   b. Whit wid du tink if some ane axed dee, [...].
      ‘What would you think if someone asked you, [...].’

Furthermore, a similar subject–object distinction for the second person singular has also been observed in Welsh English. In Wales, *thee* was used as both object and subject pronoun whereas *thou* was attested only as subject pronoun. To what extent this distinction is still in use today, however, is open for debate (Penhallurick 1991: 176-177, 2004: 105-106, 2007: 170).

(12) Welsh English (Penhallurick 1991: 176-177)
   a. Where are thee going?
   b. Pleased to see thee, mate!
   c. Hast thou got a gumboil?

In addition, a clear distinction between second person singular subject and object pronoun case forms is also retained in the dialects of the Bolton area. This variety also uses *thou* as second person singular subject form, whereas *thee* is used as second person singular object form (Shorrocks 1999: 72-80):

(13) Bolton Dialect (Shorrocks 1999: 194, 208)
   a. ‘thou couldn’t have wished for nowt nicer.’
      ‘you couldn’t have wished for anything nicer.’
   b. I’ll go agate with thee.
      ‘I’ll go home with you.’

Besides the retention of a case distinction for the second person, the Bolton dialect is remarkable for another peculiarity in its pronominal paradigm. This variety distinguishes
between stressed and unstressed forms of both subject and object forms (Shorrocks 1999: 72, 76). This distinction, although it is not as marked in Bolton as elsewhere, is very noteworthy, since such a contrast between emphatic and unemphatic forms may strongly influence the choice and distribution of pronoun case forms in many varieties of English, as we will see below (cf. Section 2.2.3).

However, when surveying the distribution and use of pronoun case forms across varieties of English, we observe not only retention but also extension phenomena. An example of a pronoun form that has extended its functional range is the use of us in several traditional varieties in the North of England and in Earlier African American English. In these varieties, us is not only used as a first person plural object pronoun but also as possessive determiner instead of our (Britain 2007b: 97; Kautzsch 2004: 351; Kortmann 2004a: 1096; Trudgill and Chambers 1991b: 7, Wakelin 1984: 81).

(14) North of England (Britain 2007b: 97)
We all take us cars to work nowadays

It should be noted that us is also used in other non-standard functions in Great Britain, two of which, in subject function as an instance of pronoun exchange (cf. Section 2.2.3) and in pronoun–noun phrase constructions (cf. Section 2.3.1), will be discussed in detail below.

A similar feature, though less pervasive than the use of object forms as possessive determiners, is the use of subject pronouns in possessive functions. In Urban African American Vernacular English, for example, the third person plural subject form they is used frequently as a possessive determiner. In fact, possessive they is often considered as one of those features that clearly distinguish Urban African American Vernacular English from European American English vernaculars (Green 2002: 102-103; Rickford 1999: 7; Wolfram 2004b: 333). Moreover, this feature is also observed in African Bahamian English, although it seems to be less prominent than in African American Vernacular English (Schneider 2004: 1113).

(15) African American Vernacular English (Green 2002: 103)
If they wanna go out and do something else with it, that’s they business.
‘If they want to go out and do something else with it, that’s their business.’

(16) Bahamian English (Reaser and Torbert 2004: 399)
They usually bring their own equipment
‘They usually bring their own equipment’
More widespread than the possessive uses of *us* and *they* is possessive *me*, i.e. the use of the first person object pronoun form *me* as a possessive determiner (Anderwald 2004: 177; cf. Section 2.1). This feature is fairly widespread across L1 varieties of English, as it is attested for six of the eight varieties of British English surveyed in Kortmann and Szmrecsanyi (2004: 1163), as well as for several North American varieties (cf. Clarke 2004, 2010; Wolfram 2004a; Kautzsch 2004). Examples for this phenomenon from Britain and North America include the following:

(17) Southeast England (Anderwald 2004: 177)
    I think me memory’s getting bad now, somehow.

(18) Newfoundland English (Clarke 2010: 83)
    Me teeth and that was drove through me lips.

However, while the phenomena discussed so far in this section can be considered either retention or extension phenomena, possessive *me* is more difficult to account for. This is due to the fact that this feature can be regarded as both a retention and an extension phenomenon. With regard to the former, some accounts assume that that “possessive *me* is a remnant of Middle English *mi/my*, which, as a very frequent and unstressed form, may not have undergone the Great Vowel Shift” (Anderwald 2004: 177). Thus, this feature can be explained by means of a phonological retention, which “reflects a pre-Great Vowel Shift” pronunciation (Clarke 2010: 40). Although this explanation seems very plausible, most accounts consider possessive *me* to be the result of an extension process whereby the object pronoun case form encroaches on the contexts and functions where Standard English requires the possessive determiner *my* (cf. Anderwald 2004: 177). This explanation is also very plausible since the extension of the functional range of pronoun forms is also attested for other pronouns in other varieties, as the discussion of possessive *us* and possessive *they* has shown. Since both explanations are equally plausible, the decision as to whether possessive *me* is better interpreted as a retention or extension phenomenon has to remain unresolved.

Except for possessive *me*, the discussion so far has centred mainly on features that are either retained from earlier case systems or on features of varieties that have even extended the use of some forms at the expense of others. Moreover, most of the phenomena addressed thus far are mainly attested in traditional L1 varieties. The following features, however, are much more difficult to classify and to account for, since most often no
explanation is offered for these variation phenomena. Moreover, the following features are observed in varieties that are themselves difficult to classify, since they are somewhere between the mainly traditional varieties dealt with so far and the family of varieties to be discussed in the next section, i.e. pidgins and creoles. Though some of the following varieties, such as Hawai’i Creole, are still considered to be pidgin and creole varieties today, the status of others is rather controversial, as in the cases of St. Helena English and Butler English (cf. e.g. Hancock 1991 vs. Wilson and Mesthrie 2004 for St. Helena; Mühlhäusler 1978: 15 vs. Kortmann and Szmrecsanyi 2004: 1145 for Butler English). What can be safely assumed for these varieties is, however, that all of them have either a pidgin and creole background or have at least been influenced by them. The ‘intermediate’ status of these varieties is also reflected by their use of pronoun case forms and the distribution thereof. Their pronominal paradigms are different from those of prototypical pidgin and creole varieties and are actually more similar to those of L1 varieties in that they do not exhibit as much variation from the standard paradigm as clear instances of basilectal pidgins such as Tok Pisin and Bislama (Crowley 2004: 684; Sakoda and Siegel 2004: 765; Smith 2004: 723; Wilson and Mesthrie 2004: 1013). Moreover, these varieties exhibit instances of pronoun case variation that cannot be allocated to superficially similar features in either the traditional L1 varieties discussed above or in the pidgin and creole varieties discussed in Section 2.2.2. Thus, a brief look at the pronominal paradigms of these ‘intermediate’ varieties suffices to show how difficult they actually are to classify.

The first variety to be discussed here is Hawai’i Creole. In this variety, object pronoun forms may be used in subject pronoun contexts. This object for subject form usage can be observed both in the subject position of declarative sentences and in the subject position of interrogative sentences. In Hawai’i Creole interrogative sentences, object forms are even used consistently instead of subject forms (cf. (19a)–(19b)) (Sakoda and Siegel 2004: 765-766).

(19) Hawai’i Creole (Sakoda and Siegel 2004: 766)
   a. hr sick.
      ‘she’s sick.’
   b. Hu him?
      ‘who is he?’

Considering the fact that this variety, unlike many other pidgins and creoles (cf. Section 2.2.2), possesses several sets of pronouns, including subject, object, possessive and reflexive
ones, these instances of variation, which in the case of pronominal case assignment in interrogative sentences are systematic, are even more remarkable (Sakoda and Siegel 2004: 765-767).

A similar variation phenomenon, though less systematic and basically confined to the first person plural, can be observed in St. Helena English:

(20) St. Helena English (Wilson and Mesthrie 2004: 1012-1013)
   a. Us is round there.
   b. Us go look.
   c. And then we goes along and we spread out.

Like Hawai’i Creole, this variety also possesses a pronominal paradigm that is in principle quite similar to that of Standard English. However, in St. Helena English, the first person plural object pronoun form *us* is used not only in object, but also in subject contexts (H Hancock 1991: 21; Wilson and Mesthrie 2004: 1013). As example (20c) illustrates, *we* is also used in St. Helena English in subject position. Hence, this feature is reminiscent not only of the possessive use of *us* discussed above, but also of the systematic use of object pronoun case forms in subject contexts known by the name of “pronoun exchange” (cf. Wagner 2004: 158-159; Section 2.2.3). Since it is not clear, however, whether the use of object forms in subject contexts in St. Helena English is functionally equivalent to the phenomenon of pronoun exchange, this feature is listed here and not in Section 2.2.3, which specifically focuses on functionally motivated reanalysis processes of pronoun case forms across varieties of English.

While St. Helena English uses the first person plural object pronoun in subject contexts, the opposite scenario, i.e. the extended use of the subject form *we*, can be observed in Liberian Settler English. In this variety, the first person plural form *we* can be used not only as subject pronoun, but also as an object pronoun form and even as a first person plural possessive determiner (Singler 2004: 889). According to the literature, the distribution of non-standard *we* in Present-Day Liberian Settler English suggests that it has been in use for a very long time. Whether it has its origins in Liberia or whether the early settlers transported it from the US to Liberia remains, however, unclear (Singler 2004: 889). Possible evidence in favour of a transportation hypothesis is the possessive use of *they* in African American English discussed earlier in this section, which is a similar functional extension of a subject pronoun form to the expense of other forms. Examples for the
Liberian Settler English use of *we* as a subject pronoun, an object pronoun and a possessive determiner instead of *our* are provided below:

(21) Liberian Settler English (Singler 2004: 889)
   a. When we done make we farm, we n’t know nothing about sell, we keep it, to have to eat.
      ‘After we made our farm, we didn’t think about selling [the produce]; we kept it so that
      we would have something to eat.
   b. Our people didn’t learn we how to swim.

The final feature to be discussed in this section is the use and distribution of first person singular pronoun forms in Butler English, which is a variety of English spoken on the Indian subcontinent and which is marked by a considerable extent of variability (e.g. Sebba 1997: 124). In this variety, the possessive determiner *my*, the subject form *I* and the object form *me* are used interchangeably (Hosali 2004: 1035; Sebba 1997: 125-126). However, scholars posit several explanations for this remarkable degree of variability. Hosali (2004), for example, accounts for the interchangeable use of pronoun case forms by means of “simplification” (2004: 1035). Sebba (1997), in contrast, comes to the conclusion that “individual speakers seem to behave idiosyncratically” (1997: 126). The considerable extent of variation observed in the use of pronoun case forms and the different explanations put forward to account for it makes it difficult to arrive at any robust generalisations for this variety (cf. Sebba 1997: 124-126). Examples of the interchangeable use of first person pronoun and possessive determiner forms are given below:

(22) Butler English (Sebba 1997: 126)
   a. My working spencer’s officer
      ‘I was working for …’
   b. Me not drinking madam.
      ‘I do not drink …’
   c. I age eh about fifty-one.
      ‘My age eh is about …’

So far, this survey has discussed case assignment systems and features that have varied between retention, extension and more or less idiosyncratic variation. The following section examines pronoun case distribution in pidgin and creole varieties, which also exhibit a wide scope of variation marked by the extremes of simplification and innovation.
2.2.2 Pronoun Case in Pidgins and Creoles: Emerging Pronominal Paradigms between Simplification and Innovation

In the preceding section, we briefly touched upon the use of pronoun case forms in some pidgin and creole varieties. This section will be completely dedicated to the use and distribution of pronoun case forms in this group of English varieties. The following examples taken from British Creole, Tok Pisin, Solomon Islands Pijin and Bislama exemplify what may often be considered typical instances of pronoun case use and distribution in English pidgin and creole varieties (Crowley 2004; Jourdan 2004; Sebba 2004, 2007; Smith 2004):

(23) British Creole (Sebba 2004: 202-207)
   a. Him deh ah jail.
      ‘He is in jail.’
   b. Phone Lefty, tell him seh we ready fe him now.
      ‘... tell him that we’re ready for him now.
   c. Me go a de airport.
      ‘I go to the airport.’
   d. Nuh tell me she, you nuh recognise yuh husband sister!
      ‘Don’t tell me you don’t recognize your husband’s sister!’

(24) Tok Pisin (Smith 2004: 730, 733, 738)
   a. em i laik go long gad em
      she/he PRED likes/is about to go to the garden
      ‘She/He likes/is about to go to the garden.’
   b. mi givim han long em
      ‘I have him/her a (helping) hand’
   c. mi kukim pinis
      I cook-TRANS COMPLETIVE
      ‘I have cooked it’
   d. yu yet yu les lo mi
      you FOCUS/REFLEX you tired of me
      ‘It’s you that’s tired of me.’

(25) Solomon Islands Pijin (Jourdan 2004: 706, 715)
   a. Disfela big ren ia renim mi tumas.
      ‘This big rain storm (rained on me) drenched me.’
   b. Mi wakem gaden blong mifela finis.
      ‘I have completed my work in our garden.’

(26) Bislama (Crowley 2004: 698-700)
   a. Hem i holem rop i taet.
      ‘She/He held the rope tightly.’
   b. Mi talem lo hem se bae mi kam.
      ‘I told him/her that I would come.’
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

c. Mi kam from yu bin singaotem mi.
   ‘I came because you called me.’

As the examples in (23)–(26) illustrate, the pronominal case paradigms in several pidgin and creole varieties have been simplified or reduced in that many of these varieties do not distinguish distinct forms for personal pronouns in subject and object function (Crowley 2004; Jourdan 2004; Kortmann et al. 2004; Mühlhäusler 1997; Sebba 2004, 2007; Smith 2004). The absence of a morphologically encoded distinction between subject and object pronoun case forms in some pidgins and creoles has even led to the generalisation that creoles do not exhibit case distinctions in pronouns at all (Burridge 2004: 1131; Escure 2004: 541). Such an assumption is, however, oversimplified if one takes the following examples from Jamaican Creole into consideration:

(27) Jamaican Creole
   a. Luk hou a krievm im krievm. (Patrick 2004: 421)
      ‘See how greedy she is!’
   b. She only do half day work fi come fi follow him go a airport. (Sistren 1987: 103)
      ‘She only worked half a day in order come here to follow him (going) to the airport.’

While some studies simply state that Jamaican Creole does not possess an overtly marked case system (Bailey 1966: 146; Radford 1997: 182, 205), other studies draw a more differentiated picture of this variety, acknowledging that the presence or absence of case marking strongly correlates with the level of acquisition and that in mesolectal varieties, not only of Jamaican but also other creoles, case distinctions can be observed (e.g. Huber 2004: 199; Patrick 2004: 428; Sebba 1997: 211; James and Youssef 2004: 466). Thus, a total lack of case distinctions can only be assumed to hold for the most basilectal forms of some creoles rather than for creoles in general.

The assumption that pidgins and creoles generally do not exhibit pronominal case distinctions is also challenged by the fact that the Australian Torres Strait Creole, the Surinamese Creoles and the Creoles of Trinidad and Tobago, for example, exhibit a reduced set of pronoun case distinctions even in their basilectal forms (Malcolm 2004; Shnukal 1988; Winford and Migge 2004; James and Youssef 2004). The Surinamese Eastern Maroon Creole employs a distinction exclusively for third person singular subject and object pronoun forms (Winford and Migge 2004: 508-509). In a similar vein, Torres Strait Creole distinguishes between first person singular subject and object pronoun forms but employs no morphologically encoded distinction for any other person (Shnukal 1988: 30-33). Examples
for such a reduced set of pronominal case distinctions from Eastern Maroon Creole and from Torres Strait Creole are given below:

(28) Eastern Maroon Creole (Winford and Migge 2004: 504, 510, 486, 507)
   a. Ef a feni en, da a o boo.  
      If she find it then she FUT breathe  
      ‘If she finds it (French papers), she’ll be happy/relieved.’
   b. A dati o kii en.  
      FOC DEM FUT kill her  
      ‘That will kill her.’
   c. Mi e wasi.  
      IMPFV wash  
      ‘I am washing (myself).’
   d. A tja wan bii kon gi mi.  
      she carry a beer come give me  
      ‘She brought me a beer.’

(29) Torres Strait Creole (Shnukal 1988: 58, 77)
   a. Ai wete yu ya we briz.  
      ‘I’ve been waiting for you here on the bridge.’
   b. Samting i apen ene mi.  
      ‘Something happened to me.’
   c. Em bin ran go sanbis.  
      ‘She/he ran to the shore.’
   d. Sun em i kama ut, ai pole em.  
      ‘As soon as she/he left, I followed her/him.’

Furthermore, there are pidgin and creole varieties that are even closer to the Standard English varieties in terms of pronominal case distinctions than the varieties just addressed. Cases in point are Kamtok, which distinguishes subject and object forms for the first and third person singular and the third person plural, and Gullah, which exhibits pronominal subject–object distinctions for as many persons – though not for as many genders – as the standard varieties of English do (Ayafor 2004: 916; Mufwene 2004: 362-363). Hence, it is more accurate to speak of variable degrees of simplification or reduction that may be observed in the pronominal paradigms of some pidgins and creoles rather than of a general absence of pronoun case distinctions in this group of varieties (Burridge 2004: 1131; Escure 2004: 541).

However, when surveying the distribution and use of pronoun case forms in pidgins and creoles, we cannot only observe trends towards simplification or reduction, but also quite the opposite, i.e. extension phenomena of the Standard English paradigm. An often-
The cited example of the extension of the Standard English pronominal case marking system is the pronominal paradigm of Samoan Plantation Pidgin English (cf. Table 4).

<table>
<thead>
<tr>
<th>PERSON</th>
<th>SUBJECT</th>
<th>OBJECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st SINGULAR</td>
<td>mi</td>
<td>(bilong) mi</td>
</tr>
<tr>
<td>2nd SINGULAR</td>
<td>yu</td>
<td>(bilong) yu</td>
</tr>
<tr>
<td>3rd SINGULAR</td>
<td>em, him, hi</td>
<td>(bilong) em (him)</td>
</tr>
<tr>
<td>1st PLURAL</td>
<td>mi ol</td>
<td>(bilong) as</td>
</tr>
<tr>
<td>2nd PLURAL</td>
<td>yu ol</td>
<td>(bilong) yu ol</td>
</tr>
<tr>
<td>3rd PLURAL</td>
<td>emol, himol</td>
<td>(bilong) dem</td>
</tr>
</tbody>
</table>

Table 4: Personal Pronouns in Samoan Plantation Pidgin English (Mühlhäusler 1997: 148)

According to Mühlhäusler (1983), Samoan Plantation Pidgin English distinguishes between subject and object forms throughout the whole pronominal paradigm, although there also seems to be a tendency to use the same set of pronouns in both object and subject functions (1983: 59). If his observations are correct, this means that Samoan Plantation Pidgin English not only distinguishes as many forms as Standard English but even extends the Standard English paradigm by additionally distinguishing between second person subject and object forms in both singular and plural. Thus, Samoan Plantation Pidgin English is one of the most remarkable varieties discussed so far, as it possesses object forms for each person in the pronominal paradigm. This seems even more striking considering the fact that it is a pidgin variety. Regrettably, Mühlhäusler does not provide any clear instances of these case distinctions for the second person, either for the singular or for the plural. In his examples (Mühlhäusler 1983), *bilong + pronoun* is actually most often used in a possessive and not an object function. The former usage as possessive is, however, not surprising, as a very similar or even identical construction is also attested for geographically neighbouring pidgin and creole varieties such as *blo(ng) + PRONOUN* in Bislama, *blong + PRONOUN* in Solomon Islands Pijin and *(bi)long + PRONOUN* in Tok Pisin (Crowley 2004: 684; Jourdan 2004: 707; Smith 2004: 733). The object use of this construction, however, seems to be restricted exclusively to Samoan Plantation Pidgin English, which becomes even more exceptional if we
reconsider the fact that the neighbouring varieties just mentioned – Tok Pisin, Bislama and Solomon Islands Pijin – lack case marking altogether, as has been discussed. Unfortunately, despite its very exceptional status, this feature does not occur frequently in Mühlhäusler’s examples. In fact, Mühlhäusler (1983) only provides one example of *bilong + PRONOUN* that could be interpreted as object case form (cf. (30)). However *bilong us* in (30) could also be interpreted as a possessive form:

(30) Samoan Plantation Pidgin English (Mühlhäusler 1983: 66)

Mi stap long Fiji WAN FAIV YIA BIPO. Mi go long ples mekim suga bilong as. Plenti Indian fella wokim de. Plenti Yuropin I wokim long suga...

‘I was in Fiji fifteen years ago. I went to this place to cut sugar for us. Many Indians were working there. Many Europeans were employed in the sugar industry.’

or:

‘I was in Fiji fifteen years ago. I went to this place to cut our sugar. Many Indians were working there. Many Europeans were employed in the sugar industry.’

In order to assess how stable this object case marking is, especially for the second persons, it would be desirable to devote more future work on Samoan Plantation Pidgin English to this particular aspect, since this seems to be one of its most remarkable characteristics not only when compared to other pidgin and creole varieties but within the whole paradigm of English varieties.

While Samoan Plantation Pidgin is assumed to extend the pronominal paradigm by adding case distinctions that are not or no longer available in Standard English, other pidgin and creole varieties exhibit another type of expansion or innovation strategy in their pronominal systems. British Creole, Ghanaian Pidgin English, Nigerian Pidgin English, the Surinamese Creoles and the Creoles of Trinidad and Tobago, for example, distinguish emphatic from unemphatic personal pronoun forms (cf. Faráclas 1996, 2004; Huber 1999, 2004; Sutcliffe 1984; Winford and Migge 2004; James and Youssef 2004). In some varieties, such as in British Creole, this distinction is indicated simply by differences in pitch (Sutcliffe 1984: 228). In other varieties, however, the difference between emphatic and non-emphatic pronoun forms is much more pronounced and results in a marked distinction between emphatic “free” and non-emphatic “bound” pronoun forms, which will now be exemplified with the help of Ghanaian Pidgin English (cf. Huber 1999: 197, 199). The Ghanaian Pidgin pronominal paradigm has been described as follows:
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

<table>
<thead>
<tr>
<th><strong>SUBJECT PRONOUNS (FREE)</strong></th>
<th></th>
<th><strong>SUBJECT PRONOUNS (BOUND)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>SINGULAR</td>
<td>PLURAL</td>
</tr>
<tr>
<td>1st</td>
<td>mi</td>
<td>wi</td>
</tr>
<tr>
<td>2nd</td>
<td>jù</td>
<td>jù</td>
</tr>
<tr>
<td>3rd</td>
<td>in ([in ~ i])</td>
<td>dèm ([dèm ~ dè])</td>
</tr>
</tbody>
</table>

*Table 5: Subject Pronouns in Ghanaian Pidgin English (Huber 1999: 197)*

<table>
<thead>
<tr>
<th><strong>OBJECT PRONOUNS (FREE)</strong></th>
<th></th>
<th><strong>OBJECT PRONOUNS (BOUND)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>SINGULAR</td>
<td>PLURAL</td>
</tr>
<tr>
<td>1st</td>
<td>mi</td>
<td>wi, ìs, ès, ìs</td>
</tr>
<tr>
<td>2nd</td>
<td>jù</td>
<td>jù</td>
</tr>
<tr>
<td>3rd</td>
<td>am</td>
<td>dèm ([dèm ~ dè])</td>
</tr>
</tbody>
</table>

*Table 6: Object Pronouns in Ghanaian Pidgin English (Huber 1999: 199)*

In Ghanaian Pidgin English, free subject pronouns are marked with an acute accent and hence bear a high tone. They cannot directly precede a verb, but occur in emphatic or contrastive contexts such as the focus position of a cleft sentence, whereas the low-toned or unemphatic bound pronoun forms always precede the verb slot (Huber 1999: 195-200, 2004: 871-872). Thus, the third person pronoun functioning as the subject of the cleft sentence in (31) is a bound form of the third person singular, whereas the focal pronoun of the cleft is a free form of the third person plural (cf. (31)):

(31) Ghanaian Pidgin English (Huber 1999: 185)

```
ì bì dèm giv
3 Sg/Subject/Bound COP 3 Pl/Subject/Free give
ès som moni
1 Pl/Object/Bound some money

‘it was they who gave us money’
```

In analogy to the two sets of subject pronouns, there are also two sets of object pronouns in Ghanaian Pidgin English. Bound object pronouns are also low-toned and typically found in non-emphatic contexts, and in most cases immediately follow verbs, whereas their free counterparts bear a high tone and occur in all other positions, as for example the object in comparative constructions or as heads of relative clauses (Huber 1999: 198).
Of course, the questions which arise are how the distinction between emphatic and non-emphatic forms is related to pronoun case forms across varieties of English and why this is relevant for the current survey and particularly for the distribution of pronoun case forms in subject predicative complements. First of all, this marked distinction between particularly stressed and unstressed pronoun forms is highly interesting, since the distinction between free and bound forms may be clearly marked in Ghanaian Pidgin. Table 5 and Table 6 indicate, for example, that a threefold distinction between subject free, subject bound and object free/bound pronoun forms is observable for the third person singular in this variety. Secondly, there may be a much stronger contrast between free and bound, i.e. between emphatic and non-emphatic, pronouns than between subject and object forms, as can be deduced from the forms for the first person singular (cf. Table 5 and Table 6). Thirdly, the distinction between emphatic and non-emphatic or between free and bound contexts is very important for the distribution of pronoun case forms in other varieties. This is due to the fact that the distinction between emphatic and non-emphatic contexts may be responsible for the decision as to which pronoun or case form is employed in which context, as we will see in the next section (cf. 2.2.3). Fourthly, the distinction between free and bound or between emphatic and non-emphatic forms and contexts may also have consequences for the distribution of pronoun case forms in those contexts that are the main focus of this study, i.e. contexts that seem to admit a choice between subject and object pronoun case forms. For example, possible differences between it-clefts and it BE sentences in terms of pronoun case choice may be explained by this distinction between free and bound or emphatic and non-emphatic contexts. Although pronouns in both clause-types appear in a postverbal position after it and a form of BE, the distinction introduced above would predict different pronoun form distributions for the two sentence types, as is partly indicated by the example in (31). Whether there are actually differences observable in these two contexts will be thoroughly discussed later.

In sum, the preceding two sections have shown that the use and distribution of pronoun case forms in English constitutes a cline, or even multidimensional continuum, exhibiting very marked cross-varietal differences. Varieties that lack overt case marking altogether in their basilectal forms are towards one end of this pronominal case marking cline. Cases in point are Tok Pisin and Bislama (e.g. Crowley 2004; Jourdan 2004; Sebba 2004, 2007; Smith 2004). However, the varieties exhibiting the most complex case marking
systems are not the Standard varieties as might be expected but also other non-standard varieties of English. First of all, this group of varieties with very complex case-marking systems includes the pidgin and creole varieties that have expanded the standard paradigm, such as Samoan Plantation Pidgin and Ghanaian Pidgin English (Huber 1999: 193-199, 2004: 871-872; Mühlhäusler 1983: 59, 1997: 148-149; Wales 1996: 92). Moreover, this group of varieties contains also those traditional varieties of English which have retained more case distinctions than the Standard English paradigm. Some of them, such as the Bolton dialect, may in addition even possess different sets of emphatic and non-emphatic pronoun forms (Beal 1993: 205, 2004: 118; Shorrocks 1999: 72-80; Wakelin 1984: 79). Thus, the discussion so far has not only demonstrated that there are marked cross-varietal differences in the use of pronoun case forms, but also that the distinction of emphatic and non-emphatic pronoun forms plays an important role in several varieties of English.

2.2.3 Pronoun Case across Varieties: The Reanalysis and Re-functionalisation of Pronominal Case Forms and Paradigms

After having surveyed the distribution of pronoun case forms in many varieties of English including traditional L1 varieties as well as pidgins and creoles, this study now focuses on a set of non-standard phenomena which have in common that they are based on a re-functionalisation process or reanalysis of existing pronoun case forms. Since this study assumes that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements, the following section is of central importance for the approach adopted in this study (cf. Chapter 4).

The first reanalysis process to be mentioned in this section can be observed in Appalachian English. In this variety, the historic pronoun form hit has not only been retained but also re-functionalised as an emphatic variant of the third person neuter pronoun. Hence, hit alternates with it in this variety (e.g. Montgomery 2004: 262; van Gelderen 2006: 166). As far as the distribution of these two forms is concerned, Appalachian English uses hit mainly in stressed positions, particularly as a subject form, as is illustrated by the example below (cf. (32)) (e.g. Montgomery 2004: 262):
And he followed the noise of the bear down a ways as far down as he heared it, and hit was so dark he couldn’t see any further, and he got to feeling with his gun to see if he could feel the bulk of the bear, and he couldn’t feel anything.

Moreover, a morphologically encoded contrast between emphatic and non-emphatic third person pronoun neuter forms can also be found in the traditional variety of English spoken in East Anglia:

(33) East Anglia (Poussa 1997: 691)

When you cut the corn and stack it and that heat theself dry, that’s a proper heat, that’s a natural heat. But when that’s dried on a dryer, that’s dried and then that’s – As that go along the dryer there’s the cool blowers what blow the cool air into it, that cool it. I don’t think that’s dry the same as that do when that dry theself.

As can be seen from the examples in (33), emphasis or stress also plays an important role in distinguishing third person singular neuter forms in this variety: It is only used in unstressed contexts whereas that/thass corresponds to stressed Standard English it (Britain 2007b: 96; Kortmann 2004a: 1097; Peitsara 1996: 293-294; Poussa 1997: 691-698; Trudgill: 2004: 146-147). Interestingly, this distinction in usage between that/thass and it in East Anglia has even been put forward as a case contrast between these two forms in earlier descriptions of this phenomenon. With regard to these variants it has been noted that “[…] it occurs only as an object pronoun, with third-person neuter singular subjects being indicated by that” (Trudgill and Chambers 1991b: 8). Indeed, this assumption is confirmed by the example given above where it is used only as object form (cf. ((33))). Furthermore, a distinction between a stressed and an unstressed variant of the third person plural can also be observed in this variety; they is usually used in stressed contexts, whereas thee is used in unstressed contexts (Kortmann 2004a: 1097).

The distinction between emphatic and unemphatic third person neuter forms is, however, not restricted to L1 varieties of English. In a similar vein, Fiji English distinguishes between two different forms for the third person singular neuter, or more precisely, third person ‘inanimate’ in this variety (Mugler and Tent 2004: 775). The two forms in question are (the) thing and it in Fiji English.

(34) Fiji English

a. The thing had been prepared right from April 19. (Siegel 1989: 55)
b. When I come back from school I bring it to you, the thing at home. (Thomas 1991: 37)

---

As might have already been inferred from the examples in (34), *the* thing is normally used as the subject of the sentence or where Standard English would use a noun. Compared to *it*, *the* thing has been considered as the “stronger” form of the two (Kelly 1975: 29-31; Mugler and Tent 2004: 774). Thus, we observe a situation quite similar to that found in East Anglia and Appalachian English.

While the former reanalysis phenomena have been more or less idiosyncratic and restricted to single varieties, there are also re-functionalisation processes that can be observed across several varieties of English in a similar way. The most prominent reanalysis or re-functionalisation phenomenon in the domain of pronominal case forms is known as “pronoun exchange” and is attested in several traditional varieties in England as well as in the variety of English spoken in Newfoundland (e.g. Britain 2007b: 96; Clarke 2004: 313; Ihalainen 1991: 106-107; Kortmann 2004a: 1097-1098; Wagner 2004: 157-159; Wakelin 1984: 81). The phenomenon of pronoun exchange can be described as the use of subject pronoun case forms in contexts and positions in which Standard English only and obligatorily uses object pronoun case forms (cf. Kortmann 2004a: 1097; Wagner 2004: 157; Section 2.1). Thus, subject pronoun forms may be used, for example, as objects of transitive verbs (cf. (35a)) or as objects of prepositions (cf. (35b)):

(35) Southwest England (Wagner 2002: 2)
   a. Don’t know if you’ve met he or no?
   b. There was a lot of they down there.

Moreover, the use of object pronoun case forms in contexts in which Standard English normally uses only subject pronoun case forms is likewise attested, as the examples below demonstrate:

(36) Southwest England (Ihalainen 1991: 105-106)
   a. Her do live by the pub, don’ er?
   b. We don’t know, do us?

However, the use of non-coordinated object pronoun case forms in positions and functions in which Standard English obligatorily uses subject pronoun case forms seems to be subject to much more restrictions (Clarke 2004: 313; Ihalainen 1991: 106-107; Trudgill 2004: 147; Wagner 2004: 158). In fact, each variety in which pronoun exchange is attested exhibits particular characteristics with regard to the set of pronouns that can be exchanged and with
regard to the overall frequency and likelihood of exchanging certain forms, as the subsequent discussion will show.

Wagner (2004), for example, reports for the Southwest of England that subject pronouns are used much more often in object positions than uncoordinated object pronouns in subject positions (2004: 158). Furthermore, Wagner (2004) notes that regions with a high share of subject pronouns in object positions very often exhibit a low share of exchanged object pronouns, and vice versa. In addition, Wagner concludes that the use of subject forms in object pronoun contexts seems to diffuse from West Cornwall eastwards (2004: 158-159). As a possible explanation for this phenomenon, Wagner observes a general tendency in colloquial varieties of English to use subject forms in object contexts, such as in the notorious between you and I, which may promote such a spread (Wagner 2004: 158-159).

In East Anglia, only subject pronouns can be exchanged and the southwestern usage of him, her and us as subjects does not seem to have been possible. Only the use of he, she, we, and they as objects has been clearly attested (Trudgill 2004: 147-148).

(37) East Anglia (Trudgill 2004: 147-148)

Tha’s where the’re gooin’, are they? Pas’ the mill,
Along the fiel’ path leadin’ tard the woods;
I’ll give he what for some day, that I will [...].

In the Southeast of England, on the other hand, the exact opposite is attested, namely the use of object pronoun forms in subject functions (Edwards 1993: 229):

(38) Southeast England (Edwards 1993: 229)

a. Him and her are the ones you should pick.
b. Them what you like should come.

It should be noted, however, that the latter example in (38b) given by Edwards (1993) could also be a non-standard demonstrative pronoun and not a personal pronoun form. Hence, this example is rather inappropriate to illustrate this object for subject pronoun usage pattern, which is, however, very interesting since it constitutes the reversal of the patterns attested in the Southwest of England and in East Anglia.

In contrast to Ihalainen’s assumption that pronoun exchange is confined to southern and western dialects of English (1994: 231), this phenomenon is also attested for the North of England, though restricted to the first person plural (Beal 1993: 205-206; 2004: 117-119; Britain 2007b: 96).
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

(39) North of England (Beal 1993: 205)
   a. They beat we four nil!
   b. Us’ll do it.

Furthermore, pronoun exchange can also be observed in the variety of English spoken in Newfoundland, where its presence is attributed to the influence of settlers from the Southwest of England (e.g. Clarke 2004; 313; Paddock 1991, 1994; Schneider 2004: 1113; Shorrocks 1992: 439-441; Wales 1996: 89-90).

(40) Newfoundland English
   a. And dere was ‘Melier (‘Amelia’) next to she (i.e. ‘in age’) (Clarke 2004: 313)
   b. Give that to I (Newhook 2002: 25-26)

While some studies demonstrate that the use of subject pronouns in object positions is still fairly widespread, especially among male speakers (Newhook 2002: 59-63), the use of object pronouns in subject positions seems to be much less frequent today (Clarke 2004: 312; Paddock 1994: 260).

Having discussed where and in what ways pronoun exchange occurs in different varieties of English, the reasons why this is done at all still have to be addressed. The central notion for the understanding of pronoun exchange is ‘emphasis’ (e.g. Kortmann 2004a: 1097; Wakelin 1984: 81). According to the body of literature, subject pronoun case forms are used when the respective pronoun form is particularly emphasised, irrespective of their position in the clause. Object pronoun case forms, in contrast, are used whenever the respective pronoun form is not emphasised (Elworthy 1877: 32-39; Kortmann 2004a: 1097; Kruisinga 1905: 35-36; Trudgill 2004: 147; Wagner 2004: 158; Wakelin 1984: 81; Wright 1905: 271). This would explain the exchange of pronoun forms for all varieties discussed above except for the exchange phenomenon observed in the Southeast of England: There, the opposite seems to be the case, as the emphatic pronouns in south-eastern non-standard English take the same form as the Standard English object forms (Edwards 1993: 229). Unfortunately, the authors often do not clarify what exactly is meant by the notion of ‘emphasis’ and the contrast between emphatic and non-emphatic pronouns, i.e. whether it is a matter of phonological or pragmatic prominence (Wagner 2002: 8). Thus, it is not surprising that this traditional explanation has been modified and expanded to claim that subject forms in object positions are not exclusively restricted to emphatic, i.e. phonologically stressed, contexts. Subject pronouns in object slots are also attested in positions that are rather phonologically unstressed but that express a contrast or that have a
contrastive function. Hence, contrastive stress is even today one of those factors favouring subject forms in object positions (Ihalainen 1991: 106-107; Shorrocks 1992: 440-441). Some authors even consider the phenomenon of pronoun exchange as a systematic re-functionalisation process in which the former case contrasts are reanalysed to mark the Focus of a clause in some varieties of English (cf. Paddock 1994). This interpretation of this phenomenon is indeed very interesting and it will be further discussed in Chapter 4, since it is of central relevance to the present study. Moreover, it has been remarked by some authors that the syntactic contexts in which the pronouns occur seem also to influence the choice of the case forms. There are certain favourable contexts, such as the position adjacent to a verb and the slot following a preposition, in which pronoun exchange is most likely to occur (e.g. Elworthy 1877: 33; Kruisinga 1905: 35-38; Rogers 1979: 34; Story et al. 1990: xxvi; Wagner 2002: 2, 8, 11, 16-20).

Despite the differences in occurrence, definition and explanation, the remarkable thing about the phenomenon of pronoun exchange is that the standard pronominal paradigm has been reanalysed and re-functionalised in that – depending on the terminology of the respective account – stress, Focus, emphasis, contrast and syntactical context influence the distribution of pronoun case forms. Thus, new pragmatic functions were allocated to available pronoun forms that had lost their original syntactic functions due to the emergence of a fixed SVO word order and the considerable structural changes the English language has gone through since Old English (e.g. Shorrock 1992: 440-441; Paddock 1991: 36-37, 1994: 256-263; Rogers 1979: 35; Wagner 2004: 158; Wales 1996: 91). This also ties in with the observations made for Ghanaian Pidgin English in Section 2.2.2, namely that the differences between emphatic and non-emphatic pronouns may be more pronounced than the contrast between subject and object forms and that certain syntactic contexts favour certain pronoun forms.

In sum, the phenomena discussed in this section have shown that there seems to be a close interaction between the degree of emphasis, stress or Focus and the choice of the pronoun form or pronoun case forms. In particular, the morphological encoding of a contrast between emphatic and unemphatic, stressed and unstressed or focussed and unfocussed pronoun forms seems to be a desideratum of many varieties of English. This is also in line with the observations made in the preceding two sections, where we have also already
observed that the encoding of a contrast between emphatic and non-emphatic forms may be very important.

2.3 Contexts Permitting Variability in the Use of Pronoun Case Forms in Standard English and across Varieties of English

So far, we have only discussed the Standard English paradigm of pronominal case forms and their distribution (cf. Section 2.1) as well as clear instances of non-standard case forms and their distribution (cf. Section 2.2). Now, we will shift our attention to those contexts in which a competition between subject and object pronoun forms can be observed. These contexts are also responsible for a controversial discussion and for alternative theories accounting for the variation in pronoun case distribution observed in these environments in English. However, to maintain the survey character of this chapter, the discussion will be confined in the following sections to the mere introduction of the most important environments in which subject and object forms compete with each other in Section 2.3.1 as well as to their treatment in the variationist literature (cf. Section 2.3.2). The different theories accounting for the distribution of pronoun case forms and for the contexts exhibiting variability will be addressed and discussed in Chapter 3.

2.3.1 Variability in the Use of Pronoun Case Forms: The Environments

Even in Standard English, a considerable amount of variation with regard to pronominal case selection can be observed (e.g. Huddleston and Pullum 2002: 458-459). The result of this kind of variation is a considerable uncertainty among both native and non-native speakers of English regarding the choice of the correct pronoun case form (cf. Section 1.1). Today, this uncertainty can be observed, especially on the Internet, not only in different question and answer forums, as example (41) illustrates, but also in many other contexts such as (42), a reader’s comment on a newspaper article:

(41) Which is grammatically correct: It is me? or It is I? Which do you say? (.uk/it is me/15.07.2008)

(42) [...] someone has to say it and this time it is me (or I). (.au/it is me/14.07.2008)
In the following, the most important constructions in which subject and object pronouns often alternate according to the body of literature are introduced and discussed (Biber et al. 1999: 335-340; Huddleston and Pullum 2002: 458-467; Jespersen and Haislund 1949: 220-281; Quirk et al. 1985: 335-339; Quinn 2005a: 201-309).

(43) Subject Predicative Complement: *it BE* Sentence
   a. Who did this? – It was he/him.
   b. Who is the best student in class? – It is I/me.

(44) Subject Predicative Complement: *it*-Cleft
   a. It was he/him who kissed you.
   b. It is they/them who have failed the exam.

The first contexts to be discussed are of course subject predicative complements. Following *it* and a form of *BE*, both subject and object pronoun forms are attested. This applies to clauses in which the pronoun form is a simple subject predicative, i.e. *it BE* sentences, like (43) and to clauses in which the pronoun functions as the focal pronoun of an *it*-cleft construction, as in (44). Although these contexts are indeed very similar, there are differences in the pronominal case choice observable in the literature. Whereas the use of subject forms as predicative without being followed by a dependent clause is often considered to be very formal or even pedantic, the use of subject forms as focal pronoun in *it*-clefts is more frequent (Biber et al. 1999: 335; Erdmann 1978: 75-78; Huddleston and Pullum 2002: 459; Maier 2013). In fact, in constructions like those represented in (43), the object forms are clearly predominant despite the prescriptive grammar tradition calling for subject forms in these contexts (Biber et al. 1999: 335; Quirk et al. 1985: 336). Furthermore, pronoun case variation is also observed in predicative constructions with *this or that* rather than *it* as their subject pronouns (Huddleston and Pullum 2002: 459). However, these instances are less prototypical, which is also established by the fact that they are discussed much less often in the literature than those subject predicative complements that are the focus of this study, i.e. those following *it* and a form of *BE* (e.g. Biber et al. 1999: 335; Quirk et al. 1985: 336).

(45) *Than*-Comparatives
   a. Jane is younger than I/me.
   b. Our competitors are just better than we/us.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

(46) Comparative as
   a. Anna bought the same dress as I/me.
   b. They are by no means as good as we/us.

Variation in the use of pronoun case forms can also be observed for pronouns following as and than. According to traditional grammars, the pronoun forms in (45) and (46) are subjects of elliptical clauses and thus are supposed to occur in the same case form as when the finite verb form is present, i.e. the subject form (e.g. Huddleston and Pullum 2002: 460). Whether pronouns following than and as can indeed be interpreted as elliptical clauses also depends on whether than and as are interpreted as conjunctions or as prepositions. Conjunctions have no influence on pronoun case choice and consequently would require a subject form, while prepositions are obligatorily succeeded by an object pronoun form (Huddleston and Pullum 2002 460; Jespersen 1933: 133-138; Quirk et al. 1985: 337). This, however, is in itself a difficult question to answer, as there is no conclusive answer available (Huddleston and Pullum 2002: 460, 1113-1118; Jespersen and Haislund 1949: 227-232). Notwithstanding these definitory problems with regard to the statuses of than and as, the use of object forms in these contexts is extremely widespread and ubiquitous in colloquial speech (Biber et al. 1999: 336-337; Huddleston and Pullum 2002: 460; Jespersen 1933: 133). The widespread acceptance of object forms after than and as is also witnessed by other authors, who consider the prepositional use of than as having “become, or are on their way to becoming, standard or marginally standard” (Trudgill 1984b: 43).

(47) Pronoun–Noun Phrase Constructions
   a. We/Us British don’t like that.
   b. We/Us people are the sovereign!

As is illustrated in (47), variation with the first person plural can even be observed in non-coordinated subjects of finite sentences. Although we is considered the norm, both regional varieties and colloquial varieties of Standard English exhibit the object form us in contexts where the pronoun form is subject to an appositive modifier (Huddleston and Pullum 2002: 447, 459; Wales 1996: 100). In these contexts, the first person plural pronoun form is modified by an integrated dependent in the noun phrase structure, which is usually semantically restrictive in that it restricts the denotation of the nominal head and typically provides identifying information. Furthermore, this construction is different from supplementary appositions in that the head and its appositive are not pronounced separately but belong to the same intonation unit (Huddleston and Pullum 2002: 447).
(48) Subjects of Gerund-Participials in Adjunct Function (Huddleston and Pullum 2002: 460)
   a. We were in Greville’s office, I sitting in his swivel chair behind the vast expanse of desk, Anette sorting yesterday’ roughly heaped higgledy-piggledy papers back into the drawers and files that had earlier contained them.
   b. He could think of a few himself, I expect, him being so much in the business already.

(49) Subjects of Gerund-Participials in Complement Function (Huddleston and Pullum 2002: 1190-1193)
   a. I remember his/him reading my mail.
   b. I resented their/them going without me.
   c. No one objected to him/his joining the party.

A further environment in which variability in the usage of pronoun case forms can be observed is the subject position of gerund-participials of which we can distinguish two subclasses. The first is gerund-participials functioning as adjuncts or supplements to a clause, as shown in (48). In these contexts, variation between subject and object forms can be observed, with the object form being the marginal alternative in informal style and the subject form the more common option. The dominance of the subject forms, which is a surprising finding taking the preference of object forms in other contexts into account, can be attributed to the fact that this construction itself is of a rather formal nature. Thus, the preference of the subject forms is considered to be in line with the formal character of subject form usage attested for it BE sentences and comparatives discussed above (Huddleston and Pullum 2002: 459-460, 1191-1192). Moreover, alternation of pronoun case forms can also be observed in contexts in which the non-finite clause is functioning as a complement, as shown in (49). In these contexts, the pronouns may not alternate between subject and object pronoun forms but instead between possessive and object forms. According to the literature, the choice of one form over the other may be influenced by a number of variables such as the degree of formality, the type of the noun phrase and the matrix construction. Interestingly, in this context it is again not the object form but the alternant, which in this case is the possessive form, that is considered the more formal variant of the two (Huddleston and Pullum 2002: 467-468, 1192-1193).

(50) Subjects of (other) Verbless Clauses
   a. She is a bore, he/him a party animal.
   b. What? He/Him a traitor?

However, we can observe alternating pronoun case forms not only in gerund constructions but also in other verbless clauses – or small clauses as they are called in some accounts (Huddleston and Pullum 2002: 460-461; Jespersen and Haislund 1949: 239; Quinn
2005a: 262-272). The sentences above in (50a) and (50b), though verbless, contain a subject, i.e. the pronoun form, and a predicative element. What distinguishes them from their corresponding full-fledged counterparts is the absence of a finite verb form, i.e. the copula verb BE (Haegeman and Guéron 1999: 109; Huddleston and Pullum 2002: 461). In these contexts, object forms are used in more colloquial or informal contexts than their subject counterparts (Huddleston and Pullum 2002: 461; Quirk et al. 1985: 844).

(51) Complements of Exclusive but
a. No one but he/him can break my heart.
b. Nobody but he/him can save us now.
c. Nobody helped Jim but me.
d. I love nobody but her.

Another context in which alternating pronoun forms are attested is the noun phrase slot following so-called “exclusive but”, the meaning of which can be paraphrased as ‘except’ (cf. Huddleston and Pullum 2002: 461). Since for except the same usage patterns are attested as for but as far as the distribution of pronoun case forms is concerned (Quirk et al. 1985: 339; Swan 2005: 429), both but and except are discussed together in this paragraph. In these contexts, indefinite pronouns such as nobody, everyone, all, etc. preceding but or except are followed by a pronoun form. These pronoun forms may surface either as subject or object form, also depending on whether but and except are considered conjunctions or prepositions, a situation which is reminiscent of that observed in comparative than and as discussed above. Although the subject forms here are considered the more formal variant, they are normally restricted to instances in which but/except + PRONOUN immediately follows the subject as in (51a) and (51b). In (51c) and (51d), the subject form does not seem to be an admissible option, but may be considered as an instance of hypercorrection (Huddleston and Pullum 2002: 461; Quirk et al. 1985: 339). However, the view that the degree of formality or the position of the pronoun influence or co-determine the distribution of pronoun case forms following but is not universally accepted. Some authors generally rule out the use of subject forms after but irrespective of the degree of formality or the position of the pronoun (e.g. Swan 2005: 429). Thus, there is dissent even among scholars as to which contexts exhibit or allow for alternating pronoun case forms and which do not. However, this is certainly not the only context for which it is unclear whether or not it is variable in terms of pronoun case distribution, as we will see in the subsequent discussion of independent pronouns, left and right dislocations and modified pronouns.
Independent Pronouns

a. What did you do? – I/Me?
b. Who tried to rob you? – He/Him!

The next contexts discussed in this study which allow for variability in the use of pronoun case forms are so-called “independent pronouns” or “stranded subjects”, i.e. pronouns which surface on their own as very brief responses to questions, especially in conversations, as illustrated in (52) (e.g. Biber et al. 1999: 339; Greenbaum 1996a: 171; Quinn 2005a: 228-233; Quirk et al. 1985: 337; Swan 2005: 428). In terms of their syntactic status, independent pronouns have been analysed as either ellipted clauses (e.g. Huddleston and Pullum 2002: 461; Quirk et al. 1985: 337) or as non-integrated or non-sentential noun phrases (e.g. Biber et al. 1999: 339). What all accounts discussing this phenomenon have in common, however, is that they state a clear preference for object forms as independent pronouns. In some accounts (Kjellmer 1986: 445-446; Quirk et al. 1972: 230), this preference for object forms has been explained in terms of formality, whereas other accounts, without any reference to stylistic differences, consider independent pronouns as functions in which object forms normally occur (e.g. Biber et al. 1999: 339; Greenbaum 1996a: 171; Jespersen 1933: 136; Kruisinga and Erades 1911 [1960]: 439). Some accounts go even further by explicitly ruling out the use of subject forms altogether in this function, unless they are directly succeeded by a finite verb (e.g. Swan 2005: 428). Thus, although this context has traditionally been considered to allow for pronoun case variation – at least in earlier times (e.g. Quinn 2005a: 228-233) – it is not clear to what extent this context still exhibits variability.

Left and Right Dislocations

a. I/Me, I usually don’t like cucumbers.
b. I have always defended you, I/me.

Pronoun case forms in left and right dislocations (53) have also been traditionally discussed in the context of variability of pronoun case distribution (e.g. Erdmann 1978: 68-69; Jespersen and Haislund 1949: 223-225; Quinn 2005a: 214-228). However, left and right dislocations are not only in this respect but also in others similar to isolated pronouns discussed above. To begin with, the pronouns in left and right dislocations are clause-peripheral and hence also not integrated in the clause structure they are preceding or following. The syntactic similarity of both independent pronouns and left and right dislocations is also acknowledged by the fact that some authors even discuss these
phenomena under the same heading (e.g. Biber et al. 1999: 339). Moreover, a clear preference for object pronoun forms is observable in left and right dislocations, as was the case for independent pronouns (Biber et al. 1999: 339; Quinn 2005a: 214-228). Furthermore, although alternating case forms in left and right dislocations have been attested in both past and present (e.g. Jespersen and Haislund 1949: 223-225; Quinn 2005a: 214-228), there are also accounts that rule out the use of subject pronoun forms in these contexts. Huddleston and Pullum, for example, consider these left and right dislocations “characteristic of informal style” and discuss them in the context of “constructions where accusative is obligatory” (2002: 461-462). Hence, we are dealing once again with a context in which the variability of the pronominal case assignment is rather controversial, at least nowadays.

(54) Modified Pronouns
   a. Poor little I/me. (Jespersen and Haislund 1949: 134)
   b. Poor/Clever/Lucky (old) you! (Quirk et al. 1985: 352)

A similar situation to that of independent pronouns and left and right dislocations can be observed for pronouns that are modified by an adjective, as in (54). Although there are other ways to modify pronouns, this type of modification is also frequently mentioned in the discussion of contexts that permit variability in the use of pronoun case forms (e.g. Jespersen and Haislund 1949: 134; Quinn 2005a: 288-289). As was the case before, however, many accounts claim that object forms are the only possible option in this construction (e.g. Greenbaum 1996a: 171; Quirk et al. 1985: 352; Swan 2005: 429), and other accounts even explicitly preclude the use of subject forms from this context in Present-Day English (Huddleston and Pullum 2002: 430). Thus, it is rather controversial to what extent these contexts exhibit or permit variation in the use of pronoun case forms.

(55) Coordinated Noun Phrases
   a. Jane and I/me split up last year.
   b. He/Him and Katherine are on holiday.
   c. Between you and me/I, this is the stuff I really love.
   d. This was really an amazing experience for Henry and I/me.

Without revealing too much of the subsequent discussion of the different theories accounting for the distribution and variation of pronoun case forms in Chapter 3, as well as the factors influencing this distribution (cf. Chapter 6), many contemporary grammars of English agree, as we have seen above, that one of the central factors accounting for the alternation between subject and object forms, especially in (43)–(51), is the degree of
formality. While object forms are favoured in informal registers and situations, the subject forms are still associated with more formal styles and registers (Biber et al. 1999: 335-340; Huddleston and Pullum 2002: 459-467; Quirk et al. 1985: 336-338). Though the variation of pronoun case forms observed in coordinated noun phrases is also, among other factors, supposed to be determined by the level of formality, coordinated noun phrases deserve an extra comment. In the examples in (55), two different kinds of variation between subject and object forms can be observed in the coordinated noun phrases at hand. In (55a) and (55b), object forms are used for subject forms, and in (55c) and (55d) subject forms are used for object forms. Although it is often considered non-standard or even stigmatised (Huddleston and Pullum 2002: 463), the use of object pronoun forms in coordinated subjects is common in Present-Day English, particularly in conversation, and thus is often considered as a matter of style and formality (Biber et al. 1999: 337-338; Kerswill 2007: 35; Quirk et al. 1985: 338).

The use of subject pronouns in coordinated objects is, in turn, closely related to the use of object forms instead of subject forms in coordinated subjects in that it is related to the stigmatism attached to object forms in coordinated subjects like those in (55a) and (55b). Due to the prescriptive bias against object forms in subject coordinations, people may overgeneralise this avoidance strategy to include coordinated noun phrases in positions that would require object pronoun forms. Hence, this use of subject forms in coordinations in object positions has often been labelled as instance of hypercorrection (e.g. Bauer 2002: 107; Quirk et al. 1985: 338). It has been remarked, however, that coordinate constructions in object positions with final I, as in (56a), have become so frequent that they must be considered a feature of Standard English. Thus, the label hypercorrection should be confined to instances like (56b) and (56c) (Huddleston and Pullum 2002: 463).

(56) Subject Forms in Coordinated Objects (Huddleston and Pullum 2002: 463)
   a. The present was supposed to represent Helen and I, that was the problem.
   b. There’s a tendency for he and I to clash.
   c. They’ve invited the Smiths and we to lunch.

However, the same thing could be said about the use of object forms in coordinated subjects: They, too, are common and widely accepted by a large number of speakers in Present-Day English and yet their usage is considered to be non-standard or even “reprehensible” (Huddleston and Pullum 2002: 463; Kerswill 2007: 35; Quirk et al. 1985: 338). This difficulty in assessing and defining the standardness or non-standardness of using
a particular pronoun case form in one of the pronoun case variation contexts when compared to another actually very similar one illustrates the intermediate status of these phenomena between clear instances of the standard system and clear non-standard patterns of pronoun case distribution. Therefore, it seems both reasonable and necessary to approach these contexts exhibiting variability from both the Standard English perspective, as represented by contemporary grammar books, and the variationist perspective, in order to illuminate these contexts and the possible factors influencing the distribution of pronoun case forms as good as possible.

In general, coordinated noun phrases have so far received the most scholarly attention among those constructions that exhibit variation in the choice of pronoun case forms (e.g. Angermeyer and Singler 2003; Parker, Riley and Meyer 1988; Quattlebaum 1994). This is probably because of the high frequency of this construction compared to other constructions exhibiting variation in the use of pronoun case forms (Quinn 2009). Hence, it is not surprising that this construction is also the context that is most widely discussed across varieties of English.

2.3.2 Variable Usage of Pronoun Case Forms across Varieties of English

To begin with, one very prominent example of the environments exhibiting variability in the use of pronoun case forms is that of the *it* BE sentences, i.e. simple pronominal complements following *it* and a form of BE.

(57) Subject Predicative Complement: *it* BE Sentence

Who ate the cake? – It was she/her.

In this context, the use of subject pronoun forms is often perceived as very formal, even in Standard English (Biber et al. 1999: 335-336; Huddleston and Pullum 2002: 459; Trudgill 1984b: 43). Therefore, it is not surprising that this feature is seldom discussed in the body of literature on varieties of English, and that in most of the variationist studies that actually mention this context, the same strong preference for object forms in *it* BE sentences is attested as in the Standard varieties (Harris 1993: 146-147; Maier 2013; Quinn 2009: 41-42; Shorrocks 1999: 78; Wagner 2004: 158). One of the few accounts that actually quantitatively examines the distribution of pronoun case forms in *it* BE sentences is the account given in
Biber et al. (1999: 335-336). Unfortunately, their analysis only distinguishes between different registers and persons and makes no distinction between American and British English (Biber et al. 1999: 15-28). Another noteworthy exception to this general scantiness of quantitative studies analysing the distribution of pronoun case forms in *it BE* sentences across varieties of English is Maier (2013). This study notes significant differences in the distribution of pronoun case forms in this context between American and British English, with the British data exhibiting a significantly higher share of subject pronoun forms in *it BE* sentences than the American data (cf. Maier 2013). The lack of quantitative analyses tackling this issue, particularly across varieties of English, is, however, partly due to the rare occurrence of *it BE* sentences in available corpora, since huge amounts of data are required to obtain a sufficient number of tokens to quantify this phenomenon (Quinn 2009: 41).

A comparable situation holds for the other context subsumed under the heading of subject predicative complements, i.e. the focal pronoun position in *it*-clefts. In Quinn’s (2009) survey, the few attested tokens do not allow for the establishment of a trend with regard to regional differences in the distribution of pronoun case forms in *it*-clefts (Quinn 2009: 42). The results presented by Biber et al. (1999), however, indicate marked differences between spoken and written language as well as with regard to the factors register and relative pronoun, which calls for further research on this issue (Biber et al. 1999: 335-336).

Maier (2013) also finds significant differences in the use of pronoun case forms in *it*-clefts between British and American English. This study shows that subject forms are significantly more likely to be used in the American *it*-cleft data than in the British *it*-cleft data, particularly in the spoken subsets of the examined corpora (Maier 2013). In general, however, the distribution of pronoun case forms in the focal position of *it*-clefts is by and large rather neglected in the body of literature discussing varieties of English. Responsible for this fact may be again the difficult data situation and the unclear status of the competing forms in terms of their standardness or non-standardness, i.e. whether the use of one form or the other is a decision of ‘right’ or ‘wrong’, or merely a matter of style and formality (Biber et al. 1999: 335; Huddleston and Pullum 2002: 459-463; Quirk et al. 1985: 336-339.).
More widely discussed than the pronoun choice in *it*-clefts, though still not extensively, is the choice of pronoun case forms in *than*-comparatives. Despite the formerly prescriptive predilection of the subject forms in this construction (Treble and Vallins 1961: 41), at least in informal English, *than* is overwhelmingly interpreted as a preposition followed by an object form rather than a conjunction introducing an elliptical clause followed by a subject pronoun form (e.g. Collins and Peters 2004: 605; Quirk et al. 1985: 335). The growing acceptance of *than* as a preposition is also attested by other authors awarding object forms after *than* the status of “marginally standard” or even “becoming standard” (Trudgill 1984b: 43), which very nicely illustrates the in-between status of these constructions exhibiting variability in pronoun case distribution between standard and non-standard features. From a variationist perspective, the distribution of pronoun case forms in *than*-comparatives is best discussed and documented for Australian English, which is one of the very few varieties for which this issue is discussed at all (e.g. Collins and Peters 2004: 605). Elicitation studies show a stronger preference for the object forms following *than* in Australian English than in comparable data from the UK. Furthermore, *than* followed by an object form is also more readily associated with spoken language than with formal writing in Australian English, a distinction that can also be observed in the Standard varieties of English (Biber et al. 1999: 336; Burridge 2004: 1118; Collins and Peters 2004: 605; Quirk et al. 1985: 337).

(59) Australian English (Collins and Peters 2004: 605)
   a. He had drunk far more than I and he was at least forty years older.
   b. The statue had become a boy some years older than me.

In addition to the accounts on Australian English, the strong preference of the object forms over the subject forms in *than*-comparatives has also been observed in traditional British English dialects. In the dialect of the Bolton area, for example, the preference for the object form in comparatives is so strong that the use of subject forms is ruled out altogether unless they are followed by a finite verb (Shorrocks 1999: 58 fn2). Moreover, the bias in favour of the object form in *than*-comparatives in Australian English has not been considered a distinctive local pattern but rather as a part of a global trend towards a telic development in terms of pronoun case distribution (Burridge 2004: 1118; Collins and Peters 2004: 605; Wales 1996: 107; Section 3.2). The clear preference for the object forms and the correlation of the pronoun form choice with the degree of formality in the elicitation studies could not,
however, be confirmed by a more recent corpus-based study also analysing pronoun-case forms in *than*-comparatives in the varieties of English in Australia and the UK. The results of this study show case variation in all written subsets of the analysed corpora. Moreover, the results indicate that the selection of case forms correlates neither with the syntactic properties nor with the formality of the texts in which they surface, which could, however, be partly attributed to the small size of the data sample (Quinn 2009: 42). In an earlier study on Standard English, a clear preference for the object forms was attested in conversations, whereas in the fiction subset, the distribution of subject and object forms was more balanced, though still with a majority of object forms (Biber et al. 1999: 337). Again we have to conclude that hardly any quantitative analyses, particularly in the field of variationist linguistics, tackle the issue of pronominal case marking in *than*-comparatives. This may again be attributed to the unclear status of the standardness or non-standardness of using a particular pronoun case form in this context and to the difficult data situation (Quinn 2009: 41-46; Trudgill 1984b: 43).

While the contexts discussed thus far have barely been considered in the variationist literature, coordinate noun phrases have had much more attention paid to them. Indeed, this is the most widely attested and discussed context exhibiting variability in the use of pronoun case forms across varieties of English, particularly with regard to the use of object pronoun forms in coordinated subjects. The use of *me* instead of *I* in coordinate subjects is, for example, the second most widely attested non-standard feature surveyed in Kortmann and Szmrecsanyi (2004: 1157). This feature is attested for all British, American, Caribbean, Australian, and Asian varieties as well as for seven of the nine African varieties examined in their survey. The only world region where it is not attested is the Pacific (Kortmann and Szmrecsanyi 2004: 1162-1183). Interestingly, although this feature is obviously attested for most varieties surveyed in Kortmann and Szmrecsanyi’s global synopsis (2004: 1142-1202), it is far less frequently discussed in the individual chapters of Kortmann et al. (2004). For instance, only one of the eight British and four of the nine American varieties attesting this feature in the global synopsis actually discuss it in their corresponding chapters, i.e. the chapters on the dialects in the North of England, Colloquial American English, Appalachian English, the rural and ethnic varieties of in the Southeast, and Urban African Vernacular English (cf. Beal 2004; Murray and Simon 2004; Montgomery 2004; Wolfram 2004a, 2004b). This discrepancy between the conscious documentation of this feature in only a few
varieties and the passive acknowledgement of this phenomenon in the vast majority of surveyed varieties further illustrates the in-between status of these contexts allowing for variable pronoun usage in the interface between the standardness and clear non-standardness in English. Those chapters in Kortmann et al. (2004) that actually discuss object forms in coordinated subjects are also noteworthy for another reason: They do not confine the discussion of this phenomenon exclusively to the first person singular object form use in coordinated subjects as Kortmann and Szmrecsanyi (2004) do. These chapters also either attest the use of object pronouns in coordinate subjects for other persons (cf. Beal 2004; Burridge 2004; Hundt, Hay and Gordon 2004; McCormick 2004: 999; Murray and Simon 2004; Montgomery 2004; Wolfram 2004a, 2004b) or at least do not preclude the use of object forms for other persons from this construction (Malcolm 2004: 673), as the following examples illustrate:

(60) North of England (Beal 2004: 117)
   a. Me and my mam and dad are going out for a meal.
   b. Him and me were there.

(61) Appalachian English (Montgomery 2004: 262)
   a. So me and four cousins began right then and there to lay our plans to go.
   b. Him and them dogs killed that bear.

(62) New Zealand English (Hundt, Hay and Gordon 2004: 586)
   a. [T]hat was a hard case eh cos me and my mate were at the hospital.
   b. [N]o wonder her and I look blurry eyed at school.

The impression that this phenomenon, i.e. the use of object pronouns in coordinated subjects, is exclusively discussed from a variationist perspective in Kortmann et al. (2004) should, however, be avoided. This feature is also discussed by many other authors and is attested for many varieties other than those explicitly mentioned above (e.g. Beal 1993: 206; Harris 1993: 146-147; Johnston 2007: 118; Shorrocks 1999: 78; Quinn 2009). The reason this discussion has so far focussed mainly on the phenomenon’s discussion in Kortmann et al. (2004) is that no other handbook offers a commensurate amount of cross-varietal information on this topic. However, by the discrepancy between the infrequent discussion of this feature in the individual chapters and the common attestation of it in the global

---

6 Why Kortmann and Szmrecsanyi confined their survey to me instead of I in coordinated subjects and did not include other personal pronoun forms is not made clear in their chapter. They simply state that they picked “[...] all the usual suspects known from survey articles on grammatical properties of (individual groups of) non-standard varieties of English, with a slight bias towards features observed in L1 varieties” (2004: 1145).
synopsis, this book also clearly illustrates the somehow neglected status of this context as well as all other environments exhibiting variation in terms of pronoun case distribution in the framework of variationist linguistics.

As is the case with most handbooks on varieties of English, the accounts mentioned above very often only discuss the presence or absence of a particular feature in a given variety. Although the importance and value of these handbook descriptions is not to be questioned, this often simplified binary distinction between the presence or absence of a feature unfortunately does not allow any conclusions with regard to a more fine-grained analysis of the distribution of the attested or unattested features (Siemund, Maier and Schweinberger 2009: 20). That is, it is not possible to quantify that feature α is more frequent in variety A than in variety B. However, as the majority of varietal differences are not categorical but quantitative in nature, only a fine-grained analysis, based on a sufficient amount of data, can identify the functional, semantic and pragmatic patterns underlying these distributional differences (Kortmann 2006: 603). This is particularly true for the construction at hand because, although it is widely distributed, its actual status in the interface between standard and non-standard is still open for debate. Nevertheless, there are hardly any studies trying to quantify the differences in the use of object forms in coordinated subjects across varieties of English, which is, however, partly due to the existing databases, as we will see below (cf. Chapter 7). The same holds for the constructions that are the main focus of this study, i.e. it-clefts and it BE sentences.

A noteworthy exception to this research gap is Quinn (2009). Her survey of pronoun case distribution covering Australian, New Zealand, American and British English identifies remarkable regional differences in terms of object pronoun use in coordinated subjects, especially in the spoken subsets of her data. Her British English data exhibit the highest share of object form use in coordinated subjects, followed by American and New Zealand English and complemented by Australian English with the lowest share of object forms in coordinated subjects. Although the share of first person singular object forms is higher than that of third person object forms in coordinated subjects, the observed differences between the individual varieties remain stable across the two observed pronouns (Quinn 2009: 35-41). Thus, in spite of the drawback of the low frequency of this construction in her databases, Quinn’s study offers valuable information on the distribution of pronoun case forms in coordinated subjects across a set of different varieties of English (Quinn 2009).
So far, we have only discussed the use of object forms in coordinated subjects and have not made mention of its ‘sibling’ construction, the use of subject pronoun forms in coordinated objects. This is simply because this phenomenon is not nearly as extensively discussed in most handbooks on varieties of English as the use of object forms in coordinated subjects (cf. Britain 2007a; Kortmann et al. 2004; Milroy and Milroy 1993; Trudgill 1984a; Trudgill and Hannah 2002). Only a few chapters in Kortmann et al. (2004) make explicit mention of the use of subject forms in coordinated objects. In one case, this is not even done in its own right but in the context of pronoun exchange (Wagner 2004). What is interesting, though, is that this account notes a “general tendency in colloquial English to use subject forms in non-subject functions, e.g. after prepositions (between you and I)” (Wagner 2004: 159). In another case, the use of subject forms in coordinated objects is described as a typical feature of the vernacular varieties of Australia and New Zealand (Burridge 2004: 1118).

(63)    Australian and New Zealand Vernacular English (Ibid.)

    a. He gave it to Fred and me.
    b. He gave it to Fred and I.

Still, it has to be remarked that this is not done in the individual chapters on Australian and New Zealand English but in the synopsis summarising the morphological and syntactic variation in this world region (cf. Burridge 2004). Hence, we observe a similar situation as with the use of object forms in coordinated subjects, namely that although the presence of a feature is widely acknowledged, it is not extensively discussed in the individual chapters on the given varieties (cf. Burridge 2004; Hundt, Hay and Gordon 2004; Collins and Peters 2004; Pawley 2004; Simpson 2004). The incidental mention of subject pronouns in coordinated objects quoted above (Wagner 2004: 159) is, however, rather exemplary for the discussion of subject pronouns in coordinated objects in the variationist literature. Like other studies (cf. Bauer 2002: 107; Collins 1989: 146, Quinn 2009: 36; Shorrocks 1999: 79), Wagner’s article mentions the notorious example of between you and I in particular and acknowledges its widespread use (Wagner 2004: 159). This is in line with an Australian English study reporting an 80 per cent rate of acceptability for this construction in formal contexts (Collins 1989: 146) which, in turn, has led to the assumption that the use of between you and I may eventually supersede the use of between you and me and hence may become the predominant variant in the future (Bauer 2002: 107). Furthermore, the increasing spread
and acceptance of *between you and I* is not only witnessed in the variationist literature but also, as we have seen before (cf. Section 2.3.1), in well-established reference grammars of English, which award this construction even the status of Standard English (Huddleston and Pullum 2002: 463). With the exception of the construction *between you and I*, however, not much is said about the use of subject pronoun forms in coordinated objects in the relevant handbooks and studies discussing varieties of English. Again, a noteworthy exception is Quinn (2009). Her results, though generalised over a wide range of different syntactic contexts and drawn from rather few instances, indicate not only that there seem to be remarkable regional differences in the use of subject forms in non-subject contexts but also that the use of subject forms in non-subject contexts seems to be more widespread in New Zealand and Australian English than in British and American English according to her databases (Quinn 2009: 36-41).

As introduced in Section 2.3.1, alternation in the use of pronoun case forms can even be observed in non-coordinated subjects of finite sentences. This is true for contexts in which the first person plural pronoun form is subject to an appositive modifier (Huddleston and Pullum 2002: 447, 459; Wales 1996: 100). This construction has also been simply called pronoun–noun phrase construction, or Pr–NP for short (e.g. Quinn 2005a, 2009), and examples for this construction from different varieties include the following:

(64)  Southeast England (Anderwald 2004: 178)
Us kids used to pinch the sweets like hell.

(65)  Australian Vernacular English (Pawley 2004: 635)
I remember us two boys was workin in the paddock.

Again, this feature is attested not only for regional varieties, such as Australian Vernacular English and the English in the Southeast of England (cf. Anderwald 2004; Pawley 2004), but also for colloquial varieties of Standard English (Huddleston and Pullum 2002; Pawley 2004; Wales 1996: 100). In addition, something similar has also been observed for the third person plural, namely the use of *them* in combination with a noun phrase. This latter feature is even more widespread across varieties of English, especially in L1 varieties, as it is attested in all American and Australian and in seven of the eight British varieties surveyed in Kortmann and Szmrecsanyi (2004: 1162, 1166, 1173).

(66)  Bolton Dialect (Shorrocks: 1999: 79)
Us Three went
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

(67) Irish English (Harris 1993: 145)
   Them two fellas was hit.

(68) Appalachian English (Montgomery 2004: 264)
   Them looks a whole lot steeper and taller than they did in my young days.

Although this usage of the third person plural form has been considered by some authors as to be on a par with the use of the first person plural form in Pronoun–Noun Phrase constructions (e.g. Quinn 2005a, 2009), this is not really the case because in these contexts, the third person plural forms are substitutes for the Standard English plural demonstrative. This assessment of the demonstrative status of these third person plural forms is also echoed in large parts of the variationist literature (e.g. Kortmann et al. 2004; Milroy and Milroy 1993; Britain 2007a). Furthermore, it has been ascertained that they is not usually found in this construction, so that there is only variation between them and the competing demonstrative pronoun (Anderwald 2004: 178; Quirk et al. 1985: 352). Furthermore, due to the case and number syncretism of the second person, variation in this context seems to be confined to the first person plural (Anderwald 2004: 178). Although the scope of variation in this Pr–NP construction is severely restricted and although the phenomenon is attested for many varieties of English including both Standard and non-standard varieties, there are hardly any quantitative studies examining this phenomenon (cf. Quinn 2009).

A further environment in which variable usage of pronoun case forms is observed is the subject position of gerund-participials. As noted previously, if the gerund participle is in adjunct position (cf. (48a) and (48b)), the case form of its subject may vary between subject and object form. If the gerund participle is, however, not an adjunct but filling an argument slot in a clause, the pronouns forms may alternate between object and possessive forms (cf. (49a) and (49b)). The latter variation pattern, i.e. object instead of possessive forms, is also attested for a number of varieties, such as Australian English and the dialect spoken in the Bolton area.

(69) Australian English (Burridge 2004: 1118)
   He was angry at me scoring a goal.

(70) Bolton Dialect (Shorrocks 1999: 79)
   It’s no use thee going down, is it?

In both varieties, a strong preference for the object forms is attested not only for this construction but also for many other contexts that permit a choice between different pronoun forms or case forms (Burridge 2004: 1118; Shorrocks 1999: 79). For Australian
English, a preference for the object form over the possessive form is attested even for the written varieties, which contrasts with the findings for British and American English (cf. Collins and Peters 2004: 605). Furthermore, with regard to the use of pronoun forms in participial gerunds, it should be stated that in Irish English, even the use of subject forms in these contexts is observed (Pietsch 2007).

(71) Irish English (Pietsch 2007: 175)
   a. If an Irishman goes to drive horses or Bollocks here after he coming out from home, he might as well go whistle a gig to a milestone.
   b. I recd. Your letter about the first of Jany. last and would have written an answer to you ere now were it for I being paying Michl. Moores passage as required by you.

Although a considerable amount of literature on Irish English grammar has accumulated in recent years, the use of subject forms in participials has not been widely discussed in the literature (cf. Pietsch 2007: 166). The same is also true for the competition of possessive and object forms in this context, for which hardly any cross-varietal quantitative studies exist (cf. Collins and Peters 2004: 605-606). This, unfortunate as it is, is in line with what we have observed thus far for the other contexts with variable pronoun form usage.

(72) Irish English (Harris 1993: 147)
   Who’s there? – Me.

(73) Bolton Dialect (Shorrocks 1999: 77)
   Who wants one? – Me.

Another variable context of pronoun form usage which is mentioned in the variationist literature but which is still far from being extensively discussed or even quantitatively assessed across different varieties is the case of so called “isolated” pronouns (Harris 1993: 147) or “stranded” pronouns (Quirk et al. 1985: 337). In Standard English, the object pronoun forms are regularly used and the same preference is attested for the few varieties in which this phenomenon is discussed (Biber et al. 1999: 339; Harris 1993: 147; Shorrocks 1999: 77).

(74) Left Dislocation in Scottish English (Johnston 2007: 118)
   Me, I had chased’t.

(75) Right Dislocation in English in the Channel Islands (Ramisch 2007: 179)
   There was a few [crystal sets]. My brother-in-law had one. But we didn’t have any, us.

(76) Right Dislocation in Chicano English (Bailey and Santa Ana 2004: 380)
   I don’t think I had a teacher that I didn’t really like him.
The next contexts to be discussed are also clause-peripheral usages of pronouns forms or, in other words, left and right dislocations of pronoun forms (cf. Biber et al. 1999: 339). As mentioned above, these constructions are considered to be non-standard by some authors and even the variability of left and right dislocations in terms of pronoun case distribution has been called into question, since the object form is considered to be the sole option in these contexts (Huddleston and Pullum 2002: 462). Hence, it is not surprising to also find a strong preference for object case forms in the variationist literature addressing these phenomena (e.g. Beal 1993: 210, 2004: 119; Johnston 2007: 118; Ramisch 2007: 178; Shorrocks 1999: 79). In contrast to other environments allowing for variability in terms of pronoun case usage, the phenomena of left and right dislocation are well attested across many varieties of English, including both traditional L1 varieties such as Scottish English as well as L2 varieties such as Chicano English (e.g. Bailey and Santa Ana 2004: 380; Miller 2004: 66-69; Montgomery 2004: 279; Schneider 2004: 1110-1111). That left and right dislocations receive a lot of attention compared to other variable contexts may be due to the fact that they are considered to be rather informal (Huddleston and Pullum 2002: 462). Hence, they may be perceived as more non-standard than other contexts, such as it BE sentences, it-clefts or than-comparatives. Nevertheless, variationist linguistics still lacks studies examining this phenomenon quantitatively across an array of different varieties.

In sum, it has to be concluded that not all the contexts which exhibit variation in the use of pronoun case forms are actually discussed in the relevant variationist literature (cf. Section 2.3.1). If variation in the distribution of pronoun case forms in a certain context is discussed in the description of a certain variety, it most often happens rather superficially as a singular phenomenon along with a series of other particularities of the variety at hand. Hence, there are hardly any studies systematically comparing and analysing different contexts exhibiting pronoun case variation across varieties of English, neither quantitatively nor qualitatively. Thus, the distribution of pronoun case forms in these variable contexts has, from a variationist perspective, been rather neglected. This is because of the difficult data situation, since many of the contexts in which variable pronoun form use is attested occur extremely infrequently (cf. Quinn 2009). In addition, these contexts may also have been neglected or overlooked due their intermediate or ‘in-between’ status in the interface between standardness and non-standardness. The ‘intermediateness’ of these contexts exhibiting variability in the use of pronoun case forms has become obvious especially in the
inconsistent classification of subject pronoun usage in coordinated objects – in some instances being labelled as standard and in others as non-standard – and the classification of the use of object forms as “becoming, standard or marginally standard” (Huddleston and Pullum 2002: 463; Trudgill 1984b: 43), which may finally result in the assumption that these contexts are simply too standard-like to discuss them in the context of the non-standard features of a given variety.

2.4 The Distribution of Pronoun Case Forms in English: Interim Summary

To conclude, this chapter has shown that due to the dramatic changes English underwent from Old English to Middle English, virtually all case inflections and case forms were lost. The former functions of case have been taken over by other means such as prepositions and word order and only very few exponents of the once complex and productive morphological case system have been retained in Present-Day English. The clearest remnants of the former English case system are preserved in the paradigm of personal pronouns, where we can distinguish between object and subject pronoun case forms.

According to the conventions of Standard English grammar, subject pronouns are obligatorily used in non-inverted subjects of finite sentences, inverted subjects in direct questions, inverted subjects in questions tags and subjects of finite subordinate clauses, while object pronoun forms are obligatorily used as direct and indirect objects of verbs, objects of prepositions and subjects of non-finite infinitival clauses introduced by for.

While this distribution seems to be quite simple in theory, the cross-varietal surveys in Sections 2.2.1 and 2.2.2 demonstrate that reality turns out to be much more complicated. In particular in 2.2.1, it has been shown that some varieties of English have retained earlier forms of the pronominal case paradigms, other varieties have extended the functional range of some pronoun forms to cover those of others and that other varieties allow for varying degrees of variation which may range even to more or less idiosyncratic usage patterns of pronoun case forms. The survey has also demonstrated that pidgin and creoles are marked not only by simplification in their pronominal paradigms but also by a high degree of innovation, such as the explicit marking of emphatic and non-emphatic pronoun case forms.
The survey in Section 2.2.3 has clearly pointed out that the encoding of an emphatic–unemphatic, focussed–unfocussed or stressed–unstressed contrast is a desideratum of many varieties of English, which can be consistently observed across many different and unrelated varieties. That this observation is of central importance for the distribution of pronoun case forms in subject predicative complements, i.e. *it BE* sentences and *it*-clefs, will be more thoroughly discussed in Chapter 4.

Furthermore, this chapter has also shown in Section 2.3 that except for subject predicative complements, English possesses many other contexts which permit a choice between subject and object pronoun case forms. These contexts allowing for variability in the use of pronoun case forms have in common that they receive comparatively little attention in the variationist body of literature (cf. Section 2.3.2). As a possible reason, this chapter has identified the insecure status of these phenomena in the interface of clear standardness and non-standardness, since some of them may be considered as being simply too normal to be intriguing for variationist research. Moreover, another feature that is shared by the bulk of contexts allowing for variability is the fact that there have hardly been any quantitative studies, let alone multivariate ones, that try to examine the distribution of pronoun case forms systematically in these contexts in one or more varieties of English.

Although there are few quantitative studies analysing variable contexts in general and subject predicative complements in particular, which is partly due to the difficult data situation (cf. Chapter 7), the surveys in Sections 2.2 and 2.3 have identified several possible factors which may influence the distribution of pronoun case forms in these contexts, such as the degree of formality, the person, the difference between *it*-clefs and *it BE* sentences or the difference between emphatic and unemphatic contexts. These possible factors will be thoroughly discussed, particularly with regard to subject predicative complements in Chapter 6.
3 Previous Approaches to the Distribution of Pronoun Case Forms in English

So far, Jespersen’s (1933: 132) evaluation with regard to the difficulty of describing and explaining the distribution of pronoun forms in Present-Day English has been fully borne out. The discussion in Section 2.1 has shown that there is not even a general consensus on the actual number of case distinctions in Present-Day English. Moreover, the survey of the distribution of pronoun case forms across varieties of English in Section 2.2 showed that the respective inventories, distributions and usage of pronoun case forms may vary tremendously from variety to variety. In Sections 2.3.1 and 2.3.2, the situation became even more complex since the surveys in those sections demonstrated that Present-Day English exhibits a considerable number of contexts in which the distribution of pronoun forms may vary – both in Standard English and across varieties of English.

In light of this intricate situation, it should not come as a surprise that linguistic theory has put forward different approaches to account for the distribution of pronoun case forms in Present-Day English and the variation observed therein. In this chapter, a survey of the most important of these approaches is provided. Section 3.1 begins with a very brief survey and discussion of two influential formal approaches to case and pronoun case distribution in English. Then, Section 3.2 discusses a very popular approach that mainly focuses on the pronoun’s position relative to the finite verb in a clause in order to account for the distribution of pronoun case forms. Furthermore, the account dealt with in Section 3.3 has also enjoyed considerable popularity in linguistic theorising. However, its focus is not on the position of a pronoun but on its assumed membership in a “weak” or “strong” pronoun class. Then, Section 3.4 provides a brief survey of some functional accounts and of the factors and variables that have been identified by these accounts as possibly influencing the distribution of pronoun case forms in contexts allowing for variability in general and subject predicative complements in particular. Finally, Section 3.5 summarises the major findings.
3.1 Formal Approaches to the Distribution of Pronoun Case Forms

Although the outlook of this study is functional and hence this study cannot do justice to the full scope of work on case that has been conducted in the formal theoretical framework, it still seems beneficial to briefly introduce two current standard models of case assignment and pronoun case distribution (Haegeman 1993; Haegeman and Guéron 1999; Radford 2004, 2009). It is helpful to have a brief glance at prototypical formal models of case assignment because some proponents of other accounts, which will be introduced in the subsequent sections, distance themselves from these standard formal models (cf. Sections 3.2–3.4). Thus, the reasoning underlying their models may also become clearer if the models from which they distinguish themselves are briefly outlined. However, due to the general perspective of this study and in view of the vast body of literature on case that has accumulated in the formal framework, this introduction has to remain cursory. For a more detailed discussion of particular formal models, the interested reader is referred to Quinn (2005a: 26-64), who offers a detailed survey, discussion and evaluation of some current formal models of case, or to single studies of case carried out in the formal linguistic framework (e.g. Burzio 2000; Chomsky 1993, 1995; Kiparsky 1997; Wunderlich 1997).

To begin with, it should be noted that much of the formal syntactic literature still postulates an intact case system for Present-Day English (e.g. Chomsky 1993; van Gelderen 2010: 83-85; Haegeman and Guéron 1999: 127-145; Quinn 2005a: 26-64; Radford 2009: 119-125). The clearest indications of this assumed case system in Present-Day English are personal pronouns (cf. Section 2.1). They differ morphologically from both nouns and other pronouns in that they have, at least partially, different nominative, i.e. subject, accusative, i.e. object, and genitive, i.e. possessive, forms whereas nouns only have a common case form for the nominative and the accusative cases and a distinct genitive form marked by ‘s (Haegeman and Guéron 1999: 129; Radford 2004: 38). Although overt case marking is severely restricted, mainstream formal linguistic theory proposes that not only personal pronouns but indeed all noun phrases must be marked for case. For the latter, however, case marking may be abstract in contrast to overt, i.e. morphologically not visible (cf. Haegeman and Guéron 1999: 133).
With regard to the distribution of subject and object pronoun forms, or nominative and accusative pronoun forms as they are called in these approaches, the following is proposed by standard generative approaches, such as those put forward by Haegeman (1993: 140-182) and Haegeman and Guéron (1999: 128–133):

(77) a. He/*him likes her/*she.  
    b. He expects them to invite his girlfriend.  
    c. For them to invite not his girlfriend was totally out of place.  
    d. I consider him a total moron.  
    e. They listened to us/*we.

As can be inferred from the examples in (77a)–(77e), the nominative or subject pronoun forms are used as subjects of finite clauses. As illustrated in Figure 1, they interact with the finite inflection, the so-called I(NFL) which is marked for tense and agreement. Depending on the respective formal account, these agreement features are responsible for “assigning” or “licensing” the case forms to the respective noun phrase or pronoun (Haegeman and Guéron 1999: 129).

![Figure 1: The Distribution of Pronoun Case Forms in Government and Binding Theory (adapted from Haegeman and Guéron 1999: 129)](image)

While the assignment or licensing of nominative case by the agreement relation between the subject of a finite clause and the agreement features in I(NFL) is relatively straightforward, the distribution of accusative case is more complicated since it is realised not only on the object of verbs (77a) and the complements of a prepositions (77e), but also
on the subjects of non-finite clauses (77c) and the subjects of small clauses (77d) (cf. Haegeman and Guéron 1999: 128-129).

With regard to objects of transitive verbs, it is assumed that the verb “case-marks” or licenses the case of its object noun phrases (Haegeman and Guéron 1999: 130). For English, this means that the transitive verb assigns or licenses accusative case to its object or objects. A similar relationship holds for prepositional objects. In the case of prepositional objects, the preposition governs its complement and case-marks or licenses the case of its object, which is, in English, also the accusative case. In general, accusative case is associated with verbs and prepositions, or to use the terminology of this approach with V(erb) and P(reposition), which is why V and P are called “case assigners” (Haegeman and Guéron 1999: 129-131).

For the accusative forms in (77b)–(77d), however, the explanation as to why this case form is used is more complicated simply because the accusative forms of these sentences are also technically subject noun phrases of non-finite sentences. In (77b), the accusative form *them* is the subject of a non-finite clause, which is itself the complement of the transitive verb *expect*. In (77c), the accusative form *them* is both the subject of a non-finite clause and the complement of the preposition *for*. Finally in (77d), the accusative pronoun *him* functions as the subject of a small clause, which is itself also the complement of the verb *consider*. Thus, in each of these sentences, the accusative pronoun is on the one hand the subject of a non-finite clause and on the other hand the complement of a verb or preposition or of a case-assigner V or P. These case-assigners “govern” the pronoun forms and they are responsible for the assignment or licensing of accusative case to the pronoun forms in (77b)–(77d) (Haegeman and Guéron 1999: 130-131).

Thus, it can be noted that this standard generative account proposes that nominative case is assigned by an agreement relation between the subject of the finite clause and the agreement features in I(NFL), whereas accusative case is assigned by the government relation between either a verb or a preposition and its complement (Haegeman 1993: 162; Haegeman and Guéron 1999: 145).

The second formal account to be briefly introduced here is the Minimalist one, the goal of which is to use as little theoretical machinery as possible (Radford 2009: 124). In the Minimalist approach to case distribution, two concepts are particularly important: the relation “c-command” and the “Earliness Principle” (Radford 2009: 120). With respect to the former, it can be stated that a constituent X c-commands a constituent Y if X is no lower than
Y in the structure of the syntactic tree. This means that X is either higher up in the structure or the two are at the same level (Radford 2004: 327). The second important concept is the Earliness Principle which determines the implementation of linguistic applications and according to which operations apply as early as possible in a derivation, i.e. in the formation of a word, phrase or other linguistic structure (Radford 2009: 120).

(78) a. (ø)We see them.
   b. She went with him to the cinema.
   c. She would very much prefer for/me to see her.
   d. He may think that/that she is beautiful.

In sentences such as (78c) and (78d), the complementisers for and that can have either an overt spellout or a zero spellout, the latter of which are represented as for and that in (78c) and (78d). With regard to the distribution of accusative case, the Minimalist approach assumes that accusative case is assigned to a noun phrase or a pronoun if it is c-commanded by a transitive head. Traditionally, items which assign accusative case are termed transitive, like the transitive verb in (78a), the transitive preposition in (78b) or the transitive complementiser in (78c) (Radford 2009: 120). As a consequence of the Earliness Principle, case is assigned to a noun phrase by the closest case-assigner, which in (78a) is see, in (78b) with, and in (78c) for for (Radford 2009: 120-124).

However, in order to arrive at a uniform account of case assignment in general and pronoun case distribution in particular, the Minimalist approach proposes not only that accusative case is assigned to a noun phrase by means of c-commanding but also that a similar process is responsible for the assignment of nominative case. Hence, a noun or pronoun is assigned nominative case if it is c-commanded by an intransitive finite complementiser such as that/that in (78d) or the null complementiser ø in (78a). This means that according to this account all complete clauses and sentences are governed by a complementiser. If a clause, however, contains no overt complementiser such as (78a), the minimalist approach assumes that this clause is still introduced by a null complementiser, i.e., by one that is not morphologically encoded (Radford 2004: 112, 2009: 122). Therefore, for the sentence (78d) for example, minimalism assumes the following underlying structure (cf. Radford 2009: 122):
As we can see from Figure 2, the overall clause is introduced by a null finite complementiser and it is this complementiser that determines the case of the third person singular masculine pronoun, which is in the present case nominative. In analogy, it can be stated that the finite complementiser *that* assigns nominative case to the third person singular feminine pronoun (cf. Figure 2). It is important to note that in English, all finite complementisers assigning nominative case are intransitive, whereas all finite complementisers assigning accusative case are transitive, as has already been noted above (Radford 2009: 122). Thus, we can conclude that according to the Minimalist approach, accusative case is assigned to a noun phrase if it is c-commanded by a transitive head and that nominative case is assigned to a noun phrase if it is c-commanded by an intransitive finite complementiser (Radford 2009: 124).

As noted above, the Minimalist approach tries to use only the theoretical machinery that is absolutely necessary, since it aims to develop a constrained version of Universal Grammar utilising only those concepts that are conceptually indispensable. This aim leads to the different explanation when compared to the standard generative account outlined above. Furthermore, the reduction of the theoretical apparatus also results in the fact that traditional grammatical concepts are no longer maintained. Whereas the former account still uses the concepts of subject and object (e.g. Haegeman and Guéron 1999: 129), these grammatical relations are not employed in the Minimalist approach. This is due to the fact...
that the traditional equation of accusative case with object status is – among others – inadequate because accusative case is also used for case marking on non-finite subjects as the examples in (77c) and (78c) illustrate. Thus, the concept of c-command is used in this linguistic approach, but neither the concept of “subjecthood” nor of “objecthood” (Radford 2009: 125).

Despite the differences in terminology, modes of explanation and concepts, these two formal models of case assignment share common ground in very important aspects. First of all, both approaches assume that not only personal pronouns but rather all noun phrases in English are marked for case either overtly or abstractly due to the universal nature of this phenomenon (e.g. Haegeman and Guéron 1999: 144; Radford 2009: 119). In both approaches, this marking is necessitated by structural case mechanisms (e.g. Haegeman and Guéron 1999: 144; Radford 2009: 119-125). Furthermore, what both accounts also have in common and what is of immediate relevance for the present study is the fact that none of them addresses the contexts exhibiting variation in pronoun case usage. Thus, neither of these approaches can account for the variation observed in the distribution of pronoun case forms in subject predicative complements, nor in any other contexts exhibiting variability in the use of pronoun case forms introduced and discussed in Sections 2.3.1 and 2.3.2.

This discrepancy between theory and observation has not gone unnoticed by the many linguists working in formal frameworks. As a consequence, some formal linguists have tried to combine their respective formal approach with other linguistic models or explanations in order to be able to account for the variation in the use of pronoun case forms attested in many contexts (cf. Section 2.3). Examples of this strategy are, on the one hand, Parker, Reily and Meyer (1988), who try to account for variation in the use of pronoun case forms in coordinated noun phrases by combining the principles of Government and Binding Theory (Chomsky 1981) with those of a Functional approach to grammar (Kuno 1987). On the other hand, Parrott (2007) tries to explain variation in the distribution of pronoun case forms mainly with the help of a Minimalist approach combined with a variationist approach in the Labovian tradition. A final example of an account trying to reconcile formal linguistic theory with the variation attested in authentic data is offered by Sobin (1997). This approach postulates the existence of “grammatical viruses”, which are supposed to be grammar-external and which are deemed responsible for the variation observed in the distribution of pronoun case forms in many of the contexts discussed in Section 2.3 (Sobin 1997). While
these latter, mainly combinatorial, approaches represent one possible strategy pursued by some formal linguists to reconcile the observed variation with the linguistic framework they are working in, another strategy simply consists of advocating for another approach than a purely case-based one to account for the distribution of pronoun case forms in Present-Day English.

3.2 ‘Positional’ Approaches to the Distribution of Pronoun Case Forms

Due to the fact that standard formal approaches discussing the distribution of pronoun case forms fail to account for the variation observed in many contexts, such as those that are the focus of this study (cf. Section 1.2), the assumption that the distribution of pronominal case forms in English is subject to underlying case mechanisms has met with a lot of opposition, even within the formalist framework (e.g. Emonds 1976: 197, 1985: 220-224, 237-241; Klima 1964: 1-8; Parrott 2007: 254-276). Emonds, for example, argues that “the remnants of case found on English pronouns should not be generated by the mechanisms for morphological case” (1985: 220). The reason for this assumption is that “a true case-marking rule has to be ‘healthy’, or it ceases to be a case-marking rule at all” (Emonds 1985: 220). “Healthy” means that a case-marking rule has to be morphologically transparent, i.e. productive. In Present-Day English, however, this is not the case, since morphologically distinct case forms have only been retained for the personal pronouns I–me, she–her, he–him, we–us and they–them (cf. section 2.1). Thus, the few case distinctions maintained in Present-Day English are therefore not sufficient to postulate a case system. The same is also true, for example, for the few case distinctions still maintained in Spanish (Emonds 1985: 222).

Alternatively, the position of the pronoun form in a clause – particularly in relation to the finite verb – has been considered to be an extremely important, if not the decisive, factor in determining the distribution of the different pronoun forms (e.g. Burridge 2004: 1118; Emonds 1976, 1985; Hudson 1995; Jespersen 1933: 136; Jespersen and Haislund 1949: 241; Klima 1964: 17). A very good example to illustrate this ‘positional’ approach⁷ to the

---

⁷ Henceforth, the terms ‘positional account(s)’ or ‘positional approach(es)’ will be used to refer to the accounts discussed in this section.
distribution of pronoun case forms in English is Quirk et al. (1985: 337-339). On the one hand, Quirk et al. (1985) still employ a functional or structural explanation for the distribution of pronoun case forms in Present-Day English. Thus, “subjective” forms are used as subjects, subject complements and subjects of ellipted clauses, whereas “objective” forms are used as objects of verbs and prepositions. However, Quirk et al. restrict this structural or functional explanation explicitly to the rather formal varieties of English (Quirk et al. 1985: 336-337). For the vernacular or less formal varieties of English, on the other hand, they employ a clearly positional account. In this positional account, the clause is divided into “subject territory”, i.e. the preverbal subject position immediately preceding the finite verb, in which subject pronoun forms are used, and “object territory”, including all noun phrase positions apart from that directly preceding the finite verb, where object forms are used (Quirk et al. 1985: 337). Thus, these informal varieties have not retained the traditional case distinctions but instead use the object pronoun forms as unmarked default forms “in the absence of positive reasons for using the subjective pronoun variant” (Quirk et al. 1985: 338). Furthermore, the notion of object territory is even extended to cover coordinated noun phrases, since it is claimed that the coordination of a noun phrase distances the subject from the subsequent finite verb to a certain degree. By doing so, this approach can explain the use of object pronoun forms in many contexts exhibiting variability (cf. Section 2.3), including, for example, coordinated noun phrases in subject position (80a), it BE sentences (80b), it-clefts (80c), independent pronouns (80d) and than-comparatives (80e) (Quirk et al. 1985: 337-338).

79) Pronoun Forms in “Subject Territory”
   a. She is a nice lady.
   b. They have just bought a new house.

80) Pronoun Forms in “Object Territory”
   a. Me and Jodie are a couple.
   b. It is him.
   c. It is her who broke my heart.
   d. Who is it? – Me.
   e. Jane is more intelligent than him.

Interestingly, although this approach has found many advocates even recently (e.g. Burridge 2004; Quirk et al. 1985), it has a very long history. For example, nearly eighty years ago, Jespersen (1933) already noted that
“ [...] the natural tendency in English has been towards a state in which the nominative of pronouns is used only where it is clearly the subject, and where this is shown by close proximity (generally position immediately before) a verb, while the objective is used everywhere else “(1933: 136).

However, the earliest proponent of such a positional approach is very likely Cooper (1685). As early as the late 17th century, he stated in his Grammatica linguae Anglica: “l, thou, he, she, we, ye, they, verbis anteponuntur, me, thee, him, her, us, you, them, postponuntur verbis & præpositionibus” (Cooper 1685 [1968]: 121).

As outlined above, this positional account can explain the use of object pronouns in many contexts exhibiting variability in the use of pronoun form usage, including coordinated noun phrases in subject position (80a), predicative complements in it BE sentences (80b), independent pronouns (80c) and focal pronouns in it-clefts (80d). Quite naturally, however, the question which also arises is how the proponents of such an approach can account for the opposite scenario, i.e. the use of subject pronoun forms in object territory. In general, the most frequently mentioned reason given for this kind of variation in the use of pronoun case forms is circumscribed with the notions of style and formality (e.g. Emonds 1985: 238, Klima 1964: 1-5; Quirk et al. 1985: 337). Emonds (1985), for example, attributes the use of subject pronouns in object territory to a so-called “prestige subject pronoun usage in Modern English” which is analysed in his account as

“ [...] a mixture of some minor local rules (e.g. “use I after a coordinate conjunction”), avoidance of a range of constructions, over-corrections, formulas (It is I vs. ?Mary couldn’t be he in the play), and constant semi-conscious sociologically determined self-correction” (1985: 238).

As indicated above, Quirk et al. (1985) distinguish between formal and informal varieties of English. In the formal ones, subject pronouns may not only be used in the position and function of subjects, but also as subject complements and stranded subjects. Hence, in formal varieties and registers, the distribution of pronoun forms is still determined – at least partly – by functional considerations, which can explain why subject forms may still be used in object territory. In informal varieties, however, the distinction between subject and object territory is crucial and former case distinctions are not upheld any longer (Quirk

---

8 “I, thou, he, she, we, ye, they, are placed before the verb, me, thee, him, her, us, you, them, are placed after verbs and prepositions [translation, mine]” (Cooper 1685 [1968]: 121).
et al. 1985: 336-338). Furthermore, as was the case with Emonds’ account, Quirk et al. also acknowledge a strong prescriptive inclination towards the use of subject forms to which phenomena such as hypercorrect uses of coordinated noun phrases with subject forms in object territory can be attributed (1985: 338). Instances of such hypercorrections include the notorious *between you and I* coordination and coordinated noun phrases in object position such as *It would be great for he and I to work together*. Another factor that is deemed to promote the use of subject pronoun forms in coordinated objects appears to be politeness, since the sequence *NP and I* seems to be perceived as more polite than the use of *NP and me* (Quirk et al. 1985: 338).

A final remark concerning the positional approach to pronoun case distribution pertains to its range of applications, since this aspect is also of immediate relevance to the present study. As noted above, Quirk et al. apply the positional approach to the informal or vernacular varieties of English (1985: 337-338). Another positional account, however, goes one step further by postulating a “general trend in English towards case selection [being] dictated by position rather than function – the nominative is largely confined to clause-initial preverbal position; accusative appears elsewhere” (Burridge 2004: 1118). Thus, this latter approach seems not only to assume that the factor position is the main or even sole determinant of the distribution of pronoun case forms in English, but also that this is generally true. Hence, this positional explanation is deemed applicable to all varieties of English, not only to both formal and informal ones, but also to geographic or regional varieties of English. Whether such an assumed homogenous situation can be observed across varieties of English will be examined in chapters to come – at least for the distribution of pronoun case forms in subject predicative complements.
3.3 The ‘Weak vs. Strong’ Pronoun Distinction

The approach discussed in this section trying to account for pronoun case distribution in English can also look back on a long scholarly tradition. Although it makes predictions that are very similar to those of the positional approach with regard to the distribution of pronoun forms (cf. Section 3.2), its focus is more on the distinction between different classes of pronoun forms than on delimiting specific areas or territories in which certain pronoun forms are supposed to occur:

“The real difference between ‘I’ and ‘me’ is that ‘I’ is an inseparable prefix used to form finite verbs, while ‘me’ is an independent or absolute pronoun, which can be used without a verb to follow. These distinctions are carried out in vulgar English as strictly as in French, where the distinction between ‘je’ and the absolute ‘moi’ is rigidly enforced” (Sweet 1875: 495).

As this quotation of Sweet (1875) indicates, proponents of this account see striking similarities between the distributions of pronoun case forms in French and English. In order to better grasp these analogies, a few brief remarks on the distribution of pronoun forms in French will clarify this line of thinking. In French, there is a fundamental distinction between conjunct and disjunct pronoun forms, which is particularly important for subject pronouns. The French conjunct subject pronouns – particularly the distinct forms je, tu, il and ils – can only be used when they appear in close contextual adjacency to the finite verb (cf. (81)). Their disjunct counterparts moi, toi, lui and eux, however, can occur in a much wider range of contexts. These contexts include – among others – the focus positions of it-clefts (82a), left dislocations (82b), independent pronouns (82c), comparatives (82d), it BE sentences (82e) and coordinated subjects (82f) (e.g. L’Hullier 1999: 488-509; Menuet 1997: 182-193; Weinrich 1982: 81-89). These environments correspond very well to those in which variation has amply been attested for English (cf. Section 2.3), as not only the mere contexts but also the English translations of the examples in (82) illustrate.

81) Weak Pronoun Contexts in French
a. Je ferme la porte.
   ‘I close the door.’
Hence, the proponents of the present approach tend to distinguish “weak” from “strong” or “clitic” from “non-clitic” pronouns in English in order to account for the distribution of pronoun case forms in analogy to the differentiation of those classes in Romance languages such as French (e.g. Harris 1981; Kjellmer 1986; Quinn 2005a, 2005b; Shorrocks 1992). Therefore, this approach is referred to as the ‘weak vs. strong’ pronoun approach in the present study.

As a consequence of the supposed analogy between English and French, this approach assumes for the less formal or vernacular varieties of English that

“[...] the distinction between I and me is precisely that between ‘conjunctive’ and ‘disjunctive’, in the sense of the term used by grammarians of French, ‘clitic’ and ‘non-clitic’ to use more widely current terms” (Harris 1981: 18-19).

While the former account is mainly restricted in its scope to the first person singular (cf. Harris 1981), other authors extend this weak vs. strong distinction to the rest of the pronominal paradigm, at the same time confining its range of application to certain varieties of English. With regard to the Farnworth dialect spoken in the Northwest of England and “to some degree in popular English, too” Shorrocks (1992) states that

“[...] the subjective or nominative forms of the personal pronouns are not “strong” enough to support a modifier; they cannot be coordinated nor can they carry the emphasis associated with the imperative, vocative, or unattached status. They seem to be on their way to becoming clitics” (Shorrocks 1992: 435).
This last issue raised by Shorrocks (1992), namely the dynamic nature of the transition of subject pronouns towards a set of clitic pronouns, is a characteristic of this weak vs. strong pronoun approach that is also reflected in other accounts (e.g. Harris 1981; Kjellmer 1986) and will be further addressed below. With regard to the expected or predicted distribution of pronoun forms, this approach distinguishing between “weak” and “strong” pronoun forms is indeed very similar to the positional account: The “weak” English subject pronoun forms are more or less restricted to the noun phrase positions adjacent to the finite verb, which corresponds to the subject territory of the positional approach, whereas the English “strong” object pronouns are expected to occur everywhere else, which also matches the object territory introduced in the former account (cf. Section 3.2).

In view of these similarities, it is not surprising to find accounts trying to reconcile or even integrate these two concepts or dichotomies, i.e. strong vs. weak and subject territory vs. object territory into one approach. Quinn (2005a, 2005b) refers to both dichotomies in her account, in which she tries to explain the distribution and variation of pronoun case forms in Present-Day English. In fact, due to her general formal linguistic outlook, her approach combines elements of all approaches introduced and discussed thus far (Quinn 2005a, 2005b, 2008).

According to her constraint-based approach, the “morphosyntactic status of the pronoun”, i.e. the distinction between weak and strong pronouns has substantial repercussions on the distribution of pronoun case forms in Present-Day English (Quinn 2005a: 380-383, 2005b: 4-16). In particular, it is proposed that weak pronouns invariably occur in their subject case form when they are used as uncoordinated and unmodified subjects of a finite clause, whereas weak pronouns always appear in their object form when they are used as unmodified and uncoordinated objects of finite clauses (e.g. Quinn 2005b: 4-11, 2008: 5). Therefore, for Quinn’s class of weak pronouns, the position of the pronoun plays a crucial role in determining the form, which is called “Positional Case checking” in her approach (e.g. Quinn 2005b: 5). In contrast, strong pronouns are to some extent immune to Positional Case influences which is why pronoun case variation is basically restricted to strong pronoun contexts (Quinn 2005b: 6, 2008: 5). Quinn’s strong pronoun contexts include coordinated noun phrases, it-clefs, than-comparatives, Pr–NP constructions, it BE sentences, left and right dislocated pronouns, independent pronouns and modified pronouns (e.g. Quinn 2005a: 2, 201-309). Again, these are the contexts that have
traditionally been associated with variation in terms of pronoun case form usage (cf. Section 2.3).

Quinn’s (e.g. 2005a, 2005b) account, however, does not simply equate weak pronouns with subject pronouns and strong pronouns are not considered to be the equivalent of object pronouns, as her proposed distribution of weak pronoun forms outlined above clearly indicates. Instead, other differences in the morphosyntactic status of the pronouns are made responsible for the distinction between strong and weak pronouns in English. In particular, it is argued that weak pronouns are base-generated in D and behave like heads and maximal projections of the DP. Strong pronouns, however, are base-generated in N and raise to D before Spell-Out, which means that a DP containing a strong pronoun is internally complex and can also contain modifiers (e.g. Quinn 2005a: 76, 2005b: 5-16, 2008: 5-6). Therefore, strong pronouns are in a way separated from the agreement-related functional heads which are associated in this account with the “Positional Case checking” constraint determining the distribution of weak pronoun forms, which leads to the partial immunity against “Positional Case checking” previously mentioned (e.g. Quinn 2005a: 76-77, 2005b: 14-15).

That some more or less abstract notion of distance also seems to play a role in this account is interesting. In this approach, it seems to be the distance between pronoun form and agreement-related functional head in a syntactic tree that is supposed to influence the choice or distribution of pronoun forms, whereas in the previous Section 3.2, the distance or separation of the pronoun from the finite verb is deemed to influence the distribution of pronoun forms in informal English (Quirk et al. 1985: 338). However, despite the fact that strong pronouns are partially immune to the “Positional Case checking” constraint (e.g. Quinn 2005b: 6), they are subject to what Quinn calls “Relative Positional coding” constraint, i.e. the relation between the case form of a pronoun and its position within a construction, which means that at least for coordinated strong pronouns, the position of occurrence exerts some influence on their surface form (e.g. Quinn 2005b: 9-10).

As a result of the interaction of several constraints and pronoun classes, this account employs basically five different sets of pronouns, namely two different sets for weak pronouns and three different sets for strong pronouns, as is illustrated in Figure 3:
With regard to the weak pronouns, “nominative case forms” are used consistently as subjects in weak pronoun contexts, whereas “objective case forms” are used as objects of verbs and prepositions in weak pronoun contexts (e.g. Quinn 2005b: 5, 15). With respect to strong pronouns occurring in coordinated noun phrases, the forms me, he, she, we, they and who are favoured in initial position, whereas I, him, her, us, them and whom are favoured in final position. Finally, in uncoordinated strong pronoun contexts, there is supposed to be a trend towards the use of the invariant forms me, him, her, us, them, who and whoever (e.g. Quinn 2005a: 381), suggesting that the “nominative forms of strong pronouns (especially strong non-1sg pronouns) are slowly being eliminated” (Quinn 2005b: 2).

The emphasis of the dynamic trend towards an invariant inventory of strong pronoun forms is one of the major characteristics of the weak vs. strong pronoun approach since it is not only proposed by Quinn (e.g. 2005a: 381a, 2005b: 2), but also by other advocates of this
approach (e.g. Harris 1981: 18; Shorrocks 1992: 435, 441). The reason that the proponents of this account stress the transitional character of the Present-Day distribution of pronoun case is the basic teleological assumption underlying this approach. It is uncontested by proponents of the weak vs. strong pronoun account that the distribution of case forms in Old English was determined by the syntactic function of the nouns and pronouns in a clause (e.g. Denison 1993: 16-22; Quinn 2005b: 3). It is also common knowledge that the transition from Old English to Modern English is marked by a considerable decline in the inflectional morphology in general and case morphology in particular (e.g. Denison 1993: 20-23). The advocates of the weak vs. strong pronoun approach, however, consider this process of attrition as ongoing since the putative endpoint of this transition process, a system of pronoun distribution like that of French, has not been reached yet (e.g. Harris 1981: 20; Kjellmer 1986: 448-449; Quinn 2005a: 381-382). According to the proponents of the weak vs. strong pronoun account, this assumed dynamic transition from a formerly case-governed pronoun system to a system of clitic and non-clitic pronouns is the most important reason why there is still variation observable in the use of pronoun case forms, particularly in strong pronoun contexts (e.g. Harris 1981: 18-20; Quinn 2005a: 382-383).

In addition to the still transitional character of this distribution of pronoun case forms in Present-Day English, another important factor responsible for the currently observed variation is the assumed prescriptive inclination towards the subject pronoun forms in some strong pronoun contexts. Hence, this bias in favour of the subject pronoun forms is supposed to hinder the ultimate transition from the status quo to a state similar or even comparable to that in French (e.g. Harris 1981: 18-20; Quinn 2005a 383). With regard to coordinated subjects, Harris (1981) states that the transition towards the use of strong or object pronoun forms “is however being constantly impeded by the pressure for ‘subjects’ to be ‘nominative’, this half-remembered pattern leading to hypercorrection and inconsistency” (1981: 19). In addition, for it-clefts he notes that “[h]ere, too, me is gaining ground but slowly and amidst uncertainty and hypercorrection” (Harris 1981: 20).

Hence, this assumed prescriptive bias promoting the use of subject pronoun forms which delays the transition to a clear weak vs. strong pronoun distinction is – in addition to the predicted distributions of pronoun case forms – another very important feature that is shared by both the positional and the weak vs. strong pronoun accounts. Finally, it is also interesting to note that the weak vs. strong pronoun approach – like the positional account –
is also not restricted to a particular linguistic tradition, but has found advocates not only in
different decades or even centuries, but also in different linguistic frameworks (e.g. Harris
1981; Quinn 2005a, 2005b; Shorrocks 1992; Sweet 1875).

3.4 Functional Approaches to the Distribution of Pronoun Case
Forms

The mainly theoretical approaches to pronoun case distribution introduced in the previous
three Sections 3.1–3.3 are different from the functional accounts briefly discussed in this
section in two very important respects. Firstly, all of the approaches discussed thus far offer
a more or less coherent and mainly moncausal explanation for the distribution of pronoun
case forms, which can be generalised over a wide range of contexts such as underlying case
mechanisms, the position of the pronoun relative to the finite verb, or the membership to a
certain pronoun class (cf. Sections 3.1–3.3).

This is not necessarily the case for the accounts dealt with in this section. Some of
them focus instead on certain contexts or constructions where variability has been observed
and thus often provide or discuss variables and explanations that are relevant particularly
for the respective context or contexts. Furthermore, some studies identify a whole slew of
possible factors that may influence the distribution of pronoun case forms in a certain
variable context, which indicates that the distribution of pronoun case forms may rather be
conceived of as a multifactorial phenomenon (e.g. Angermeyer and Singler 2003). Unfortu-
nately, there are only a few functional studies analysing the distribution of pronoun
case forms in contexts allowing for variability in general and subject predicative
complements particular. Thus, where applicable, we will broaden our scope in order find out
how pronoun case distribution and the variation observed in many contexts can be
explained from a functional linguistic perspective. Secondly, whereas many of the accounts
introduced in the previous sections are rather theoretical both in outlook and execution (e.g.
Emonds 1985; Haegeman 1993; Harris 1981; Radford 2004), many of the studies discussed in
this section rely on a sound empirical foundation on which they base their assumptions and
findings (e.g. Angermeyer and Singler 2003; Biber et al. 1999; Erdmann 1978). In the
following, this section presents a survey of the most important findings and assumptions of
those functional accounts that are immediately relevant for the present study. However, this section will be rather brief in order to avoid repetitiveness, since many of the factors and variables identified as potentially influencing the distribution of pronoun case forms will be more thoroughly discussed and motivated in Chapter 6 which introduces the variables tested in the quantitative analyses in Chapters 8–14 of this study.

As noted above, functional approaches trying to account for the distribution of pronoun case forms in Present-Day English are in general rather scarce except for the accounts provided in some comprehensive grammar books (e.g. Biber et al. 1999). Another exception to this general scantiness of functional accounts is Hudson (1995). However, this approach takes not only functional aspects but also the impact of the pronoun’s position into consideration and is therefore on the borderline between clearly functional and positional approaches (cf. Section 3.2). As stated in Section 2.1, Hudson emphatically rejects the idea that English still has any morphological case distinctions (e.g. 1995: 390). Instead, what appear to be remnants of overt morphological case marking in Present-Day English would be much better analysed in terms of very specific lexical rules or items of word classes. Hence, the pronoun forms I and me are merely considered as alternants of the first singular personal pronoun selected according to a minor lexical rule. This lexical rule responds to the factors function, i.e. subject of a tensed verb or not, and coordination, i.e. next to and or not, and thus determines the form of the pronoun (Hudson 1995: 390). Likewise, the alternants my and mine are, according to this account, chosen by another minor lexical rule depending on the presence versus absence of a common noun as complement of the possessive determiner. Thus, my is used when it functions as a possessive determinant followed by a noun phrase, whereas mine is used when it functions as a pronoun proper replacing the possessive determiner my and a noun phrase modified by it. Similarly, the possessive marker ’s is analysed as a possessive determiner which needs a preceding and not a succeeding noun phrase (Hudson 1995: 390-391). However, despite Hudson’s (1995) clear emphasis of the importance of the pronouns’ functions in a clause, his approach is also in the tradition of accounts that see the position of the pronoun in the sentence as the most important factor for the distribution of pronoun case forms in Present-Day English. Thus, according to this approach, the distribution of pronoun case forms is determined by the interaction of functional and positional factors, which, moreover, may also differ from variety to variety (cf. Hudson 1995: 379-381, 390-391).
Particularly with regard to the main focus of this study there are accounts that identify the syntactic function of the pronoun as a very important factor influencing the distribution of pronoun case forms in *it*-clefs and *it BE* sentences (Biber et al. 1999: 335-336; Erdmann 1978; Maier 2013). This contrasts, of course, with studies that have cast doubts on the impact of such functional syntactic factors (cf. Burridge 2004: 1118; Harris 1981). After analysing his corpus of 40 British novels, Erdmann (1978), for example, comes to the conclusion that the distribution of pronoun case forms is at least co-determined by the syntactic function of the pronoun in the sentence. Thus, if the personal pronoun is used in an object position, it is very likely to occur in its object form. If, however, the personal pronoun occurs in a subject predicative complement position, it can vary between subject and object form. Furthermore, this study observes that the likelihood for observing a subject form in subject predicative complements is much higher in *it*-clefs than in *it BE* sentences (Erdmann 1978: 78-79). Even within the class of *it*-clefs, there are further differences in the use of pronoun case forms observable depending on whether the focal pronoun is coreferential with the subject or an object of the following clause (e.g. Erdmann 1978: 75-78). Hence, this study concludes that the “pronominal system has changed from a morphologically determined to a syntactically determined” system (Erdmann 1978: 79). In a similar vein, Biber et al. (1999: 335-336) also note that the distribution of pronoun forms in subject predicative complements is influenced by the sentence type, i.e. *it BE* sentences vs. *it*-clef. Moreover, they observe with regard to *it*-clefs that the relative pronoun following the focal pronoun also influences the choice of the pronoun form (Biber et al. 1999: 335-336).

In addition to these syntactic and partially discourse-pragmatic factors, the degree of formality in which a certain pronoun form occurs has also been identified as possibly influencing the distribution of pronoun case forms in virtually all contexts exhibiting variability in the use of pronoun case forms. The importance or impact of this sociolinguistic factor has been acknowledged by proponents of functional accounts as well as by advocates of both positional and weak vs. strong pronoun accounts (cf. Sections 3.2 and 3.3). Indeed, most accounts agree that informal situations favour the use of object pronouns, while a high degree of formality is supposed to promote the use of subject pronoun case forms (e.g. Biber et al. 1999: 335-336; Erdmann 1978: 72; Quirk et al. 1985: 336-338; Shorrocks 1992: 436; Sobin 1997: 418-419; Sweet 1875: 495). Moreover, it is widely believed that the higher
degree of formality associated with subject pronoun forms and their putatively higher degree of politeness in combination with the prescriptive inclination towards the use of the subject pronoun forms are also responsible for so-called hypercorrect uses of pronoun forms, i.e. instances of subject forms in which object forms are expected. Cases in point are coordinated object noun phrases such as the quasi-idiomatic between you and I and sentences such as He saw John and I yesterday (Angermeyer and Singler 2003: 177; Huddleston and Pullum 2002: 463; Quirk et al. 1985: 338).

Moreover, Angermeyer and Singler (2003) identify a whole range of other sociolinguistic factors that may influence the choice of a pronoun form in contexts exhibiting variability in terms of pronoun case distribution. In their discussion of coordinate noun phrases, they identify the speaker’s sex, age and degree of education as sociolinguistic factors potentially influencing the distribution of pronoun forms (Angermeyer and Singler 2003: 184-193). Unfortunately, not all of these variables can be analysed in the present study due to the characteristics of the analysed datasets and the difficulty of obtaining enough data (cf. Chapter 7). Furthermore, syntactic as well as pragmatic features, such as familiarity of referents, usage frequency and focus, also seem to influence the distribution of pronoun forms in certain contexts (Angermeyer and Singler 2003: 197). Although Angermeyer and Singler’s (2003) study is restricted to coordinate noun phrases, it seems likely that at least some of these factors are also relevant for the choice of pronoun forms in other contexts, such as it-clefts and it BE sentences.

In light of the many other sociolinguistic factors supposed to influence the distribution of pronoun case forms, and due to the general variability of pronoun case usage attested in many contexts (cf. Section 2.3), it is more than likely that different varieties of English also exhibit different distributional patterns in the use of pronoun forms in subject predicative complements (cf. Section 2.3.2). First of all, this assumption can be attributed to the massive scope of variation that can be observed in the use of pronoun case forms across varieties of English (cf. Sections 2.2.1–2.2.3). Moreover, cross-varietal differences in the distribution of pronoun case forms have also been attested for other contexts exhibiting variability, such as coordinated noun phrases (cf. Section 2.3.2; Quinn 2009: 36-42). In fact, some scholars consider the scope of variation that can be observed in the pronoun case paradigms and the distribution of these forms in some domains across varieties as so extensive “that any pan- or polylectal approach seems fraught with difficulties, if not
foredoomed to failure” (Shorrocks 1992: 441-442). Most importantly, however, cross-varietal differences have already been observed for the contexts examined in this study, i.e. *it*-clefs and *it* BE sentences. Unfortunately, these significant differences between British and American English have only been attested by means of simple chi-square test analyses (Maier 2013), which need further refining by multivariate statistical models.

In sum, contrary to the assumptions of much of the relevant literature which tries to account for the distribution of pronoun case forms mainly by means of more or less monocausal explanations (cf. Sections 3.1–3.3), several accounts have identified a number of functional factors which seem to influence the distribution of pronoun case forms in contexts exhibiting variability in general and subject predicative complements in particular. So far there have not been many quantitative studies examining the impact of these factors, either monofactorially or multifactorially (cf. Maier 2013; Quinn 2009). Hence, this study will try to assess which factors influence the distribution of pronoun case forms in subject predicative complements and if so to what extent they do so.

### 3.5 The Distribution of Pronoun Case Forms in Linguistic Theory: Interim Summary

The present chapter has demonstrated that linguistic theorising has offered several possible explanations to account for the distribution of pronoun case forms in Present-Day English.

Section 3.1 has shown that much of the formal linguistic theory still proposes that the alternation between subject and object pronoun case forms is determined exclusively by structural case mechanisms. However, the discussion has also shown that variation or variable contexts are very often not addressed in these standard formal accounts to case marking and this has led to the fact that some formal linguists try to either combine or reconcile their respective formal model with other linguistic models in order to account for the variation observed in the use of pronoun case forms in Present-Day English or even advocate other approaches to pronoun case distribution.

One of the other accounts that has many supporters in the formal camp is the positional approach discussed in Section 3.2. This approach suggests that the distribution of pronoun case forms is determined by the position of the pronoun in the sentence relative to
the finite verb. Quirk et al., for example, divide the sentence into subject territory and object territory (1985: 337-338). According to this account the NP slot immediately preceding the finite verb is the subject territory where we expect to find subject pronouns, whereas all other NP positions in a sentence are object territory in which object pronoun forms are supposed to occur (Quirk et al. 1985: 337). Variation in this approach is mainly explained by stylistic differences. In formal varieties, the subject form is retained in more formerly nominative contexts and a general stylistic bias towards the use of subject pronouns is also postulated for more formal registers (cf. Section 3.2).

The approach trying to account for pronoun case distribution introduced in Section 3.3 makes nearly the same predictions as the positional approach. In this approach, subject pronouns are also restricted to the noun phrase position immediately preceding the verb. However, the focus of this weak vs. strong pronoun approach is more on the distinction between different classes of pronouns than on identifying certain territories in which the pronoun forms are supposed to occur. In a nutshell, this account proposes that the distributional differences can be attributed to a split in the pronominal paradigm which is due to a distinction between weak and strong pronoun forms, similar to the situation in French. Variation in the use of pronoun case forms in certain contexts is mainly explained in terms of the assumed dynamic transition from a formerly case-governed pronoun system to a system of clitic and non-clitic pronoun forms. Thus, both the positional and the weak vs. strong pronoun case approach predict that we should observe only or mainly object pronouns in both it-clefts and it BE sentences because in both sentence types the pronoun in question is either in object territory or in a strong pronoun context.

Although these approaches may initially seem very plausible, unfortunately, they do not fully comply with the empirical observations made in several functional accounts introduced and discussed in Section 3.4. In particular, it has been demonstrated that the findings of several functional studies analysing different contexts allowing for variability in the use of pronoun case forms suggest that the distribution of pronoun case forms may be influenced by several different factors such as the syntactic function, the degree of formality, the geographic provenance as well as pragmatic factors, such as familiarity or focus. As a consequence, the distribution of pronoun case forms particularly in subject predicative complements may also be considered a multifactorial phenomenon influenced or determined by several factors simultaneously.
4 A Focus-Oriented Approach to Pronoun Case Distribution in Subject Predicative Complements

The two preceding chapters have laid the necessary foundations on which the assumptions introduced in the present chapter rest, namely that the distribution of pronoun case forms in subject predicative complements is strongly influenced by pragmatic factors, particularly by the extent to which a pronoun is focussed in a clause.

So far, we have taken a rather general perspective on pronoun case forms in Present-Day English and their distribution in Standard English (cf. Section 2.1), varieties of English (cf. Section 2.2), and contexts allowing for variability in the use of pronoun case forms in Standard English and across varieties of English (cf. Sections 2.3.1 and 2.3.2).

Moreover, this study has also introduced the most important approaches explaining the distribution of pronoun case forms and accounting for the variability in terms of pronoun case usage observed in many contexts (cf. Sections 3.1–3.4). However, as noted in Section 3.4, functional empirical studies suggest that the distribution of pronoun case forms in Present-Day English may be much more complex than has been assumed thus far, since the distribution of pronoun forms may be influenced by several factors simultaneously. Thus, monocausal explanations may be not sufficient to account for the complexity of this phenomenon.

Furthermore, the cross-varietal survey in Sections 2.2.1–2.2.3 has shown that pragmatic factors play an enormously important role in the distribution of pronoun case forms in many varieties of English. However, almost none of the mainly theoretical approaches discussed in Sections 3.1–3.3 take the importance of such factors into consideration – either in general, or in contexts allowing for variability in the use of pronoun case forms, or with regard to the constructions in the focus of this study, i.e. subject predicative complements. Only some of the functional empirical accounts discussed in Section 3.4 acknowledge that pragmatic factors such as familiarity with the referents or focus may influence the distribution of pronoun case forms in contexts allowing for variability.

In the following chapters, we will narrow our scope from pronoun case distribution in English in general to the contexts that are of particular interest to the present study, i.e. the
distributions of pronoun case forms in *it*-clefts and *it BE* sentences, phenomena subsumed under the heading of subject predicative complements.

Acknowledging the strong influence of pragmatic factors on the distribution of pronoun case forms in many varieties of English (cf. Sections 2.2.1–2.2.3), this study tries to assess to what extent pronoun case distribution in subject predicative complements may also be explained by pragmatic factors, particularly with regard to Focus, as this factor seems to play a prominent role in the distribution of pronoun case forms in several contexts of different varieties of English, as has been outlined in Sections 2.2.3 and 3.4 (cf. also Paddock 1994). Thus, the present study argues for a pragmatic or Focus-oriented approach to the distribution of pronoun case forms. In particular, this study assumes, in accordance with one of the central hypotheses outlined in Section 5.2, that subject pronoun case forms are used as Focus markers in subject predicative complements. Although the strong influence of positional and syntactic criteria is also acknowledged by the account proposed here, these factors alone are probably insufficient to account for the variation observed in contexts allowing for variability in the use of pronoun case forms in general and in subject predicative complements in particular (e.g. Angermeyer and Singler 2003; Erdmann 1978; Maier 2013).

However, the pragmatic or Focus-oriented approach adopted here is not deemed so much as a substitute for the other functional approaches discussed in Section 3.4 but rather as a complement, and a necessary one, to obtain a more accurate and complete picture of the distribution of pronoun case forms particularly in *it BE* sentences and *it*-clefts. This is due to the fact that the cross-varietal surveys in Sections 2.2 and 2.3.2 and the survey of functional accounts in Section 3.4 have already identified a number of other potential factors influencing the distribution of pronoun case forms, which will be addressed in detail in Chapter 6. Thus, the distribution of pronoun case forms in subject predicative complements is also probably best considered a phenomenon that is influenced by several different factors at the same time. Nevertheless, this study assumes that Focus is a particularly influential factor in the distribution of pronoun case forms in subject predicative complements.

Before it is outlined in detail why and how such a re-functionalisation of subject pronoun case forms from case to Focus markers in subject predicative complements has come about and how such an assumption can actually be motivated (cf. Section 4.2), some general remarks on the notion and nature of Focus as applied in this study are in order (cf.
Section 4.1. Section 4.3 introduces and discusses the underlying mechanism responsible for the fact that subject pronouns and no other pronoun forms may have been reanalysed as Focus markers, and Section 4.4 provides a summary of the major assumptions of this section.

4.1 The Notion of Focus as Used in this Study

Following the traditional view of Functional Grammar, this study defines Focus or focussed constituents as the relatively most important, salient or highlighted information in a clause (Dik 1978: 130; Siewierska 2004: 159; Weinert and Miller 1996: 179). Thus, the speaker or, depending on the mode of discourse, writer may focus “on it in order to make it, and the information it carries, cognitively salient for the addressee” (Weinert and Miller 1996: 179). To acknowledge the fact that this study uses the notion Focus as defined in the framework of Functional Grammar, we also adopt the convention of spelling Focus with a capital letter in order to indicate that this study uses a distinct or particular definition of this concept (e.g. Dik 1978: 130; Siewierska 1991: 173-180). Most often, the Focus or focussed constituent denotes an element of the predication that predicates, i.e. states, something about the Topic, i.e. the entity, term, or referent the utterance is primarily about (e.g. Siewierska 1991: 149). Hence, in those accounts proposing a binary distinction of the clause into either topic and comment or theme and rheme, it is part of the comment or rheme (cf. Siewierska 1991: 174). The Topic/Focus distinction in Functional Grammar is not, however, perceived as a simple binary distinction. Although positional preferences for pragmatic functions are generally acknowledged in this approach (e.g. Siewierska 1991: 149), the Focus or focussed constituent of a clause is not simply identified by relying on the position of the constituents within a clause. Instead, Functional Grammar often uses a question and answer scheme in order to identify the Focus of a clause (cf. (83)–(85) below):

(83)  a. Who kissed you?
     b. John did.

(84)  a. What did the thief do?
     b. She stole my purse.

(85)  a. Where did you see Dürer’s Rhinoceros?
     b. I saw Dürer’s Rhinoceros when I was visiting the British Museum in London.
Since Functional Grammar assumes that the Focus in a question falls upon the question word, the Focus in an answer is supposed to be assigned to that part of the expression that provides the nucleus of the answer to the question asked, i.e. the direct response to the question word, for example *John* in (83), *stole my purse* in (84) and *the British Museum in London* in (85) (Siewierska 1991: 175). In a similar manner, it is argued that the Focus in declarative sentences can be identified as that clausal element which would constitute the actual answer if the sentence had been uttered in response to a question, as the examples in (83)–(85) illustrate (Siewierska 1991: 174-175). However, despite this relatively simple procedure, the identification of what exactly constitutes the Focus of a sentence may not always be so obvious, because the belated formulation of a question may prove difficult. In the declarative sentences in (86b) and (87b), for example, other underlying questions are easily conceivable, which may result in different focussed constituents (cf. Siewierska 1991: 175):

(86)  
a. What did the thief do with your purse?  
b. She *stole* my purse.

(87)  
a. When did you see Dürer’s Rhinoceros?  
b. I saw Dürer’s Rhinoceros *when I was visiting the British Museum in London*.

Thus, in contrast to (84a), in (86a) it is not the whole predication, but rather the verbal action *stole* that is the focussed constituent. Similarly, in (87b), it is not only the adverbial of place that is focussed, but rather the whole clause *when I was visiting the British Museum*. Hence, the definition of the term Focus as “relatively most important or salient information” (Siewierska 1991: 174) is rather fuzzy and the identification of the focussed constituents may indeed be difficult simply because it is strongly context-dependent, which is the nature of pragmatic things.

In spite of the difficulties of actually pinning down Focus, cross-linguistic and typological research has identified a number of devices by which focussed constituents may be made prominent (Dik 1989: 278):

I. Prosodic Prominence

II. Special constituent order, i.e. special positions for focussed constituents in the order of the clause

III. Special Focus markers, i.e. particles which highlight the focussed constituents in the clause
IV. Special Focus constructions, i.e. constructions which per se define a specific constituent as fulfilling Focus function

Thus, a language may use one or more of these devices to mark the Focus of a clause or sentence, given that there are one or more means of Focus marking available. In speech, the most prototypical Focus marker is, of course, prosody, although Focus and prosodic prominence should not be equated. Whereas prosodic prominence often entails Focus, the reverse is not necessarily true, because clause constituents may be marked for Focus by other means, such as special Focus constructions, without receiving prosodic prominence (Dik 1989: 278; Siewierska 1991: 175). Furthermore, Topics may also receive prosodic prominence, which means that prosodic prominence alone may not be sufficient to identify the Focus of a clause (Siewierska 1991: 149). In the written mode of discourse, prosodic prominence is of course not available and if no other Focus device is used, the identification of the focussed elements hinges solely on the reader’s assessment of the most important piece of information (Siewierska 1991: 174-175). Furthermore, there is very often a “trade-off” between the different Focus marking devices (Dik 1989: 278). If, for example, Focus is marked by a certain Focus construction, another Focus marker, such as prosodic prominence, may not be necessary to identify the focussed constituent as the relatively most important one within a clause (cf. Dik 1989: 278-279).

English, like many other languages, possesses different strategies with the help of which focussed constituents may be highlighted. Examples of Focus marking strategies available in English are provided in (88)–(92) (e.g. Dik 1989: 279; 1997: 292; Quirk et al. 1985: 604, 1365, 1384; Weinert and Miller 1996: 179).

(88) Prosodic Prominence as Focus Marker in English (e.g. Dik 1989: 279)
Peter likes SANDRA [prosodic prominence]

(89) Special Constituent Order as Focus Marker in English (e.g. Quirk et al. 1985: 1365, 1384)
Peter painted Sandra’s kitchen pink.

(90) Special Constructions as Focus Markers in English: it-Clefts (e.g. Dik 1997: 291)
It was Peter who painted Sandra’s kitchen pink.

(91) Special Focus Markers in English: Focussing Subjuncts (e.g. Quirk et al. 1985: 604)
a. Peter loves only Sandra.
b. Peter loves mainly Sandra.
Special Focus Markers in English: Subject Pronouns in Subject Predicative Complements

a. It was they who were scrutinized by Orwell, the Pilgrim Trust and other social investigators. (BNC/ACH/W_non_ac_soc_science)

b. What happened then? It’s the Ogre! I know it is she! Bel and the Ogress-one and the same! (COCA/FIC/ContempFic)

As the example in (88) illustrates, the Focus of a sentence may be indicated by prosodic prominence, i.e. by putting particular phonetic emphasis or stress on the entity in question, which, in the present case, is SANDRA. However, since focussed constituents are very often placed at the end of a clause in English (e.g. Quirk et al. 1985: 1365, 1384), as is indicated by the example given in (89), we can also observe that different Focus marking strategies may not only be subject to a certain “trade-off”, as noted above, but may also even co-occur. Thus, SANDRA in (88) receives not only prosodic prominence but also what has been called “end-focus” within a clause, since this entity occurs in the noun phrase slot that is very often associated with the Focus position in English (e.g. Quirk et al. 1985: 1363-1375, 1384). Except for prosodic prominence and special constituent order, English also possesses particular Focus constructions, such as it-clefts, which are of central importance to the present study and which highlight a certain clausal constituent as the Focus, as is indicated by the example in (90) (e.g. Biber et al. 1999: 958-960). Furthermore, English also possesses particular Focus markers (cf. (91a) and (91b)). These so-called focussing subjuncts can be used to assign Focus status to clausal constituents ranging from single terms to whole predications (Quirk et al. 1985: 604-605). Focussing subjuncts are usually divided into two classes, i.e. exclusives (cf. (91a)) and particularisers (cf. (91b)). While exclusives such as alone, exactly, only and simply confine the application of the utterance exclusively to the focussed constituent, particularisers such as largely, mainly and at least restrict the utterance’s application predominantly but not exclusively to the focussed constituent (e.g. Quirk et al. 1985: 604).

Finally, as the examples in (92a) and (92b) illustrate, this study assumes that in English, Focus may also be indicated by using subject pronoun case forms in subject predicative complements. Indeed, one of the central assumptions of this study is that subject pronoun forms may have been reanalysed as Focus markers in subject predicative complements exploiting the potential variability in terms of pronoun case usage in these contexts. How such an assumption can be motivated and accounted for is outlined in the following section.
4.2 From Case to Focus Markers: The Reanalysis of Subject Pronoun Forms as Focus Markers in Subject Predicative Complements

Due to the importance of pragmatic factors in the distribution of pronoun forms observed in many varieties of English (cf. Sections 2.2.1–2.2.3), this study also investigates whether pragmatic functions exert a certain influence on the distribution of pronoun case forms in subject predicative complements. In particular, this study is interested in whether or not subject pronouns may be used to mark Focus in subject predicative complements, or – to put it slightly differently – whether the likelihood of encountering subject pronouns in subject predicative complements increases when the noun phrase slot in which it appears is clearly focussed. This possible usage of subject forms as postverbal Focus markers may at first seem unorthodox or even counter-intuitive, bearing in mind that subject pronouns are often associated with preverbal, unstressed topical occurrences (e.g. Pietsch 2007: 167, 2009: 146). Nevertheless, there are still good reasons for assuming that the subject forms in subject predicative complements are indeed more likely to occur in more focussed positions than in less focussed positions, as in now explained in detail.

From a diachronic point of view, the possible reinterpretation of the subject pronoun forms in subject predicative complements as Focus markers can be explained as follows. It is a well-known fact that the Old English case system had more or less collapsed by the advent of Middle English. As a result, nouns lost all nominative, dative and accusative distinctions and also the former genitive is now usually considered to be a system of possessive modifiers rather than a true case category (e.g. Denison 1993: 20-21; Hudson 1995; Hollmann 2009: 317-321; Lass 2006: 51-52). Thus, different case forms have been more or less confined to the domain of personal pronouns, distinguishing subject and object forms (Hollmann 2009: 319; Lass 2006: 51-52). However, the earlier function of the different case forms in English, i.e. marking the semantic relationships between verbs and noun phrases or between heads and dependents (cf. Blake 2001: 1; Butt 2006: 4), has mostly been taken over by other means, especially by a rather rigid SVO word order and by means of prepositions (cf. Section 2.1). This means that although different pronoun case forms have been retained, they have been stripped of their original function of marking relationships between heads and dependents (e.g. Denison 1993: 21; Mustanoja 1960: 348). In historical linguistics, it is
assumed that if a language encodes a grammatical category morphologically and the grammatical distinction is lost for some reason before the morphological forms that encode it, the respective language or speech community has three different options to further deal with these morphological remnants (Lass 1990: 82):

I. The morphological forms can be discarded entirely

II. The morphological forms can be kept as marginal forms or non-functional, non-expressive residual forms

III. The morphological forms can be maintained, but instead of being relegated to some marginal phenomenon, they can be re-functionalised, perhaps even in a systematic way

The first option mentioned above describes what has happened to many pronoun case forms in English, for example the former accusative pronouns. They simply disappeared from the system over the course of time (e.g. Denison 1993: 16-23). As a consequence, some authors even go so far as to suggest that the remaining subject pronoun forms may also ultimately vanish from the pronoun system since they are much more restricted than object pronouns in their functional range (e.g. Kjellmer 1986: 49; Wales 1996: 96, 107-109). However, this seems rather unlikely, given the conserving forces of a standardised language that tend to inhibit language change (e.g. Milroy 1999: 27). Furthermore, it should not be forgotten that although the functional range of subject forms is much more confined when compared to that of their object pronoun counterparts, subject pronouns still exhibit a very high token frequency, even surpassing that of object pronoun forms in corpora such as the British National Corpus (cf. Sections 4.3 and 7.1.2).\(^9\) Indeed, this high token frequency constitutes another main obstacle to a possible abandonment of the remaining subject forms (e.g. Bybee 2010: 24-25).

The second option described above is the status often attributed to the remaining pronoun case forms in Present-Day English. As their main function, i.e. the licensing of dependency relationships, has been relocated to other linguistic strategies, the different subject and object forms are often considered to be positional variants as the discussion in Section 3.2 has shown (cf. Burridge 2004: 1118; Quirk et al. 1985: 336-338).

The third option described in Lass’ account (1990: 82) is the basis not only for the weak vs. strong pronoun approach outlined in Section 3.3 (cf. also Harris 1981), but also for

the stance taken here, namely that the subject pronouns may have been reanalysed as
Focus markers in subject predicative complements following *it* and a form of *BE*. This
assumption is corroborated by a similar kind of re-functionalisation process already
discussed in detail in Section 2.2.3, i.e. the phenomenon of pronoun exchange. In some
traditional varieties of English, most prominently in the South of England, pronoun exchange
typically occurs when subject pronoun forms are used in object positions or in all other
positions where we would expect to find object pronoun forms according to the typical
usage or distribution in Standard English (Wagner 2004: 157-158). As noted in Section 2.2.3,
although different notions such as the marking of emphasis, stress, and contrast have been
used to motivate this phenomenon, it is most remarkable that the standard pronominal
paradigm has been re-functionalised in some varieties. Hence, depending on the
terminology of the respective account, the factor stress, Focus, contrast or emphasis may
severely affect the distribution of pronoun case forms in these varieties, since new
pragmatic functions have been allocated to available pronoun forms which had lost their
original syntactic functions (e.g. Shorrocks 1992: 440-441; Paddock 1991: 36-37; 1994: 256-
263; Rogers 1979: 35; Wagner 2004: 158; or see Section 2.2.3 for detailed overview).

Interestingly, in Newfoundland English, pronoun exchange has been explicitly
interpreted as a Focus marking device, whereby “the former syntactic case distinction (of
subject versus object), wholly or partially lost in the Wessex mother dialects, was generally
replaced by a pragmatic Focus distinction in the Newfoundland dialects” (Paddock 1994:
260). In view of this development, it seems plausible that a similar re-functionalisation
process, though in an extenuated form, has taken place in subject predicative complements
following *it* and a form of *BE* even in the Standard varieties of English. Thus, similar to the
situation in Newfoundland English, subject pronoun forms may also be used in subject
predicative complements following *it* and a form of *BE* to mark focussed pronouns, thereby
taking advantage of the potential for variability in the use of pronoun case forms.

This possible reanalysis of subject pronouns as Focus markers in contexts allowing for
variability in Standard English is, however, corroborated by more than the singular
phenomenon of pronoun exchange discussed above. This assumption is also substantiated
by the robust trend observed in Section 2.2 in which pronominal paradigms across varieties
of English are prone to distinguishing between stressed and unstressed or marked and
unmarked pronoun forms, although it is often not specified what exactly is meant by the
distinction stressed–unstressed or marked–unmarked, i.e. whether the supposed differences are of a prosodic or pragmatic nature. What is, however, remarkable is that this distinction seems to be important for many varieties of English from very different backgrounds as the examples given below, discussed in detail in Section 2.2 and briefly recapitulated here for convenience, illustrate:

To begin with, a distinction between stressed and unstressed pronoun forms can be observed, for example, in rather traditional dialects such as that spoken in the Bolton area, where a distinction is made between both stressed and unstressed subject and object pronouns (Shorrocks 1999: 72, 76). Moreover, the survey in Section 2.2.3 also showed that some varieties, such as Appalachian English, Fiji English and the variety of English spoken in East Anglia, exhibit stressed–unstressed distinctions even for the third person neuter (Montgomery 2004: 262; Mugler and Tent 2004: 774; Trudgill 2004: 146-147). Furthermore, this distinction also seems to be very important for some pidgin and creole varieties. In Ghanaian Pidgin, for instance, the survey in Section 2.2.2 noted emphatic and unemphatic sets for both subject and object pronouns (Huber 1999: 197-199). As far as the distribution of these forms is concerned, emphatic subject pronouns are used in emphatic, focussed or contrastive contexts such as the focal position of it-clefts, whereas their unemphatic counterparts are used only in the noun phrase slot immediately preceding the finite verb (Huber 1999: 195). As stated in Section 2.4, the encoding of emphatic–unemphatic, stressed–unstressed or focussed–unfocussed contrasts in the respective pronominal paradigms seems to be in fact a desideratum of many varieties of English. Thus, the importance attributed to a distinction between more and less emphatic, stressed or focussed pronoun forms in different varieties of English suggests that a possible reanalysis of subject pronoun case forms as Focus markers in subject predicative complements seems indeed very plausible, both from a historical and a cross-varietal perspective.

The assumption of subject forms fulfilling certain pragmatic functions in subject predicative complements, i.e. particularly that of Focus marking, is also supported to some extent by the literature discussing contexts exhibiting variability in the use of pronoun case forms, although this is often done indirectly (e.g. Hopkins 1975: 33; Quinn 2005a: 246). The potential influence of this factor is acknowledged most explicitly in Wales (1996), who states that “the pedantry of the subjective case variant can be exploited for conscious effect, and, preceded by an intensifier provides a useful end-stress” (1996: 95). She illustrates her
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

statement, which is compatible with the stance taken here, with the following example (Wales 1996: 95):

(93) ‘Oh look Dad, there’s John Cleese’, he said excitedly. It was, of course, he. (*The London Evening Standard, 11 February 1993*)

Moreover, Erdmann (1978: 75) notes in his section on “Focussed constructions” that subject pronoun case forms in *it BE* sentences are more likely to occur in sentences in which the pronoun in predicative complement position is co-referential with the subject in a preceding or following sentence. As an example of such a construction, he cites the following sentence (Erdmann 1978: 75):

(94) If anyone was responsible for Brian’s death, it was she. (*Huxley, Eyeless in Gaza: 378*)

The special kind of *it BE* sentences illustrated by the example in (94) will be thoroughly discussed in Section 6.1.2.5, which deals specifically with the operationalisation of the factor FOCUS. Suffice it here to say that focussed constructions have already been acknowledged by other studies as potentially exhibiting different distributions of pronoun case forms when compared to less focussed constructions (e.g. Erdmann 1978: 78; Wales 1996: 95).

Furthermore, as was already mentioned in Section 3.4, pragmatic factors in general and Focus in particular have also been identified in another context in which variability in pronoun case usage is observed, i.e. coordinated noun phrases, as a potential factor influencing the choice of the pronoun case form or “coordination pattern” (Angermeyer and Singler 2003: 197). Thus, in view of the preceding discussion, the assumption that the factor Focus may influence the distribution of pronoun case forms in *it BE* sentences and *it*-clefts seems very plausible. In fact, Focus may be a very important factor in accounting for the variability of pronoun case usage in these contexts taking into consideration the important role pragmatic functions in general and Focus in particular may have in the distribution of pronoun forms across and within varieties of English.

Finally, further support for the assumption that subject pronouns may be reanalysed as Focus markers in subject predicative complement positions can be deduced from a more recent functional approach to grammar, i.e. Construction Grammar. Depending on whether the difference between the subject and the object pronoun forms can be considered a syntactical one, Construction Grammar’s *Principle of No Synonymity* holds in the case of pronoun case variation in subject predicative complements. According to this principle, two
constructions that are syntactically different must also be semantically or pragmatically different (Goldberg 1995: 67). Since the difference between subject and object pronoun forms in the contexts in question cannot be regarded as a semantic one since the semantic meaning or content is not altered by exchanging one form for the other, the difference has to be pragmatic, provided the difference between the two morphological variants qualifies as a syntactical distinction in this framework. That Focus may play a prominent role in cases in which the Principle of No Synonymity takes effect has been observed in the context of ditransitive constructions and their prepositional paraphrase. With regard to ditransitive constructions as well, the choice of one alternant over the other seems to be strongly influenced by the degree of Focus the constituents of these constructions receive (cf. Goldberg 1995: 89-92).

4.3 From Subject Pronouns to Focus Markers: An Explanation for Choosing and Using the Subject Forms

So far, we have introduced the notion of Focus as used in this study as well as different Focus marking strategies available in Present-Day English. Section 4.2 has provided a possible explanation and motivation for this study’s assumption that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements following it and a form of BE. A very important issue that must still be addressed, however, is why subject pronoun forms and not other pronoun forms, such as possessives or reflexives (cf. Section 2.1), may be used as Focus markers in subject predicative complements.

In general, the reasons for why subject pronoun case forms may be subject to functionalisation processes and not other available options can be explained with a phenomenon called “Markiertheistemkerung”, i.e. “markedness reversal”, by Mayerthaler (1981: 48) or, to avoid the term markedness, deviation from “default expectation” by Haspelmath (2006: 62). Both authors introduce and illustrate this phenomenon with the help of a non-linguistic example, i.e. the dress code on beaches (Haspelmath 2006: 62; Mayerthaler 1981: 48). Haspelmath, translating Mayerthaler’s (1981: 48) original version from German into English, states that
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

“normally people wear bathing suits, so naked bathing is the marked case. On nudist beaches, we find markedness reversal – wearing a bathing suit becomes marked, i.e. unexpected and therefore remarkable” (Haspelmath 2006: 62).

Transferring this phenomenon from the beach to the distribution of pronoun case forms in *it-*clefts and *it BE* sentences, we can say that subject pronoun forms may be used as Focus markers in subject predicative complements precisely because they are not expected to occur in these contexts. In order to understand and assess how unexpected subject pronoun case forms in subject predicative complement or in postverbal positions actually are, let us have a brief look at the token frequencies of case-sensitive personal pronouns in the *British National Corpus* as given by Kilgariff (2006b).

<table>
<thead>
<tr>
<th>BRITISH NATIONAL CORPUS</th>
<th>SUBJECT FORM (N)</th>
<th>OBJECT FORMS (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ST SINGULAR</td>
<td>884599</td>
<td>138151</td>
</tr>
<tr>
<td>3RD SINGULAR (F)</td>
<td>380257</td>
<td>108710 (326968(^{10}))</td>
</tr>
<tr>
<td>3RD SINGULAR (M)</td>
<td>681255</td>
<td>165014</td>
</tr>
<tr>
<td>1ST PLURAL</td>
<td>358039</td>
<td>62350</td>
</tr>
<tr>
<td>3RD PLURAL</td>
<td>433441</td>
<td>173414</td>
</tr>
</tbody>
</table>

*Table 7: The Token Frequencies of Case-Sensitive Personal Pronoun Forms in the British National Corpus (Kilgariff 2006b)*

As Table 7 indicates, the subject pronoun forms exhibit a consistently and considerably higher token frequency than their object pronoun counterparts in the *British National Corpus*. For the third person feminine forms, this is true even if the token frequencies of both possible usages of *her* as object pronoun and as possessive determiner are added. Indeed, while the third person subject pronoun *they* occurs ‘only’ roughly 2.5 times as often as its object pronoun counterpart *them*, the first person singular subject pronoun *I* is used more than six times as often as its object pronoun counterpart *me*. These token frequencies are remarkable not only when considering accounts hypothesising that subject pronoun forms may disappear from the English pronominal system in the more or less distant future (e.g. Kjellmer 1986: 49; Wales 1996: 96, 107-109), but also when taking the typical functions of subject and object pronoun forms in Standard English into account. As a matter of fact,

\(^{10}\) Cumulative token number of the personal object pronoun *her* (N = 108,710) and the possessive determiner *her* (N = 218,258) (Kilgariff 2006b)
the survey in Section 2.1 has demonstrated that subject pronoun case forms are prototypically used only as subjects of finite verbs in declarative and interrogative sentences, while object pronoun forms have a much wider functional range, occurring not only as the object of verbs and prepositions but also as the subjects of non-finite sentences. In view of their rather narrow functional range, the theoretical approaches discussed in Sections 3.1–3.3, notably the positional and the weak vs. strong pronoun account, conclude that subject pronoun case forms are virtually restricted to the noun phrase slot in immediate adjacency to the finite verb.

In the approach adopted here, the strong association of the subject pronoun case forms with the subject function in a sentence and the wider functional range of the object pronoun case forms are also acknowledged. It can be even stated that the high token frequency of subject pronouns together with their prototypical function as subjects of finite clauses means that this form-function unit of subject pronouns as subjects of finite sentences is deeply cognitively “entrenched” (e.g. Schmid 2007: 118-119). Thus, the importance of the position of a pronoun or its function in a clause is hereby explicitly acknowledged.

However, while some accounts deduce a severe restriction of the subject pronouns’ range of application from this strong association (cf. Sections 3.2 and 3.3), the present study considers it to be the foundation on which the possible reanalysis of subject pronoun case forms as Focus markers rests – not only in subject predicative complements, but also in some of the pronoun exchange regions (cf. Section 2.2.3). It is precisely this strong association of subject pronoun forms with the subject function in finite clauses on which the principle of deviation from default expectations or markedness reversal operates. Because subject pronoun forms are prototypically expected to occur only as subjects of finite clauses, occurrences of subject pronouns in contexts other than that are very remarkable (e.g. Mayerthaler 1981: 48-49). In fact, probably no other class of pronouns is as strongly associated with a certain syntactic function or slot as the set of subject pronouns. This is true not only for the object pronouns, which have a much wider functional and/or distributional range than subject pronouns, but also for possessive pronouns (e.g. Huddleston and Pullum 2002: 458-472). Even reflexive pronouns are less strongly associated with a certain function or position within a clause, although they are also subject to rather strict constraints, when compared, for example, to object pronouns (e.g. Siemund, Maier and Schweinberger 2012).
This strong association also implies that because subject pronoun forms are normally used only as subjects of finite sentences, they can be re-functionalised more easily than other pronoun forms. Their typically narrow range of application means that a re-functionalisation or extension of the prototypical usage pattern to include also Focus marking bears a comparatively lower risk of ambiguity or functional clashes than does the use of other pronouns. Since object and possessive pronouns, for example, can already occur in many more contexts or noun slots, a re-functionalisation of these pronouns as Focus markers would thus be less salient and less easily recognisable than the use of subject forms. Moreover, since subject forms very often refer to the Topic, i.e. the referent the utterance primarily is about (e.g. Siewierska 1991: 149), they are per se already pragmatically salient, as they normally already refer to another pragmatically important constituent. Thus, in view of these facts, a re-functionalisation of the subject forms as Focus markers, particularly in contexts in which they are not expected, such as *it BE* sentences or *it*-clefts, becomes very plausible, and even more so given the fact that such a reanalysis has explicitly been attested for the pronoun exchange region in Newfoundland (cf. Paddock 1994). This, too, can be attributed to the same underlying principle, i.e. that of markedness reversal or deviation from default expectations.

Moreover, further evidence for such a pragmatic reanalysis comes from the examples Mayerthaler (1981) provides for his concept of “Markiertheitsumkehrung” or markedness reversal. Quoting examples from both Latin and German, he demonstrates that such a markedness reversal process may be pragmatically determined (cf. Mayerthaler 1981: 49-50).

Finally, this principle of markedness reversal or deviation from default expectations may be able to account not only for the variation observed in subject predicative complements and for the assumed pragmatic re-functionalisation of subject pronoun forms in *it*-clefts and *it BE* sentences, but may also help to account for the variability attested in other contexts, such as left dislocations or subject pronoun forms in coordinated noun phrases in object position, as will be discussed in Chapter 16. Hence, since the possible range of application for the principle of markedness reversal is by no means restricted to the constructions central to this study, it can also help to account for other variation phenomena and is therefore of general relevance (cf. Chapter 16).
While this chapter has already provided a possible explanation and motivation for the reanalysis of subject pronoun case forms as Focus markers in subject predicative complements, one very intricate issue still remains open, namely the operationalisation of this factor. This admittedly rather complex matter will be thoroughly discussed in Chapter 6 as well as the motivation and operationalisation of all other variables tested in the statistical analyses conducted in Chapters 8–14.

4.4 A Focus-Oriented Approach to Pronoun Case Distribution in Subject Predicative Complements: Interim Summary

The present chapter has introduced the approach proposed by this study, assuming that the distribution of pronoun case forms in subject predicative complements may, in addition to other parameters, be strongly influenced by pragmatic factors, particularly by the factor Focus. In particular, the present approach suggests that subject pronoun case forms may have been re-functionalised as Focus markers in subject predicative complements. Thus, this study assumes that the more focussed a particular context is, the more likely it is to observe a subject pronoun case form in subject predicative complement position.

In a first step (cf. Section 4.1), the term Focus as used in this study was introduced. Following the tenets of Functional Grammar, the concept Focus is defined as the relatively most important, salient or highlighted information in a clause (Dik 1978: 130; Siewierska 2004: 159; Weinert and Miller 1996: 179).

Although the assumption that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements may initially seem odd, this chapter has provided evidence that clearly motivates such a hypothesis (cf. Section 4.2). From a diachronic point of view, such a reanalysis can be explained by the fact that the pronoun case forms that were retained after the collapse of the Old English case system have been stripped of their original function as markers of the relationships between heads and dependents, and this loss of their original function has made a reanalysis as pragmatic markers possible. From a synchronic point of view, such a re-functionalisation also seems very plausible, since similar phenomena are attested for many regional varieties of English. A prominent example for such a reanalysis process is the phenomenon of pronoun exchange.
where former subject pronouns are also used to mark emphaticness, Focus or stress, whereas object pronouns are used for unstressed, unemphatic or unfocussed contexts. Moreover, some varieties employ an emphatic–unemphatic contrast even for third person neuter. An emphatic–unemphatic or focussed–unfocussed contrast can even be observed for pidgin and creole varieties, which indicates that pragmatic factors play an important role in the pronominal paradigms across varieties of English. Finally, other studies discussing pronoun case distribution in subject predicative complements have also acknowledged that the subject form may be used, particularly in focussed contexts, to achieve what Wales calls “end-stress” (Wales 1996: 95).

The reason that subject pronoun forms in particular may have been reanalysed as Focus markers, as opposed to other pronominal alternatives, was accounted for by a principle called markedness reversal (Mayerthaler 1981) or deviation from default expectations (Haspelmath 2006) introduced and discussed in Section 4.3. Following this principle, this study assumes that precisely because subject pronoun case forms are usually restricted to preverbal noun phrase slots, they are not expected to occur in postverbal slots. This makes them more salient in postverbal subject predicative complement positions and qualifies them as Focus markers. Moreover, it has already been proposed that this principle may in its application not be confined to subject predicative complements, but may even extend to other contexts allowing for variability in the use of pronoun case forms (cf. Chapter 16).
5 Aims, Research Questions and Central Hypotheses of this Study

After having provided comprehensive surveys sketching the distribution of pronoun case forms in English (cf. Chapter 2) as well as the most important approaches accounting for the distribution of pronoun case forms in Present-Day English (cf. Chapter 3), the last chapter introduced the approach adopted for the present study (cf. Chapter 4). Based on this essential groundwork, we are now finally in the position to formulate explicitly and outline the major objectives of this study. In Section 5.1, the main research questions for this study are introduced, while Section 5.2 more specifically converts the most important of these questions into hypotheses which will be tested in this study.

5.1 Research Questions

On a very general scale, the aim of this study is to examine the distribution of pronoun case forms in subject predicative complements, i.e. it-clefts and it BE sentences, and to identify and possibly explain which factors influence the distribution of pronoun case forms in these contexts, and for what reasons. Ideally, on the basis of the insights gained from the analysis of the distribution of subject and object pronoun case forms in subject predicative complements, this study also aims to shed more light on the distribution of pronoun case forms in Present-Day English in general and thus to obtain an enhanced understanding of these variation phenomena and the factors influencing or determining them. From these rather general objectives the following research questions derive:

Q1: How are the subject and object pronoun case forms distributed in subject predicative complements?

As outlined in Chapters 2 and 3, there are very few studies that try to quantitatively analyse the distribution of pronoun case forms in subject predicative complements, even though variation in these contexts has been attested for a very long time (e.g. Sweet 1875: 495).
This is true not only for studies with a clearly cross-varietal focus (e.g. Maier 2013; Quinn 2009), but also for studies that restrict their scope to a particular variety of English (e.g. Erdmann 1978) or to Standard English in general (e.g. Biber et al. 1999). In view of this shortcoming, the most general aim of this study is simply to find out how pronoun case forms are distributed in subject predicative complements in the corpora and Web-derived datasets (cf. Chapter 7). This objective is motivated by the fact that even the distribution of pronoun case forms in *it BE* sentences and *it*-clefs is unknown for most of the varieties examined in this study, particularly for Irish, Australian, Indian and South African English (cf. Sections 2.3 and 7.1).

Moreover, in addition to this very general and rather descriptive research objective, more specific questions, such as the following one, are addressed:

Q2: Is it true that the distribution of pronoun case forms in Present-Day English is determined mainly by the position of the pronoun in a clause?

As outlined in Chapter 3, much of the current linguistic theory assumes that the distribution of pronoun case forms in Present-Day English is determined by the position of the pronoun relative to the finite verb or by its assumed membership to a certain class of pronouns. Indeed, most of the accounts discussed in Chapter 3 assume that we are supposed to observe only or mainly object pronoun case forms in postverbal positions. Thus, the function of the pronoun in a clause is not assigned much importance, particularly when compared to other potential factors such as position or pronoun class membership (cf. Sections 3.2 and 3.3). These assumptions imply that we are expected to observe mainly or only object pronoun case forms in both *it BE* sentences and *it*-clefs. However, while *it BE* sentences and *it*-clefs are superficially quite similar, they are functionally different, as will be outlined and discussed in Sections 6.1.2.1.1 and 6.1.2.5. Thus, on the basis of the different constructions subsumed under the heading of subject predicative complements, i.e. *it BE* sentences and *it*-clefs, and the different functions of the pronouns in these sentences (cf. Sections 6.1.2.1 and 6.1.2.5), this study aims to test the validity of the assumptions voiced in much of the current linguistic theory according to which the position of the pronoun is the most important or even the sole factor in determining the form of the pronoun case form in
Present-Day English (cf. Chapter 3). In addition, this study also aims to address the following question:

Q3: Which other factors influence the distribution of pronoun case forms in subject predicative complements?

The discussion of the different mainly theoretical approaches in Chapter 3 has shown that the degree of formality or the mode of discourse is also supposed to – at least – co-influence the distribution of pronoun case forms in many contexts allowing for variability in the use of pronoun forms. Furthermore, the surveys in Chapter 2 and Section 3.4 have identified additional factors that may potentially influence the distribution of pronoun case forms in contexts allowing for variability in general and subject predicative complements in particular. Thus, it is very likely that the distribution of pronoun case forms in *it*-clefts and *it BE* sentences is influenced by several predictors simultaneously. As a consequence, the following Chapter 6 introduces, discusses and motivates those factors which are often considered relevant in the literature as possibly influencing the distribution of pronoun case forms in the contexts examined in this study. To investigate whether or not these factors actually exert an influence on the distribution of pronoun case forms in subject predicative complements is a further objective of this study.

Furthermore, this study is not only interested in the factors influencing the distribution of pronoun case forms, but also in their effect sizes:

Q4: Assuming that the distribution of pronoun case forms is influenced by several factors simultaneously, which are the strongest ones?

Closely related to the last two questions – yet not fully included and very important in its own right – is the fourth research question this study intends to address. In case that the distribution of pronoun case forms in subject predicative complements is indeed influenced by several factors, this study will examine whether there is any predictor or group of predictors that exerts a particularly strong influence on the distribution of pronoun case forms. More precisely, this study wants to find out how big the effects sizes of the significant
variables are in the analysed datasets. This means that this study wants to examine whether, for example, the degree of formality is more important than the number or person of a pronoun. This is necessary as effect sizes are an indispensable means for a proper assessment of the significance of a research result (Field, Miles and Field 2012: 58). If a particularly strong factor or factor group can be identified, it will be interesting to see how such a result relates to the assumptions of the literature accounting for the distribution of pronoun case forms in Present-Day English as well as to the assumptions of the present study (cf. Chapter 4).

Speaking of the assumptions of the present study, this study also aims to test whether the Focus-oriented approach introduced in the previous chapter holds or not:

Q5: Is it true that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements?

The surveys in Chapter 2 have suggested a strong impact of pragmatic factors on the distribution of pronoun case forms across varieties of English. Furthermore, the strong impact that pragmatic factors may exert on the distribution of pronoun case forms has also been confirmed by some studies discussing contexts exhibiting variability in the use of subject and object pronoun case forms (cf. Sections 3.4 and 4.2). As a consequence, this strong impact of pragmatic factors in different contexts and across varieties of English has lead this study to assume and adopt the Focus-oriented approach introduced, motivated and discussed in the preceding chapter (cf. Chapter 4). Therefore, this study is very much interested in testing the assumption that subject pronouns may have been reanalysed as Focus markers in subject predicative complements.

Moreover, this study also intends to discuss the cross-varietal dimension of the distribution of pronoun case forms in subject predicative complements:

Q6: Is it true that there is a general or global trend observable in the distribution of pronoun case forms in English?
In the cross-varietal surveys provided in Sections 2.2 and 2.3.2, we have observed that the distribution of pronoun case forms in and across varieties of English seems to vary tremendously. Nevertheless, we also noted in Section 3.2 that there are accounts which propose a “general trend in English towards case selection [being] dictated by position rather than function – the nominative is largely confined to clause-initial preverbal position; accusative appears elsewhere” (Burridge 2004: 1118). This quote seems to propose not only that the distribution of pronoun case forms in Present-Day English is mainly determined by the pronoun’s position rather than its function within a clause, but also that varieties of English are expected to be rather uniform as far as the distribution of pronoun case forms is concerned. Thus, the present study intends to analyse whether this assumption is true or whether marked cross-varietal differences are perhaps observable.

The last question we would like to address pertains to the methodology adopted in the present study:

Q7: Is it possible to conduct sound quantitative analyses with data obtained from Google?

As noted above, the scarcity of quantitative studies analysing the distribution of pronoun case forms in subject predicative complement contexts can partly be attributed to the fact that many of the existing corpora, particularly those which are cross-varietal, are simply too small for robust quantitative or statistical analyses, not only of *it BE* sentences and *it*-clefts, but also of most other variable contexts in the use of pronoun case forms (cf. Quinn 2009: 46). In view of the difficulty of obtaining enough data, this study not only relies on closed mega-corpora (cf. Sections 7.1.2 and 7.1.3), but also on Web-derived datasets (cf. Section 7.2), since freely accessible mega-corpora are only available for British and American English, but not for the other varieties examined in this study. Therefore, with the help of Google, Web-derived databases for British, Australian, Irish, South African and Indian English were compiled specifically for the present study in order to quantitatively analyse the distribution of pronoun case forms in subject predicative complements in varieties other than British and American English. However, such an approach has very often been met with reservations, since the use of Google data carries with it many risks and problems which must be overcome in order to arrive at reliable results (cf. Chapter 7). In view of the many reasons
and accounts arguing against the use of data obtained from commercial Web crawlers as input for linguistic studies (e.g. Kilgarriff 2006a), this study tries to assess whether or not it is possible to arrive at robust conclusions and generalisations with the help of data obtained from Google.

5.2 Central Hypotheses

Based on the research questions formulated in the preceding section and on the insights gained from Chapters 2 and 3 as well as on the assumptions derived from these insights presented in Chapter 4, this study will test the following five central hypotheses:

H1: If the distribution of pronoun case forms is solely or mainly determined by the pronoun’s position or membership to a certain pronoun class, then we will not observe significant differences in the distribution of pronoun case forms between it-clefts and it BE sentences (e.g. Burridge 2004: 1118; Harris 1981; Quirk et al. 1985: 336-338).

With the help of this hypothesis, we can test the validity of much of the current linguistic theory as pertaining to pronoun case distribution in Present-Day English. As discussed in Chapter 3, much of the relevant literature assumes that subject pronoun case forms are for the most part restricted to the noun phrase slot immediately preceding the finite verb, whether for reasons of the pronoun’s position, its pronoun class membership or due to the influence of a case-assigner. If these accounts are right, we expect to observe similar distributions for both it-clefts and it BE sentences. If, however, we note marked differences in the distributions of pronoun case forms between these two sentence types, there must be other factors that strongly influence the distribution of pronoun case forms in subject predicative complements which have so far been neglected by much of the current linguistic theory (cf. Sections 3.1–3.3).

H2: If the results of earlier studies attesting a strong impact of functional factors on the distribution of pronoun case forms are correct (e.g. Angermeyer and Singler 2003; Biber et al. 1999; Maier 2013), then this influence will also be borne out by the
results obtained from the multivariate analyses testing the variables introduced and discussed in Chapter 6.

The second hypothesis tested in this study is closely related to the first hypothesis, since it is nearly its reversal. In contrast to the assumptions of the accounts discussed in Sections 3.1–3.3, several mainly empirical accounts introduced in Section 3.4 observe that functional factors such as the syntactic function of the pronoun within a clause or its co-reference with either the subject or the object of a following dependent clause may influence the distribution of pronoun case forms in *it*-clefs and *it BE* sentences. However, since these functional accounts do not employ multivariate statistical modelling (e.g. Biber et al. 1999; Erdmann 1978; Maier 2013), this study still has to assess whether and to what extent such factors significantly affect the distribution of pronoun case forms in subject predicative complements, or whether those accounts which consider functional considerations as marginal are correct (cf. Sections 3.1–3.3).

H3: If subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements, then we will observe that the more focussed a particular subject predicative complement context is, the more likely it is to observe a subject pronoun form.

Thirdly, this study also aims to test whether the strong impact of pragmatic factors on the distribution of pronoun case forms attested for many different contexts in many varieties of English is also observed in the subject predicative complements of the varieties examined in the present study. In particular, this study tests whether the assumption discussed in Chapter 4 according to which subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements holds, or whether this claim has to be abandoned. How this hypothesis can be actually operationalised will be outlined in detail in Section 6.1.2.5.

H4: If there really is a “general trend in English towards case selection [being] dictated by position rather than by function” (Burridge 2004: 1118), then we will not observe dramatic differences between the different varieties analysed in this study. If, however, the situation is as complex and diverse as is assumed by other authors (e.g. Shorrocks 1992), we will observe substantial variability in the examined data.
Although the cross-varietal surveys in Section 2.2 have demonstrated that the distribution of pronoun case forms may be a variety-specific phenomenon, the discussion of the contexts allowing for variability in Sections 2.3.1 and 2.3.2 has also shown that these contexts do not receive much attention in the variationist literature. This can be partly attributed to the fact that many of the contexts in which variation in the use of pronoun case forms is attested are probably considered to be too Standard-like to be discussed in surveys of non-standard features of varieties of English (cf. Section 2.3.2). Hence, there are hardly any studies comparing these contexts systematically across varieties of English. Furthermore, there are also accounts which seem to assume that the distribution of pronoun case forms in such contexts is more or less the same across varieties of English (cf. Burridge 2004: 118). Thus, this study will try to determine if there is indeed a general trend observable across varieties of English with regard to the distribution of pronoun case forms in subject predicative complements or if instead there are pronounced differences between the examined varieties.

H5: If data obtained from Google can really be used as basis for quantitative studies (e.g. Hundt, Nesselhauf and Biewer 2007b; Mair 2007), then we will obtain robust results from the Web-derived datasets specifically compiled for this study.

As noted above (cf. Section 5.1), due to the scarcity of subject predicative complements with a case-sensitive pronoun in closed corpora, five Web-derived datasets were compiled specifically for this study by means of a commercial Web crawler. However, whether or not Google data provides a sensible basis for corpus linguistic studies and thus should be used as input for quantitative linguistic studies is still an issue of ongoing debate (cf. Sections 7.2.2–7.2.3). In view of the fact that so far, there are hardly any studies that attempt to systematically use and refine Google data to examine a particular morphosyntactic phenomenon by means of multivariate statistical modelling, this study tries to assess whether such analyses using Web-derived data are actually possible. In particular, it is assumed that if Google data can be used for sound quantitative studies, we should not obtain results that are contrary to the results of a more or less comparable traditional closed corpus. Thus, by having a look at the data from both the British National Corpus and the data obtained from the top-Level domain of the United Kingdom, we should, despite all the
differences between these two datasets (cf. Chapter 7), be able to say whether the results of these two datasets seem more or less coherent or if they are totally inconsistent.
6 Variables and Statistical Modelling

As the discussion in the previous chapters has shown, a remarkable amount of research has already been devoted to the distribution of pronoun case forms in Present-Day English. However, as has also become evident, there are only few studies that have tried to quantitatively assess the distribution of these forms, which is, however, partly due to the difficulty of obtaining enough data (cf. Chapters 2 and 3). In view of the general scarcity of quantitative studies in the domain of pronoun case distribution, it is therefore hardly surprising that – at least to the knowledge of the author – there are no studies to date that have tried to assess the distribution of pronoun case forms in subject predicative complement position with the help of elaborate multivariate statistical modelling. In order to bridge this research gap and to answer the research questions as well as to test the hypotheses introduced in the preceding Chapter 5, this study intends to provide a thorough quantitative analysis of pronoun case distribution in subject predicative complements in six varieties of English. In particular, the present study tries to assess the correlation between the rather unexpected use of subject pronoun forms in subject predicative complement position following it and a form of BE (cf. Chapter 3), i.e. the dependent variable (cf. Section 6.1), and a set of independent variables introduced and discussed in Section 6.1.2. After introducing the independent variables, Section 6.1.3 briefly touches upon the difficulty of reasonably classifying them. Furthermore, Section 6.2 introduces the statistical models used to analyse the data as well as key notions necessary for the comprehension of the results reported in Chapters 8–14. Finally, Section 6.3 provides an interim summary of the major issues discussed in the present chapter.

6.1 The Variables: Description and Motivation

The following sections provide an outline and motivation of both the dependent (cf. Section 6.1.1) and the independent variables (cf. Section 6.1.2). Since the independent variable Focus introduced in Section 6.1.2.5 has not yet played a prominent role in the discussion of pronoun case forms in subject predicative complements following it and a form of BE, a
more thorough discussion will be provided. Section 6.1.2.7 provides a concise summary of the variables as well as their operationalisation.

6.1.1 The Dependent Variable

In light of the preceding discussion, the dependent variable of the analyses seems quite obvious. This study assumes that subject pronoun case forms have been re-functionalised as Focus markers in *it BE* sentences and in *it*-clefts (cf. Chapter 4). Thus, in order to answer the research questions and test the hypotheses postulated and outlined in Sections 5.1 and 5.2, this study is interested in the factors that may either promote or inhibit the use of subject pronoun case forms in subject predicative complement position. Hence, the occurrence of subject pronoun forms in subject predicative complement position is the dependent variable in this study. The operationalisation of this variable is presented in Table 8:

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUBJECT CASE FORM</td>
<td>Occurrence of subject pronoun form in subject predicative complement position</td>
<td>Nominal (dummy variable)</td>
<td>0, 1</td>
<td>SUBJECT: I, he, she, we, they OBJECT: me, him, her, us, them (reference)</td>
</tr>
</tbody>
</table>

*Table 8: The Dependent Variable of the Following Analyses*

Although the choice of pronoun forms in these contexts seems clearly to be a binary one between subject and object forms, a few remarks on the use of untriggered *self*-forms, i.e. *self*-forms without a pronominal antecedent, in subject predicative complement position are in order, since the use of these *self*-forms in the contexts examined in this study has also been attested (e.g. Erdmann 1978: 72; Harris 1981: 18; Hernández 2002: 269). Most authors consider the use of untriggered *self*-forms in contexts allowing for variability in general and in subject predicative complement position in particular as an avoidance strategy by speakers and writers in order not to be forced to choose between a subject and an object pronoun case form (e.g. Erdmann 1978: 72; Harris 1981: 18; Hernández 2002: 269; Wales 1996: 97, 104, 107). However, it has also been proposed that in some varieties the use of
untriggered *self*-forms is not merely an avoidance strategy but actually a possible alternative to the use of subject and object forms, which can be attributed to substratal or adstratal influence (e.g. Filppula 1999: 85-87; Filppula, Klemola and Paulasto 2008: 176). Irish English is one of the varieties in which untriggered *self*-forms are supposed to occur in subject predicative complement position in this way, particularly in *it*-clefts (Filppula 1999: 85-87, Odlin 1997: 44). However, although untriggered *self*-forms are supposed to be a viable alternative to subject and object forms in Irish English subject predicative complements in general and *it*-clefts in particular, it has been admitted even by proponents of such a position that *it*-clefts involving *self*-forms as focal pronouns are actually very rare in Irish English (e.g. Filppula 1999: 86). In fact, corpus analyses conducted by Beal (in prog.) and by the author of this study have shown that neither the ICE-Ireland corpus nor the *Northern Irish Transcribed Corpus of Speech* contain instances of untriggered *self*-forms in subject predicative complement position. In the historical *Hamburg Corpus of Irish English* (cf. Pietsch 2009), there is only one *it*-cleft with a *self*-form as focal pronoun, although it occurs in a coordination and with a case-insensitive second person pronoun (cf. Section 7.3):

(95) But instead of your letter it is yourself and Family I expected this twelve months Back (HCIE: Brogan01)

However, if untriggered *self*-forms were indeed a common competitor to object or subject pronoun forms in subject predicative complement position, we would expect to observe them more frequently or at least observe them, particularly considering that untriggered *self*-forms in other contexts are not only frequently attested but are also significantly more widespread in Irish English than in other varieties (cf. Siemund, Maier and Schweinberger 2012). On the basis of this evidence or, depending on the perspective, due to this lack of evidence, it is not advisable to take untriggered *self*-forms as potential alternative for subject and object forms in the contexts under examination into consideration. Even more importantly, however, untriggered *self*-forms are not on a par with subject and object pronouns because they are not case-sensitive themselves but are often considered to be an avoidance strategy to evade the decision between subject and object form (e.g. Erdmann 1978: 72; Harris 1981: 18; Hernández 2002: 269). Hence, although they may possibly be alternants, their lack of case-sensitivity causes them to be unfit for a further level of the dependent variable Subject Case Form in the present study. Furthermore, *self*-forms are not genuine remnants of the former Old English case system; as complex forms,
they are based on these remnants, but are not remnants themselves and thus are not really comparable to subject and object pronoun forms (cf. Section 2.1). Moreover, they are not discussed as alternatives in discussions about the distribution of pronoun case forms in subject predicative complements in reference grammars of English (e.g. Biber et al. 1999: 335-336; Huddleston and Pullum 2002: 459). As a consequence, untriggered self-forms were treated like any other noun phrase apart from uncoordinated subject and object pronoun case forms and thus have not been further considered for this study (cf. Section 7.3). Thus, in accordance with the principle of accountability, only subject predicative complements with either an object or a subject pronoun case form following it and a form of BE were considered for the analyses of the data, since the choice between subject and object case form is considered to be a closed set of variants in this study (cf. Labov 1972: 72; Tagliamonte 2012: 9-10).

6.1.2 The Independent Variables

The independent variables consist mainly of factors put forward in the body of literature and which are supposed to influence the choice of pronoun case forms, such as CONSTRUCTION, PERSON and NUMBER. Furthermore, on the basis of the survey of pronoun case across varieties of English in Chapter 2 and the Focus-oriented approach introduced in Chapter 4, the variables FOCUS and REGIONAL VARIETY are also included in order to be able to answer the research questions and test the hypotheses outlined in Chapter 5.

6.1.2.1 Construction

6.1.2.1.1 Inter-Construcational Variation: It BE Sentences versus it-Clefts

The first independent variable to be discussed is the construction type in which the pronominal subject predicative complement following it and a form of BE occurs. This can be either an it BE sentence (e.g. Quinn 2005a), i.e. a simple pronominal subject predicative complement consisting of it, a form of BE and a pronoun form, as in (96a) or an it-cleft, as in (96b):
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

(96)  a. Who is it? – It is I/me.
      b. It is I/me who knocks on the door.

Scholarly opinion on whether the difference between these two constructions or sentence types influences the distribution of the pronoun forms is divided. On the one hand, there are studies that stress the similarity between these two constructions in terms of pronoun case distribution. According to these accounts, both contexts or constructions clearly favour the object pronoun forms. This preference can be explained either in terms of positional considerations or in terms of different pronoun classes, where the former subject pronouns have become or are in the process of becoming clitics, or in terms of a combination of both (cf. Emonds 1986: 96-100; Harris 1981: 19-20; Hopkins 1975: 32-34, fn 3; Quinn 2005a: 138-139, 242-248 and Chapter 3 for a detailed overview of these accounts).

On the other hand, there are a number of accounts that observe marked differences between the distribution of pronoun case forms in it-clefts and it BE sentences (e.g. Erdmann 1978: 75-78; Maier 2013; Wales 1996: 95-99). Biber et al. observe, for example, that “[a]ccusative forms are predominant after the copula be. However, cleft constructions in fiction and news generally have nominative forms followed by who” (1999: 335). This higher share of subject pronoun forms in the focal position of it-clefts when compared to it BE sentences is also attested by other studies (e.g. Erdmann 1978; Siemund, Maier and Schweinberger 2009; Wales 1996), and it has often been explained by the special status of the focal pronoun in it-clefts. Biber et al. note, for example, that “[t]he nominative forms are presumably felt to be more correct, as they are typically coreferential with the subject of the following subordinate clause” (1999: 336). Hence, some accounts even go so far as to state that object forms in the focal pronoun position of it-clefts are “not acceptable” (Phythian 1980, quoted in Wales 1996: 95).

In view of this controversy, it seems more than justified to include this variable into the subsequent analyses in order to determine whether or not this variable significantly influences the choice of pronoun case forms in subject predicative complements following it and a form of BE. Furthermore, this variable is important for another reason: One of the central hypotheses of this study is that – due to the cross-varietal importance of concepts such as stress, emphasis, or Focus in the distribution of pronoun case forms (cf. Section 2.2.3) – the pragmatic status or, to be more precise, the ‘degree’ of Focus may also play a role in the distribution of pronoun case forms in subject predicative complements. In view of the fact that one of the main functions of it-clefts is to highlight or focus a certain
constituent (e.g. Biber et al. 1999: 958-963; Huddleston and Pullum 2002: 1414-1427), the inclusion of this variable seems necessary also from the point of view of another variable, i.e. FOCUS, which will be further discussed in section 6.1.2.5. The operationalisation of the variable CONSTRUCTION is as follows:

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTRUCTION</td>
<td>Occurrence of the pronoun in either an it BE sentence or an it-cleft sentence</td>
<td>Nominal (dummy variable)</td>
<td>0, 1</td>
<td>CLEFT: it-Clefts [it:it BE: it BE Sentences (reference)]</td>
</tr>
</tbody>
</table>

*Table 9: The Independent Variable CONSTRUCTION of the Following Analyses*

6.1.2.1.2 Intra-Constructional Variation: The Influence of the Independent Variables in the it BE Sentences and it-CLEFTs

Besides analysing the distribution of pronoun case forms in subject predicative complements in general, we will also examine each construction type, i.e. it BE sentences and it-clefts separately: Simple descriptive and bivariate analyses of the phenomena at hand indicate marked differences in the distribution of pronoun case forms between it-clefts and it BE sentences (e.g. Biber et al. 1999: 335-336; Erdmann 1978: 75-78; Maier 2013). Thus, it will be interesting to see if and how the independent variables identified and motivated for this study as well as possible interactions between them influence the distribution of pronoun case forms in each construction type. Hence, this will yield more fine-grained results for the analysis of the distribution of pronoun case forms in subject predicative complements. Furthermore, this approach seems advisable considering the fact that not all factors are applicable to both constructions to the same extent. The factor FOCUS discussed in Section 6.1.2.5, for example, is restricted in its operationalisation to it BE sentences (cf. Section 6.1.2.5).

Furthermore, there are variables that are deemed important for the distribution of pronoun case forms in it-clefts, which are, however, not applicable to it BE sentences. The most important of these cleft-specific variables is the co-reference of the clefted element, i.e. the syntactic function of the clefted or highlighted pronoun in the following dependent clause. If the focal pronoun of an it-cleft is co-referential with the subject of the following
dependent clause, it is deemed very likely to observe subject pronoun forms. If the focal pronoun is, however, co-referential with an object of the following dependent clause, it is more likely to observe object pronoun forms (e.g. Erdmann 1978: 75-78; Huddleston and Pullum 2002: 459; Quinn 2005a: 133-142; Quirk et al. 1985: 338; Wales 1996: 96). Hence, it will be interesting to see whether or not this assumption can also be confirmed for the *it*-clefs in the datasets used for this study. The operationalisation of the variable CO-REFERENCE is given below:

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO-REFERENCE</td>
<td>Co-reference of the clefted pronoun in the dependent (cleft) clause with either the subject or an Object</td>
<td>Nominal (dummy variable)</td>
<td>0, 1</td>
<td>AS_SUBJ: clefted pronoun is co-referential with the subject of the following clause AS_OBJ: clefted pronoun is co-referential with an/the object of the following clause (reference)</td>
</tr>
</tbody>
</table>

*Table 10: The *it*-Cleft-Specific Independent Variable CO-REFERENCE of the Following Analyses*

In addition to the co-reference of the clefted element with either the subject or an object of the dependent clause, the choice of the relative pronoun is another cleft-specific variable often mentioned as possibly influencing the distribution of pronoun case forms in *it*-clefs. According to some accounts, the subject personal pronoun forms are more likely to occur if the focal or focussed pronoun is followed by the relative pronoun *who*, whereas object pronoun forms are more likely to occur when the dependent clause is introduced by *that* or a *zero* relative pronoun (e.g. Biber et al. 1999: 335-336; Wales 1996: 95-96). Furthermore, some studies assume significant interactions between the variables RELATIVE PRONOUN and MODE OF DISCOURSE (e.g. Biber et al. 1999: 336; Quirk et al. 1985: 338-339; Wales 1996: 96). However, since the choice of the relative pronoun is, of course, dependent on the syntactic function of the clefted pronoun in the following dependent clause, we can predict a high degree of multicollinearity between the two cleft-specific variables CO-REFERENCE and RELATIVE PRONOUN even before conducting any multivariate analyses (cf. Section 6.2 for a more detailed discussion of the phenomenon of multicollinearity and the problems arising from it). And indeed, pilot analyses confirm this assumption. Thus, this second cleft-specific variable is not taken into consideration for the subsequent statistical analyses.
In sum, we can record that with regard to the variable CONSTRUCTION the approach of this study is two-fold. First, this study analyses the superordinate subject predicative complements category in order to test the central research questions and hypotheses, as outlined in Chapter 5. In a second step, this study conducts more fine-grained analyses of the individual construction types, i.e. it BE sentences and it-clefts, to determine how the results obtained for the individual constructions relate to the assumptions and hypotheses of previous research as well as to the assumptions and hypotheses of this study formulated in the two preceding Chapters 4 and 5. In this manner, this study will attempt to obtain the best possible results and a maximally accurate picture of the distribution of pronoun case forms in it-clefts and it BE sentences.

6.1.2.2 Person

The second independent variable to be tested in the subsequent analyses of the different datasets is the person of the pronoun, i.e. the potential distributional differences between first and third person pronouns in subject predicative complements. For this variable, a survey of the body of literature yields a rather complex picture.

With regard to it BE sentences, Quinn, who generally notes a clear prevalence of object forms in these contexts (e.g. 2005a: 246, 2009: 41), states that “[n]on-1sg nominatives would seem to be extremely rare, and restricted to certain discourse contexts” (2005a: 135).

A slightly different picture is depicted in Wales (1996). Although this account also asserts a clear dominance of object pronoun forms in it BE sentences, the only authentic examples that are provided to illustrate the use of subject forms in this context are third person pronouns (Wales 1996: 95):

(97) a. ‘Oh look Dad, there’s John Cleese’, he said excitedly. It was, of course, he. (The London Evening Standard, 11 February 1993)

b. Could the blunderer from the Sketch be related to the author of last week’s … piece in the Mail? Michael Kemp admitted to the Eye last week that it was indeed he. (Private Eye, 28 August 1992)

As far as the distribution of pronoun case forms in it-clefts is concerned, both accounts agree that third person subject forms are more likely to occur than first person subject forms in this context (Quinn 2005a: 135-140, 2009: 42; Wales 1996: 95-96). Thus, since there have
not been any thorough statistical analyses of the distribution of pronoun case forms in these contexts, it seems promising to assess whether or not there are indeed significant differences in the data as far as the variable PERSON is concerned. Furthermore, in view of the proposed differences between it-clefts and it BE sentences with regard to the probability of observing first or third person subject pronoun forms, it seems also promising to check whether or not there are significant interactions between the variables CONSTRUCTION and PERSON at work and if they indeed influence the distribution of pronoun case forms in subject predicative complements.

Except for these rather straightforward reasons, there are also other, more indirect, arguments that suggest that the analysis of this variable will be beneficial. From the domain of coordinated noun phrases (cf. Section 2.3), it has been reported that there are different distributions of pronoun case forms for different persons observable, particularly for the first person singular and the third person singular (cf. Angermeyer and Singler 2003: 178, 195). In addition, the survey of pronoun case forms across varieties of English, particularly the section on pidgin and creole varieties in Section 2.2.2, has also demonstrated that there may be noteworthy differences in case marking and distribution of pronoun case forms between the first and third person pronouns.

Furthermore, the potential relevance of this variable can be motivated even from a different perspective. From the work of typologists, cognitive linguists and functional linguists in general, it is a well-known fact that there are marked differences between first and third person pronouns in several important domains, such as number marking, gender marking, text frequency and cognitive accessibility (e.g. Ariel 2008: 44-53; Croft 2003: 160-162; Langacker 1991: 306-309; Siewierska 2004: 5-8, 46). These differences may even include differences in case marking (Siewierska 2004: 5-8). Moreover, first and third person pronouns differ from each other insofar as their prototypical referential status is concerned. On the one hand, first person pronouns always denote referents that are actively involved in a communicative situation. Thus, they are more prototypically deictic than third person pronouns because the referents of first person pronouns are, as speaker(s), immediately related to the persons involved in the respective communicative situation. On the other hand, although third person pronouns can also be used deictically, their prototypical use is phoric, i.e. mainly referring back to referents introduced in the prior discourse (e.g. Siewierska 2004: 7). Therefore, the relevance of the first person pronouns and their
referents for a current communicative situation is automatically given by the use of them, which identifies them and their referents as important information for the hearers and readers. The referents of third persons, however, are less easily accessible for hearers and readers, since the set of potential referents is of course much bigger and thus their referents are less salient. This crucial difference between first and third person pronouns is also evident in Langacker’s empathy hierarchy, a variation of Silverstein’s (1976) definiteness hierarchy, which ranks entities or referents in the world according to how likely it is that they receive our attention and empathy (Langacker 1991: 307):

    speaker > hearer > human > animal > physical object > abstract entity

Thus, the referents of first person pronouns are most likely to receive our interest and empathy, while the referents of third person pronouns are less likely to receive our attention and empathy (Langacker 1991: 307; Schmid 2007: 132-133). Moreover, these differences may also bear out on the grammatical structure (Langacker 1991: 307). Bearing in mind that this study assumes that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complement contexts, these differences between first and third person pronouns may be important, since these differences are of a pragmatic as well as a grammatical nature. Thus, it is possible that the pragmatic differences inherent in the pronominal paradigm itself may also influence the distribution of pronoun case forms in subject predicative complements.

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
</table>
| PERSON        | Occurrence of a pronoun in a subject predicative complement construction in either first or third person | Nominal (dummy variable) | 0, 1 | FIRST: I, me, we us  
               |             |               |        | THIRD: he, him, she, her, they them (reference) |

Table 11: The Independent Variable PERSON of the Following Analyses
6.1.2.3 Number

Another variable put forth in the literature as potentially influential in the choice of pronoun case forms after *it* and a form of *BE* is the number of the pronoun, i.e. the difference between singular and plural pronoun forms (e.g. Sobin 1997; Quinn 2005a).

However, compared to other possible predictors, this variable has so far received relatively little attention. Indeed, many studies focus on the distribution of singular pronoun forms, which is often already indicated by the title of the respective studies (cf. Erdmann 1978; Harris 1981). A good example of this bias in favour of the singular pronoun forms is Biber et al. (1999). Although their account is one of the few that actually try to quantify the distribution of pronoun case forms in subject predicative complements, particularly for *it*-clefts and *it BE* sentences, they unfortunately restrict their analysis to first and third person singular forms and do not take possible differences between singular and plural pronouns into consideration (cf. Biber et al. 1999: 335-336).

Furthermore, those few accounts that actually discuss potential differences in the distribution of pronoun case forms between singular and plural forms in subject predicative complements, are again at variance. With regard to *it BE* sentences, there are accounts that not only state that singular pronouns are more likely to occur in their subject form but even explicitly rule out the possibility of plural subject forms occurring in this construction (Sobin 1997: 334). For *it*-clefts, the use of plural subject forms is not generally ruled out, but is still deemed questionable (Sobin 1997: 334). However, these claims have not been corroborated by other studies to date (e.g. Quinn 2005a: 135; Wales 1996: 96). This may be due in part to the fact that some studies seem to consider the differences between first and third person as the more noteworthy ones (e.g. Quinn 2005a, 2009; Wales 1996).

Although Sobin (1997) tries to explain the distribution of pronoun case forms by means of grammatical viruses (cf. Section 3.1), another possible explanation for this assumed difference between singular and plural pronoun forms is also conceivable. In analogy to the distinction between first and third person pronoun case forms, possible differences between singular and plural in the distribution of pronoun case forms in subject predicative complements could also be accounted for in principle in terms of an implicational hierarchy (cf. Croft 2003: 126-132).

(99)  Croft’s number hierarchy (Croft 2003: 126):

  singular < plural < dual < trial/paucal
Thus, possible differences between singular and plural pronouns in the distribution of pronoun case forms could also be due to the typological differences between these two number categories, much like those outlined in the previous section for the person distinctions (cf. Section 6.1.2.2).

In view of these diverging positions, it may prove illuminative to include the variable NUMBER in this analysis to assess the actual influence of this variable on the distribution of pronoun case forms in subject predicative complements. Furthermore, this inclusion will also help to determine whether or not there are any interactions with other variables, such as the sentence type in which the pronoun occurs (cf. Sobin 1997).

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER</td>
<td>Occurrence of a pronoun in a subject predicative complement construction in either singular or plural</td>
<td>Nominal (dummy variable)</td>
<td>0, 1</td>
<td>SINGULAR: I, me, he, him, she, her PLURAL: we, us, they, them (reference)</td>
</tr>
</tbody>
</table>

*Table 12: The Independent Variable NUMBER of the Following Analyses*

### 6.1.2.4 Mode of Discourse

Probably the most frequently mentioned factor as potentially influencing the distribution of pronoun case forms in subject predicative complements is the degree of formality. As was already addressed in Chapter 3, most accounts discussing the distribution of pronoun case forms assume that the distinction between formal and informal language exerts a strong influence on the distribution of pronoun case forms (e.g. Angermeyer and Singler 2003: 108-187; Biber et al. 1999: 335-336; Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Wales 1996: 91-108). Again, a good example to illustrate the importance attributed to this variable is Quirk et al. (1985). Their approach has already been discussed in Section 3.2 as an example of what has been called the positional approach, but this is not the whole truth, since the distinction between formal and informal usage is another crucial factor to explain variability in the use of pronoun case forms in their account (Quirk et al. 1985: 336-339).
They begin their description of the distribution of pronoun case forms in English by stating that

“[t]he choice between subjective and objective case forms is made on the basis of a pronoun’s function in the clause. As their name implies, subjective personal pronouns function as subject and sometimes as object complement; objective personal pronouns function as object, prepositional complement and sometimes as subject complement” (Quirk et al. 1985: 336).

However, in the course of their discussion, they are forced to admit that this structural explanation is restricted to the more formal varieties of English:

“This purely structural explanation, however, is not the whole story. To account for the general pattern of pronoun usage in informal style, it is reasonable to say that the traditional case distinctions do not operate here any more than they do with the genitive [...]. Instead, there is a broad division of the finite clause into ‘SUBJECT TERRITORY’ (the preverbal subject position) and ‘OBJECT TERRITORY’ (which includes all noun phrase positions apart from hat immediately preceding the verb)” (Quirk et al. 1985: 337).

Thus, depending on the degree of formality, Quirk et al. (1985) provide two different explanations for the distribution of pronoun case forms in Present-Day English: a structural one for the formal registers and a positional one for the informal registers. In a similar vein, other authors have also noted the distinction between formal and informal varieties as an important factor influencing the distribution of pronoun case forms in subject predicative complements (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Sweet 1875: 495).

Hence, this study tests whether different modes of discourse, i.e. spoken data vs. written data vs. computer-mediated communication, or CMC for short, exhibit different distributions in terms of pronoun case forms.

Assessing potential differences in formality on the basis of this coarse-grained mode of discourse distinction may, however, evoke some scepticism at first because it has been shown that there is no clear division between all spoken and all written language. Indeed, the most marked differences may not be observed simply between speech and writing but between planned and unplanned speech and writing (e.g. Biber 1988; Miller 2006: 671-673). In addition, recent corpus studies have demonstrated that fine-grained data-driven analyses may also reveal remarkable sub-register differences with regard to the distribution of certain
linguistic phenomena which have so far been rather neglected (e.g. Gries 2006, 2010; Sand 2004). However, although such a fine-grained data-driven register analysis would in principle be desirable, the situation in this study is much more complex (cf. Section 7.2.4.3). Whereas Gries (2006), for example, analyses the distribution of relatively frequent features such as present perfect and ditransitive constructions in homogenous corpora that were compiled in a controlled top-down fashion, this study deals with extremely low-frequency phenomena in heterogeneous datasets, some of which are compiled and also classified in a bottom-up fashion (cf. Section 7.2.4.3). Thus, a data-driven analysis would meet with a lot of difficulties. This is true not only for the BNC and the COCA but also and particularly for the Internet datasets (cf. Chapter 7). The Internet data is subdivided into more than 20 different subgenres, which in turn could easily be further subdivided into more-fine-grained subclasses (cf. Section 7.2.4.3). For some datasets, particularly the Irish and Indian ones, a data-driven analysis would yield a subdivision into such small data groups that the result could hardly be considered representative of the respective registers and text types, as to allow for reliable generalisations (cf. also Section 7.5). Thus, in order to ensure a certain degree of comparability between the different datasets and the results obtained from them, a basic categorisation seems to be more favourable in the present case.

Furthermore, this analysis of formality based on a coarse-grained mode of discourse distinction can also be motivated by other factors. Although the formal–informal and the spoken–written distinctions can be conceived of as parts of a multidimensional continuum with overlapping areas (e.g. Crystal 2011: 19), it is still assumed that the “spoken register is generally more informal than the written register [… ” (McEnery, Xiao and Tono 2006: 265). This is not only due to the mode of transmission, but also to differences in other crucial aspects in which speech and writing prototypically differ from each other, such as the degree of interactiveness and online production, the intended audience, the dialectal domain, the main communicative purpose and the shared immediate situation (e.g. Biber et al. 1999: 16).

Moreover, the basic mode of discourse distinction employed for the present study is also reflected in the relevant literature (cf. Hopkins 1975: 28; Wales 1996: 101). Other very important reasons for using a simple mode of discourse differentiation to assess differences in formality are the results presented by Biber et al. (1999: 336) for the distribution of pronoun case forms in subject predicative complements following it and a form of BE. The results they obtain from the Longman Corpus of Spoken and Written English show –
particularly for *it*-clefs – marked differences in the distributions of pronoun case forms between the spoken and the written data. However, the two written registers distinguished in Biber et al.’s study differ from each other not as considerably as they differ from the spoken data (1999: 336). In addition, these results are corroborated by another rather coarse-grained study of pronoun case distribution in subject predicative complements. Particularly with regard to *it*-clefs, Maier (2013) observes significant differences in the distribution of pronoun case forms between the spoken and the written subsets of both the *British National Corpus* and the *Corpus of Contemporary American English* (Maier 2013). These findings also suggest possible interactions between the mode of discourse and the construction in which a pronoun occurs. Finally, this distinction between spoken and written discourse is not only employed in the context of subject predicative complements (e.g. Biber et al. 1999; Hopkins 1975; Wales 1996), but also in studies addressing other contexts exhibiting variability in the use of pronoun case forms, such as coordinate noun phrases (e.g. Angermeyer and Singler 2003: 181).

Although the discussion of this variable has so far focussed only on the distinction between two modes of discourse, the nature of the data used for this study necessitates – depending on the respective dataset – a distinction of up to three different discourse modes. For the *British National Corpus* and the *Corpus of Contemporary American English*, this study employs a traditional binary distinction between written and spoken data, whereas the Web-derived datasets require a tripartite distinction between spoken data, written data and CMC data, i.e. the text-based computer-mediated data comprising genres such as blogs, forum posts, etc. (cf. Section 7.2.4.3.3) that are closely linked to the emergence of the Internet as a new mass media and means of mass communication (e.g. Crystal 2011; Herring 2011). In the relevant literature, it has become traditional to consider CMC as a discourse mode of its own or even as a kind of middle ground or hybrid between the spoken and the written modes of discourse (e.g. Crystal 2011: 19-35; Herring 2010b). From early on, scholars assigned labels such as “written speech” (Maynor 1994) or “visible speech” (Colomb and Simutis 1996) to CMC in order to point out that it shares some features with both traditional modes of discourse but at the same time cannot be identified with either one of them (Crystal 2011: 32; Herring 2010a). With regard to the intermediateness of the CMC mode it has been pointed out that CMC
“[...] shares numerous characteristics with (informal) spoken conversation – e.g., typographic practices that imitate spoken prosody; discourse produced in chunks that resemble ‘intonation units’ [...] ; turn-taking; topic development via step-wise moves [...] ; ‘conversational code-switching [...] – and moreover that CMC fulfils many of the same social functions as spoken conversation” (Herring 2010a).

In addition, CMC may also differ from the other two modes of discourse with regard to grammar and morphology. These differences affect primarily the domains of orthography, morphology and syntax and include an overall higher likelihood of exhibiting non-standard or variety-specific features when compared to the traditional written mode of discourse (Herring 2011). In view of these possible differences, it is justified to assume that there may also be differences in the distribution of pronoun case forms in subject predicative complements observable between the CMC data and the other two modes of discourse. Thus, the distinction of different modes of discourse to assess potential differences with regard to formality in the distribution of pronoun case forms seems not only justified but also very promising considering the importance much of the literature has attributed to this factor. Moreover, it will be interesting to see whether there are indeed interactions with other variables as suggested by former coarse-grained studies (Biber et al. 1999: 336; Maier 2013). Finally, the operationalisation of this variable looks as follow:

<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODE OF DISCOURSE</td>
<td>Occurrence of a pronoun in a subject predicative complement construction in either spoken or computer-mediated or written discourse</td>
<td>Nominal (dummy variable)</td>
<td>0, 1</td>
<td>SPOK: spoken discourse CMC: computer-mediated discourse WRIT: written discourse (reference)</td>
</tr>
</tbody>
</table>

Table 13: The Independent Variable MODE OF DISCOURSE of the Following Analyses
6.1.2.5 Focus

In accordance with the Focus-oriented approach introduced in Chapter 4 and the central hypotheses outlined in Chapter 5, this study also tries to analyse to what extent the distribution of pronoun case forms can be explained in terms of pragmatic functions, particularly with regard to Focus, since this seems to play a prominent role in the pronoun case distribution in several contexts as has been outlined, for example, in Sections 2.2.3 and 3.4 (cf. also Paddock 1994). The motivation and explanation for how and why subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements has already been thoroughly dealt with in Chapter 4.

However, despite all evidence for the assumption that Focus may strongly influence the choice of pronoun case forms in subject predicative complements, a very important issue still has to be addressed, namely the operationalisation of this variable. In Section 4.1, it has been stated that the definition of Focus as the “relatively most salient or important information” is vague (e.g. Siewierska 1991: 174) and that in written contexts, much depends on the reader’s assessment, since one of the most important focussing devices, i.e. prosodic prominence, is not available (Siewierska 1991: 175). In view of these difficulties, many quantitative studies in other subject areas where Focus is believed to play a role have refrained from analysing pragmatic factors, although the analysis of such an influence in future research has been suggested (e.g. Lohmann 2011: 31). Moreover, this difficult point of departure is further complicated by the fact that nearly all pronouns in subject predicative complement positions can be considered as focussed, both in it BE sentences and it-clefts. In it BE sentences, the postverbal pronoun form can clearly be interpreted as the “relatively most salient or important information” (e.g. Siewierska 1991: 174). In fact, it has even been referred to as “focus pronoun” in analogy to it-clefts (Quinn 2005a: 242). In it-clefts, the personal pronoun following it and a form of BE is by definition focussed, as the main function of the construction is to highlight or focus the clefted constituent (e.g. Biber et al. 1999: 959; Hedberg 1990: 7; Weinert and Miller 1996). Thus, in view of these facts, the operationalisation of this variable seems Gordian.

However, despite these difficulties, it should still be possible to operationalise this variable, since different constructions exhibit different degrees of Focus. This is due to the fact that there are special focussing constructions, the main function of which is to highlight a certain clausal constituent as being relatively the most important one (e.g. Siewierska
Hence, these special Focus constructions should exhibit differences in the distribution of pronoun case forms when compared to their unmarked counterparts if the hypothesis that Focus marking influences the distribution of pronoun forms is correct.

A first step to approach this factor is to distinguish between *it*-clefts and *it* BE sentences. Although the focal pronouns of *it*-clefts and *it* BE sentences seem superficially to be quite similar, these two constructions differ not only syntactically but also pragmatically from one another, which will be exploited for the subsequent analyses. It is widely acknowledged that *it*-clefts are a special focussing device both in speech and writing used to mark a certain constituent as particularly prominent (e.g. Biber et al. 1999: 959-963; Collins 1991: 1-6; Dik 1997: 291; Quirk et al. 1985: 1383-1387). By breaking up a single clause (Biber et al. 1999: 958), cleft sentences make use of what Quirk et al. (1985: 1384) call “end-focus within an SVC clause” to highlight a particular constituent. Although this “end-focus” can also be assumed for the pronouns in *it* BE sentences, there are still very important differences between the pronouns in each sentence type. Firstly, *it*-clefts are formed deliberately in order to achieve this particular “end-focus” effect. Hence, the speaker or writer – by forming the *it*-cleft – explicitly wants the pronoun to be focussed. This is evident in (100a) and (100b) below, where the same sentences could be easily expressed by means of a declarative sentence conveying the same propositional content. However, in each case, an *it*-cleft is chosen to highlight that it was the person referred to by the focussed pronoun and no one else who is responsible for the action described in the dependent clause in (100a) and (100b). In the *it* BE sentences in (101a) and (101b), on the other hand, this “end-focus” results more or less automatically from the required constituent order.

If Focus influences the choice of pronoun forms in subject predicative complements, one can expect to observe different distributions for these two constructions because *it*-clefts are more deliberately formed in order to highlight a certain constituent. Hence it is likely that
other highlighting devices, i.e. the use of subject pronoun forms as Focus markers, might also occur more frequently in this particularly focussed context (cf. Section 4.1).

A second important factor that corroborates the assumption that the pronominal subject predicative complements in it-clefts differ from those in it BE sentences in terms of their pragmatic status is the “Janus-like status” of the focal pronoun in it-clefts (Quirk et al. 1985: 338). Quirk at al. use this term in order to illustrate the uncertainty about the grammatical function of the focal pronoun in it-clefts being both subject complement with respect to it and BE and also prototypically subject or object with regard to the following verb in the dependent clause (Quirk et al. 1985: 338). However, this Janus-like status can be assumed in grammatical as well as pragmatic terms. While the focal pronoun clearly constitutes the Focus of the main clause, i.e. the relatively most important piece of information, it can also often be considered as the Topic of the following dependent clause, i.e. the entity the utterance or the dependent clause is primarily about (Siewierska 1991: 149). This becomes particularly apparent in “contact clauses” (Filppula 2004: 84), i.e. subject relative clauses with a zero relativiser, such as those given in (102):

(102) a. I don't want you here. It's Gerry I want. It 's him should've been coming home tonight. " (BNC/CEH/W_fict_prose)
b. What do you think? PORE: Again, I just wonder if it was they found themselves high and dry. (COCA/SPOK/CNN_King)

Although examples like these have been excluded from the statistical analyses (cf. Section 7.3), the sentences in (102) illustrate this Janus-like status of the focal pronoun – both in syntactic and pragmatic terms – very nicely because the focal pronoun functions as the crucial hinge that connects both clauses. Thus, the pronoun in it-clefts is not only focussed but also has an outstandingly prominent status both in syntactic and in pragmatic terms.

Thirdly, the focal pronouns in it BE sentences and it-clefts may very often – though not always – be distinguished in terms of the type or quality of Focus they bear. Accounts of Functional Grammar acknowledge a primary split in the domain of Focus, i.e. a distinction between contrastive and completive Focus (e.g. Dik 1989: 282-285; Siewierska 1991: 176-177, 2004: 160-162). On the one hand, completive or non-contrastive Focus often designates information that is intended to fill a gap in the pragmatic information of the addressee. This prototypically also includes the Focus determined on the basis of the question and answer tests illustrated in Section 4.1. Contrastive Focus, on the other hand, is considered to apply to information that the speaker or writer considers to be directly opposed to a set of
possible alternatives that may be presupposed by the addressee (Siewierska 1991: 177-178).

With regard to subject predicative complements, it BE sentences can very often be associated with completive Focus while it-clefts can be more readily associated with contrastive Focus as illustrated in (103) and (104).

(103) a. "Alain?" At the sound of his enraged voice Jenna's legs almost gave way beneath her. There was a feeling of relief -- that was true -- that it was Alain and not some midnight prowler, but also there was a wonderful floating feeling of joy just to hear his voice, no matter how angry he sounded. "Of course it is me!" he roared. (BNC/HGD/W_fict_prose)
b. Who is it that's raised people's taxes? It's him. (COCA/SPOK/Ind_Limbaugh)

(104) a. Freed also said that it was he rather than Duna, who had taken a 90% stake in Hungary's largest circulation English-language [...]. (BNC/CSV/W_non_ac_tech_engin)
b. SCALIA: It was Al Gore who made it a judicial question. It was he who brought it into the Florida courts. We didn't go looking for trouble. (COCA/SPOK/CBS_Sixty)

However, although the examples in (103) and (104) may suggest that It BE sentences can be correlated with completive Focus and it-clefts with contrastive Focus, such a rigid binary distinction would be too simplistic. As we will see below, there are also it BE sentences that can be clearly considered as bearing contrastive Focus. Moreover, with regard to it-clefts, it is widely acknowledged that they are not expressing contrastive Focus per se and that indeed a remarkable proportion of it-clefts does not bear contrastive Focus (e.g. Weinert and Miler 1996: 199-205). However, despite the considerable share of non-contrastive instances (e.g. Weinert and Miller 1996: 200), it-clefts are still considered to be “typically contrastive” (Biber et al. 1999: 962). This is also corroborated by the fact that special constructions overtly encoding contrastive Focus in other languages tend to be translated with the help of it-clefts (Siewierska 1991: 176-177). Prototypical examples of it BE sentences, however, rather seem to employ completive Focus. Good indicators supporting this view are the examples given in common English grammar books:

(105) a. A: Who is there? B: It's I/me. (Quirk et al. 1985: 337)
b. Who spoke first? Was it all sure it was he? (Biber et al. 1999: 35)
c. Who said that? ~ (It was) her. (Swan 2005: 428)

Following the definitions given above, the examples in (105) can be associated with completive rather than with contrastive Focus. Furthermore, the citation of these examples in reference grammars of English marks them as somewhat prototypical for this construction. Thus, although there are exceptions, it can be assumed that it BE sentences
and *it*-clefs also tend to differ in the quality of Focus that is assigned to their respective focal pronouns.

In view of the preceding discussion, it is assumed that focal pronouns of *it*-clefs should be prototypically more prominent or more focussed than their counterparts in *it BE* sentences and thus may exhibit different distributions of pronoun case forms. However, although the assumed difference between *it*-clefs and *it BE* sentences in the use of pronoun case forms is a necessary condition to attribute a certain influence on pronoun case assignment in subject predicative complement position to the concept of Focus, a simple binary distinction between *it*-clefs and *it BE* sentences is not sufficient. The potential differences could also be due to the syntactic differences between these two constructions, i.e. they could be caused by the syntactic “Janus-like status” of the pronouns in *it*-clefs (Quirk et al. 1985: 338) and thus this variable here would more or less coincide with the **Construction** variable outlined in Section 6.1.2.1. Hence, in order to really unravel the potential influence Focus might have on the distribution of pronoun case forms, a further subdivision is required, namely between “normal” *it BE* sentences and particularly emphatic or focussed ones. Although the concept of Focus is itself very difficult to pin down, and despite the fact that the identification of Focus in a clause without the help of prosodic prominence is even more difficult, the identification of *it BE* sentences with particularly focussed pronouns should still be possible by looking for sentence structures in which the pronoun form is made particularly prominent. Thus, as a sufficient condition, a subdivision is needed between unmarked, i.e. “normal”, *it BE* sentences and highlighted, i.e. clauses with a particularly focussed pronoun forms, in order to assess the actual influence of the factor **Focus** on the distribution of pronoun case forms in subject predicative complements.

Before this study identifies and describes types of what are particularly focussed *it BE* sentences, it has to be remarked that the subsequent distinction does not imply that only *it BE* sentences of these categories can be considered as being particularly focussed or as receiving special Focus. Rather, it is suggested that the two classes described below are usually more focussed than prototypical examples of *it BE* sentences (cf. (103)) due to a number of peculiarities, which are outlined below. Hence, simply for the sake of being able to operationalise the variable **Focus**, this study restricts the notion of particularly focussed *it BE* sentences to the subclasses of *it BE* sentences introduced and discussed below.
For the category of marked \textit{it} \textit{BE} sentences, two clause types or subtypes can be isolated in which the personal pronoun can be considered as particularly focussed. Examples for the first clause type in this category of particularly focussed \textit{it} \textit{BE} sentences are given in (106):

(106) a. And if there is anyone who made it possible for Tony Blair to soar, it is she. (COCA/MAG/Bazaar)

b. If anybody owns that character, it’s Marlon Brando. If anybody can do what they like with it, it’s him. (BNC/ACP/W-pop_lore)

c. That is Jamie Forrester. If anyone is a replacement for Wallace it is he (due to him making Strach look TALL). (BNC/J1H/W_email)

In (106), the clauses in which the focal pronouns occur are syntactically \textit{it} \textit{BE} sentences. However, they exhibit certain properties which set them apart from other \textit{it} \textit{BE} sentences, such as those in (103) and (105). The first characteristic that clearly distinguishes this type of marked \textit{it} \textit{BE} sentences from the more prototypical, unmarked ones in (103) and (105) is their syntactic structure. Sentences of the first particularly focussed category are preceded by an \textit{if}-clause in which the subject is either a non-assertive pronoun or a noun phrase modified by a non-assertive determiner which is then followed by an \textit{it} \textit{BE} sentence serving as the matrix clause. Very rarely, assertive pronouns or determiners may be used instead of non-assertive ones in this type of particularly focussed \textit{it} \textit{BE} sentence. Another particularity with regard to their sentence structure is that the clausal sequence of the sentences in (106) seems more or less fixed:

(107) a. And if there is anyone who made it possible for Tony Blair to soar, it is she. (COCA/MAG/Bazaar)

b. If anybody owns that character, it’s Marlon Brando. If anybody can do what they like with it, it’s him. (BNC/ACP/W-pop_lore)

c. That is Jamie Forrester. If anyone is a replacement for Wallace it is he (due to him making Strach look TALL). (BNC/J1H/W_email)

(108) a. If anybody can do what they like with it, it’s him. (BNC/ACP/W-pop_lore)

b. If anybody, it’s him who can do what they like with it.

c. If anybody can do what they like with it.

(109) a. If anyone is a replacement for Wallace it is he (due to him making Strach look TALL) (BNC/J1H/W_email)

b. If it is he (due to him making Strach look TALL), if anyone is a replacement for Wallace.

c. If anyone is a replacement for Wallace.

Although subordinate clauses may in principle be ordered initially, finally or even in the middle of their superordinate clauses (e.g. Quirk et al. 1985: 1037), this does not seem to be possible – at least to the same degree – with these marked \textit{it} \textit{BE} sentences as the somewhat
awkward re-orderings of these sentences in (107b)–(109b) indicate. It seems that the if-conditions have to precede the it BE sentences as a kind of scene-setting topic that lays the foundation for the interpretation of the following it BE sentence (cf. Diessel 2004: 168), because postposing the complete if-clause after the it BE sentence does not seem to achieve the same effect as preposing it. However, the following example illustrates that this more or less fixed structure cannot solely be attributed to the combination of if-clause and non-assertive pronoun or modifier:

(110) a. If anyone ever says that, pretend not to hear (Quirk et al. 1985: 784)
    b. Pretend not to hear, if anyone ever says that.

Thus, it seems that these focussed it BE sentences in (106) also seem to differ from normal if-clauses as illustrated in (110). A possible explanation for these peculiarities distinguishing these marked it BE sentences both from unmarked ones and from more prototypical if-sentences as in (110) can be deduced from the reformulations of the clauses (107a)–(109a) in (107c)–(109c). In (107c)–(109c), if followed by the non-assertive pronoun or determiner is interpreted as an elliptical conditional clause, a frequently occurring grammatical phenomenon11, and the remaining sentence parts are rephrased or reinterpreted as it-cleft. When comparing the sentences in (107b)–(109b) to those in (107c)–(109c), the latter ones seem to be the more appropriate reformulations of the sentences in (107a)–(109a). Thus, these particularly marked it BE sentences in (107a)–(109a) could be interpreted as a blend of if-clauses and it-clefts, the cleft sentence of which precedes the focussed element. Although the dependent cleft clause in these particularly focussed it BE sentences seems to have syntactically merged with the ellipted if-clause, the predicate of the if-clause still seems to be more ‘closely’ related to the focussed pronoun of the it BE sentence than to the subject of the if-clause, as the comparison of (107b)–(109b) with (107c)–(109c) clearly indicates. Hence, these marked it BE sentences could be interpreted as reversed it-clefts, the dependent clauses of which have fused with the preceding scene-setting elliptic if-clauses. That it is possible, in general, for a focussed constituent to be preceded by a dependent clause is a well-known fact from the class of wh-clefts (e.g. Biber et al. 1999: 959-963; Collins 2006). Thus, the particularly focussed status of the pronouns in these marked it BE sentences can be partly explained by their similarities to both wh-clefts and it-clefts.

However, this first class of marked *it BE* sentences also exhibits other properties that makes them similar to *it*-clefts in some respects.

Firstly, in more pragmatic terms, it can be noted that the focal pronouns *she* in (107a) and *him* in (108a), for example, are not only the Focus of the matrix clause but they can also be considered as the Topic of the preceding subordinate clauses, i.e. the entity the utterance is primarily about (cf. Siewierska 1991: 149). In both sentences, the predication in the *if*-clause structure is more or less ruled out for anybody but the person identified by the focal pronoun in the *it BE* sentence.

Secondly, the first type of particularly focussed *it BE* sentences represented by the examples in (106) also resembles *it*-clefts with regard to the quality of Focus that it denotes. As stated above, prototypical *it BE* sentences can often be associated with completive Focus, i.e. information intended to fill a gap in the information of the addressee. This is also partly the case in the examples in (106). In (106a), for example, *it is she* can on the one hand be clearly considered as the completive Focus of the more or less implicit question *Who made it possible for Tony Blair to soar?*. On the other hand, the Focus in this type of *it BE* sentences can also be considered as clearly contrastive, as the focussed pronoun is directly opposed to a range of alternatives presupposed by the *if*-clause and the non-assertive pronoun in the subordinate clause (Siewierska 1991: 177), which becomes particularly apparent when considering the reformulation in (107c). Thus, *she* and *him* in (106a) and (106b), for example, are clearly contrasted with the potential alternatives evoked by *anyone* and *anybody* in their respective dependent *if*-clauses. Hence, this is another aspect which seems to distinguish this particularly focussed type of *it BE* sentence from the more prototypical unmarked type (cf. (103) and (105)).

Thirdly, these particularly focussed *it BE* sentences are also similar to *it*-clefts in that they resemble what other studies term “reduced cleft sentence”, i.e. clefts with omitted cleft clauses (cf. Weinert and Miller 1996: 201). Although reduced cleft sentences are not a category employed in this study, since they are very hard to distinguish from ordinary *it BE* sentences, this notion is still interesting in the contexts of particularly focussed *it BE* sentences. According to Weinert and Miller (1996: 201-202), reduced clefts can refer to more or less explicit entities, the majority of them is contrastive and the presupposed material for the possible contrast evoked by the (reduced) cleft is usually explicitly present in the preceding discourse which makes the omission of the cleft clause, i.e. the dependent
clause, possible. These characteristics apply to the class of particularly focussed \textit{it BE} sentences in (106) very well because they also refer to explicit entities, for example Cherie Blair, Marlon Brando and Jamie Forrester in (106a)–(106c), and they can be considered as contrastive due to the preceding discourse. Thus, the examples in (106) can be interpreted not only as a kind of blended reversed cleft sentence as discussed above (cf. (107c)–(109c)), but they can also be interpreted as the abridged versions of the following full-fledged \textit{it}-cleft sentences:

\begin{equation}
\text{(111) a. } \text{And if there is anyone who made it possible for Tony Blair to soar, it is she who did it.} \\
\text{b. } \text{If anybody owns that character, it's Marlon Brando. If anybody can do what they like with it, it's him who can.} \\
\text{c. That is Jamie Forrester. If anyone is a replacement for Wallace it is he who is (due to him making Strach look TALL).}
\end{equation}

Whether this class of particularly focussed \textit{it BE} sentences is now better interpreted as either abridged or as blended reversed cleft sentences is not crucial for the purpose here. Either way, there are striking parallels between this class of particularly focussed \textit{it BE} sentences and \textit{it}-cleft sentences which clearly set them apart from the “normal” class of \textit{it BE} sentences illustrated in (103) and which evidently justify a subdivision between particularly focussed (cf. (106)) and prototypical \textit{it BE} sentences (cf. (103) and (105)).

Furthermore, the special status of the pronoun in this type of \textit{it BE} sentences has also been observed by other studies (cf. Erdmann 1978). It has been noted not only that these sentences exhibit different distributions of pronoun case forms than other \textit{it BE} sentences, but also that “the subjective case form is preferably used in sentences where the pronoun in predicate position is marked as the subject in a preceding or following sentence” (Erdmann 1978: 75). As examples illustrating this category, the following sentences are quoted (Erdmann 1978: 75):

\begin{equation}
\text{(112) a. } \text{‘... if anybody deserves one it is he.’ (Mitford, The blessing: 176)} \\
\text{b. } \text{‘If anyone was responsible for Brian’s death, it was she. (Huxley, Eyeless in Gaza: 378)}
\end{equation}

Therefore, this group of marked \textit{it BE} sentences seems to be a good candidate for assessing whether Focus influences the distribution of pronoun case forms in subject predicative complements following \textit{it} and a form of \textit{BE}.

After having thoroughly discussed the first class of particularly focussed \textit{it BE} sentences, the second category will now be introduced. This second type of \textit{it BE} sentence in which the personal pronoun can be considered as particularly focussed is illustrated in (113):
(113) a. Only the man directly in front of them examined them, however; the rest continued with their work. Yi -- for surely it was he -- did not speak for some moments. His tongue touched his lips several times as he surveyed the four visitors, as if he was tasting the air. (COCA/FIC/BkSF:AlchimistsJournal)
b. Louis Napoleon (President 1848–51; Emperor 1852–70) --; for it was he --; was the nephew of Napoleon Bonaparte (First Consul 1799–1804; Emperor 1804–14): the nephew bathed in the reflected glory of the uncle. (BNC/ACP/W_pop_lore)
c. It's not surprising, then, that Peter Kelly - for it is he - has a background in catering. As a young man he worked as a chef at Marfield House in Co. Wexford. (ie/it is he/14.07.2008)

Again, this second type of particularly focussed it BE sentences can be distinguished from the class of unmarked it BE sentences by its syntactic peculiarities. First of all, this sentence type is singled out from more prototypical it BE sentences by the fact that is introduced by the conjunction for, typically used for introducing the reason, cause or detailed proof for something.12 Secondly and even more importantly, the position of these it BE sentences introduced by for within the sentences in which they are embedded is syntactically very unusual. According to Quirk et al. (1985: 922, 1106-1107), clauses introduced with for are not permitted either initially or medially; they have to follow another clause, a rule which is clearly violated by the sentences in (113). The it BE sentences introduced by for in (113) immediately follow the subject of the superordinate clause, i.e. clause-medially, which clearly marks the exceptional character of these clauses. Thirdly, these it BE sentences also partially resemble what Quirk et al. call “partial, weak, non-restrictive” apposition (1985: 1305), as they index co-reference with the subject of the superordinate clause, i.e. Yi, Louis Napoleon and Peter Kelly in (113a)–(113c) above, and immediately follow as syntactic insertions after the subjects of the matrix clauses. However, despite these superficial similarities, the it BE sentences in (113) cannot be considered full-fledged appositions, because they fail to meet central requirements for this category, such as the possibility to omit each appositive separately without affecting the acceptability of the sentence (Quirk et al. 1985: 1302). Hence, this second class of marked it BE sentences is already very notable from a structural or syntactic point of view.

From a pragmatic point of view, this second class of particularly focussed it BE sentences is also very remarkable. Examples of this class seem to differ from rather prototypical it BE sentences exemplified in (103) and (105) both in their basic function and also in the type of Focus they seem to exhibit. As discussed above, prototypical it BE

sentences often supply information that is needed to “fill a gap in the pragmatic information of the addressees” (Siewierska 1991: 178). Thus, typical examples of *it BE* sentences are often illustrated as answers to questions, as shown by the grammar book examples in (105). This is rather different for this second class of marked *it BE* sentences in (113). In these instances, the information supplied by the *it BE* sentence is already given in the superordinate clause, which means that there is no missing information which has to be supplied. Thus, it is unlikely that filling pragmatic gaps is the main or sole function of the *it BE* sentences in (113). Instead, they could be reaffirming the importance of the Topic of the sentence of the matrix clause, i.e. the person or entity the predication predicates something about (Siewierska 1991: 149). Hence, similar to *it*-clefs, the speakers or writers of these particularly focussed *it BE* sentences seem to deliberately exploit the “end-focus within an SVC clause” (Quirk et al. 1985: 1384) to highlight what they deem to be the most important piece of information, which in each of the cases in (113) is the subject of the superordinate clause.

The interpretation of these *it BE* sentences as Focus marking devices seems also justified when considering the fact that these sentences could even be omitted entirely without rendering the sentences incomplete or ungrammatical. Thus, the function of this type of *it BE* sentence can be considered to be clearly pragmatically motivated.

This characterisation of these sentences as Focus marking devices or devices which provide emphatic reaffirmation of the subject of the matrix clause is to some extent also underlined by the introduction of these sentences by the conjunction *for*, typically used to cite reasons, causes or detailed proof for something.13 Apart from expressing causality, this conjunction also evokes a certain degree of contrast by more or less explicitly expressing that the following predication can be made only because it is the person in question and no one else who is the subject and/or Topic of the superordinate clause. This contrastive interpretation of the Focus in this second class of particularly focussed *it BE* sentences can also be motivated by the fact that it is hard to conceive of a question to which the *it BE* sentences in (113) could be an appropriate answer (cf. Section 4.1). Finally, this characterisation of this clause type as being particularly focussed is also corroborated by a similar type of *it*-clefs which is also introduced by *for*:

---
Sometimes in her dream she was struggling against the undertow and snatching Billy by the tail of his shirt and knocking the water out of his lungs. For it was she who had saved the child and not some stranger who happened along. (COCA/FIC/VirginiaQRev)

The third force is the headteachers, for it is they who have the most decisive influence on what it is or is not safe to do or say in a school. (BNC/CKS/W_non_ac_polit_law_edu)

Thus, when comparing the *it BE* sentences in (113) to the *it*-clefts in (114), some accounts would probably consider the former to be reduced *it*-clefts (cf. Weinert and Miller 1996: 201-202). Although the notion of reduced cleft is not supported in this study, the characteristics attributed to them, i.e. reference to more or less explicit entities, a frequently contrastive use and the presence of the contrast-evoking material in the preceding discourse (Weinert and Miller 1996: 201-202) also applies to the two classes of particularly focussed *it BE* sentences distinguished in this study. This, in turn, corroborates their special status within the domain of *it BE* sentences.

In view of the preceding discussion, it seems safe to assume that both classes of particularly focussed *it BE* sentences are clearly distinct from more prototypical *it BE* sentences both syntactically and – what is more important – pragmatically. Therefore, they seem to be good candidates for assessing whether Focus influences the choice of pronoun case forms in subject predicative complements. Although it would have been desirable to analyse whether there are even differences between these two different classes of particularly focussed *it BE* sentences, this is not possible since they are extremely rare in large databases. Even in the dataset that yields the highest number of *it BE* sentences in this study, i.e. the *Corpus of Contemporary American English* (cf. Section 9.2), only 81 of the 2595 *it BE* sentences fall into these two classes described above. Therefore, these two classes are collapsed to form one single class of particularly focussed *it BE* sentences in the subsequent analyses in order to assess whether or not this variable has a significant effect on the distribution of pronoun case forms in the examined varieties.

To sum up, the operationalisation of the variable Focus consists of a necessary and a sufficient condition. The necessary condition states that if Focus does indeed influence the distribution of pronoun case forms and if focussed pronouns are more likely to surface in their subject forms, we expect to observe a higher proportion of subject pronoun case forms in *it*-clefts than in *it BE* sentences. This necessary condition is already covered by the Construction variable described in Section 6.1.2.1. The sufficient condition, however, states that the analysis of the data should also yield a higher share of subject pronoun forms in particularly focussed *it BE* sentences than in normal ones, if hypothesis H3 is correct.
according to which subject pronoun forms are used as postverbal Focus markers in subject predicative complements. Therefore, a further distinction between “normal” *it BE* sentences and particularly focussed ones, as described above, is introduced. In a nutshell, these particularly focussed *it BE* sentences are both syntactically and pragmatically different from prototypical *it BE* sentences and in some respects are more similar to clefts (cf. Weinert and Miller 1996: 201-202). However, since one crucial component of *it*-clefts is missing in these two subtypes discussed above, i.e. the dependent clause, we consider these sentences to be *it BE* sentences, though particularly focussed ones. Either way, these particularly focussed *it BE* sentences are different from normal *it BE* sentences and different from normal *it*-clefts but take a kind of intermediate position. Thus, they seem to be ideal for operationalising the variable *Focus*.

In the subsequent chapters, unless otherwise stated, the variable name *Focus* will refer to the “sufficient” condition of this variable, i.e. the distinction between unmarked and particularly focussed *it BE* sentences, as the necessary condition is already covered by the *Construction* variable and its parameter value *Cleft*. The sufficient condition will be operationalised as follows:

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
<th>Variable Type</th>
<th>Levels</th>
<th>Definition of the Parameter Values</th>
</tr>
</thead>
</table>
| *Focus*       | Occurrence of a pronoun in either a particularly focussed *it BE* sentence or a prototypical, i.e., normal *it BE* sentence | Nominal (dummy variable) | 0, 1 | *Focus*: pronoun occurs in one of the two classes of particularly focussed *it BE* sentence  
*Normal*: pronoun occurs in an *it BE* sentence that cannot be assigned to the two classes of particularly focussed *it BE* sentences (reference) |

*Table 14: The Independent Variable *Focus* of the Following Analyses*

6.1.2.6 Regional Variety

After having introduced and discussed parameters that may influence the distribution of pronoun case forms within geographical varieties (cf. Sections 6.1.2.1–6.1.2.5), this study also aims to test whether regional varieties of English differ from each other in their use of pronoun case forms in *it BE* sentences in *it*-clefts (cf. Section 5.2).
In view of the enormous amount of variation in the use of pronoun case forms attested across varieties of English (Cf. Sections 2.2), it does not come as a surprise that cross-varietal differences in the use of pronoun case forms in *it BE* sentences and *it*-clefs have already been attested (cf. Section 2.3.2). With regard to *it BE* sentences, Maier (2013) observes significant differences between American and British English with the British data exhibiting a significantly higher share of subject pronoun forms than the American data (cf. Maier 2013). With regard to *it*-clefs, this study also attests significant differences in the use of pronoun case forms between British and American English. However, while Maier (2013) only uses bivariate statistical tests, the present study relies on multifactorial regression models that will yield even more fine-grained results and thus will also be able to assess whether significant interactions with other variables are at work.

Considering the already attested differences between American and British English, it is very likely that cross-varietal differences will also be observed for the other varieties discussed in this study, i.e. Australian, Irish, South African and Indian English. Thus, the inclusion of the variable *REGIONAL VARIETY* is justified and considering the extent of variation observed in the pronominal paradigms of varieties of English more than reasonable (cf. Chapter 2). The reasons these six varieties, i.e. British, American, Australian, Irish, South African and Indian English, are examined in this study are the following. If it is assumed that there are marked cross-varietal differences in the distribution of pronoun case forms (cf. H4 Section 5.2), these six varieties seem to be appropriate study subjects since they represent very important regional varieties, each of which exhibits many morphosyntactic particularities (e.g. Britain 2007a; Filppula 1999; Kachru, Kachru and Nelson 2006; Kortmann et al. 2004; Peters, Collins and Smith 2009; Sailaja 2009; Schneider 2007). Thus, it is likely that these varieties also exhibit particularities with regard to the distribution of pronoun case forms in subject predicative complements. Furthermore, if it is assumed that there are indeed any general or even global trends observable as far as the distribution of pronoun case forms is concerned (cf. Burridge 2004: 1118), they may be identified with the help of these six regional varieties (cf. H4; Section 5.2), since this selection represents varieties from five different continents.

However, the variable *REGIONAL VARIETY* cannot be as easily implemented as the preceding variables into the statistical modelling. This is due to the heterogeneity of the datasets used for this study (cf. Section 7.2.4.3). Leaving aside for a moment the fact that
comparability in corpus linguistics is in itself a very complicated issue (cf. Section 7.5), this study deals with very different datasets which are not easily comparable.

The Web-derived datasets, for example, may differ considerably in their size and composition from each other, which makes a direct comparison of the data obtained in the different top-level domains rather difficult (cf. Sections 7.2.4.1–7.2.4.3 for further details). In a similar vein, a direct comparison between the closed corpora is also not easily possible because there are also notable differences between the BNC and the COCA (cf. Sections 7.1.2 and 7.1.3 for further details). Moreover, a comparison of the Web-derived datasets with the closed corpora is also not easily possible because the former include three modes of discourse whereas the closed corpora only contain two (cf. Section 6.1.2.4). As a consequence, instead of transferring the data points of the different datasets into one big statistical model, we will analyse each dataset separately in Chapters 8–14 in order to find out if the variables discussed in the preceding Sections 6.1.2.1–6.1.2.5 influence the distribution of pronoun case forms in subject predicative complements and, if so, to what extent. However, salient cross-varietal trends and differences should still come to the fore even though each dataset is analysed separately. Indeed, this approach seems more methodologically appropriate than collapsing the different datasets into one big model given the differences of the datasets as well as the general difficulty of comparing corpora (cf. Chapter 7).

6.1.2.7 Summary of the Independent Variables
In sum, Table 15 given below presents a concise overview of the independent variables used for the multivariate analyses conducted in Chapters 8–14. However, as outlined in Section 6.1.2.1, not all variables are applicable to *it*-clefts and *it BE* sentences and the superordinate category of subject predicative complements to the same extent. The *CO-REFERENCE* variable, for example, is only relevant for the subset of *it*-clefts, whereas the *FOCUS* variable is mainly relevant for *it BE* sentences and the superordinate category of subject predicative complements. Moreover, as discussed in Section 6.1.2.6, the variable *REGIONAL VARIETY* is not considered in the multivariate analyses at all due to the characteristics of the different datasets (cf. Chapter 7).
<table>
<thead>
<tr>
<th>VARIABLE NAME</th>
<th>DESCRIPTION</th>
<th>VARIABLE TYPE</th>
<th>LEVELS</th>
<th>DEFINITION OF THE PARAMETER VALUES</th>
</tr>
</thead>
</table>
| CONSTRUCTION  | Occurrence of the pronoun in either an *it BE* sentence or an *it*-cleft sentence | Nominal (dummy variable) | 0, 1  | **CLEFT**: *it*-clefts  
*BE*: *it* BE Sentences (reference) |
| CO-REFERENCE  | Co-reference of the clefted pronoun in the dependent (cleft) clause with either the subject or an Object | Nominal (dummy variable) | 0, 1  | **As_Subj**: clefted pronoun is co-referential with the subject of the following clause  
**As_Obj**: clefted pronoun is co-referential with an/the object of the following clause (reference) |
| PERSON        | Occurrence of a pronoun in a subject predicative complement construction in either first or third person | Nominal (dummy variable) | 0, 1  | **FIRST**: I, me, we us  
**THIRD**: he, him, she, her, they them (reference) |
| NUMBER        | Occurrence of a pronoun in a subject predicative complement construction in either singular or plural | Nominal (dummy variable) | 0, 1  | **SINGULAR**: I, me, he, him, she, her  
**PLURAL**: we, us, they, them (reference) |
| MODE OF DISCOURSE | Occurrence of a pronoun in a subject predicative complement construction in either spoken or computer-mediated or written discourse | Nominal (dummy variable) | 0, 1  | **SPOK**: spoken discourse  
**CMC**: computer-mediated discourse  
**WRIT**: written discourse (reference) |
| FOCUS         | Occurrence of a pronoun in either a particularly focussed *it BE* sentence or a prototypical, i.e., normal *it BE* sentence | Nominal (dummy variable) | 0, 1  | **FOCUS**: pronoun occurs in one of the two classes of particularly focussed *it* *BE* sentence  
**NORMAL**: pronoun occurs in an *it* *BE* sentence that cannot be assigned to the two classes of particularly focussed *it* *BE* sentences (reference) |

*Table 15: Overview of the Independent Variables Tested in the Multivariate Analyses*
6.1.3 The Independent Variables: The Difficulty of their Classification and Implications for their Interpretation

The attentive reader may have noticed that we have simply introduced the independent variables in the preceding Sections 6.1.2.1–6.1.2.6 without even attempting to systematise or classify them. However, this procedure contrasts with many quantitative studies that try to arrange their independent variables neatly along well-established linguistic levels of description, such as phonetics and phonology, semantics and morphosyntax to mention but a few (e.g. Gries 2003: 48-61; Lohmann 2011: 28-62). Moreover, this also does not really conform to the aims formulated in Section 5.1 according to which this study tries to investigate which factor or set of factors most strongly influences the distribution of pronoun case forms (cf. Q4 in Section 5.1).

In order to accommodate the aims postulated in Section 5.1, we can try to arrange the independent variables according to such well-established linguistic levels of description. Thus, MODE OF DISCOURSE (cf. Section 6.1.2.4) and REGIONAL VARIETY (cf. Section 6.1.2.6) can be labelled as sociolinguistic variables (cf. Labov 1972), and FOCUS can be classified as pragmatic variable. In addition, the variables PERSON and NUMBER can be assigned to the group of morphosyntactic or grammatical variables. While this seems straightforward so far, the variables CONSTRUCTION and CO-REFERENCE, however, do not lend themselves as easily to a classification as the previous ones. It could be asked, for example, whether these variables are mainly syntactic or primarily pragmatic or whether it may be best to avoid a definite decision in favour of one of these possible categories by simply employing a third one, such as the label discourse-pragmatic, which could reconcile both former options. This difficulty of deciding between different categories for these variables is simply due to the fact that the CONSTRUCTION variable, i.e. the distinction between it-clefts and it BE sentences, for example, is not only syntactically relevant but also pragmatically (cf. Sections 6.1.2.1 and 6.1.2.5). The same could be argued to hold for the CO-REFERENCE variable in Section 6.1.2.1.2.

Moreover, taking the Focus-oriented approach into consideration (cf. Chapter 4), even the PERSON and the MODE OF DISCOURSE variables can suddenly become less straightforward as far as their assignment to a certain linguistic level of analysis is concerned. With regard to the former, the discussion in Section 6.1.2.2 has demonstrated that the first and the third persons differ from each other in many respects and notably with regard to their pragmatic status. Thus, it may be difficult to determine which differences on
which linguistic level may be accountable for possible differences in the distributions of pronoun case forms. Likewise, the discussion of the different focussing devices in Section 4.1 has also shown that different modes of discourse possess different focussing strategies. Prosodic prominence, for example, is only available in spoken discourse, which means that different modes of discourse may use different strategies to highlight pragmatically salient elements. Thus, even such supposedly straightforward variables suddenly become less clear-cut when considered from a pragmatic point of view.

In view of the difficulties of arriving at an unambiguous and tenable categorisation, this study does not employ a rigid a priori classification of the analysed variables. Instead, this study first examines which variables turn out to be significant and then tries to assess whether a variable or a set of variables can be identified which turns out to be particularly important in accounting for the distribution of pronoun case forms in subject predicative complements. Based on these results, we can then try to relate our findings to the assumptions voiced both in the body of literature and in the present study (cf. Chapters 3 and 4).

### 6.2 The Statistical Design of the Study

As this study assumes that the occurrence of either a subject or an object pronoun case form in subject predicative complement position may be influenced by several independent factors either inhibiting or promoting the use of one or the other pronoun form in the analysed contexts, the data will be subjected to multivariate statistical analyses. The advantages of multivariate statistical models in general over simpler bi- or tri-variate statistical methods are widely acknowledged in quantitative variationist studies and have been thoroughly discussed elsewhere (e.g. Bresnan et al. 2007; Eid, Gollwitzer and Schmitt 2010: 264-265; Gries 2003: 31-43; Lohmann 2011: 69-70; Szmrecsaynyi 2006: 53-59). Thus, a detailed account of why multivariate analyses, where applicable, are clearly preferable to bi- or tri-variate ones is not deemed necessary here. Instead, the interested reader is referred to the studies quoted above.

For this study in particular, binary logistic regression models are applied to assess how the occurrence of subject pronoun case forms is influenced by one or more of the
independent factors discussed in Section 6.1.2 as well as the possible interactions thereof. In general, binary logistic regression models are very advantageous and very common in quantitative linguistic studies because they allow a prediction of the occurrence rate of a dependent variable based on the potential influence of a set of independent variables. Furthermore, this statistical technique is capable not only of determining the actual influence of each independent factor but also its direction. This means that binary logistic regressions state whether an independent variable inhibits or promotes the occurrence of a certain parameter value of the dependent variable. In the present case, this means that this method can clearly state which factors foster and which factors inhibit the likelihood of observing subject forms in subject predicative complements. In addition, binary logistic regressions also report how much of the variation in the analysed dataset is explained by the applied statistical model, i.e. by the choice of the independent variables potentially influencing the parameter values of the dependent variable. Finally, binary logistic regressions also determine the accuracy with which the applied model predicts the observed occurrences of the linguistic alternants or variants in the respective data (e.g. Hilpert 2008: 403-406; Pampel 2000: 1-35; Szmrecsanyi 2006: 53-56).

The present study conducts the binary logistic regression models using R, a programming language offering tremendously sophisticated tools for statistical analysis. In particular, the logistic regression model (lrm) function is used to conduct the analyses. If necessary, for example for certain model diagnostic calculations, the corresponding general linear models (glm) are also calculated in order to complement the lrm models.

With regard to model building, this study aims at complying with the Principle of Parsimony, which is also often called ‘Occam’s razor’. According to this principle, which reads in Latin as “entia non sunt multiplicanda praeter necessitatem” (quoted from Gries 2008: 30), the simplest explanation is always preferred over more complex ones, in cases where there are two or more equally good explanations for a certain phenomenon (cf. Crawley 2007: 325). Thus, this study pursues the goal of obtaining the best minimal-adequate models for the data containing, if possible, only significant variables. This approach is, however, applied not only to achieve a certain explanatory simplicity and elegance, but also to avoid the serious statistical problem of model overfitting, which might arise when too many insignificant variables are included in a statistical model. On the one hand, model

---

overfitting may turn out to be problematic because the more non-significant independent variables maintained in a regression model, the higher the probability that actually significant factors are presented as being non-significant. This is due to the fact that their real effects cannot be estimated precisely enough because of the noise included in the model. On the other hand, model overfitting may also result in estimating a factor as being statistically significant, although its correlation with the dependent variable is the result of pure chance (e.g. Backhaus et al. 2008: 84). Thus, in order to avoid these problems and to achieve the best and simultaneously simplest possible explanations for the data, this study applies a stepwise approach to obtain the best minimal-adequate regression models.

This means that this study starts the analysis of each dataset with maximal models containing all potential factors influencing the dependent variable introduced in Section 6.1.2 as well as all simple interactions between the independent variables (cf. Crawley 2007: 235-239). Then, a stepwise approach of variable elimination is applied in order to achieve the best possible model fit (e.g. Gries 2008: 284-294). However, as the elimination of potential factors as well as the stepwise approach in general is not uncontroversial (cf. Johnson 2010), only interactions with p>0.5 are excluded stepwise. From then on, each exclusion step of a non-significant interaction or variable is tested and verified by means of backwards bootstrap validations\textsuperscript{15} to ensure that the final models are indeed the best minimal-adequate ones. Furthermore, bootstrapping also prevents the model not only from containing redundant non-significant factors but also from pushing the pursuit of minimal-adequacy too far by excluding more variables than is recommended according to the bootstrapping validation processes (cf. Baayen 2008: 193-194, 204-205; Crawley 2007: 325-329; Johnson 2010).

In addition, another correction measure is used to ensure the best possible fit of the statistical models conducted in this study and to prevent model overfitting. This additional means is the “Penalized Maximum Likelihood Estimation” (Baayen 2008: 205-206). This correction is necessary because even when including only coefficients in the model building process that fit the data as closely as possible, it is still more or less inevitable that noise is also fitted. However, data points with extreme values resulting from noise are assigned the same weight as regular data points by the statistic models. Hence, the models as a whole may be influenced by these extreme values and exhibit results that are also too extreme. In

\textsuperscript{15} Bootstrap validation is executed with R’s validate function. The number of bootstrap runs performed in each bootstrap validation is, in analogy to Baayen (2008: 204-205), 200.
replication experiments of the original study, however, it is less probable that these extreme values will be reproduced. As a consequence, due to the diminished probability of observing these extreme values, the values of the coefficients expected in these replication studies are closer to zero than in the original model – a fact or observation that has been called “shrinkage” (Baayen 2008: 205). The phenomenon of shrinkage can now be exploited to further fit the applied models. This is done by employing the Penalized Maximum Likelihood Estimation which calculates and then implements a penalty factor into the preliminary statistical model which shrinks the whole model towards zero. By doing this, large values for coefficients are diminished which at the same time also considerably decreases the probability of model overfitting. Furthermore, the predictions produced by these penalized models may also be more accurate for the prediction of unseen data (Baayen 2008: 205). As the best possible penalty factor cannot be estimated in advance, several different penalty factors must be taken into consideration. R’s Design package fortunately offers a function called pentrace that estimates the best factor according to which a particular model is subjected to this shrinkage towards zero (Baayen 2008: 205-206).

A further very important aspect with regard to regression modelling is the issue of multicollinearity (cf. Section 6.1.2.1.2). One of the fundamental tenets of regression modelling is the assumption that the independent variables must not strongly correlate with each other. Most notably, there must not be a linear correlation between two independent variables, since this would constitute a case of perfect multicollinearity. Multicollinearity is to be avoided since a high degree leads to unreliable regression coefficients and unstable results (Backhaus et al. 2008: 87-89; Crawley 2007: 448; Szmrecsanyi 2006: 54). However, all empirical data exhibit a certain degree of multicollinearity. Therefore, the models used for this study may be affected by this phenomenon (e.g. Backhaus et al. 2008). As a consequence, multicollinearity measures for all binary logistic regression models and independent variables are reported in Appendix A. In particular, Appendix A reports the Variance Inflation Factors, since they are a common measure for multicollinearity (e.g. Lohmann 2011: 78; Peukert 2012; Szmrecsanyi 2006: 54). If necessary, the Variance Inflation Factors were complemented by another measure for multicollinearity, the so-called Condition Indices (e.g. Schlittgen 2004: 267; Wollschläger 2010: 233). This is necessary since there is no consensus in the relevant literature as to which value of the Variance Inflation Factors constitutes a critical threshold the exceedance of which indicates multicollinearity.
Hence, the Condition Indices as further quality criterion may clarify whether a factor or interaction with a rather high Variance Inflation Factor indicates multicollinearity or not (cf. Field 2009: 224; O’Brien 2007; Szmrecsanyi 2006: 215; Wollschläger 2010: 223; and see Appendix A for further details).

As logistic regression models are probably among the most current and most widely used multivariate statistic tools used in linguistics (e.g. Hilpert 2008: 403-406; Lohmann 2011: 76-80; Peukert 2012; Szmrecsanyi 2006: 53), it is not necessary to introduce and discuss this statistical technique in detail here. Instead, the interested reader is referred to more comprehensive introductions into regression modelling (e.g. Backhaus et al. 2008: 52-111; Field 2009: 264-315; Pampel 2000: 1-35). The present study confines itself to the brief introduction and explanation of those terms and concepts which are of immediate relevance for the understanding of the reported results of the models conducted in this study:

**Significance values of the independent variables:** The significance values reported for each factor will reveal whether the independent variables discussed and motivated in Section 6.1.2 as well as possible interactions thereof indeed significantly influence the parameter values of the dependent variable, i.e. whether they significantly influence the choice of pronoun case forms in subject predicative complements following *it* and a form of *BE* or whether their influence is attributable to mere chance. In general, a factor is considered to significantly influence a dependent variable if its p-value is less than or equal to five per cent, i.e. *p*≤0.05. If this is the case, it is admissible to reject the null hypothesis, i.e. the assumption that the given factor does not influence the choice of pronoun case forms in subject predicative complement position. This threshold is often called “the five per cent level” and p-values falling below this value are often indicated simply by an asterisk “*” (Butler 1985: 71). Whereas p-values at the five per cent level are merely called “significant”, there is also a “one per cent level” and a “0.1 per cent level” (Butler 1985: 71). P-values at the one per cent level, *p*≤0.01, are labelled as “highly significant” and are simply indicated by two asterisks “**”. P-values at the 0.1 per cent level, *p*≤0.001, are called “very highly significant” and are often abbreviated with three asterisks “***” (Butler 1985: 71). Although values of *p*≥0.05 are not normally considered statistically significant, p-values ≤0.1 may still be of some relevance and deserve further consideration since the five per cent level is after all an arbitrary demarcation. This is particularly true if the aim of the study is to look for suggestive evidence motivating further research on a given variable. P-values ≤0.1 are often called “marginally
significant” and this significance level is often indicated simply by “+” (cf. Butler 1985: 71; Lohmann 2011: 79).

**Effect sizes:** Since p-values merely indicate whether or not a given independent variable significantly influences the dependent variable, this study also provides the effect sizes of the individual factors in the minimal models. Effect sizes provide further meaningful information since they report a measure of the actual extent to which a factor in question influences the dependent variable (e.g. Field 2009: 785). Furthermore, effect sizes also indicate the direction of this influence, i.e. whether a given factor promotes or inhibits the occurrence of a given parameter value of the dependent variable. Similar to other studies (e.g. Lohmann 2011), this study reports effect sizes by means of two different scale units, i.e. regression coefficients and odds ratios.

Regression coefficients are given simply as coefficients in the tables summarising the most important regression results in the respective Chapters 8–14 below. As a simple rule of thumb, it can be said that the more a certain regression coefficient deviates from zero, the higher the effect of the respective independent variable on the outcome of the dependent variable. Values of regression coefficients may range from $-\infty$ to $+\infty$. Negative values of regression coefficients indicate that the presence of a certain independent variable makes a certain outcome of the dependent variable less likely. A positive regression coefficient, however, indicates an increased likelihood of a specific parameter value in the presence of the respective independent variable (cf. Gries 2008: 289). Although their interpretation seems straightforward, regression coefficients are not standardised and they unfortunately do not allow cross-model comparisons. In order to allow for cross-model comparisons, this study also reports odds ratios which are standardised effect size measures.

Odds ratios are the second effect size unit of measurement employed in the current study. Odds are defined in this context as “the probability of an event occurring divided by the probability of that event not occurring” (Field 2009: 790). Per definition, odds ratios range from 0 to $+\infty$ and they indicate the numerical factor by which the odds of a certain event such as a specific parameter value of the dependent variable – in our case the occurrence of a subject pronoun form – is multiplied by the presence or absence of a certain categorical independent variable (cf. Field 2009: 270-271; Lohmann 2011: 80; Szmrecsanyi 2006: 58). With regard to their interpretation, odds ratios with a value greater than 1 promote the outcome of a certain value of the dependent variable, whereas odds ratios
smaller than 1 indicate that the presence of a certain independent variable decreases the likelihood of a certain value of the dependent variable. If the odds ratios take the value 1, the independent variable has no effect at all on the outcome or parameter value of the dependent variable (Field 2009: 270-271; Pampel 2000: 21-23; Szmrecsanyi 2006: 58). By means of odds ratios, the percentage can be calculated by which a given independent variable inhibits or promotes a certain outcome of the dependent variable (cf. Gries 2008: 289).

Significance of the model: In addition to the significance values and effect sizes of the independent variables, this study also reports the likelihood ratio chi-square values of the models as a whole (cf. Gries 2009: 296, 305). This is done in order to determine whether or not the sum of the included independent variables significantly influences the outcome of the dependent variable and hence to clearly assess whether or not the overall model is statistically significant (cf. Gries 2009: 291-303; Szmrecsanyi 2006: 56-58).

Variance explained by the model: In order to determine whether or not the applied models are “substantially significant”, this study also reports the Nagelkerke’s $R^2$ value of each regression model (cf. Szmrecsanyi 2006: 55). A Nagelkerke’s $R^2$ value ranges per definition from 0 to 1. This value specifies the share of variation in the parameter values of the dependent variable that is explained by the independent variables in the applied regression model (cf. Backhaus et al. 2008: 282-283; Szmrecsanyi 2006: 55). This means that an $R^2$ value of 0 implies that there is no correlation between the dependent and the independent variables, whereas an $R^2$ value of 1 signals a perfect 100 per cent explanation of the observed variance in the values of the dependent variable by means of the independent variables included in a certain regression model (e.g. Backhaus et al. 2008: 263-265, 282-283; Szmrecsanyi 2006: 55). Thus, an $R^2$ value of 0.64 indicates, for example, that the applied model can explain 64 per cent of the encountered variation in the dependent variable. As in similar studies, a model is deemed “substantially significant” if its $R^2 \geq 0.05$. Hence, the independent variables are expected to explain at least five per cent of the observed variance in the dependent variable in order to render the model “substantially significant” (Szmrecsanyi 2006: 55).

---

16 To calculate the percentage by which a given independent variable inhibits or promotes a certain outcome of the dependent variable, the following formula can be applied: \( (\text{ODDS RATIO} - 1) \times 100 = \text{percentage} \) (cf. Gries 2008: 289).

17 For alternative views see Backhaus et al. (2008: 270).
Predictive power of the model: In order to assess how well the applied regression models fare in correctly classifying the data, this study also determines the percentage rate of correctly predicted cases of each individual model (% correctly predicted) compared to the values of the respective baseline model (% baseline). The percentage rates of correctly predicted cases indicate the exact number of correctly predicted parameter values of the dependent variable in the underlying model. For the present study, this means that the percentage value reports the number of instances in which the applied regression model correctly predicts the occurrence of subject and object forms as observed in the data. A percentage share of 84.44 per cent, for example, indicates that the applied model correctly predicts the outcome of the dependent variable in 84.44 per cent of all instances (cf. Lohmann 2011: 79; Szmrecsanyi 2006: 54-55). This share is then compared to the percentage rate of the baseline model which states the accuracy of a model in classifying the data when none of the independent variables discussed and motivated above is taken into consideration and hence the outcome of the prediction is the result of mere chance. For the present study, we take as a baseline model the distribution of pronoun case forms as observed in the data, and consider it to be the result of pure chance. By comparing the percentage value of the baseline model to the value for the predictive accuracy of the applied regression model, it can be seen how much better the regression model fares in correctly classifying the data than the baseline model does (cf. Szmrecsanyi 2006: 54-55). Whereas the baseline models in the present study are based on empirical grounds, theoretically motivated baseline models are also conceivable (cf. Lohmann 2011: 79). For example, following the discussion in Sections 3.2 and 3.3, it could be argued that neither the positional nor the approach distinguishing between weak and strong pronoun classes expects the occurrence of subject forms in subject predicative complement position. Hence, a theory-driven baseline model could be assumed predicting a percentage of null for the occurrence of subject pronoun forms in the analysed contexts. Yet, due to the overall empirical outlook of this study, the data-driven baseline model is clearly preferred in the present study.

Classificatory accuracy: Finally, this study also reports – though for reasons of space only in Appendix B – values referring to the classificatory accuracy of the applied models. In particular, this study will report the C statistics and Somer’s $D_{xy}$ for each of the regression models (Gries 2008: 289). The measuring unit simply referred to as C reports the extent to
which the predicted probability and the actual observation agree. This measure unit ranges from 0.5 to 1. In cases where this measure unit takes the value 0.5, predictions are considered to be random. C values of 1, however, indicate a perfect prediction of the applied model (Baayen 2008: 204; Peng, Lee and Ingersoll 2002: 6-8). If the values of C are greater than 0.8, this is usually considered a reliable sign that the applied model has a good predictive ability. Furthermore, this study will report Somers’s $D_{xy}$ in Appendix B. This value constitutes a rank correlation between predicted probabilities and ranges from 0, i.e. indicating complete randomness, to 1, i.e. reporting an ideal prediction (Baayen 2008: 204).

6.3 Variables and Statistical Modelling: Interim Summary

This chapter has introduced the dependent and independent variables used to examine the distribution of pronoun case forms as well as the statistical method applied to analyse the data.

Section 6.1.1 introduced the occurrence of subject pronoun case forms as the dependent variable of this study. This is due to the fact that much of the current linguistic literature assumes that subject pronoun case forms are basically restricted to the noun phrase slot immediately preceding the finite verb. Hence, the use of subject pronoun case forms in $it$-clefts and $it$ $BE$ sentences is, according to these accounts, rather unexpected (cf. Sections 3.2 and 3.3). However, this study assumes that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements (cf. Chapter 4). Thus, this study aims to test not only which factors promote or inhibit the use of subject pronoun case forms and to what extent, but also whether the Focus-oriented approach adopted in this study can be substantiated (cf. Chapters 4 and 5).

In Section 6.1.2, we presented and motivated the independent variables used in this study. According to the relevant literature and complying with the central hypotheses outlined in Chapter 5, we identified the variables $CONSTRUCTION$, $CO-REFERENCE$, $PERSON$, $NUMBER$, $MODE$ $OF$ $DISCOURSE$, $FOCUS$ and $REGIONAL$ $VARIETY$ as potentially influencing the distribution of pronoun case forms in subject predicative complements in general and $it$-clefts and $it$ $BE$ sentences in particular. However, as discussed in Sections 6.1.2.1 and 6.1.2.5, not all variables are applicable to all subject predicative complements to the same extent.
Whereas **CO-REFERENCE** is restricted to the subclass of *it*-clefts (cf. Section 6.1.2.1.2), the variable **FOCUS** is mainly relevant for *it BE* sentences in its operationalisation (cf. Section 6.1.2.5). Moreover, due to the heterogeneity of the datasets, the variable **REGIONAL VARIETY** can only be examined indirectly by comparing the models which use data from distinct regional varieties (cf. Sections 6.1.2.6 and 7.5). In Section 6.1.2.1.2, we also noted that the empirical approach of this study consists of two major steps. Firstly, we conduct a multivariate analysis of all subject predicative complements following *it* and form of *BE* for each of the respective datasets in order to be able to test the general hypotheses and research questions of this study as outlined in Chapter 5. Secondly, we then execute more fine-grained analyses of the individual construction types to assess the possible influence of construction-specific factors and to determine how the results obtained for the individual constructions relate to the assumptions and hypotheses of previous research as well as to the assumptions and hypotheses of this study as formulated in Chapters 4 and 5.

Section 6.1.3 then briefly discussed the difficulty of reasonably classifying the independent variables. While many quantitative studies try to arrange their independent variables in an orderly fashion along well-established linguistic levels of description, this is not advisable in the present study, since most, if not all, variables can be also be interpreted from the Focus-oriented perspective adopted in this study (cf. Chapter 4), which makes an a priori classification of the variables rather more complicated than helpful. Therefore, this study first examines which variables significantly influence the distribution of pronoun case forms and then tries to determine if there is a variable or a set of variables that turns out to be particularly important in accounting for the distribution of pronoun case forms in subject predicative complements.

Finally Section 6.2 introduced the multifactorial statistical method, i.e. binary logistic regression modelling, with the help of which we can examine and explain the distribution of pronoun case forms in subject predicative complements. Moreover, this section outlined the measures taken to arrive at the best possible model fit and hence also at the best possible results. In addition, Section 6.2 also briefly explored the key notions, terms and concepts necessary to understand the results obtained from the regression models presented and discussed in Chapters 8–14.
7 Data Used for this Study

The following chapter presents and describes the different datasets used for the quantitative analyses in this study in order to assess the distribution of pronoun case forms in subject predicative complements in and across varieties of English. This chapter can be very broadly subdivided into six different sections. In Section 7.1, the different corpora used for this study are briefly introduced and discussed. As the corpus data are complemented by data from the World Wide Web, a detailed discussion of the possibilities, advantages and disadvantages, and various attitudes towards the use of Web data found in previous research as well as an account of the preliminary considerations, methods of data compilation, data processing and data cleaning particular to the Web data is provided in Section 7.2. Section 7.3, outlines and discusses the editing and further processing of the data irrespective of their source, i.e. of both corpus and Web-derived data. In particular, this section describes and motivates those constructions which had to be excluded from the subsequent statistical analyses. Section 7.4 provides a very concise overview of the token numbers obtained for the different corpora and Web-derived datasets that form the basis for the subsequent statistical analyses, the results of which are reported in Chapters 8–14. Then, Section 7.5 briefly discusses the notions of representativeness, balancedness and comparability and their implications for the different datasets examined in this study. Finally, Section 7.6 gives a brief overview of the major insights obtained from this chapter. In general, the exact ways of data gathering, processing and classification conducted in this study are – to the knowledge of the author – unprecedented. Thus, this chapter offers a meticulous and very thorough description of the study’s databases to ensure a high degree of methodological transparency and also to allow for replication and follow-up studies using the same or a similar way of data compilation.

7.1 Corpus Data

Since this study draws on different data sources, Sections 7.1.1–7.1.3 describe and discuss the different corpora used for this study. Section 7.1.1, briefly introduces and discusses the *International Corpus of English* as a potential data source for the analysis of pronoun case
forms in subject predicative complements. Next, Section 7.1.2 comments on the *British National Corpus* as a data source for this study, while Section 7.1.3 introduces and discusses the *Contemporary Corpus of American English* as this study’s data source for American English.

In general, if not specified otherwise, following Biber et al. (1999: 335-336, 1134), searches were conducted for all instances of *it* + *is/was/’s* + *I/me/he/him/she/her/we/us/they/them* occurring in the corpora discussed below, which yielded all potential instances of *it* *BE* sentences and *it-clefts* with a case-sensitive focal pronoun.

### 7.1.1 The International Corpus of English

Due to this study’s aim to analyse the distribution of pronoun case forms in and across varieties of English, the *International Corpus of English* (henceforth ICE), seems to be the ideal starting point. This is because the ICE project started 1990 with the intent of establishing databases for comparative studies across varieties of English.\(^\text{18}\)

Unfortunately, pilot analyses by the author of the relevant components of this corpus, i.e. ICE Great Britain, ICE India and ICE Ireland as well as analyses by Beal (in prog.) have shown that the ICE family of corpora is much too small to allow for a robust multivariate statistical analysis of the distribution of pronoun forms in subject predicative complements.

This finding, i.e. the very low numbers of pronominal subject predicative complements following *it* and a form of *BE*, has also been observed and confirmed by other studies (Quinn 2009). Analysing the spoken components of ICE New Zealand and ICE Australia, as well as other corpora with a size comparable to those of the different ICE components such as the *Freiburg Brown Corpus of American English*, the *Freiburg-LOB Corpus of British English* and the *Australian Corpus of English*, Quinn observes that “[t]here are hardly any tokens of *it* *BE* [...] and other identificational constructions [...]” (2009: 42) in the analysed corpora and that “*[i]t-clefts are fairly rare overall [...]” (2009: 43). In view of these low figures, Quinn concludes that

\(^{18}\) cf. [http://ice-corpora.net/ice/design.htm](http://ice-corpora.net/ice/design.htm), October 17, 2011, and [http://ice-corpora.net/ice/index.htm](http://ice-corpora.net/ice/index.htm), October 17, 2011.
“[…] there are considerable drawbacks to basing an investigation of pronoun variants entirely on corpus data. Constructions that give rise to pronoun case variation are comparatively rare in conversation and even written texts, so that it is difficult to obtain a sufficient number of comparable pronoun tokens in the full range of variable contexts” (Quinn 2009: 36).

In light of this insufficient number of occurrences of the relevant constructions in the International Corpus of English, this study has to rely on considerably larger databases, which are introduced and discussed in the following Sections 7.1.2–7.1.3 and 7.2.

7.1.2 The British National Corpus

Since the ICE corpora are too small to allow for robust quantitative analyses of the phenomena examined in this study, the logical next step is to turn to much larger data sources. In order to analyse the distribution of subject predicative complements in British English, this study relies to a large extent on the data provided by the British National Corpus (henceforth BNC). The BNC consists of more than 4000 text files, which amounts to a total token count of roughly 100 million words (e.g. Aston and Burnard 1998: 28-41; Leech, Rayson and Wilson 2001: 1). Furthermore, the BNC was designed to characterise contemporary spoken and written British English. This means that no imaginative texts included in the corpus were written before 1960, and no informative texts included in the BNC were written before 1975. All spoken data included were recorded in or after 1991 and more than 90 per cent of the texts included in the corpus cover the period from 1985 to 1994 (Burnard 2007; Leech, Rayson and Wilson 2001: 1). As far as the proportion of spoken and written data in the corpus is concerned, the BNC consists of approximately 90 per cent written and 10 per cent spoken data. Although the share of spoken data seems at first rather modest, it should not be forgotten that the spoken subset of the BNC is still ten times larger than the complete British ICE component, which makes it an invaluable treasure trove of spoken British English (Burnard 2007).

However, it is not only because of its mere size that the BNC been labelled “the ideal case of a representative standard mega-corpus [translation GM]” (Mukherjee 2009: 19).19

---

19 The original quote from Mukherjee (2009: 19) is as follows: “[…] dass der BNC inzwischen als der Idealfall eines repräsentativen Standard-Megakorpus gilt [...]”.
The BNC is also distinguished from most other corpora by the broad range and multitude of different text types, communicative situations and contexts represented by it as well as by the carefully crafted sampling procedures, methods and criteria which set it apart from most other corpora (e.g. Burnard 2007; Mukherjee 2009: 19-20). As a consequence, the extraordinary status of this corpus among all other available corpora is widely acknowledged (e.g. Meyer 2002: 138; Mukherjee 2009: 19-20). Meyer (2002) concludes, for example, that “[t]he British National Corpus is the culmination of all the knowledge we have gained since the 1960s about what makes a good corpus” (2002: 138). Hence, the BNC is a very suitable database for the present study, since it should be large enough to allow for the analysis of the distribution of subject predicative complements in British English. Furthermore, in view of the reference character of the BNC, this corpus also serves as a kind of yardstick for the data classification of the data obtained from the Internet (cf. Section 7.2.4.3).

Since there are different versions of the BNC available, it has to be noted that this study uses Mark Davies’ (2004–) online version of the BNC, which is accessible through Brigham Young University’s Website: http://corpus.byu.edu/bnc.20

7.1.3 The Corpus of Contemporary American English

In order to be able to analyse the distribution of pronominal case forms in subject predicate complements after it and a form of BE in American English, this study furthermore draws on the data supplied by the Corpus of Contemporary American English (henceforth COCA) (Davies 2008–). As is the case with the version of the BNC used for the current study, the COCA, which is available online at http://corpus.byu.edu/coca, is hosted at Brigham Young University and was created by Mark Davies (2008–).21

The COCA currently consists of 520 million words. This token count is only preliminary as this corpus is being continuously expanded. In particular, this means that the COCA consists of 20 million words for each year covering the period from 1990 to 2015 and new data is continuously added once or twice a year. As far as its composition is concerned, the COCA contains equal proportions of spoken data, academic texts, fiction and texts taken

---


from newspapers and popular magazines. Thus, the share of spoken data amounts to 20 per cent compared to an 80 per cent share of written data in this corpus.\footnote{Cf. Ibid: http://corpus.byu.edu/coca, August 14, 2016.}

The COCA is not only more than five times larger than the BNC and contains proportionally twice as much spoken data than the BNC, but there are also noteworthy qualitative differences between these two corpora that have to be taken into consideration.

A first importance difference concerns the age of the data included in these two corpora. While the data of the COCA currently represents the period from 1990 to 2015, the BNC data is older on average, since it includes data from 1960 to 1994. Although these different periods represented by both corpora may suggest at first glance that both corpora actually represent different generations of speakers and writers, a closer look at the BNC data reveals that no spoken BNC data is from before 1991 and the vast majority, i.e. 93 per cent, of the BNC texts are taken from the period between 1985 and 1994 (Burnard 2007; Leech, Rayson and Wilson 2001: 1). Thus, the intersection of the data of both corpora in terms of their age is – despite all differences – still considerable.

With regard to the spoken data, the COCA data consists of unscripted conversations transcribed from TV and radio shows.\footnote{Cf. http://corpus2.byu.edu/coca/compare-bnc.asp, October 21, 2011.} The BNC, however, represents a much wider spectrum of spoken data. On the one hand, it does comprise what is called “context-governed” data such as transcripts of meetings, debates, seminars and radio programmes, all of which is very similar to the spoken data of the COCA. On the other hand, the BNC also contains a large share of what is called a “demographic component”, i.e. informal conversations recorded by and from a socially stratified sample of informants who are distinguished according to several social parameters such as sex, age-group, geographic region and social class, accounting for more than four million words (Aston and Burnard 1998: 31-33). Hence, the BNC represents a much more diversified spectrum of spoken sub-genres than the COCA.\footnote{Cf. http://corpus2.byu.edu/coca/compare-bnc.asp, October 21, 2011.}

There are, however, also qualitative differences in the written data observable between these two corpora. Whereas the written data of the COCA is rather evenly distributed across the genres fiction, popular magazines, newspaper and academic texts, the spectrum of the written BNC data is again much broader, since it includes in addition to the genres included in the COCA a wide range of other texts, such as brochures, manuals,
leaflets, advertisements and letters, in order to fulfil the corpus’ major aim, i.e. providing “a microcosm of British English in its entirety” (Burnard 2007).

Another very notable difference between the BNC and the COCA pertains to the overall quality of the data, since the COCA does not seem to be as carefully crafted as the BNC. Indeed, there is a certain trade-off observable in the COCA between the sheer amount and the currentness of the data on the one hand and their quality on the other hand. A careful analysis of the data provided by the COCA yields a notable number of doublets (115), data points that are clearly not American (116), and instances that are neither contemporary nor American (117). These instances are excluded from the further statistical analyses of the data.

(115)  a.  [...] I heard a key turn in the lock. I knew right away it was him by the way his boot heels shuffled along the floor. (COCA/FIC/LiteraryRev)
    b.  [...] I heard a key turn in the lock. I knew right away it was him by the way his boot heels shuffled along the floor. (COCA/FIC/LiteraryRev) [Doublet/Repetition]

(116)  a.  “[…] is –’ Of course, it’s me; of course, it ’s me that would give this up,” Beckham said. (COCA/NEWS/Chicago) [David Beckham/British footballer]
    b.  […] Chill, Winston, it’s me and Charlie can see it ’s me, so what’s the problem? # WINSTON # The problem is, Willy […] (COCA/FIC/Mov:LockStockTwo) [British screenplay]

(117)   […] seeing you coming. Othello: (to Iago) I do believe it was he. Desdemona: How now, my lord! I have been talking with your […] (COCA/FIC/Read) [William Shakespeare]

Whereas instances such as (116a), i.e. quotes of foreigners in newspaper articles, are certainly very difficult to control for in the compilation process and therefore have to be more or less accepted, the other instances in (115)–(117) are more easily avoidable. Particularly problematic are not the doublets in (115), which can be identified and excluded rather easily with the help of office software, but the inclusion of literary works and screenplays from outside the US, which are much more difficult to detect. The British screenplay from which example (116b) is taken consists of nearly twenty thousand words and is, by the way, explicitly associated with Cockney English (Kurz 2006: 141-160). The motivation for the inclusion of such data in a corpus intending to represent contemporary American English remains obscure. Thus, in view of these potentially problematic data

27 Indeed, a quick look at the list of all texts included in the COCA, which can be downloaded as .xls-file, shows that not only British screenplays such as the one quoted above but also British novels such as
points, much more caution has to be exercised when analysing data from the COCA than when analysing BNC data.

However, despite all the differences from the BNC and the additional effort and caution required to analyse the data, the COCA is still certainly the best possible data source for the analysis of the distribution of pronoun case forms in subject predicative complements after *it* and a form of *BE* in American English. This is due to its sheer amount of data and because of the fact that it is the largest publicly accessible corpus of American English.28

One last important remark regarding the use of the COCA data in this study concerns the date of the data compilation. As has been mentioned above, the COCA is a dynamic corpus, which is continuously being expanded. Therefore, it is important to note that the data compilation for this study took place on 23 July 2010. This means, that this study is not based on the currently available 520 million word version of this corpus but only on the 400 million word version of the COCA.

### 7.2 Internet Data

In order to study what has been called “purposeful language behavior”, corpus linguists require large amounts of authentically produced written and spoken texts (Bernardini, Baroni and Evert 2006: 9). Although the 100 million words of the BNC and the 520 million words of the COCA are enough for many quantitative studies, there may be purposes for which these amounts of data are not sufficient (Hundt, Nesselhauf and Biewer 2007b: 1; Kilgariff and Grafenstette 2003: 4). Furthermore, for most varieties of English, there are no mega-corpora comparable to the BNC and the COCA available and the limitations of smaller corpora such as the different ICE components have already been addressed in Section 7.1.1. In view of this difficult data situation, it seems self-evident that the World Wide Web has become an increasingly interesting data source for linguists simply due to the more or less infinite amount of linguistic data that it offers (e.g. Hundt, Nesselhauf and Biewer 2007a; Baroni and Bernardini 2006; Bernardini, Baroni and Evert 2006: 9). Due to the scarcity of subject predicative complements in existing corpora allowing for the comparison of different

---

varieties of English, this study also makes use of the data the Internet can offer, particularly for those varieties for which no freely available closed mega-corpora exist. Thus, Web-derived datasets for Australian, Irish, South African and Indian English were compiled in order to be able to examine the distribution of pronoun case forms in it-clefts and it BE sentences in these varieties. Moreover, a Web-derived dataset for British English was also compiled in order to assess, by relating the Web data to the BNC data, whether it is possible to obtain meaningful quantitative results from data obtained from Google (cf. Chapter 5).

However, before a detailed outline of the methodology of data collection and refining processes applied in this study is offered in Section 7.2.4, Section 7.2.1 briefly addresses the fundamental question of whether or not the Web can be considered a corpus at all and introduces the most important methodological approaches related to the use of the Web as a corpus linguistic resource. Since the issue of using Web data for linguistic research is not uncontroversial, Section 7.2.2 provides an overview of the most severe problems that might arise when using Web data. Section 7.2.3 highlights the opportunities and possibilities the use of Web data can offer despite the afore-mentioned problems, and also addresses why in the future, the Internet will be an indispensable linguistic data source both for analysing linguistic varieties and for answering many linguistic questions for which large amounts of linguistic data are needed.

7.2.1 The World Wide Web as a Corpus Linguistic Resource: An Introduction

Whether or not the World Wide Web and the linguistic data obtained from it can be considered to constitute a viable corpus-linguistic resource depends, first of all, on the question of whether a rather narrow or a rather broad definition of the concept ‘corpus’ is applied (Kilgariff and Grafenstette 2003: 1-2). A rather narrow definition is exemplified by McEnery, Xiao and Tono’s conception stating “that a corpus is a collection of (1) machine-readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a language or language variety” (2006: 5). In contrast, a much more inclusive definition of corpus is employed, for example, in the field of Natural Language Processing, as is illustrated by the following quote from Manning and Schütze (1999):
“In Statistical NLP, one commonly receives as a corpus a certain amount of data from a certain domain of interest, without having any say in how it is constructed. In such cases, having more training data is normally more useful than any concerns of balance, and one should simply use all text that is there” (Manning and Schütze 1999: 120).

Whereas the former definition sets great store by the qualitative aspects of the compiled data, the latter definition attaches more importance to the quantitative side of corpora, which is also encapsulated in the view that “more data is better data” (Bernardini, Baroni and Evert 2006: 9). Due to these marked differences with regard to the definition of even the most basic concepts, it is not surprising that there are also different opinions with regard to the acceptability of using Web data for corpus linguistic studies. Scholars who are more inclined towards the narrow definition of corpora and who put great emphasis on concepts like “authenticity” and “representativeness” tend to distrust results obtained from the Web (e.g. Leech 2007). However, this rather narrow conception of corpora has been criticised as mixing two different issues, i.e. the question of “What is a corpus?” with the question “What constitutes a good corpus?” (Kilgariff and Grafenstette 2003: 2). In view of the seeming arbitrariness of corpus compilation and the lack of clarity with regard to the expected requirements, such as the notion of representativeness, which will be further elaborated on in Section 7.5, Kilgariff and Grafenstette advocate a very inclusive definition of corpora stating that “a corpus is a collection of texts when considered as an object of language or literary study” (2003: 2). This means that “[t]he answer to the question ‘is the web a corpus?’ is yes” (Kilgariff and Grafenstette 2003: 2). This study follows Kilgariff and Grafenstette’s (2003) line of reasoning. Considering that even the most basic concepts in corpus linguistics, such as the notion of a ‘corpus’ itself and the requirements corpora are supposed to meet, are still an issue of ongoing controversy, it is not surprising that the attitudes towards the use of Web data as a source for linguistic studies are also rather mixed, as we will see below (cf. Sections 7.2.2 and 7.2.3). Indeed, they tend to oscillate between exaltation and mere execration (Mair 2007: 235).

Despite the still very controversial status of using the World Wide Web as linguistic resource, very different approaches to the use of the data offered by the Internet have been put forward, of which the following four are probably the most influential ones (Bernardini, Baroni and Evert 2006: 13).
Probably the most ambitious way to use the Web as corpus linguistic resource are attempts to create a kind of “mini-Web or mega-corpus” based on the World Wide Web which can be subjected to language research and which is supposed to contain elements of both traditional corpora, e.g. part-of-speech and lemma tagging, and of the Internet, e.g. the tremendous mass and the recency of data (Bernardini, Baroni and Evert 2006: 13-14). An example of such an enterprise is the ukWaC, a two billion word lemmatised and part-of-speech-tagged British English corpus. This corpus was compiled from texts taken from the .uk top-level domain and seems to offer reliable results when compared to traditionally compiled corpora such as the BNC (Ferraresi et al. 2008). Similar corpora also applying this “mini-Web” concept are also available for other languages, such as German and Italian (Lee 2010: 116).

The second method of using the Web as a corpus is the “Web as Corpus proper” approach (Bernardini, Baroni and Evert 2006: 13). The aim of this approach is primarily to investigate the nature of the Web and the language used on it. Just as the COCA intends to represent American English, it is possible to conceive of a corpus depicting a synchronic snapshot of Web English, in order to shed light on language usage on the Web (Bernardini, Baroni and Evert 2006: 13).

The third major approach to the World Wide Web as corpus linguistic resource is to use it as a source for the compilation of so-called “Web-derived” corpora. In this method, finite corpora from texts derived, downloaded and saved from Web-sources are compiled in order to generate stable offline corpora that allow exhaustive, replicable analyses with the help of the usual corpus linguistic tools and software (e.g. Hundt, Nesselhauf and Biewer 2007b: 1-5; Mukherjee 2009: 63; Ueyama 2006: 99). This approach allows researchers to control what is included in the corpus data, and these corpora can usually be accessed by standard corpus software and hence allow a deep level of analysis just as “normal” corpora do. Thus, this method is probably the most widely acknowledged and least controversial way to use the World Wide Web as a corpus linguistic resource because it avoids many of the problems associated with Web data discussed in the next section (e.g. Hundt, Nesselhauf and Biewer 2007b; Kilgariff 2006a: 148; Lüdeling, Evert and Baroni 2007: 18). This ‘Web-derived’ corpora approach is not only similar to traditional corpora, but also to the first

---

29 Henceforth, whenever this study refers to this particular approach to derive corpora or datasets from the Web, this study encloses ‘Web-derived’ in simple quotation marks. This notion must not be confused with the neutral notion Web-derived without quotation marks, as in Web-derived dataset (cf. Chapters 10–14), which simply states that a certain dataset was derived from the Web.
approach outlined above, the “mini-Web or mega-corpus” method. The main differences between the first, i.e. the “mini-Web or mega-corpus”, and the third, i.e. the ‘Web-derived’ corpus, approaches are the scope and the size of the compiled corpora. Whereas examples of the former such as the ukWaC are extremely large, consisting of more than two billion words, and attempt to serve as heterogeneous general-purpose corpora comparable to traditional balanced corpora (Ferraresi et al. 2008), the scope and size of what is typically called ‘Web-derived’ corpus is most often much more restricted. A case in point for this approach is the 172 million word CNN transcripts corpus consisting of transcripts from the years 2000 to 2004 downloaded from the archive of the news channel CNN, which offers an enormous amount of spoken data (Hoffmann 2007: 76-77). Indeed, this approach to compiling ‘Web-derived’ corpora is often seen as the most promising one because it offers all the advantages of traditional corpora. At the same time, it also accommodates the fact that the World Wide Web will be the only source to obtain a reasonable amount of data for many varieties of English in the foreseeable future (Hundt, Nesselhauf and Biewer 2007b: 4).

The fourth major approach to use the Web as a corpus linguistic resource is not only the simplest and probably also the most widely used method, but at the same time also the most controversial one. This approach involves only querying for search terms or search phrases using commercial search engines such as Google, Bing or Yahoo. This approach has been called “Web as corpus surrogate” (Bernardini, Baroni and Evert 2006: 13), “Hunting” (Fletcher 2007: 28-29) or – depending on whether the intent of the queries is primarily heuristic or systematic – “data sniffing” or “data testing” (Hundt, Nesselhauf and Bieber 2007b: 2). However, most often this approach is simply referred to as “Web as Corpus” approach30 (e.g. Fletcher 2007: 28; Kilgariff and Grafenstette 2003; Mair 227: 235). In this approach, the Web is used as a corpus and search engines are employed instead of the usual concordancing software. This approach is often applied in contexts where no other traditional closed corpora exist or where existing traditional corpora are too small to meet the requirements of the underlying research question (e.g. Bernardini, Baroni and Evert 2006: 9). However, this approach is also highly controversial since the quality and reliability of the results obtained from this approach are often called into question (e.g. Kilgariff 2006a).

30 Henceforth, to indicate that this approach refers to a certain method of using Web data or the Web as corpus, this study encloses ‘Web as Corpus’ in simple quotation marks, whenever it refers to this particular approach.
7.2.2 The World Wide Web as a Corpus Linguistic Resource: Problems and Pitfalls

By its mere presence, most scholars acknowledge the Web as a corpus linguistic resource, yet there is a considerable amount of scepticism towards it because “it is not the corpus linguists would have compiled” (Mair 2007: 236). Particularly critical voices include the objection towards the use of Web data in general, i.e. “caveat googleator!” (Joseph 2004: 382), and the warnings of the use of data obtained from the ‘Web as Corpus’ approach in particular, because “Googleology is Bad Science” (Kilgariff 2006a). The reasons for these objections to the use of Web data, particularly to the data gained from the ‘Web as Corpus’ approach, are a number of potential problems that can arise from the use of Web data, the most eminent of which will now be discussed in detail.

Firstly, the size of the underlying corpus is deemed to be a problem. When using the ‘Web as Corpus’ approach, the size of the corpus, i.e. the token count of the Web, is unknown, due to the dynamic and ever-changing nature of the Web (e.g. Hundt, Nesselhauf and Biewer 2007b: 3; Mair 2007: 244). The exact token counts of a corpus should, however, be known (e.g. Lüdeling, Evert and Baroni 2007: 14), because otherwise basic means of quantitative analyses – particularly those such as normalization which require metric scaling – cannot be performed.

However, a much more severe potential problem concerns the overall quality of the data obtained from the ‘Web as Corpus’ approach. While more neutral accounts label Web data simply as “heterogeneous” (Volk 2002: 9), other judgements cast upon the ‘Web as Corpus’ approach and the data obtained from it include labels such as “noisy” (e.g. Keller and Lapata 2003: 460), “ramshackle” (Bernardini, Baroni and Evert 2006: 23), “dirty” (e.g. Kilgariff and Grafenstette 2003: 9; Mair 2007: 239) or even “corrupted” ( Cotterill 2010: 584). The reasons that many scholars are sceptical with regard to the quality of the Web data can be subsumed under the broad notions of origin, originality, and composition of the retrieved data.

Problems with regard to the origin of the data include the question of the regional provenance of the data. Since the majority of English Web pages originates from non-native speakers of English (Rosenbach 2007: 168), it is important to assess the actual origin of the data, particularly when studying varieties of English as in the present study. Indeed, tracing
back the actual regional provenance of the data is very difficult, although it is possible to restrict the queries to a particular national top-level domain, such as .za for South Africa, .in for India or .ie for Ireland (Leech 2007: 145). However, restricting the search to a certain top-level domain is considered to be rather unreliable and at best “a rough guide” to the regional origin of the obtained data (Fletcher 2007: 36; Lüdeling, Evert and Baroni 2007: 15). Another important aspect with regard to the origin of the data concerns the authorship of the data. As noted, most English language Web pages are authored by non-native speakers of English and “[a]s lingua franca of the digital frontier, English is both the source and target of contamination […]” (Fletcher 2007: 36). This means that apparently interesting features or phenomena observed in Web-derived data may actually be performance errors of non-native users (e.g. Joseph 2004: 382; Rosenbach 2007: 167-168). As a consequence, using qualitative evidence from Web-derived data may be problematic because the origin and the originality of the data is often unclear. In view of this difficulty, a very careful analysis is required to separate the chaff from the wheat, or the trash from the really interesting features (Rosenbach 2007: 168-169). In addition, the hit counts or estimates provided by the search engines can only give an indication and do not prove per se the prevalence of a certain linguistic variant or form (Fletcher 2007: 36). A further difficulty with regard to the origin of the data simply concerns the date of writing. Very often, it is not only unclear where and by whom a text on the Internet was written, but also when it was written (Leech 2007: 145; Lüdeling Evert and Baroni 2007: 15).

In terms of composition, researchers are faced with the problem that the data obtained from the ‘Web as Corpus’ approach contain a considerable amount of “junk” (Hundt Nesselhauf and Biewer 2007b: 3). Some search engines such as Google will return pages that do not include the search term. This may occur if the search term is contained in a link to a page but not on the Web page itself (Keller and Lapata 2003: 468). Moreover, the data derived from the ‘Web as Corpus’ approach may be abundant in duplicates or doublets, i.e. repetitions of earlier hits (Lüdeling, Evert and Baroni 2007: 14). Furthermore, even if this junk or noise is removed, the composition of the remaining data is also difficult to assess. Since the data derived from the ‘Web as Corpus’ approach do not contain any metadata, it is difficult to evaluate which text types and genres they contain. Thus, it is often difficult or even impossible without closer inspection to determine the mode of discourse, i.e. spoken versus written versus computer-mediated communication, the text type and the regional
and social origins of the texts (Hundt, Nesselhauf and Biewer 2007b: 3; Lüdeling Evert and Baroni 2007: 15). Moreover, the Internet is considered not to be a “balanced” corpus, with regard either to text types or to subject areas, since there is no oversight for what is posted on the Internet and only limited control over what is returned by a search query (e.g. Cotterill 2010: 584; Ueyama 2006: 99; Volk 2002: 9). Hence, some kinds of language data which are particularly interesting for linguistic analysis, such as private discourse and everyday communication, are per definition rather scarce on the Internet (Leech 2007: 144).

Another issue often identified with the use of data derived from the ‘Web as Corpus’ approach is the problem of representativeness (Mukherjee 2009: 60). Although commercial search engines offer easy access to enormous samples of language, it is unclear whose language use the World Wide Web represents and how representative this language use actually is. Thus, the Web and the data derived from it can, according to some scholars, hardly be considered representative of general language use (Mukherjee 2009: 60; Leech 2007: 144-145; cf. Section 7.5).

Apart from the size, the quality and the representativeness of the data, a further serious concern pertains to the replicability of the data obtained from the ‘Web as Corpus’ approach. As a consequence of the dynamic nature of the World Wide Web and its continuous expansion, it is not possible to replicate and reproduce the results obtained from an earlier query when using a commercial search engine, since the hit counts and the content obtained may vary from time to time (e.g. Lüdeling, Evert and Baroni 2007: 11; Kilgariff 2006a: 148; Lee 2010: 115). Thus, the replication and verification of the results obtained by other scholars is not possible, which diminishes the credibility of the World Wide Web as a scientific data source (Fletcher 2007: 37; Mukherjee 2009: 60).

A further issue that has been identified as problematic in the ‘Web as Corpus’ approach is the lack of exhaustiveness of the data obtained. When relying on a commercial search engine, it is unclear which parts of the Internet are searched for the query because this depends on the search algorithms of the respective search engine (e.g. Mukherjee 2009: 60). Also dependent on the search algorithms of the search engines are the ranking and indexing practices applied for each query. As the standard practice of commercial search engines is to index only the first 100,000 words on each Web page, and since search engines only provide the first 1000 or 5000 hits, the ranking of the search results is, of course, a crucial matter. Although the ranking of search hits seems to rely on the respective link
popularity, the exact practice remains obscure because the search algorithms on which the ranking of the results is based are, of course, well-kept company secrets (Fletcher 2007: 33-37; Kilgariff and Grafenstette 2003: 12; Kilgariff 2006a: 148). Furthermore, it is unclear how the respective search engines deal with duplicates and spam and how this affects the hit counts of a search engine (Kilgariff 2006a: 148). In general, very little is known about how search engines arrive at their results (Mair 2007: 244). As far as the obtained hits are concerned, it is also crucial to note that the search hits are for pages and not for actual instances of the search term (Kilgariff 2006a: 147). Finally, exhaustiveness of the results derived from the ‘Web as Corpus’ approach is constrained by the fact that commercial search engines restrict the number of queries and hits per query per user per day (Kilgariff 2006a: 147). In view of these restrictions to exhaustiveness, it is recommended to consider the hit counts merely as indications since the numbers provided by a search engine cannot actually prove the pervasiveness of a given form or its predominance compared to an alternant (e.g. Fletcher 2007: 36).

The ‘Web as Corpus’ approach has also often been criticised for its limitations in terms of practicability. As commercial search engines are not linguistic tools in the first place, many important corpus linguistic features such as wild cards, part-of-speech tagging, Key Word in Context formats, etc. are available only to a very limited extent or not available at all (e.g. Kilgariff 2006a: 147; Lüdeling Evert and Baroni 2007: 13; Mukherjee 2009: 60; Rohdenburg 2007: 205). Another feature of commercial search engines that has severe repercussions on the practicability of using the ‘Web as Corpus’ approach is the fact many search engines ignore most or even all of the punctuation marks in the retrieved hits and also generally conduct case-insensitive queries, even if the search term is in quotation marks. Moreover, the auto-correction function of the search engine may also influence the obtained results because it may query for another search string than the one entered, for example ignoring or excluding certain orthographic variants. As a consequence, this may result in a considerable amount of so-called “false positives” such as tokens in which the entered search string crosses a phrase, clause or even sentence boundary (cf. Section 7.2.4.2). This, of course, may strongly influence or even skew the results of a quantitative analysis (Keller and Lapata 2003: 468; Lüdeling Evert and Baroni 2007: 12; Rosenbach 2007: 171). Even if reasonable results are obtained, it has been remarked that commercial search engines do often not provide enough context for each hit in order to ensure an
unambiguous interpretation of the obtained instance (Kilgariff and Grafenstette 2003: 12). Thus, the elimination of ambiguous hits, false hits and duplicates has been considered to be “prohibitively time-consuming” (Lüdeling, Evert and Baroni 2007: 14). Indeed, it has been concluded that all the advantages of easily accessible Web data simply vanish when all the limitations imposed by the use of a commercial search engine have to be overcome (Kilgariff 2006a: 147).

A further factor influencing the data derived from the ‘Web as Corpus’ approach is the commercial interest of the search engines. Since commercial search engines provide services in return for advertisement fees, researchers applying the ‘Web as Corpus’ approach have to accept the phenomenon of “paid positioning”, i.e. the high ranking of Websites paying advertising fees to the search engine, by which search engines try to direct their users towards the Websites of their clients. Hence, more interesting results – at least from a linguistic point of view – may receive a lower ranking or no ranking at all among the first 1000 hits, which tend to favour the advertising clients (Fletcher 2007: 30). Closely related to this issue is the “inbuilt local bias” of commercial search engines (Hundt, Nesselhauf and Biewer 2007b: 3). This means that search engines privilege hits that are closer to the location of the respective user. By guessing from the user’s IP address their location, commercial search engines rank and display results and advertisements by the presumed geographic proximity to the user’s home. Thus, depending on the geographic location of the researcher, she or he will obtain different results. While this is of course a viable marketing strategy, it may be problematic for linguistic analysis (Fletcher 2007: 30; Hundt, Nesselhauf and Biewer 2007b: 3). An additional factor influencing the data obtained from a commercial search engine is that search engines compile individual user profiles. Since researchers do not use the Internet only for language queries but also for personal matters, this also leads to a certain degree of idiosyncrasy in the results obtained from the ‘Web as Corpus’ approach (Fletcher 2007: 30-31; Hundt, Nesselhauf and Biewer 2007b: 3).

In view of these potential problems, quantitative analyses by means of the ‘Web as Corpus’ approach have often been deemed problematic. In particular, the low level of precision, the low recall, i.e. the high share of missed correct items, and the simple fact that commercial search engines are not developed for linguistic searches are deemed to severely affect the quality of quantitative analyses (Lüdeling, Evert and Baroni 2007: 12; Rosenbach 2007: 171). Hence, when simply relying on hit counts, the obtained frequencies are rather
unreliable, since they are influenced by many additional factors for which it is hard to control (Kilgariff and Grafenstette 2003: 12).

7.2.3 The World Wide Web as a Corpus Linguistic Resource: Prospects and Opportunities

The most obvious advantage of the World Wide Web over any traditional closed language corpus is the virtually infinite amount of data it can offer. Its data stock is not only inestimably larger than that of any other existing corpus, but it also offers an enormous wealth of different text types, registers and varieties which have so far not been sufficiently documented in closed corpora or which have so far not been documented at all (e.g. Bernardini, Baroni and Evert 2006: 9; Mair 2007: 233; Mukherjee 2009: 59; Rosenbach 2007: 167). The enormous amount of data offered by the World Wide Web becomes even more attractive in light of the fact that for some areas of linguistic research, even mega-corpora such as the BNC or COCA are not big enough. Examples include studies in lexicography, morphological productivity or language change (Hundt, Nesselhauf and Biewer 2007b: 1; Mair 2007: 235). In view of the limitations of existing closed corpora, it seems therefore to be a logical consequence to tap this enormous text collection for linguistic research in order to – at least – complement the data obtained from closed corpora (cf. Hundt, Nesselhauf and Biewer 2007b: 2; Mair 2007: 235; Mukherjee 2009: 62).

Another feature that makes the World Wide Web a very interesting corpus linguistic resource is its immediate and easy availability and accessibility. In general, Web-derived databases, including ‘Web-derived’ corpora and data obtained from the ‘Web as Corpus’ approach, are much more economically and often more quickly compiled than closed corpora are (Hundt, Nesselhauf and Biewer 2007b: 3; Ueyama 2006: 99). This is attributable to the fact that large amounts of Web data are freely available and easily accessible by means of an ordinary computer. Moreover, this is also because the Web data are already computerised and do not have to be transcribed (Fletcher 2007: 27; Mukherjee 2009: 59). Hence, despite all of the problems discussed in Section 7.2.2, the easy accessibility and ready availability of Web data makes the World Wide Web a very attractive linguistic resource, since it is possible to search billions of words for a specified search string within seconds and because search engine hit counts provide at least very crude but also very quickly obtained
frequency estimates (e.g. Cotterill 2010: 584; Kilgariff 2006a: 147). In addition, it has been argued that commercial search engines – and among them particularly Google – are so popular among linguists because of their user-friendliness, their familiar user interface and the neat presentation of the obtained results (Bernardini, Baroni and Evert 2006: 37).

Very closely related to the ready availability and the easy accessibility of Web-derived data is the fact that the data offered by the World Wide Web is very often “fresher” and more spontaneous than that of closed corpora (Fletcher 2007: 26; Lüdeling, Evert and Baroni 2007: 19). Indeed, the data included in traditionally compiled closed corpora are often quickly dated, sometimes even on the very day they are published (Hundt, Nesselhauf and Biewer 2007b: 2; Volk 2002: 2). Hence, the recency of the data on the Internet also makes it very interesting for many fields of linguistic research.

Closely related to the “freshness” of Web data is an aspect that has been mentioned as a main argument against the use of Web data, i.e. the constant change of the Internet and the resulting problems of replication. However, this problem is not particular to the use of either Web data or data obtained from the ‘Web as Corpus’ approach:

“A result computed today may not be exactly reproducible tomorrow. But, as (Kilgariff 2001) [sic] notes, this is [the] same for the water in any river and nobody will conclude that investigating water molecules is therefore senseless. We will have to learn to fish in the waters of the web” (Volk 2002: 9).

Indeed, the problem of replicability as well as most other challenges posed by the use of Web data and the use of the ‘Web as Corpus’ approach discussed in Section 7.2.2 offer the opportunity to critically address or re-assess fundamental issues and concepts that are relevant not only for studies using Web data but also for every other corpus-based study. This re-evaluation encompasses concepts such as the mode of compilation, the balance, the representativeness and the comparability of the data (e.g. Hundt, Nesselhauf and Biewer 2007b: 4; Leech 2007; cf. Section 7.5). In any case, a positive by-product of the noisiness of the Web data is that all issues concerning the quality, origin and composition of the data have become very prominent (cf. Bernardini, Baroni and Evert 2006: 19). These issues and questions, however, are relevant for empirical data in general and are often simply taken for granted when using traditional closed corpora without actually critically questioning them.

A further important aspect that makes the World Wide Web very attractive for corpus linguistic analysis is the linguistic diversity of the Internet. Languages and varieties of
languages for which no or no sufficiently large corpora have yet been compiled can be accessed and hence analysed using Web data (Fletcher 2007: 27; Hundt, Nesselhauf and Biewer 2007b: 1-5). By restricting the search to certain top-level domains, it is possible to distinguish between different languages and even varieties of languages (e.g. Rosenbach 2007: 177). Although it was stated in Section 7.2.2 that the restriction to a certain top-level domain without further data cleaning is only a “rough guide” to the regional origin of the data (Fletcher 2007: 36), having a rough guide is still much better than having no guide at all. As a matter of fact, corpus linguistics has – with the exception of the International Corpus of English – mainly focused on the so-called Inner-Circle varieties of English, i.e. varieties in regions where English is spoken as the dominant L1. For many other varieties of English, particularly the less thoroughly researched ones such as Fiji English or St. Helena English, there are very often no corpora yet available that allow for cross-varietal comparisons. Moreover, it is deemed rather unlikely that such corpora will be compiled in the foreseeable future (Hundt, Nesselhauf and Biewer 2007b: 1). However, analyses of low-frequency phenomena such as those that are the main focus of this study very often require even mega-corpora as database in order to yield robust quantitative results. Corpora of that size, however, are only publicly available for a very few varieties. Hence, if linguists want to analyse or examine varieties for which no corpora or no sufficiently large corpora exist, they have to tap other or additional sources (Hundt, Nesselhauf and Biewer 2007b: 1).

Furthermore, it has been shown that with regard to the distribution of certain morphosyntactic variants that have been observed and attested in closed corpora, regional preferences are replicated on the World Wide Web (Mair 2007). In line with these insights from variationist studies are observations from CMC research. They note that despite the overall global character of the Web, many forums have a national rather than an international audience and that there are also many localisation effects and not only trends towards an increased globalisation on the Internet (Herring 2010b).

In addition, Web data are also very interesting for corpus linguistic studies due to the wide range of different text types which are represented on the Web and which can be easily retrieved by using a commercial search engine. Furthermore, only the World Wide Web allows for studies of the language of the Web itself. With the expansion of the World Wide Web, new text types have emerged and gained importance such as weblogs, Internet forums and e-mails, to mention but a few. These have not only become an object of study...
themselves but are also more or less exclusively available on the Internet (e.g. Fletcher 2007: 27; Leech 2007: 133). As a consequence, the Web offers the possibility to study traditional written and spoken modes of discourse and to examine new communication forms and text types which bear traits of both traditional written and spoken modes of communication (e.g. Crystal 2011; Hundt, Nesselhauf and Biewer 2007b: 1). In view of its wide range of text types and modes of discourse, the Web can be considered even more comprehensive in depicting the diversity and complexity of communication forms and modes than any available closed corpus can ever be. This alone makes the Internet and the data obtained from it an enormously attractive object of study not only for corpus linguists but for linguistics in general.

Finally, the World Wide Web and the ‘Web as Corpus’ approach permit a quick and easy realisation of many quantitative analyses of phenomena for which closed corpora, particular those representing different varieties of English, are simply too small. Although the results of these analyses may be affected by the problems associated with Web-derived data and the ‘Web as Corpus’ approach discussed in Section 7.2.2, the Internet still offers “revealing results, which can be confirmed by replication” (Leech 2007: 145). Indeed, studies by Keller and Lapata (2003), Mair (2006, 2007) and Rohdenburg (2007) show independently from one another that at least for English, the results and frequencies obtained from the Web correlate fairly well with those obtained from closed electronic corpora. Hence, in view of these parallels, even rather critical voices come to acknowledge that the results obtained from the ‘Web as Corpus’ approach and those obtained from traditional corpora may complement each other very well, particularly when analysing low frequency phenomena (Mukherjee 2009: 62). If there are no corpora to complement the ‘Web as Corpus’ data, scholars have to decide for themselves whether they are willing to rely on a “rough guide” (Fletcher 2007: 36; Meyer et al. 2003: 247) or whether they can or want to do without the opportunities the World Wide Web and the ‘Web as Corpus’ approach can offer (Rosenbach 2007: 182).

In sum, the advantages and the potential for linguistic analyses offered by Google and other commercial Web crawlers are widely acknowledged (e.g. Rohdenburg 2007: 205-206). Whenever the requirements of a certain research question exceed the possibilities of closed corpora, it can be very useful to tap the vast amount of already digitalised texts that the Internet offers (Mair 2007: 233; Rohdenburg 2007: 205-206). Hence, when the starting point
for a study is a certain linguistic phenomenon, the Web may offer a very good option to test the underlying hypotheses. This is particularly true for those contexts in which traditional closed corpora cannot provide enough tokens to allow for a quantitative analysis of the phenomenon in question (Rosenbach 2007: 170, 182). Thus, in many cases, existing closed corpora have to be complemented by other data sources. And indeed, the Internet has been considered particularly suitable for complementing existing corpora (e.g. Hundt, Nesselhauf and Biewer 2007b: 4; Mair 2007: 235; Mukherjee 2009: 62). In view of the data scarcity for many varieties of English and for the study of many linguistic phenomena, it has been affirmed that “the web is and probably will be one of our best sources of information” (Hundt, Nesselhauf and Biewer 2007b: 3).

7.2.4 The Internet Data Collection for this Study

We have already noted that subject predicative complements are so rare in closed corpora that most corpora allowing for cross-varietal comparisons do not contain enough data to allow for a quantitative analysis of these phenomena (cf. Section 7.1.1; Quinn 2009). As a consequence, this study uses both closed mega-corpora, i.e. the BNC and the COCA, and a set of prospectively compiled Web-derived datasets as databases for the quantitative analyses in Chapters 8–14. The motivation for this study’s reliance on Web data should have become apparent in the preceding section. Although the use of data obtained from the Web is not without problems, Section 7.2.3 has pointed out that data obtained from the Web may be the only option for examining low-frequency phenomena, such as it-clefts and it BE sentences, in varieties for which no closed mega-corpora comparable to the BNC or the COCA exist. Moreover, several studies have demonstrated that regional distribution patterns of morphosyntactic features which can be observed in closed corpora are replicated by top-level domain-sensitive Web queries (e.g. Mair 2007; Rohdenburg 2007). Thus, this study assumes that if there are indeed cross-varietal differences in the distribution of pronoun case forms in subject predicative complements, they should also be reflected in the Web-derived datasets compiled for the respective regions or top-level domains.

The following Sections 7.2.4.1–7.2.4.3 outline the methodology applied for the collection of the Web data used for this study. Furthermore, important repercussions of the applied methodology on the results of this study will also be addressed.
7.2.4.1 The Collection of the Raw Data

The approach used for this study combines elements of both the ‘Web-derived’ corpus and the ‘Web as Corpus’ approaches introduced and discussed in Section 7.2.1. The reasons for this course of action are the following. On the one hand, the ‘Web-derived’ corpus approach offers many advantages since it allows the researcher to control for exactly what is included in the corpus data and results in finite, sampled corpora, thus evading many of the problems typically associated with Web data. On the other hand, the data are most often confined to a particular text type or genre such as newspapers (cf. Schilk 2011) or TV-transcripts (cf. Hoffmann 2007). Since this study intends to take different modes of discourse into consideration (cf. Section 6.1.2.4), this study uses elements of both accounts because the ‘Web as Corpus’ approach – or simply the use of search engines – retrieves a much wider spectrum of different text types, genres and also modes of discourse than the ‘Web-derived’ corpus approach typically does (cf. Section 7.2.1). This is not to say that the approach adopted in this study is superior to those of previous studies with different approaches, but simply that “[d]ifferent phenomena and different research questions will require different tools and methods” (Rosenbach 2007: 182).

Before it is outlined exactly how the data have been obtained, some general remarks are in order as to how the Web data collection for this study differs from that of the corpus data discussed in 7.1. First of all, the Web queries for this study were conducted only for it is + I/me/he/him/we/us/they/them. This means that in contrast to the corpus data (cf. Section 7.1), the Web queries are restricted to only one form of BE, i.e. is. This restriction is due to the sheer amount of data retrieved by the Web queries and the enormous effort required to clean, refine and categorise the data points, which is discussed in detail in Sections 7.2.4.2 and 7.2.4.3. Hence, taking into consideration all possible forms of BE would have been beyond the scope of this study. The reason why the form is was favoured over was and ‘s is simply that the combination of it is + PRONOUN yielded considerably more hits than the combinations of it’s + PRONOUN and of it was + PRONOUN. Another very important difference between the Web data and the corpus data used for this study is that the third person feminine pronouns she and her were not considered in the collection of the Web data. The exclusion of the third person feminine pronouns is a consequence of the syncretism of the object/possessive forms her. This syncretism results not only in too many false positives, which in this case means many possessive uses of her, but also in low recall,
i.e. a very low number of correct hits compared to the other pronoun pairs at hand (Lüdeling Evert and Baroni 2007: 12). Hence, the feminine pronoun forms were not considered since their inclusion might have severely skewed the results. Furthermore, pronominal gender has not been mentioned thus far in the discussion of pronoun case variation in subject predicative complements as a factor potentially influencing the distribution of pronoun case forms (cf. Section 6.1.2). Thus, it seems legitimate to neglect this variable in the collection of the Web data for this study and for the subsequent analyses.

How exactly the Web data used for this study was obtained is now explained in detail. First of all, the respective search string it is + PRONOUN was entered into the Google Advanced Search box called “this exact wording or phrase” at the following Website: http://www.google.com/advanced_search?H=en. Next, the default set up was changed from ten to 100 displayed hits per page in order to get as many hits as possible displayed on one page. Then, it was determined in the “Language” search box that all displayed pages should be in English. Finally, the search results were restricted to the respective top-level domains relevant for this study. These are .uk for the United Kingdom, .au for Australia, .ie. for Ireland, .za for South Africa and .in for India. So far, this approach is analogous to that employed by other scholars (e.g. Mair 2007; Rosenbach 2007).

Then, in order to conserve the obtained results and to clean, classify and verify the data, each of the results pages retrieved by Google was saved as an html file. Since Google restricts the displayed results for each query to a maximum of 1000 hits (cf. Fletcher 2007: 37), this means that for each combination of it is + PRONOUN in each top-level domain, a finite sample of maximally 1000 hits could be obtained, i.e. a maximum of ten html files per search string were saved. In most cases, however, the number of obtained hits was considerably lower. The saving of these results or html files basically led to the compilation of a “home-made” corpus, at least according to Kilgariff and Grafenstette’s (2003: 2) definition of the concept of corpus (cf. Section 7.2.1). Furthermore, saving the results helps to avoid the momentariness of the results, since Google results are virtually irreplicable (cf. Section 7.2.2). Thus, a ‘Web-derived’ database by means of the ‘Web as Corpus’ approach was compiled. Moreover, by simply using all results retrieved by Google for each search string, a certain degree of exhaustiveness was achieved. With regard to the unknown token size of the respective top-level domain, this should not be a problem for the subsequent analyses for the following reasons. As argued in Section 6.1.1, this study assumes that the
choice between pronoun forms in subject predicative complements following *it* and a form of *BE* is a binary one, i.e. a choice between subject and object pronoun forms. Hence, all things being equal, the same ratios in the distribution of pronoun case forms should be observed across the different top-level domains. Furthermore, since this study uses binary logistic regression models for the statistical analyses and not tests that require knowledge of the population size (cf. Section 6.2), the exact token number of the analysed top-level domains is not necessarily needed to arrive at robust results. With regard to the lack of control of what is included in the hits retrieved by Google, the most severe effects of this limited control is made up for by the subsequent data cleaning and quality control of each and every one of the obtained data points (cf. Section 7.2.4.2).

In addition, the data collection was conducted in a very limited period of time, i.e. on the 14th and 15th of July 2008. This was done in order to obtain a synchronic snapshot of the data in the respective top-level domains. The data could not be collected on a single day because of the usage restriction posed by commercial search engines. As noted above (cf. Kilgariff 2006a: 147), commercial Web crawlers such as Google or Yahoo limit the number of queries per user per day. Hence, the data collection had to take place on two subsequent days, because it was not possible to do it on a single day. These usage restrictions are also the main reason that Google was used as the search engine. Other search engines are even more restrictive in terms of displayed search hits per user per day, as Figure 4 illustrates:
Although it would have been desirable to conduct the data collection in a time frame as small as possible, it is rather unlikely that this time lag of one day significantly influences the outcome of this study.

In view of the shortcomings of commercial search engines, the question might arise as to why a so-called Web-concordancer, i.e. a linguistically refined search engine, was not used to conduct the data collection for the present study (e.g. Bernadini, Baroni and Evert 2006: 11). The reasons, however, such a tool was not used are the following. Firstly, Web-concordancers such as WebCorp are themselves dependent on commercial search engines (e.g. Mukherjee 2009: 61, Lüdeling Evert and Baroni: 2007: 16). Thus, they actually inherit some of the most serious problems of the ‘Web as Corpus’ approach from the search engines on which they rely. In the case of WebCorp, for example, the search for abstract search strings is severely restricted. It is also unclear which parts of the Web are searched by the Web-concordancer and the results are also never exactly replicable (Lüdeling, Evert and

---

31 This screenshot was made by the author when trying to collect data for the search string it is me (cf. 7.2.4.1) with the help of yahoo.com on 4th August 2008. The screenshot illustrates that this was not possible due to the more severe restrictions of yahoo.com.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Baroni 2007: 16; Rosenbach 2007: 176-177, 182-183). In addition, the program is considered to be rather slow and susceptible to breakdowns (e.g. Rosenbach 2007: 182-183). Secondly, Web-concordancers yield considerably fewer results than the commercial search engines from which they draw their hits (e.g. Lüdeling, Evert and Baroni 2007: 17). This is a serious problem, particularly for the analysis of low-frequency phenomena in smaller top-level domains. In the present case, even Google returns rather few hits for potential subject predicative complements in some of the analysed regional domains, as will be demonstrated below (cf. Section 7.4). Thirdly and most importantly, however, Web-concordancers do not solve the problem of thoroughly cleaning, verifying and classifying the obtained raw data. Hence, in the present case, commercial search engines seem to be the better choice.

7.2.4.2 Refining the Raw Data: The Cleaning and Verification of the Data

Although the task of cleaning data obtained from the ‘Web as Corpus’ approach is “prohibitively time-consuming” (Lüdeling, Evert and Baroni: 2007: 13), there is a simple rule that “the better it is done, the better the outcomes” (Kilgariff 2006a: 149).

In a first step of refining the crude data obtained from the ‘Web as Corpus’ approach, the html files were uploaded into the concordancing program MonoConc Pro 2.2. By activating the Sentence mode function which filters out search strings that are interrupted by sentence boundaries, a considerable amount of junk can be removed. Examples for false hits identified and filtered out by this simple procedure include the following:

(118) Indeed it is. I think a bounce is probably near, but we might see some selling in the meantime. (ie/it is l/14.07.2008)
(119) This antipodean superstar makes the programme what it is. He is not one of those presenters who is so bad he is good. (.uk/it is he/15.07.2008)

In the next step, with the help of office software, instances in which the search string it is + PRONOUN is interrupted by commas are also filtered out:

(120) It is, I believe, doomed to be obsolescence. (.au/it is he/14.07.2008)
(121) It is, he says, “a dream come true” for any archivist and representative of what he regards as a very forward thinking approach at GROS. (.uk/it is he/15.07.2008)

It is not a coincidence that the examples in (118)–(121) illustrate instances in which subject pronoun forms are affected, since they are more likely to occur in clause-initial position and hence are more strongly affected by the fact that commercial search engines ignore
punctuation marks, even if the predefined search string is set in quotation marks (cf. Section 7.2.2). Unfortunately, this practice leads to a substantial number of false positives (cf. Keller and Lapata 2003: 468). The immediate repercussion of the search engine’s removal of punctuation marks is therefore a certain skewing of the results. Compared to the corpus data of the BNC and of the COCA, a smaller proportion of subject pronoun forms in subject predicative complements following it is can be expected. This is due to the fact that while the Web-derived data for this study consists of finite samples of maximally 1000 Google hits, these finite samples suffer from a larger share of false hits due to the search engine’s ignorance of punctuation marks. Nevertheless, this skewing effect should not be a fundamental problem in the present study because we can predict its outcome, i.e. a lower share of subject pronoun forms in subject predicative complements in the Internet data when compared to the corpus data. In particular, the non-consideration of punctuation marks particularly influences the data for the first person singular, which suffers from the highest share of false positives due to the disregard of punctuation marks. As outlined in Section 7.2.2, the data may also be influenced by the local bias, the individual user profile and the overall commercial character of the search engine used for the data collection. However, since the same search engine and the same desktop computer were used to collect the data for this study, these effects should influence the data from the different top-level domains to the same extent. Thus, although these skewing effects make a comparison of the Web data to the corpus data even more complicated (cf. Section 7.5), we should still obtain meaningful results because we can predict the outcome of the most severe skewing effect, i.e. the removal of punctuation marks, and keep other important potentially interfering variables, such as the local bias and the individual user profile, constant for all Web-derived data.

After this first rough and ready elimination of many irrelevant structures and false positives, the more elaborate and sophisticated stages of the data cleaning were conducted. In order to verify that the data are indeed from the region indicated by the top-level domain in which they occur, each of the remaining data points was googled again to assess whether it can actually be assigned to the country or region in question. This was done by augmenting the it is + PRONOUN search string of each data point with some idiosyncratic context, often simply by searching for the whole or parts of the clause in which the subject predicative complement occurs. Then, the phrase was set in quotation marks and a Google
query without a top-level domain restriction was conducted for each and every one of these contextually expanded data points. If more than one search hit for an expanded data point was obtained, it was tried to determine where this data point originally came from. If it was not possible to ascertain its origin, the data point was excluded from further analysis. In general, data points were excluded when their origin could not be determined or if they were attributable to a person who was not from the respective region or country. This means that even quotes of, for example, French footballers or British politicians that occurred in Irish or Australian newspaper articles were excluded. In a similar vein, data points that occurred on Websites which could not clearly be attributed to a certain person, company, corporation, club or institution and hence the origin of which remained obscure were not considered for further analysis. This practice was administered restrictively so that in cases of doubt the data were always excluded. Moreover, the data was carefully examined in order to identify and exclude duplicates and near-duplicates such as slightly altered newswire texts. Indeed, much labour and time was devoted to this quality control in order to obtain the best possible data (cf. Kilgariff 2006a: 148). By applying this time-consuming procedure, every effort was taken to minimise the quality problem often ascribed to Web-derived data (cf. Section 7.2.2). Like in the BNC (cf. Burnard 2007), data points that were clearly older than 1975 were likewise excluded. The same is true for data points that occurred on URLs that could not be accessed during this verification and cleaning process. Possible reasons for this inaccessibility are, for example, that the URL has become inactive or that the Website is password-protected.

7.2.4.3 Refining the Raw Data: The Classification of the Data

Another laborious but necessary stage that immediately accompanied the verification process of the Web-derived data was their classification in terms of text type and mode of discourse. Since Web-derived data do not usually contain any metadata (cf. Lüdeling Evert and Baroni 2007: 15), this information had to be added manually in order to allow for a mode of discourse analysis as described in Section 6.1.2.4. The basis for the mode of discourse classification adopted in this study forms, however, a classification of the respective text types. This classification, both in terms of text type and mode of discourse, is
far from being trivial because of the peculiarities of Web-derived data compilations, particularly when compared to traditional closed corpora.

On the one hand, closed corpora usually start with a predefined corpus design which meticulously determines the modes of discourse, the conversation and text types, the different registers and the exact amount of data to be included for each of these categories in a corpus (cf. Nelson 1996: 29-33). Hence, by applying such a top-down approach, compilers of closed corpora cannot only control how much of which types of data are included in the respective corpora, but can also try to include only prototypical or at least uncontroversial specimina of their predetermined text and conversation types in order to match the established categories. Furthermore, compilers of closed corpora can also predetermine levels of text type granularity, i.e. the degree of fine-graining according to which they want to distinguish their text types and the types of conversations they want to include in their corpora. Depending on how coarse- or fine-grained the distinctions of these text types are, and depending on how many of these text types are included within a corpus, corpus analyses of a given feature may even exhibit considerably more marked differences across different text types within a variety than within a certain text type across a set of different varieties (e.g. Sand 2004: 294-295).

On the other hand, the situation is very different for Web-derived datasets, particularly for data obtained from the ‘Web as Corpus’ approach, since we have to apply not a top-down, but a bottom-up process of data classification for these datasets. This is due to the fact that we have very little influence on which kinds of data and how much of it are actually retrieved by the commercial search engine used for the data collection. This also implies that we have no influence either on the prototypicality of the text types we are confronted with or on the level of granularity according to which these texts can or should be distinguished, which results in many cases in which it is extremely difficult to convincingly motivate the assignment of a text to a certain category. Moreover, this difficult situation is further complicated by the fact that many Websites may contain different text types or even different modes of discourse. A prototypical example for such a text type or even mode of discourse mix on a Webpage is a news report or editorial in an online newspaper or journal with a comment section at the bottom of the page (cf. also Ueyama 2006: 107). In view of these challenges, it has been concluded that
“[w]hile the extensive literature on text classification [...] is certainly relevant, it most often starts from a given set of categories and cannot be readily applied to the situation where the categories are not known in advance” (Kilgariff and Grafenstette 2003: 11).

As already mentioned in Section 6.1.2.4, a very fine-grained text type classification would also result in very small data groups for some top-level domains the result of which could hardly be considered representative of the respective text or conversation types, and therefore would not allow for robust generalisations. In view of all these obstacles, a rather basic mode of discourse distinction seems to be the best way to circumvent many of these problems.

However, in order to arrive at a mode of discourse distinction, at least a very coarse-grained text type classification is indispensable. As a consequence, this study follows Ueyama (2006: 107) in adopting a heuristic approach for classifying the obtained Web data. This means that for the current study, appropriate samples of the data from the different top-level domains were preliminarily inspected in order to assess what kinds of data are included in the samples (cf. Ueyama 2006). Furthermore, the text types are often hinted at by the headlines, names or URLs of the respective Web pages. As a result of this pilot-classification, the data has been grouped into 26 broad categories, which could be further broken down into more fine-grained classes, across the three different modes of discourse. In general, the mode of discourse distinctions applied for this study are based on the conventions used for the BNC (Burnard 2007) and on the insights of CMC research (e.g. Crystal 2011; Domsch 2009; Herring 2007, 2010 a, 2010b). For the text type classification underlying the mode of discourse distinctions, recourse may also be taken to the design of the ICE corpora where that is further illuminating (cf. Nelson 1996). Again, it should be stressed that the following text type classification is tentative and heuristic in nature and was conducted solely in order to arrive at a mode of discourse differentiation necessary for the quantitative analyses of this study (cf. Section 6.1.2.4). By no means does this study challenge the usefulness of literature on text classification (e.g. Manning and Schütze 1999: 575-608).
7.2.4.3.1 The Classification of the Data: The Spoken Data

Although it might seem rather unexpected to obtain “spoken” data from the ‘Web as Corpus’ approach, the commercial search engine used for this study actually retrieved a considerable amount of transcripts of originally spoken data. These data can be assigned mainly to what is called “context-governed” data in the BNC, such as transcripts of meetings, debates, seminars and radio programmes (Aston and Burnard 1998: 31-33). Where applicable, the following classification is loosely based on that of the BNC (cf. Burnard 2007) and the ICE (cf. Nelson 1996). The different text types distinguished in the spoken mode of discourse in this study are the following:

* Court-transcripts: This category consists of several kinds of transcripts from legal proceedings, such as testimonies and examinations. In the South Africa data, this category also includes the records of the Truth and Reconciliation Commission, a court-like institution established to overcome the crimes of the Apartheid regime.32

* Interview-transcripts: This class of transcripts is more heterogeneous. It includes a wide spectrum of interviews such as interview transcripts of certain persons and even celebrities for newspapers or journals, as well as interview transcripts conducted for and included in academic papers or official reports as well as transcripts of press conferences.

* Political-transcripts: This cover term comprises different types of transcripts from clearly political institutional settings, such as transcripts of regional and national parliament debates, records of parliament hearings, transcripts of commissions of enquiry and the like.

* Sermons: This class includes sermons of different religions and religious denominations. Sermons are granted a class of their own since they occur rather frequently in the Web-derived datasets compiled for his study. Moreover, this class of texts also receives a separate category in closed corpora such as the BNC (cf. Burnard 2007).

* Speeches: This category includes a wide range of speeches from very different institutional settings. In particular, this class contains political speeches, speeches from academic settings, business meetings and conferences. Hence, this category is much more inclusive than the corresponding categories of the BNC (cf. Burnard 2007).

Broadcast-transcripts: This class encompasses transcripts of both radio and TV shows as published on the Internet. However, since many of these transcripts are interviews, it was sometimes difficult to decide whether a text is to be considered primarily an interview or a Broadcast-transcript. Hence, there are sometimes borderline cases in both classes distinguished here that could have been assigned to either category.

Miscellaneous-transcripts: Finally, this class contains those data points occurring in spoken text types that could not readily be assigned to one of the above categories and that are too infrequent to constitute a class of their own. Cases in point are, for example, lecture transcripts and phone transcripts. Some of the texts included in this class could also be assigned to other categories. Lectures, for example, are very often monologue-like and hence are very similar to speeches. However, they are not classified as speeches since the institutional setting is more specific and also because they are not subsumed under such a category in the BNC (Burnard 2007).

7.2.4.3.2 The Classification of the Data: The Written Data

Even though the number of categories distinguished for the written data in this study is greater than for the spoken mode of discourse, the categories are still more inclusive and rather broad in order to acknowledge the higher degree of heterogeneity of the written data with regard to the different text types to be considered. In addition, while most of the spoken data corresponds to the more formal “context-governed” data in the BNC (Aston and Burnard 1998: 31-33), the written data retrieved by Google is less uniform in terms of formality, which also contributes to the internal heterogeneity of the text categories postulated here. Since the classification of the written text types in the BNC is itself very broad (cf. Aston and Burnard 1998: 28-41), reference may also be made to the corpus design of the ICE where necessary in order to better illustrate the differences of the text categories applied here (cf. Nelson 1996). The different text categories distinguished for the written mode of discourse in this study are the following:

Academic texts/articles: This category includes academic articles and texts from very heterogeneous subject areas and research fields such as theology, philosophy, medical sciences, law, and gender studies, to mention but a few. Furthermore, student essays and papers such as term papers and different kinds of theses are also included in this category if they are clearly from a tertiary education background.
Advert-Info: Comprised in this class are again a wide variety of texts. Exemplars of this category not only provide information on a certain issue, company, corporation or person, but also have the intention of advertising, in a very broad sense, the issue, company, corporation or person in question. Examples for text types subsumed under this heading are company details, advertising or promotional texts, job advertisements, campaign advertising or lonely-hearts ads.

Article: This category contains a wide spectrum of informative news reports and articles on a multitude of different topics as well as non-academic articles on certain more or less specific topics. If one wanted to compare this category to that of closed corpora, it could be stated that this class as applied here includes the ICE categories, Popular Writing, Reportage and parts of the Instructional Writing section (cf. Nelson 1996: 29). Thus, the category Article as used here is again very broad.

Editorial-article: In this domain, editorials, commentaries and also essays the main foci of which are more persuasive than informative are integrated to form a category of their own (cf. Nelson 1996: 29).

Info: This is one of the largest and most heterogeneous categories used in this study. It comprises texts the least common denominator of which is “providing information”, which allows of course the inclusion of a very broad range of different texts and makes it a kind of default category. This class includes, in analogy to Ueyama’s (2006) info class, texts providing “information that pertain to initiatives, events, resources, and projects related to a certain topic without commercial or educational purposes” (Ueyama 2006: 109). The category used here is, however, much more inclusive than Ueyama’s info class since it also includes his categories njews, containing “non-journalistic news, such as community pages” (2006: 108) and his category personal, consisting of “personal homepages not created through a blog service” and hence being “less interactive” (2006: 109), rather providing information than allowing for discussions. Furthermore, this class also contains a rather broad variety of texts not necessarily represented in Ueyama’s classification scheme (2006: 108-110), such as instruction manuals, guidelines, club info pages, etc. The heterogeneity of the texts included in this category indicates the tremendous diversity of the data retrieved by the adopted ‘Web as Corpus’ approach.

FAQ/Info: This category subsumes data points that are presented in the form of FAQs, i.e. frequently asked questions, Q&As, i.e. questions and answers, etc. Although this class
basically constitutes a subclass of the *Info* category discussed above, it seems reasonable to distinguish it from the former: On the one hand, this text type has a clearly distinct form because the information in this text type is always provided in a strict question and answer format. On the other hand, this particular mode of presenting information occurs at such a high rate not only in the present data but also in the datasets of similar studies (Ueyama 2006: 109) that a categorisation of these data points in a separate class is justified.

*Legal text:* This group contains a wide range of legal texts such as solicitor’s letters, court opinions, petitions and even texts of laws. Although this kind of text is not normally distinguished explicitly in closed corpora, such as the BNC or the ICE, the distinction of this category in the present study seems reasonable given the frequent occurrence of texts from this particular subject area across the different top-level domains tapped for this study.

*Literary text:* Although the title of this class may suggest a narrower focus, this group actually encompasses a very wide range of fictional, literary and poetic texts. In particular, this class is composed of heterogeneous texts such as poems, dramas, lyrics of pop songs and many other different kinds of fictional texts including novels, short stories, fairy tales, etc. Thus, this category more or less corresponds to the class of *imaginative* texts in the BNC or to the *Creative* texts included in the ICE components (cf. Burnard 2007; Nelson 1996: 30)

*Quote in article:* This category subsumes data points that occur, for the most part, in *Articles* and *Academic texts/articles* as discussed above but are clearly marked as quotations trying to represent actual direct speech. Due to the fact that these data points are attempting to reproduce spoken language, it could be argued that these data points should be assigned to the spoken mode of discourse rather than to the written one. However, it is unclear to what extent this “spoken” data has been adjusted to meet the editorial and stylistic requirements of a given newspaper or journal. In addition, since these quotations are clearly parts of written texts, they are included in the written data. Furthermore, traditional closed corpora also consider this type of data as “written” since they do not normally distinguish between “spoken” or “written” contents of an article (cf. Burnard 2007; Nelson 1996). Indeed, the major reason for making this category was a methodological one because this class was very useful for the quality control of the data. This group of data points contained a disproportionally high share of quotes from people who were not from the top-level domain or region in question. Moreover, this category also helped to identify doublets and near-doublets probably caused by slightly altered newswire texts, which were also excluded.
Religious text: This text type also does not often receive a category of its own – either in closed corpora (cf. Burnard 2007; Nelson 1996) or in Web-derived corpora (cf. Ueyama 2006: 106). However, in view of the frequent occurrence of texts with a clearly religious or spiritual background in all top-level domains considered for the present study, the determination to include a distinct religious category seems justified. In this class, a broad range of texts from many different religions, religious denominations and quasi-religious groups and beliefs ranging from Indian Bhagvans to Irish Catholics are subsumed. Unfortunately, this text type is often difficult to distinguish from other similar text types, some of which are to be assigned to other modes of discourse, such as theological academic articles, blogs with a clearly religious message and sermons.

Report: This category is not always easy to distinguish from the Info and Academic texts/articles domains introduced above. The chief criterion for this category is the post hoc presentation or recounting of the relevant information, for example, after the respective event took place or after the presented results were obtained and in a form that cannot be considered appropriate for Academic texts/articles. Again, this class is very heterogeneous since it includes reports of conferences or business meetings on the one hand and rather personal or subjective field reports on a certain topic or experience on the other. Furthermore, this category as used here also includes diaries. This subcategory is difficult to deal with because these texts are difficult to distinguish from blogs and thus could be assigned to the CMC mode (cf. Section 7.2.4.3.3). However, Web diaries were only considered blogs if they clearly exhibited features of Computer-Mediated Communication such as the possibility of leaving comments on the Website or were called blogs rather than diaries (cf. Herring 2010a).

Official reports: This group of texts represents a subclass of the Report category. It seems again reasonable to distinguish this class from the former since this group of texts is more or less restricted to official government, parliament or legal reports, which may be different in terms of style and register from personal or even very subjective field reports written by an individual on a certain event or experience. Hence, a distinction or subdivision of the Report category seems advisable, in order not to overburden it. Finally, instances of this category occur regularly in the datasets compiled for the different top-level domains in this study.

Review: The fact that this text type has received a category of its own is due to the frequent occurrence of this text type in the data samples. Moreover, the distinction of this category is
also attributable to the fact that this text type falls into two different modes of discourse (cf. Domsch 2009), which further complicates the text type and mode of discourse classification. The class of texts discussed in this category comprises the types of reviews which can typically be found in newspapers and other print media, such as book, film, music, concert, theatre or exhibition reviews (cf. Domsch 2009: 227). It does not include consumer or customer reviews on Amazon.co.uk or similar Websites. This latter kind of review is referred to as a “computer-mediated sub genre” (Domsch 2009: 233) in the relevant literature and is consequently assigned to the CMC mode of discourse discussed below (cf. Section 7.2.4.3.3).

Teaching materials: Subsumed under this heading are a wide range of texts and materials specifically designed for teaching or teaching situations. These texts have in common that they clearly pursue instructional purposes in the sense of conveying knowledge. Examples include exam texts, language quizzes or PowerPoint presentations (cf. Ueyama 2006: 107).

7.2.4.3.3 The Classification of the Data: The Computer-Mediated Communication Data

Finally, a coarse-grained text type classification is also needed for the data of the third mode of discourse distinguished in this study, i.e. Computer-Mediated Communication or CMC for short. Although the emergence of this discourse mode is a rather recent phenomenon, it has already become widely accepted to situate this discourse mode in the interface of written and spoken language (e.g. Crystal 2011: 19-35; Herring 2010a). However, this middle ground represents a vast continuum in and of itself (Herring 2010a). Hence, it is not surprising that a considerably degree of heterogeneity can be observed in the different CMC categories distinguished for this study. For the text type classification underlying the mode of discourse distinction in this study, the following categories are distinguished for the CMC data:

Blog: This category contains data points that occur in weblogs, or blogs for short, except for data points included in the corresponding comment sections. Comments are grouped under the Post/Comment category discussed below. Hence, this class is restricted to the blog postings or entries of the blog authors and does not include the responses to them (cf. Herring 2010a). The main reason for keeping the blog texts apart from the responses evoked by them is the fact that the latter are often considered much more “conversational” than the former (Herring 2010a). Indeed, the actual blog texts are often very similar to personal reports, on the one hand, and editorials or commentaries, on the other hand, both of which
are assigned to the written mode of discourse. Hence, it is reasonable to distinguish blog texts from blog comments since the same is done for articles and editorials and the responses given to them. The blog texts alone exhibit a very high degree of diversity in terms of form and content, as do all CMC categories distinguished in this study. This very broad range of texts includes not only very personal texts which are, in form and content, clearly informal and similar to diaries, but also many journalistic blogs from the Websites of major newspapers such as the British *The Guardian* or the Australian *The Age* which are, in form and content, similar to editorials or commentaries (cf. also Domsch 2009: 229).

**E-mail**: This class contains all data points from texts that are clearly identifiable as e-mails, which is, of course, another well-established CMC text type (e.g. Herring 2010a). However, the Google-based data collection conducted for this study retrieved rather few instances of this category.

**Forum post**: This category contains all posts in forums or on bulletin board sites on which Internet users asynchronously communicate with each other, often in the form of a “topic-comment structure” (cf. Ueyama 2006: 108). In addition, this class includes mailing lists, which bear a resemblance to e-mails and could be assigned to that category as well. Since mailing lists are, however, in a way the predecessors of forums, focussing on an exchange of information and opinions within a certain community, they are considered to be more similar to forum posts than to e-mails in the present study and hence are included in this category. As was the case with blogs, this category exhibits again a high degree of internal heterogeneity both in terms of content and formality since there are Internet forums available for basically all kinds of topics and all kinds of audiences. A phenomenon that is particularly often encountered in this text type is what has been called “document-internal duplication” (Bernardini, Baroni and Evert 2006: 23). This means that other forum users may copy a certain chunk of language and paste it into their responses. These instances were excluded from further analysis in accordance with the restrictive duplication policy adopted for this study.

**Post/Comment**: This class contains all instances of reader comments and responses both to blogs and to all kinds of articles, editorials or commentaries. Although it is often difficult to distinguish this class from blogs on the one hand and forum posts on the other, it still seems sensible to distinguish this category from the former two. Firstly, this can be justified by the sheer amount of data points. This category is the second largest CMC subdomain
distinguished in this study; only Forum posts have more occurrences. Secondly, the posts included in this category have in common that they are responses to another text or blog posting, which forms a kind of shared background – a feature that is also shared by many forum threads and posts. These articles or blogs often decisively influence not only the topic but also the major focus or point of view of the discussion. Thirdly, many of these posts are comments on articles and blogs being published on the Websites of major newspapers or journals, some of which require the full name of the poster. Thus, many posts included in this category can be considered to be more public than a post in a thread of a special interest forum simply because the potential readership is more general. As a consequence, there may also be stylistic differences between forum posts included in the previous category and the posts and comments of Internet users subsumed in this class.

**CMC Review:** This category includes those reviews typical of Computer-Mediated Communication: mainly consumer reviews from Websites such as Amazon.co.uk, Ciao.co.uk or Barfinder.com.au (cf. Domsch 2009). Although these texts display many similarities to the traditional Review category discussed in the written mode of discourse section, the reviews discussed here exhibit certain features that clearly set them apart from those discussed above. To begin with, there are virtually no restrictions on the potential authorship of CMC reviews. While traditional written and published reviews are often authored by experts of the respective subjects or objects, anyone can write a consumer review on Amazon (cf. Domsch 2009). A further notable difference is the diversity of reviewed objects and subjects included in this domain. In the data used for this study, this CMC subclass includes not only consumer reviews of books, films, CDs and restaurants, all of which can also be found on the feature pages of newspapers, but also reviews of dog collars, tea kettles, escort services and even whole cities. Furthermore, these two important differences between traditional reviews and CMC reviews are also connected to a third one, i.e. the lower degree of distance between the authors and the readers of this type of review. The fact that these reviews are often written by ordinary people in an ordinary language style makes them seem trustworthy to many people because they are deemed to represent opinions similar to those of the respective reader (Domsch 2009: 226). A final difference between this kind of review and that of the traditional written mode is that these reviews are subject to reviewing themselves. Very often, readers can comment on how “helpful” they consider a review and
thus can influence the rating of the review and even of the reviewer on the respective Website (Domsch 2009: 234).
7.3 Data Points Excluded from the Statistical Analysis

Following Biber et al. (1999: 335-336, 1134), this study analysed all instances of the search strings \textit{it + is/was'/s + I/me/he/him/she/her/we/us/they/them} occurring in both the BNC and the COCA, which yielded all potential instances of \textit{it} \textit{BE} sentences and \textit{it}-clefts with a case-sensitive focal pronoun (cf. Section 7.1). In a similar vein, the Web-queries for this study were conducted for the search strings \textit{it is + l/me/he/him/we/us/they/them} to obtain the potential candidates for \textit{it} \textit{BE} sentences and \textit{it}-clefts for these search strings in the respective top-level domains (cf. 7.2.4.1 for further details). However, in order to obtain only the clear instances of \textit{it} \textit{BE} sentences and \textit{it}-clefts, all ambiguous constructions and those irrelevant for this study were excluded from further analysis. As this data refinement had to be conducted for both the corpus and the Web-derived datasets used for this study, this exclusion process was uniformly executed on all datasets. Database-specific refinement and exclusion measures have already been discussed in the relevant sections (cf. 7.1.3, 7.2.4.1 and 7.2.4.2). Constructions excluded from this study are the following.\textsuperscript{33}

First of all, all constructions were excluded which exhibited the respective search string as defined above – \textit{it} followed by a form of \textit{BE} and complemented by a case-sensitive pronoun form – but in which the constituents of the search string were not part of the same clause:

\begin{verbatim}
(122) Search String Split in Different Clauses
  a. Whatever it is I’ll need a lot of prayer. (BNC/CC8/W_misc)
  b. The only person who can fix it is him, so unless he cuts down big time, you’re better off without him. (.ie/it is him/14.07.2008)
\end{verbatim}

Moreover, sentences were excluded in which it was not possible to decide whether the search string represented an \textit{it}-cleft or \textit{it} \textit{BE} sentence, or whether the search string could also be interpreted as something else. Examples for this category of ambiguous cases include the following:

\textsuperscript{33} The following section is an extended and more comprehensive version of the corresponding section in Maier (2013).
In addition, instances in which the pronoun forms were used as either demonstrative or possessive pronouns were also excluded from further analysis.

(124) Demonstrative Use of Pronouns
a. It’s them damn calories, though. (COCA/1993/FIC/SouthwestRev)
b. We got a whole teacher got a whole lot of letters for some piece all about collecting cans for getting and it was them letters (BNC/HVB/S_interview_oral_history)

(125) Possessive Use of Pronouns
a. If it is me residential address; i have just [...]. (.au/it is me/14.07.2008)
b. The main reason is that it is me and my daughter’s favourite radio [...] (.uk/it is me/15.07.2008)

Also not considered were constructions which, although they are relevant in the discussion of contexts allowing for variability in the use of pronoun case forms in English (cf. Section 2.3), are not clear instances of either it BE sentences or it-clefts (e.g. Huddleston and Pullum 2002: 458-467; Quinn 2005a: 2). On the one hand, examples in this category include sentences in which the pronoun form is followed by a noun phrase. In these instances, it is normally the first person plural pronoun form which is modified by an integrated dependent in the noun phrase structure. This modifier is usually semantically restrictive in that it limits the denotation of the nominal head and typically provides identifying information (Huddleston and Pullum 2002: 447; cf. Section 2.3.1):

(126) Pr–NP Constructions
a. I believe that if any country has a good chance to fix the real problems, without yielding to the status quo of the international community, it is us Australians (no patriotism intended). (.au/it is us/14.07.2008)
b. I say it is we traditional healers. (.za/it is we/15.07.2008)

On the other hand, this category includes constructions in which the pronoun form is followed by a gerund participle and may act as its subject (Huddleston and Pullum 2002: 1190-1193; cf. Section 2.3.1).

(127) Pronoun + –ing Form
a. He said when I see the wind moving the leaves, it is him waving to me and I should smile his gorilla smile when that happens. (COCA/2006/SPOK/NPR_Saturday)
b. Foster, whose own career started at the tender age of 3, discovered 9-year-old Adam in a New York classroom: "It was him wanting to be something that drew me," she says. (BNC/ED3/W_pop_lore)

Although sentences such as (127a) are semantically similar to *it*-clefts, these gerund constructions are difficult to analyse. This is due to the fact that very often, it is not clear whether the personal pronoun form should be interpreted as the complement of *it* and *BE*, followed by a non-finite clause or whether the pronoun case form belongs instead to the following gerund, constituting a part of the non-finite clause which, as a whole, functions as the complement to *it* and the form of *BE* (Huddleston and Pullum 2002: 460, 468, 1192-1193). In the latter case, the situation is complicated further by the fact that the personal pronoun does not typically alternate between subject and object pronoun form, but instead between possessive and object form where the non-finite clause functions as a complement (cf. Section 2.3.1). Only in adjuncts can pronoun forms alternate between subject and object forms in such gerundial non-finite clauses (cf. Section 2.3.1). In (127b), for example, it is possible to replace the object pronoun with a possessive pronoun, whereas the replacement of the object form by a subject form would be unexpected (Huddleston and Pullum 2002: 460, 468, 1192-1193). In view of this unclear and ambiguous status, these sentences were excluded from further analysis.

Instances in which the pronoun form following *it* and a form of *BE* potentially serves as the subject of other non-finite constructions, such as those in (128), were also excluded from further analysis of the data. In these cases, it is also not clear whether the personal pronoun form is to be construed as the complement of *it* and *BE*, followed by a non-finite clause, or as the subject of the following non-finite clause, which functions as the complement to *it* and the form of *BE* in its entirety.

(128) Pronoun Succeeded by Other Non-Finite Constructions

a. The rate the donations are going, he could buy himself a mint place on the moon! But it does suck that it is him *personally* liable for all this if he loses. (.au/it is him/14.07.2008)

b. So it is they to decide that taking account of improved [...] (.in/it is they/14.07.2008)

Although the sentences in (129) have a finite verb form, they are still to some extent incomplete because we observe zero relative markers in subject positions. These sentences are called “contact clauses” (e.g. Filppula 2004: 84).
(129) Pronoun Forms in Contact Clauses

a. Are you wishing it was I had rolled under in the waves, not me you are, for doing as you ordered? Are you wishing it was I had rolled under in the waves, not your father’s grandson, because he (BNC/APW/W_fict_prose)

b. It is them feel like those sophisticated self-absorbed beauties with the worlds t their Gucci heels in a Helmut Newton photograph. (.ie/it is them/14.07.2008)

The reason for the exclusion of the contact clauses in (129) is that, much like the cases in (127) and (128), too much potential ambiguity results from the lacking relativiser. Due to the missing relative pronoun, it is often not clear whether the personal pronoun form following it and the form of BE in these sentences can indeed be considered to be the focal pronoun of an it-cleft or whether it should be analysed instead as the subject of the following clause (cf. Huddleston and Pullum 2002: 1055).

Moreover, all instances of sentences that are ambiguous between the pronoun–noun phrase constructions (cf. (126)) and the demonstrative use of personal pronoun forms (cf. (124)) on the one hand and it-clefts on the other hand were excluded as well:

(130) Ambiguous it-Clefts

a. However it remains a car park to this day, and it is us Members of the Oireachtas who are to blame for failing to restore Leinster Lawn to its former glory (.ie/it is us/14.07.2008)

b. That, of course, is punishing all Australians, when really it is us fatties who have eaten too much and should be targeted (.au/it is us/14.07.2008)

Similar to the constructions just mentioned, it-clefts and it BE sentences with coordinated pronouns were also excluded from further analysis of the data because coordinated noun phrases also represent a context sensitive to pronoun case variation (e.g. Angermeyer and Singler 2003; Biber et al. 1999: 337; Quirk et al. 1985: 338). Therefore, in instances of it-clefts or it BE sentences with coordinated noun phrases, it is not clear whether the selection of one pronominal case form over the other can be attributed solely to its occurrence in an it-cleft, it BE sentence, or a coordination, or even to an interplay of these factors.

(131) Coordinated Pronouns in it-Clefts

a. The plaintiff may be urgently in need of the money. It is he or she who will be incurring additional expenses as a result of the accident and whose earning power may have been impaired. (BNC/GVH/W_ac_polit_law_edu)

b. So he said it is me and him who were supposed to go (.za/it is me/15.07.2008)

(132) Coordinated Pronouns in it BE Sentences

a. Until then, it ’s he and his wife and other family members. (COCA/2003/NEWS/Houston)

b. I am an only child, so it is me and my mum really. (.uk/it is me/15.07.2008)
Additionally, the exclusion from the analysed data sample of coordinated pronouns in *it*-clefts and *it BE* sentences is justified considering that certain pronoun-pronoun combinations, particularly those involving *us* and *them*, seem to have an idiom-like status, since these coordinations may be used to contrast certain groups or people with others, such as ordinary citizens with the authorities.\(^{34}\) Due to this quasi-idiomatic status, it is unclear whether these constructions admit pronoun case variation to the same extent as uncoordinated pronouns do (e.g. Angermeyer and Singler 2003: 196, Erdmann 1978: 70). Hence, they were not considered further.

(133) Idiom-like Coordinated Pronouns

a. I’ll tell you, I’m -- I’m locked on them. And it ‘s them or us (COCA/1997/SPOK/CBS_Sixty)

b. It is them (theHave nots) and us(the haves) (.au/it is them/14.07.2008)

The same is true for pronouns that are contrasted with the help of certain prepositions such as *against* and *versus*. They, too, seem to have an idiom-like status and are also used to contrast certain groups of people.\(^{35}\)

(134) Idiom-like Prepositional Constructions

a. " What had excited him was the message --; " it ’s us against them. That gave me a charge that we could take back to our members (BNC/EG0/W_non_ac_soc_science)

b. I know Zach has no chance of winning, but he is the most entertaining atm and deserves to win IMO, but if it is him vs. Aleisha I would be happy for her to win to. (.au/it is him/14.07.2008)

Negative coordinations such as those in (135a)–(135d) were also excluded from further investigation because these constructions, although negative, can still be considered as coordinations (e.g. Huddleston and Pullum 2002: 1313).

(135) Negative Coordinations

a. Ashcroft served as a range rider, and many Navajos believed that it was he, not Washington, who directed the slaughter. (COCA/1998/ACAD/AmerIndianQ)

b. At the end of it all, it is us and not them. (.au/it is us/14.07.2008)

Furthermore, instances in which the pronoun form in the search string is modified by an adnominal *self*-form were likewise omitted. The reason for this omission is the fact that the addition of a *self*-form may, as in coordinations, influence the choice of a pronoun form,


since *self*-forms favour a subject form for an antecedent regardless of their position in the sentence (e.g. Jespersen and Haislund 1949: 225; Quinn 2005a: 299). Moreover, as discussed in Section 6.1.1, *self*-forms may occur only because their insertion helps to avoid the difficulty of choosing between the object and the subject pronoun forms in certain variable contexts (Harris 1981: 18; Hernández 2002: 269). Therefore, it is justified to exclude these instances from further analysis.

(136) Pronouns with Adnominal *self*-Intensifier
   a. Ajib decided that such riches should belong to someone who appreciated them, and that was himself. To take his older self’s wealth would not be stealing, he reasoned, because it was he himself who would receive it. (COCA/2007/FIC/FantasySciFi)
   b. It is not the hierarchy that determines a person’s religion, it is them themselves. (.ie/it is them/14.07.2008)

In addition, elements in which the pronoun forms were used either in a generic meaning (137) or in a metalinguistic context (138) were also eliminated from the subsequent analyses. The underlying reason for this exclusion is, of course, that these constructions do not allow for pronoun case variation to the same extent as regular subject predicative complements following *it* and a form of *BE*.

(137) Pronouns with a Generic Meaning
   a. I thought I made it clear it is "he" (.in/it is he/14.07.2008)
   b. A child is child, irrespective of whether it is he or she. In fact, when you see a puppy, kitten, a cub, baby lizard, you always get the same feeling of CUTE. (.in/it is he/14.07.2008)

(138) Pronouns with Metalinguistic Meaning
   a. Instead of uses of "I" let us consider uses of "It's me". (I regard "It's I" as pedantic.) (BNC/CK1/W_ac_humanities_arts)
   b. But in "Brown attacked Cameron, who he thought was wrong" – "who" is correct, because it is "he" not "him" who is considered wrong. (.uk/it is he/15.07.2008)

Finally, as outlined in Sections 7.1.3 and 7.2.4 all instances of clear doublets, near-doublets or repetitions of earlier hits were excluded as well.

The remaining data were subsequently coded to allow for the operationalisation of the dependent and independent variables motivated and discussed in Section 6.1. In particular, the data have been coded for the dependent variable SUBJECT CASE FORM as well as for the independent variables CONSTRUCTION, CO-REFERENCE, PERSON, NUMBER, MODE OF DISCOURSE and FOCUS.
7.4 The Database Used for the Statistical Analyses

The compilation and processing of the data obtained from the BNC, COCA and the five Web-derived dataset discussed in the previous Sections 7.1–7.3 has left us with a considerable number of tokens for the quantitative analysis of subject predicative complements.

<table>
<thead>
<tr>
<th>Token Number (final data)</th>
<th>BNC</th>
<th>COCA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1379</td>
<td>3887</td>
<td>5266</td>
</tr>
</tbody>
</table>

*Table 16: Token Numbers Obtained from the Closed Corpora for the Statistical Analyses*

As Table 16 indicates, the examination and processing of the closed corpora has yielded 1379 subject predicative complements with a case-sensitive focal pronoun for the BNC and 3887 subject predicative complements for the COCA. Thus, the data collection and processing of the closed corpora has produced a total of 5266 subject predicative complement tokens on which this study can rely. The token numbers obtained for these two corpora should indeed suffice for a thorough statistical analysis of the distribution of pronoun case forms in subject predicative complements in British and American English and to answer the questions and test the hypotheses postulated in Chapter 5.

<table>
<thead>
<tr>
<th>.uk</th>
<th>.au</th>
<th>.ie</th>
<th>.za</th>
<th>.in</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>6748</td>
<td>5555</td>
<td>2490</td>
<td>3003</td>
<td>1845</td>
<td>19641</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Token Number (raw data)</th>
<th>.uk</th>
<th>.au</th>
<th>.ie</th>
<th>.za</th>
<th>.in</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1926</td>
<td>1177</td>
<td>425</td>
<td>651</td>
<td>492</td>
<td>4671</td>
<td></td>
</tr>
</tbody>
</table>

*Table 17: Token Numbers Obtained from the Web-Derived Datasets for the Statistical Analyses*

Table 17 presents the results obtained for the five Web-derived datasets. In Table 17 as well as in the remainder of this study, the abbreviation .uk stands for the data obtained from the top-level domain of Great Britain (or more exactly of the United Kingdom), the abbreviation .au stands for the Australian, .ie for the Irish, .za for the South African and .in for the Indian top-level domains. For each of these top-level-domains, two different token
numbers are provided: one for the raw data and one for the final data. The raw data token numbers indicate the tokens for each of the top-level domains that were obtained after uploading the html files into the concordancing program and activating the *Sentence mode* function to filter out search strings that are interrupted by sentence boundaries (cf. Section 7.2.4.2). Hence, the numbers in this line indicate the number of tokens that had to be manually cleaned, verified and classified (cf. Sections 7.2.4.2 and 7.2.4.3). The numbers in the bottom line of Table 17 indicate the final numbers of tokens that form the empirical basis for the analyses of the distribution of pronoun case forms in subject predicative complements in the varieties discussed in Chapters 10–14. Comparing the raw data figures to the token numbers of the final data, Table 17 shows that a considerable proportion of the data had to be excluded in the course of the different processing stages (cf. Sections 7.2.4.2–7.3). On average, more than 76 per cent of the manually checked raw data in the Web-derived datasets were excluded, i.e. only 4671 tokens of the initially 19641 tokens were left after the manual cleaning, verification and classification processes, as Table 17 indicates. Moreover, Table 17 also shows that the Web-derived datasets differ considerably from each other as far as their respective sizes are concerned. While the British .uk dataset comprises 1926 subject predicative complement tokens, the Irish .ie dataset consists of only 425 tokens. This, however, can at least partly be attributed to the considerably different sizes of the underlying top-level domains. Nevertheless, Table 17 also demonstrates that the compilation of the Web-derived datasets has resulted in substantial token numbers for each of the respective varieties, which should suffice to analyse the distribution of pronoun case forms in subject predicative complements, *it*-clefs and *it BE* sentences in each of these varieties (e.g. Rese 2000: 107).

In sum, the statistical analyses of this study rely on 9937 tokens to examine the distribution of pronoun case forms in subject predicative complements, which should indeed enable us to address at least some of the questions and hypotheses introduced in Chapter 5. Appendix C, i.e. the CD-ROM accompanying this study, contains a copy of the complete data used for this study. This does not only ensure a maximum of transparency, but also allows for a replication of the presented results.

---

36 Cf. .uk: nominet “Registrations statistics”.
7.5 The Notions of Representativeness, Balancedness and Comparability of the Data and their Repercussions for the Present Study

Before we can finally turn to the analysis of the obtained data in Chapters 8–14, we have to address some notions that are not only important for this study, but that are of fundamental importance in corpus linguistics in general, i.e. the representativeness, balancedness and comparability of the datasets. These considerations are relevant for the present study because of the differences in corpus design and corpus size between the BNC and the COCA discussed in Section 7.1.3. Moreover, these notions are also very important for the discussion of the Web-derived data, particularly with regard to their comparability and their general explanatory power. This is due to the fact that particularly this latter type of data has often been harshly criticised. It has been claimed, for example, that the Web “is a textual universe of unfathomed extent and variety, but it can in no way be considered as a representative sample of language use in general” (Leech 2007: 145). In order to understand the motivation for this criticism and to assess whether or not it is justified, we must briefly introduce the concepts of representativeness, balancedness and comparability as they are used in corpus linguistics before we can address the repercussions of these concepts and considerations for the present study.

In general, the notion of representativeness expresses the goal of corpus linguists to compile corpora that constitute exact samples of what is referred to as general population. This means that any findings made from the compiled samples or corpora should be likewise observed in the general population which they intend to represent, and hence allow for robust generalisations over the general population based on the data derived from the corpus (e.g. Mukherjee 2009: 21). Such generalisations are, however, only possible if the corpora exactly map the situation found in the general population of the respective genre, variety or language they intend to represent (e.g. Manning and Schütze 1999: 119; Mukherjee 2009: 21).

Although representativeness is considered to be one of the fundamental tenets of corpus linguistics (e.g. Mukherjee 2009: 21), the exact method of how this goal can actually be achieved remains unresolved. One influential proposal on how to achieve representativeness is the stance taken by Biber (1993) and Biber, Conrad and Reppen (1998)
suggesting a “stratified” model of corpus compilation or language sampling. The proponents of this approach suggest that a corpus should include the full range of linguistic variants of a language. Thus, it should contain all varieties of the respective language, including text types and registers that may not be relevant for the vast majority of the speech community such as legal texts or academic articles. Hence, stock is taken of the whole range of text types and genres of a language and each of them has to be sampled in order to achieve representativeness in a corpus (Biber, Conrad and Reppen 1998: 247-248). Thus, the focus of this approach is to sample language varieties “in proportion to their heterogeneity rather than in proportion to their prevalence in the whole textual universe” of a given language (Leech 2007: 141).

However, this sampling method has met with much criticism, particularly from scholars advocating a “proportional” approach to representativeness:

“One of the fundamental aims of Corpus linguistics as I understand it is to show up language as it is actually attested in real life use. However, Biber seems to argue that in designing a corpus one should apply a notion of importance that is derived from a definition of culture. For lack of any means of operationalizing this criterion of relative importance in culture, this throws the door wide open to subjective judgement in the compilation of the body of data that is expected to provide solid empirical evidence for language use” (Váradi 2001: 591-592).

Consequently, in order to achieve representativeness, this proportional approach demands that text types and genres should be represented exactly in proportion to their pervasiveness in an actual language community when the aim of the sample or corpus is not only to represent a specific genre but the language in general (e.g. Schaedler 1976: 359-361; Schank 1973: 22-23; Váradi 2001).

In view of these rather divergent points of view, a middle course has also been proposed which acknowledges the notion of importance of certain text types yet is based on proportional considerations (Leech 2007: 138-140). In this account, the importance of a certain text or text type is assessed on the basis of how many receivers it has (Leech 2007: 138-139). Hence, a radio program with a potential audience of one million listeners is considered much more important than a face-to-face conversation between two people (Leech 2007: 138).
Thus, considering these very different perspectives of and approaches to the notion of representativeness in corpus linguistics, it is hardly surprising that the possibilities for determining representativeness as well as the overall usefulness of applying this concept to linguistic corpora are still subject to ongoing debate (e.g. Kilgariff and Grafenstette 2003).

In addition to representativeness, corpora are also supposed to comply with the principle of balancedness, i.e. they should be carefully crafted with regard to the actual amount of words each of the included text type supplies to a corpus (e.g. Mukherjee 2009: 22). However, the term balancedness has been considered simply a “meeker” substitute for the concept of representativeness in order to avoid the controversies associated with the latter (cf. Kilgariff and Grafenstette 2003: 10). Indeed, clear guidelines on how this desideratum can be achieved and on the basis of which methodological and theoretical considerations balancedness is supposed to be implemented are hardly ever given. As a consequence, even proponents of this concept have to admit that “balancedness is very difficult to demonstrate” (Leech 2007: 136).

Besides representativeness and balancedness, comparability is considered a “third yardstick for successful corpus building” (Leech 2007: 141). Comparability in corpus linguistics means that two or more corpora are designed in such a way that they differ in exactly one variable. This may be, for example, a temporal parameter, such as the difference between the LOB and FLOB corpora, or a regional parameter, such as the difference between the Brown and the LOB corpora or between the individual components of the ICE. To be comparable, however, it is assumed that corpora also have to meet the prerequisite of being representative – a requirement very difficult to ensure, as the preceding discussion has shown. To make matters even more complicated, these two concepts are often considered to be ultimately incompatible since comparability can often only be achieved by a trade-off in representativeness (Leech 2007: 142). Cases in point are the different components of the ICE. In order for them to be comparable, only speakers with a formal English medium education were considered. For L1 regions, this normally implies secondary education, while for L2 countries, tertiary education is often required for speakers to be considered (Greenbaum 1996c: 6). This means that particularly in countries where English serves as a second language, only a very small section of the English-speaking population is actually included in the corpus. Indeed, these speakers may only represent a disproportionally educated social elite (Schilk 2011: 43-44). This is problematic for two
reasons: Firstly, it is questionable to what extent this social elite is representative of the general English-speaking population of such an English L2 country. Secondly, it is also unclear to what extent these elitist L2 speaker samples can be compared to the less socially homogenous speaker samples of the L1 regions. Hence, although the concepts of representativeness, balancedness and comparability are certainly desirable, it remains totally unclear how they can be reasonably implemented.

In view of the preceding discussion and the unsolved problems of defining and implementing these desiderata, it is surprising that Leech denies Web-derived data both the status of representativeness and of balancedness (2007: 144-145). With regard to representativeness, he states that

“[a] search engine like Google employs algorithms which are totally mysterious to the average user [...]. Google provides nothing like a complete search of the web, and reports [...] show how unstable and inconsistent are the counts one gets from Google, at least at the present time. What we get is an enormous sample of the web, but how representative it is remains a mystery” (Leech 2007: 144).

Thus, the use of Google data is rejected particularly because the mode of data retrieval is deemed to be opaque. This judgement seems premature or even inappropriate considering not only the insecurity in the literature about what the notion of representativeness actually refers to but also when looking at how so-called “representative” (cf. Mukherjee 2009: 21) corpora, such as the Brown corpus, were compiled:

“When planning the Brown corpus, we convened a conference of such corpus-wise scholars as Randolph Quirk, Philip Gove and John B. Carroll. This group decided the size of the corpus (1000000 words), the number of texts (500, of 2000 words each), the universe (material in English, by American writers, first printed in the United States in the calendar year 1961), the subdivisions (15 genres, 9 of ‘informative prose’ and 6 of ‘imaginative prose’) and by a fascinating process of individual vote and average consensus, how many samples from each genre (ranging from 6 in science fiction to 80 in learned and scientific)” (Francis 1979: 117).

When taking this report into consideration, it remains doubtful whether the outcome of an unknown algorithm employed by a search engine is indeed less representative or more arbitrary than the sparsely documented considerations and negotiations of some “corpus-wise” scholars, particularly when regarding the fact that the motivations on which their
decisions were based are not fully retraceable and reproducible nowadays (cf. Leech 2007: 137).

With regard to the second important requirement corpora are supposed to meet, the criterion of balancedness is also claimed to be absent from Web-derived data for the following reasons:

“Can the proportional sense of a 'balanced corpus' be applied to it? It is true that the web gives access to a very wide range of genres, some of them well-established in the written medium, such as academic writing and fiction writing; others newly-evolving genres closer to speech, such as blogs. However, it is also true that the web by definition gives little or no access to private discourse, such as everyday conversation, telephone dialogues, and the like. [...] Searching with a search engine provides no access to spoken or manuscript data. There are major areas seriously underrepresented, if they are represented at all” (Leech 2007: 144).

These objections are not, however, fully conclusive. To begin with, a “proportional sense of [a] ‘balanced corpus’” (Leech 2007: 144) requires the consideration of the included texts and texts types in the compilation process exactly in “the relative frequency of occurrence of those genres in the textual universe as a whole” (Leech 2007: 136). However, this criterion is not met by any existing corpus, at least to the knowledge of the author. In addition, it is stated that spoken data is absent from Web-derived data. This claim is utterly incorrect, since the data collection conducted for this study has yielded a considerable amount of “spoken or manuscript data” (Leech 2007: 144), as the discussion of the Web-derived data in Section 7.2.4.3 has demonstrated. Furthermore, the lack of “private discourse” (Leech 2007: 144) is also not a viable objection to the use of Web data, since this kind of data is absent from many well-established corpora such as the COCA, the empirical value of which has hardly ever been seriously questioned. Furthermore, the composition of any corpus in terms of balancedness and texts included can be considered to be to some extent arbitrary, not only the composition of Web-derived data. This has been exemplified by the BNC, since the question could be posed as to why “[t]he BNC happens to include cake recipes and research papers on gastro-uterine diseases, but not car manuals or astronomy texts” (Kilgariff and Grafenstette 2003: 10).

As far as the concept of comparability is concerned, Leech also remains sceptical toward the use and usefulness of Web data. While acknowledging, in view of Mair’s (2007)
findings, that Web data may reveal rough cross-varietal contrasts, he still points out that “the web provides nothing like the exact comparability of text selection for different periods or different regions of the world” (Leech 2007: 145). However, taking into consideration the difficulties of establishing comparability, particularly if representativeness is considered a necessary prerequisite, exact comparability is hardly ever achieved by any corpus, not even by those compiled exactly to that end.

The preceding discussion has shown that although representativeness, balancedness and comparability are worthwhile desiderata, it remains unclear how and by what means they can be theoretically achieved, as well as how they can actually be implemented. Indeed, if the issues of representativeness, balancedness and comparability are taken seriously, virtually no existing corpus complies with them. Thus, Kilgariff and Grafenstette conclude that “[t]he web is not representative of anything else. But nor are other corpora, in any well-understood sense” (2003: 11). As a consequence, they not only defend Web data in terms of representativeness when compared to traditional corpora, but they even go one step further by postulating that

“[r]esearchers are obliged to look to larger data sources” [...] [because] probabilistic models of language based on very large quantities of data, even if that data is noisy, are better than ones based on estimates (using sophisticated smoothing techniques) from smaller, cleaner data sets” (Kilgariff and Grafenstette 2003: 4).

This reasoning is fully in line with the claims from language technologists arguing that “more data is better data” (Bernardini, Baroni and Evert 2006: 9). The call for more data is reasonable given the virtually infinite complexity and multi-faceted nature of language use, which can hardly be exhaustively depicted in a one or even a 100 million word corpus. If we conceive of language as a puzzle with an infinite number of pieces, we should be aware and honest about the limitations of each and every corpus or database. Each corpus can only represent a finite number of pieces, but not the overall picture. Nevertheless, each puzzle piece or corpus still helps to approximate the overall picture, although the whole image will probably never be fully revealed.

Thus, in view of the difficulties in defining these desiderata and the vastly diverging points of view concerning their usefulness, validity and possible implementation, the best way to save these concepts is to retreat to the smallest common denominator for the notions of representativeness and balancedness likely to be accepted by proponents of
maximally diverging positions. A good candidate for such a common denominator is actually offered by Leech himself, whereby he states that “the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be” (Leech 2007: 138). This view would likely be accepted even by Kilgariff and Grafenstette (2003) since such a corpus comes close to the general aim of representativeness that the observations made in corpora should likewise be observed in the general population, which they intend to depict, and allow for robust generalisations over this general population (Kilgariff and Grafenstette 2003: 8-11). If Leech’s (2007: 138) smallest common denominator is to be adopted, the datasets used for this study clearly meet the requirements of being both balanced and representative, since both the two corpora and the Web-derived data are indisputably very large and remarkably diverse. If Kilgariff and Grafenstette’s (2003) point of view is adopted, their requirements are likewise met, as the BNC, the COCA and the five top-level domains tapped for the present study constitute tremendously large databases and they can be expected to yield robust and generalisable results. Thus, the databases used for this study, both the closed corpora and the five Web-derived datasets, can clearly be considered as both balanced and representative of their respective varieties.

However, the comparability of the different datasets used for this study cannot be as readily assumed as their representativeness and their balancedness can. As outlined in Section 7.2.4.1, the compilation of the Web-derived datasets differs exactly in one parameter, i.e. the top-level domain to which the queries were restricted. Although it could be argued that the Web-derived datasets differ only in this variable and are thus comparable, it could be objected that the data included in the different top-level domain-based datasets may differ also considerably in composition with regard to the amount of data included for each of the discourse modes or text types in the respective datasets. These potential differences in composition, in turn, make a direct comparison of the data obtained from the different top-level domains rather difficult. Similarly, a direct comparison between the closed corpora is also impossible without further consideration (cf. Section 7.1.3). In addition, a direct comparison between the Web-derived datasets and the closed corpora is also not advisable because of the differences between them. For example, while the Web-derived datasets include three different modes of discourse, the closed corpora used for this study only contain two different discourse modes. As a consequence, instead of lumping the
different databases together into one big statistical model, we will analyse each dataset separately in the subsequent Chapters 8–14 in order to assess whether the independent variables motivated and discussed in Chapter 6 actually influence the distribution of pronoun case forms in subject predicative complements (cf. Section 6.1.1). Moreover, cross-varietal trends and differences in the distribution of pronoun case forms in subject predicative complements should still become observable given that they are salient enough (cf. Section 6.1.2.6).

7.6 Data Used for this Study: Interim Summary

The preceding chapter has introduced and discussed the different datasets that form the empirical foundation of this study.

Section 7.1 has presented the different corpora used in this study to analyse the distribution of pronoun case forms in and across varieties of English. Although the respective components of the International Corpus of English seemed to be the ideal starting point, they turned out to be too small to allow for a meaningful quantitative analysis. Hence, this study had to tap bigger corpora. In particular, Section 7.1.2 has introduced the British National Corpus as a data source for British English and Section 7.2.2 has presented the Corpus of Contemporary American English as a database for American English.

While these mega-corpora form suitable databases for the analysis of both British and American English, there are unfortunately no freely available corpora of similar extent for the other varieties examined in this study. Hence, this study relies on databases obtained from the World Wide Web for the analysis of these varieties.

As a consequence, Section 7.2.1 has provided a concise introduction of the four most common approaches to using the Internet as a linguistic resource, i.e. the ‘mini-Web or mega-corpus’ approach, the ‘Web as Corpus proper’ approach, the ‘Web-derived’ corpus approach and the ‘Web as Corpus’ approach. As the use of Web data, particularly that obtained from this last approach, is often considered problematic, and since the approach adopted in this study combines elements of both the ‘Web-derived’ corpus and of the ‘Web as Corpus’ approach, Section 7.2.2 provided a discussion of the most frequently mentioned problems associated with data obtained from Google.
Section 7.2.3 also demonstrated that Web-derived databases offer many opportunities for corpus linguistic studies. In particular, the more or less infinite amount of data, the immediate and easy availability, the currentness of the data and the linguistic diversity make the World Wide Web and the language data supplied by it extremely attractive for linguistic studies. Moreover, due to the fact there are no or no sufficiently large corpora available for many varieties of English, it has been concluded that the use of Web data is more or less inevitable (cf. Section 7.2.3). This is particularly true since several studies have already demonstrated that regional preferences with regard to the distribution of certain morphosyntactic phenomena observed in closed corpora are replicated by top-level-sensitive Web queries (cf. Section 7.2.3).

Considering these opportunities, this study also relies partly on Web-derived datasets for the analysis of the distribution of pronoun case forms. How the Web-derived datasets for this study were compiled, cleaned and verified was thoroughly discussed in Sections 7.2.4.1 and 7.2.4.2. Then, Section 7.2.4.3 introduced the bottom-up scheme applied to classify the data for a mode of discourse analysis.

After discussing the data points that had to be excluded from the statistical analyses in Section 7.3, Section 7.4 provided a very brief overview of the data eventually forming the basis of the empirical analyses conducted in the following Chapters 8–14. In sum, this study will rely on 9937 subject predicative complements. More than half of them, i.e. 5266, were obtained from the closed corpora, while a total of 4671 subject predicative complements were obtained from the Web-derived datasets.

Finally, Section 7.5 briefly addressed the notions of representativeness, balancedness and comparability of datasets in corpus linguistics, since some studies deny that Web-derived datasets have these characteristics. The discussion of these concepts has shown that Web-derived datasets are actually as representative and balanced as any other – closed – corpus. In contrast, the notion of comparability has turned out to be less readily applicable to the datasets used for this study. However, this problem is not restricted to the databases used for this study; if taken seriously, it extends even to databases that have been specifically compiled to be comparable to each other. As a consequence this study analyses each database separately, since pervasive cross-varietal trends and differences in the distribution of pronoun case forms should still become observable in case they are salient enough.
8 The Distribution of Pronoun Case Forms in the British National Corpus

This chapter presents the results of the multivariate analysis of the British National Corpus (BNC). As outlined in Section 6.1.2.1.2, this study begins with the presentation of the results obtained for the analysis of all subject predicative complements following *it* and a form of *BE* irrespective of the particular construction (cf. Section 8.1). Subsequently, Section 8.2 focuses exclusively on the distribution and analysis of pronoun case forms in *it BE* sentences in the BNC. Then, Section 8.3 examines the distribution of pronoun case forms in *it*-clefts and Section 8.4 concludes the chapter with a brief interim summary of the major results.

8.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the British National Corpus

As outlined in Sections 7.1–7.3, the data of each corpus and dataset were processed and refined in more or less the same way. Similarly, all datasets have been tested – if possible – for the influence of the same factors that were motivated and discussed in Chapter 6. Thus, a reiteration of the applied method of data processing and of the motivation for the different variables is not deemed necessary here. Instead, we will start each section by presenting the results for the respective construction in the underlying corpus or dataset directly.

After processing the raw data and excluding all irrelevant constructions (cf. Section 7.3), we are left with 1379 tokens of subject predicative complements following *it* and a form of *BE* in the BNC, which form the data basis of the regression modelling, as illustrated in Table 18. As Table 18 also indicates, nearly half of the tokens, i.e. a share of 46.27 per cent, have a subject form as the pronominal complement after *it* and a form of *BE*, whereas 53.73 per cent of all subject predicative complements exhibit an object pronoun case form.
Without entering prematurely into the discussion of the results, it is remarkable that we observe such a high share of subject pronoun forms in this context bearing in mind the mainly theoretical accounts which assume or predict very high or even nearly exclusive shares of object forms in subject predicative complements (e.g. Harris 1981: 17-20; Quinn 2005a: 138, 242-248). The numbers given in Table 18 further demonstrate how rare subject predicative complements actually are in traditional closed corpora. With only roughly 14 subject predicative complements per million words, it is not surprising that studies relying on rather small corpora, such as the ICE components, cannot produce substantial quantitative results to account for the distribution of pronoun case forms in these contexts (cf. Quinn 2009).

Turning now to the multivariate analysis of the BNC data, we have to bear in mind that this study – conforming to the Principle of Parsimony – aims to apply minimal-adequate models to the data (cf. Section 6.2). Consequently, Table 19 below, as well as all subsequent tables displaying results of regression models in this study, only includes significant variables unless bootstrap validation tells us otherwise (cf. Section 6.2).

For the regression model applied here, this means that the only main factor discussed in Section 6.1.2 that is not included in the regression model is \textit{SINGULAR}. \textit{SINGULAR} is used to operationalise the \textit{NUMBER} variable (cf. Section 6.1.2.3), i.e. the influence that the distinction between singular and plural pronoun forms may have on the use of pronoun case forms (cf. Table 19). However, since \textit{SINGULAR} has been eliminated, differences in number do not influence the choice of pronoun case forms in this corpus. While the non-significance of this factor is in line with the observations of Quinn (2005a: 135) and Wales (1996: 96), this finding contradicts the assumptions of other accounts that assert that differences in number influence the distribution of pronoun case forms (Sobin 1997: 334). In addition to the variable \textit{NUMBER}, several interactions were also eliminated from the final regression model.
In the interest of saving space, however, interactions eliminated from the minimal-adequate models are provided in footnotes throughout this study.37

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>4.64</td>
<td>***</td>
<td>103.39</td>
</tr>
<tr>
<td>FIRST</td>
<td>-0.49</td>
<td>n.s.</td>
<td>0.61</td>
</tr>
<tr>
<td>SPOK</td>
<td>0.66</td>
<td>+</td>
<td>1.93</td>
</tr>
<tr>
<td>FOCUS</td>
<td>2.32</td>
<td>***</td>
<td>10.16</td>
</tr>
<tr>
<td>SPOK*CLEFT</td>
<td>-3.53</td>
<td>***</td>
<td>0.03</td>
</tr>
<tr>
<td>CLEFT*FIRST</td>
<td>-1.59</td>
<td>***</td>
<td>0.20</td>
</tr>
</tbody>
</table>

| MODEL $\chi^2$ | 903.32 (***)
| R$^2$          | 0.64       |
| % CORRECTLY PREDICTED | 84.41 |
| % BASELINE    | 53.73      |
| N             | 1379       |

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 19: Subject Predicative Complements in the BNC: Logistic Regression Results

To start with the significant variables, the factor CLEFT which is used to operationalise the CONSTRUCTION variable (cf. Section 6.1.2.1) is very highly significant, as Table 19 clearly demonstrates. Furthermore, its rather high odds ratio of 103.39 indicates that it exerts an enormously strong influence on the choice of pronoun case forms in subject predicative complements, clearly promoting the use of subject forms in this context in the BNC. By converting this odds ratio into a per cent value, we can say that the chances of observing a subject pronoun case form in a subject predicative complement following it and a form of BE rise by 10239 per cent in the BNC if this subject predicative complement is part of an it-cleft. Thus, this result clearly corroborates accounts arguing for marked differences in the distribution of pronoun case forms between it-clefts and it BE sentences due to the functional differences of these two constructions (e.g. Biber et al. 1999: 335-336; Erdmann 1978; Maier 2013). At the same time, however, this finding seriously undermines accounts stressing the similarity of pronoun case distribution in it-clefts and it BE sentences and which

---

37 The interactions eliminated from the minimal-adequate model are the following: SPOK*FIRST, SPOK*SINGULAR, SPOK*FOCUS, CLEFT*SINGULAR, FIRST*SINGULAR, FIRST*FOCUS, SINGULAR*FOCUS.
mainly assume that the distribution of pronoun case forms is solely governed either by the position of the pronoun relative to the finite verb or by the pronoun’s membership to a certain weak or strong pronoun class or a combination of these factors (e.g. Emonds 1986: 96-100; Harris 1981: 19-20; Quinn 2005a: 138-139, 242-248). Thus, the additional functional and pragmatic load of the pronoun in *it*-clefts as discussed in Sections 6.1.2.1 and 6.1.2.5 seems to promote the use of subject pronoun case forms in this context in the BNC.

The second variable retained in the final regression model is PERSON, which is operationalised with the factor FIRST (cf. Section 6.1.2.2). By itself, however, FIRST is not significant; it only becomes significant in combination with another factor, which will be discussed below. Thus, although the body of literature assumes differences within both *it BE* sentences and *it*-clefts with regard to the likelihood of subject pronoun case forms occurring in the first or third person (Quinn 2005a: 135-140, 246; Wales 1996: 95-96), the BNC data do not exhibit significant differences in the distribution of first and third person pronoun forms in the more general subject predicative complement category.

The third variable maintained in the final regression model is the MODE OF DISCOURSE. This variable is operationalised with the help of its parameter value SPOK, which determines the difference between written and spoken data and hence assesses whether the degree of formality influences the choice of pronoun case forms in subject predicative complements (cf. Section 6.1.2.4). As can be inferred from Table 19, this factor also yields very interesting results. Although SPOK is only marginally significant, it is still remarkable to note that the outcome of this factor does not correspond to the expectations of the relevant literature. While the literature on pronoun case distribution in subject predicative complements disagrees on various points (cf. Sections 3.1–3.4), there seems to be an unequivocal consensus that the more formal a communicative situation is, the more likely it is that a subject pronoun form will occur (e.g. Biber et al. 1999: 335-336; Harris 1981: 18-19; Huddleston and Pullum 2002: 459). However, the BNC data suggest quite the opposite, as the odds ratio of 1.93 indicates. Although SPOK is only marginally significant, the spoken mode of discourse tends to promote rather than inhibit the use of subject case forms in subject predicative complements, which is indeed a fairly unexpected finding.

The fourth main effect maintained in the minimal-adequate regression model is FOCUS (cf. Section 6.1.2.5). As can be inferred from Table 19, this factor is not only very highly significant, but also promotes the use of subject pronoun case forms in subject predicative
complements, as its odds ratio of 10.16 indicates. Converting this odds ratio into a percentage value, the strong influence of Focus becomes even more explicit, since the chances of a subject pronoun case form occurring increase by 916 per cent if a pronoun occurs in a context that has been identified as being particularly focussed (cf. Section 6.1.2.5). Hence, the BNC data evidently corroborates hypothesis H3 that subject pronoun case forms in subject predicative complements may have been reanalysed as a Focus marking device exploiting the variability in this context (cf. Section and 4.2).

In addition to these main effects, two interactions also turn out to be significant. The first very highly significant interaction is Spok*Cleft, exhibiting an odds ratio of 0.03. This means that both interacting factors, Spok and Cleft, strongly antagonise each other and that the probability of observing a subject pronoun form in a subject predicative complement context diminishes considerably in the BNC, i.e. by 97 per cent, if these factors interact. This indicates that it-cLEFTs with a subject form as focal pronoun are much rarer in the spoken than in the written data of the BNC. Hence, accounts asserting or attesting an inhibiting influence of spoken or less formal language on the likelihood of observing subject forms in subject predicative complements seem to be right, at least as far as it-cLEFTs are concerned (cf. e.g. Biber et al. 1999: 335-336; Maier 2013; Quirk et al. 1985: 337-338). Moreover, this finding also confirms results of bivariate analyses which attest a considerably lower share of subject pronoun forms in the spoken than in the written data it-cLEFT data of the BNC (cf. Maier 2013).

The second very highly significant interaction observed in the BNC data is that between the factors First and Cleft. As was the case with the latter interaction, when working in tandem First and CLEFT also clearly inhibit the use of subject pronoun forms in subject predicative complements in the BNC, which becomes evident from the interaction’s odds ratio of 0.20. Thus, in cases where these two factors interact, the likelihood of observing a subject form shrinks by 80 per cent, which means that it-cLEFTs with a first person subject pronoun form as focal pronoun are less frequent than their third person counterparts. This finding is very much in line with the findings of other studies which report a higher probability for third person pronouns than for first person pronouns to occur in their subject form as focal pronouns of it-cLEFTs (e.g. Quinn 2005a: 135-136; Wales 1996: 95-96).
As far as the overall significance of the applied regression model is concerned, we note that there is a very highly significant correlation between the retained independent variables and the dependent one, i.e. the choice of pronoun case forms in subject predicative complement position, as the model chi-square value ($\chi^2 = 903.32$, $p<0.001$) in Table 19 clearly demonstrates. With regard to the variance explained, we note that the applied binary logistic regression model fares rather well with a Nagelkerke’s $R^2$ value of 0.64, accounting for 64 per cent of the observed variation in the dependent variable. Furthermore, the model is able to predict 84.41 per cent of all tokens or – to be more precise – of all linguistic choices of the writers and speakers contributing to the BNC data correctly. Thus, the model used here improves the baseline model, i.e. the assumption that the distribution of pronoun case forms in subject predicative complements in the BNC is based on pure chance, by more than 30 per cent, which indicates a very high predictive power for the model.

In sum, the results obtained here indicate that functional factors, i.e. CLEFT and FOCUS, strongly influence the choice of pronoun case forms in subject predicative complements in the BNC. This challenges assumptions suggesting that positional considerations or the membership to a certain pronoun class are the decisive or even sole factors in the distribution of pronoun case forms (e.g. Burridge 2004; Harris 1981). In addition, other variables assumed to influence the choice of pronoun case forms such as NUMBER and to some extent PERSON turned out to be non-significant, whereas the impact of SPOK – though only being marginally significant – seems to affect the choice of pronoun case forms contrarily to the expectations of the relevant literature. Furthermore, these results also indicate that it BE sentences and it-clefts may be influenced rather differently by certain predictors. This has become particularly apparent through the two very highly significant interactions retained in the minimal-adequate model (cf. Table 19). These differences seem to corroborate accounts which acknowledge marked differences in pronoun case distribution between these two constructions (e.g. Biber et al. 1999: 335-336; Erdmann 1978: 75-78). In view of these differences, it seems advisable to look separately at the two construction types, i.e. it BE sentences and it-clefts, in order to assess exactly which factors influence the distribution of pronoun case forms in which construction, and to what extent.
8.2 The Distribution of Pronoun Case Forms in \textit{it BE} Sentences in the British National Corpus

Of the 1379 subject predicative complements identified in the BNC (cf. Section 7.4 and 8.1), 569 tokens belong to the group \textit{it BE} sentences. As can be inferred from Table 20, there are marked differences in the distribution of pronoun case forms between the more general subject predicative complement category (cf. Table 18) and the distribution of pronoun case forms in \textit{it BE} sentences. In the case of \textit{it BE} sentences, the share of subject pronoun case forms in predicative complement position accounts for only eight per cent compared to a share of over 46 per cent of subject forms observed for the superordinate category of subject predicative complements (cf. Table 20 and Table 18 in Section 8.1).

<table>
<thead>
<tr>
<th>IT BE SENTENCES (BNC)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>569 (100%)</td>
<td>46 (8.08%)</td>
<td>523 (91.92%)</td>
</tr>
</tbody>
</table>

\textit{Table 20: it BE Sentences in the BNC: The Token Numbers}

The distribution of pronoun forms in the \textit{it BE} sentences of the BNC depicted in Table 20 is, however, to a greater or lesser extent in line with observations made in other accounts. Biber et al. (1999) observe an even smaller share of subject pronoun case forms after \textit{it} and \textit{BE} in their dataset, ranging from five per cent to less than 2.5 per cent depending on the register. However, their results are a difficult basis for comparison since they not only state their results as rough rounding values rather than as precise decimal numbers, they also do not distinguish between the British and American corpus data (cf. Biber et al. 1999: 24-26, 335-336). While Biber et al.’s findings are therefore difficult to relate to the BNC results presented and discussed here, the proportions reported in Table 20 match the findings of Erdmann (1978). This latter study, which is based on a corpus of British novels, reports a 9.59 per cent share of subject pronoun case forms for \textit{it BE} sentences compared to a share of object pronoun forms accounting for roughly 90.41 per cent all \textit{it BE} sentences in the data (cf. Erdmann 1978: 75).
As far as the multivariate analysis of the *it BE* sentences in the BNC is concerned, Table 21 shows the results obtained from the minimal-adequate model applied to the data. Again, since the binary regression model used to obtain the results in Table 21 only assesses the distribution of the pronoun case forms in *it BE* sentences, it should be borne in mind that the starting model for this dataset did not include all the factors motivated and discussed in Chapter 6. Thus, CLEFT, operationalising the distinction between *it*-clefts and *it BE* sentences, and the cleft-specific factor AS_SUBJ were not included in the maximal model since they are simply not applicable.

Apart from these non-applicable factors, other factors discussed in Section 6.1.2 are also not retained in the minimal-adequate model because they were eliminated in the model-building process (cf. Table 21). The first of these eliminated factors is FIRST operationalising the variable PERSON, i.e. the possible differences in the distribution of pronoun case forms between first and third person pronouns (cf. Section 6.1.2.2). The non-significance of FIRST is unexpected given the assumptions of the relevant literature. While some accounts state that “[n]on-1sg nominatives would seem to be extremely rare, and restricted to certain discourse contexts” (Quinn 2005a: 246), other accounts do not explicitly state an increased probability for a certain person to occur in their subject forms yet provide only third person singular examples, suggesting a certain preference of third person subject forms (e.g. Wales 1996: 95). Hence, it is interesting to note that the distinction between first

---

**Table 21: *it BE* Sentences in the BNC: Logistic Regression Results**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPOK</td>
<td>0.67</td>
<td>+</td>
<td>1.95</td>
</tr>
<tr>
<td>FOCUS</td>
<td>2.43</td>
<td>***</td>
<td>11.40</td>
</tr>
<tr>
<td>Model $\chi^2$</td>
<td></td>
<td>22.82 (***)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>% Correctly Predicted</td>
<td></td>
<td>92.09</td>
<td></td>
</tr>
<tr>
<td>% Baseline</td>
<td></td>
<td>91.92</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>569</td>
<td></td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

---

38 Except for the main effects discussed in text, the following interactions were eliminated in the model-building process: SPOK*FIRST, SPOK*SINGULAR, SPOK*FOCUS, FIRST*SINGULAR, FIRST*FOCUS and SINGULAR*FOCUS.
and third person pronouns seems to play no role at all for the *it BE* sentences in the BNC data.

Another factor which was eliminated in the model-building process is *SINGULAR*, which operationalises the variable *NUMBER*, i.e. the potential differences in the distribution of pronoun case forms between singular and plural pronoun forms (cf. Section 6.1.2.3). Although the possibility of plural subject forms occurring in *it BE* sentences has been explicitly ruled out (Sobin 1997: 334), this claim cannot be confirmed by the findings for the *it BE* sentences in the BNC data.

Turning now to the factors that are maintained in the minimal-adequate model, we observe that *SPOK* – though again only marginally significant – in tendency promotes the use of subject pronoun forms in *it BE* sentences, as can be inferred from its odds ratio of 1.63. Thus, this factor, operationalising the difference between the spoken and the written BNC data, behaves contrary to the expectations voiced in the literature, as was the case for the domain of subject predicative complements in Section 8.1. Indeed, observing subject pronoun forms as complements of *it BE* sentences in spoken language seems to be fairly unexpected according to many previous accounts (e.g. Biber et al. 1999: 335-336; Harris 1981: 18; Huddleston and Pullum 2002: 459). Hence, to attest the exact opposite is a very remarkable finding, since it challenges one of the major assumptions of previous research, i.e. that the degree of formality (e.g. Quirk et al. 1985: 337-338) or the distinction between spoken and written language (e.g. Biber et al. 1999: 335-336; Harris 1981: 18) is a decisive factor in inhibiting the use of subject pronoun forms in the distribution of pronoun case forms in *it BE* sentences. Furthermore, this finding underlines not only the need for analysing the distribution of pronoun case forms in subject predicative complements by means of multivariate statistics, but also the necessity of the quantitative testing and verification of postulated hypotheses in general.

Apart from the distinction between written and spoken data, the only other factor retained in the minimal-adequate model presented in Table 21 is *FOCUS*, distinguishing particularly focussed from ‘normal’ or prototypical *it BE* sentences (cf. Section 6.1.2.5). This factor is remarkable not only because it is the only very highly significant predictor retained in the final logistic regression model but also because of the strong effect it exerts on the distribution of pronoun case forms in *it BE* sentences in the BNC. This factor’s odds ratio is 11.40, making it the strongest effect in this model. Translating this value into a percentage,
we note that the probability of observing a subject pronoun form increases by 1040 per cent if the pronoun is used in a focussed rather than in a normal or prototypical it BE sentence (cf. Section 6.1.2.5). Thus, this finding supports the hypothesis that subject pronoun case forms in subject predicative complements in general and in it BE sentences in particular may have been reanalysed as a Focus marking device. This finding is in line not only with the situation found in many other varieties of English (cf. Section 2.2.3), but also with the results presented in Section 8.1 on the more general subject predicative complement category in the BNC.

With regard to the overall significance of the minimal-adequate model, we note that the model as a whole is very highly significant ($\chi^2 = 22.82, p<0.001$) (cf. Table 21). The variance explained by the model is, however, rather modest. The regression model applied here can only account for nine per cent of the observed variance in the dependent variable (Nagelkerke’s $R^2 = 0.09$). Yet, this is still enough to qualify the model as substantially significant (cf. Szmrecsanyi 2006: 55). In view of the modest share of variance explained by the model, the predictive power of the model is also rather moderate when compared to the baseline model. While the latter is able to account for 91.92 per cent, the minimal-adequate regression model applied here correctly predicts the outcome of the dependent variable in 92.09 per cent of all observations, thus improving the baseline model further, though only by 0.17 per cent.

To sum up, the statistical analysis of the it BE sentences in the BNC has produced very remarkable results. First of all, variables that are deemed to significantly influence the choice of pronoun case forms in this context according to the relevant literature, such as NUMBER and PERSON, turned out to be non-significant. Secondly, the factor SPOK influences the distribution of pronoun forms – if at all – differently than predicted by earlier accounts, since it rather promotes than inhibits the use of subject pronoun forms in it BE sentences in the BNC. Furthermore, the most important predictor in the domain of it BE sentences turned out to be the factor FOCUS. Indeed, it is the only very highly significant factor retained in the final regression model, which is very remarkable given the fact that this variable has so far been mainly neglected in the discussion of pronoun case distribution in it BE sentences. Hence, the analysis of this construction not only strongly corroborates the hypotheses that pragmatic and functional factors in general play a very important role in the distribution of pronoun case forms (cf. Section 5.2; H2 and H3), but simultaneously also weakens accounts
opting solely for positional considerations or the membership to a certain pronoun class to be the decisive factors (cf. H1).

### 8.3 The Distribution of Pronoun Case Forms in *it*-Clefts in the British National Corpus

As illustrated in Table 22, 810 instances of *it*-clefts with a case-sensitive focal pronoun were identified in the BNC. Furthermore, Table 22 also indicates that there are pronounced differences in the distribution of pronoun case forms between the *it*-clefts and the *it BE* sentences in the BNC. While the share of subject pronoun case forms in *it BE* sentences only accounts for roughly eight per cent of all pronominal complements (cf. Table 20), we observe nearly the opposite for the *it*-clefts found in the BNC, where subject pronoun forms are used in more than 73 per cent of all *it*-cleft tokens. These very different distributional patterns explain why the superordinate category, i.e. the subject predicative complements discussed in Section 8.1, exhibits a comparatively balanced distribution of subject and object pronoun forms (cf. Table 18). However, these marked differences also justify a separate analysis of these two sentence types given their distinctness. Moreover, this distribution confirms the observations made in Section 6.1.2.1.1 about the very highly significant difference in the distribution of pronoun case forms between *it*-clefts and *it BE* sentences which clearly corroborates accounts that highlight the differences in the distribution of pronoun case forms between these two constructions (e.g. Biber et al. 1999: 335-336; Erdmann 1978; Maier 2013).

<table>
<thead>
<tr>
<th>IT-CLEFTS (BNC)</th>
<th><strong>TOTAL N</strong></th>
<th><strong>SUBJECT FORMS N</strong></th>
<th><strong>OBJECT FORMS N</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>IT-CLEFTS (BNC)</td>
<td>810 (100%)</td>
<td>592 (73.09%)</td>
<td>218 (26.91%)</td>
</tr>
</tbody>
</table>

*Table 22: it-CLEFTS in the BNC: The Token Numbers*

Although these results are difficult to relate to Biber et al. (1999: 336), since Biber et al. do not distinguish between British and American data and also employ a mode of presenting their results which is sometimes difficult to interpret, the results obtained for the BNC can be compared rather well to those of Erdmann (1978). In his study, Erdmann
observes a share of subject pronoun case forms as focal pronouns of *it*-clefs amounting to 72.03 per cent of all tokens in his corpus of British novels (Erdmann 1978: 76). As discussed elsewhere (Maier 2013), this ratio corresponds to the share of subject pronoun case forms observed for the *it*-clefs in the written subset of the BNC data.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>As_SUBJ</td>
<td>3.34</td>
<td>***</td>
<td>28.35</td>
</tr>
<tr>
<td>FIRST</td>
<td>-2.48</td>
<td>***</td>
<td>0.08</td>
</tr>
<tr>
<td>SPOK</td>
<td>-3.10</td>
<td>***</td>
<td>0.04</td>
</tr>
<tr>
<td>MODEL Χ²</td>
<td></td>
<td>361.19 (***)</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td></td>
<td>83.35</td>
<td></td>
</tr>
<tr>
<td>% BASELINE</td>
<td></td>
<td>73.09</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>810</td>
<td></td>
</tr>
</tbody>
</table>

* + significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 23: *it*-Clefs in the BNC: Logistic Regression Results

The results of the statistical analysis of the *it*-clefs in the BNC are illustrated in Table 23. As discussed in Section 6.1.2.5, the variable FOCUS is operationalised with the help of particularly focussed and normal *it* BE sentences. Thus, it is not applicable here and hence was not included in the maximal model. Furthermore, the factor SINGULAR was eliminated in the course of the model-building process because the difference between singular and plural pronouns does not significantly influence the choice of pronoun case forms in the *it*-clefs of the BNC. This finding is in contrast to the expectations of the literature according to which the use of plural subject forms in *it*-clefs is rather unexpected (cf. Sobin 1997: 334). Furthermore, all possible interactions between the main effects were likewise eliminated in the model-building process.³⁹

³⁹ The interactions eliminated from the minimal-adequate model are: SPOK*FIRST, SPOK*SINGULAR, SPOK*As_SUBJ, FIRST*SINGULAR, FIRST*As_SUBJ, SINGULAR*As_SUBJ.
results for As_Subj in Table 23, this factor strongly fosters the use of subject pronoun case forms in it-clefts. Thus, in clauses in which the focal pronoun of the it-cleft is co-referential with the subject of the relative-like dependent clause, the odds of observing a subject pronoun form are 2735 per cent higher than in it-clefts in which the focal pronoun is co-referential with an object of the dependent clause. This is in line with the observations and expectations of other authors attesting a higher share of subject pronoun forms in it-clefts in which the focal pronoun is co-referential with the subject of the dependent clause (e.g. Erdmann 1978: 76-77; Huddleston and Pullum 2002: 459; Quinn 2005a: 133). Hence, the “Janus-like status” (Quirk et al. 1085: 338) of these pronoun case forms, i.e. being a subject predicative complement and a co-referent with the subject of the subsequent clause at the same time, does indeed have a strong effect on the distribution of pronoun case forms in it-clefts. Furthermore, this observation is consistent with the results obtained thus far from the BNC in that basically all functionally motivated factors with syntactic and/or pragmatic functions, such as Cleft, Focus and As_Subj, exert a very strong influence on the distribution of pronoun case forms and all of them clearly promote the use of subject case forms in the analysed contexts (cf. Sections 8.1–8.3).

As already indicated by the interaction Cleft*First in Section 8.1, the second very highly significant factor retained in the minimal-adequate regression model analysing the it-cleft data of the BNC in Table 23 is First. With an odds ratio of 0.08, First clearly inhibits the use of subject pronoun case forms in it-clefts. This means that first person pronouns are much less likely to occur in their subject forms than third person pronouns in it-clefts. Hence, this factor influences the choice of pronoun case forms exactly as predicted by the body of literature (e.g. Quinn 2005a: 135; Wales 1996: 95-96). A possible explanation for why third person pronouns are more likely to be used in their subject form than first person pronouns is difficult to motivate from a purely structural point of view. However, this difference between first and third person pronouns may be explained in terms of pragmatic functions or Focus marking, i.e. the identification of the relatively most important piece of information. As outlined in Section 6.1.2.2, there are fundamental differences in several important respects between the first and the second person on the one hand and the third person pronoun on the other. Since first and second person pronouns always denote referents that are actively involved in a communicative situation, they are more prototypically deictic than third person pronouns because their referents – either as the
speaker(s) or the hearer(s) – are immediately tied to the persons involved in the respective communicative situation. Moreover, they are also most likely to receive our attention and our empathy (e.g. Langacker 1991: 307; Siewierska 2004: 5-8). Third person pronouns can be used deictically as well. However, they are prototypically used phorically, i.e. referring back or forth to a referent or referents mentioned or introduced in the prior or subsequent discourse (e.g. Siewierska 2004: 7). Therefore, the relevance of the first and second person pronouns, as well as their referents, to a current communicative situation is automatically given by the mere use of these pronouns which identifies them and their referents as important or salient information for both the speaker and the hearer. Third person pronouns, however, may need some additional Focus marking in order to highlight their relevance for the immediate communicative situation and to clearly identify the referent referred to by the pronoun. If these differences between first and third person pronouns coincide with the factor CLEFT, it may be the case that the combination of two pragmatically salient factors, i.e. first person pronouns occurring in the focal position of an it-cleft, may be considered emphasised enough to identify the referent or referents of the first person pronoun as the most salient or important piece of information in a clause. This “trade-off between the various coding devices” (Dik 1989: 278) could explain why third person pronouns are more likely to receive additional Focus by means of subject pronoun case forms in the present context (cf. Sections 4.1 and 6.1.2.2). Moreover, this explanation is fully consistent with the Focus-oriented approach outlined in Chapter 4.

As has also already been indicated by the respective interaction in the regression model in Section 8.1, Table 23 shows that SPOK very highly significantly inhibits the use of subject pronoun forms in it-clefts in the BNC. Its odds ratio of 0.04 shows that the likelihood of observing a subject form in it-clefts in the spoken data is considerably lower than it is in the written data. This means that the odds ratio for SPOK is even lower than that of the factor FIRST. The inhibiting force of SPOK observed for it-clefts is very interesting, since this factor – though only marginally significant – tends at least to have a contrary effect in the domain of it BE sentences, as observed in Section 8.2. Thus, as far as it-clefts are concerned, SPOK does indeed influence the distribution of pronoun case forms as predicted by the literature (e.g. Biber et al. 1999: 335-336; Harris 1981: 19).

The question which arises is why it-clefts and it BE sentences behave so differently with regard to the factor SPOK. A possible explanation could again be the interaction of several
Focus marking devices. As discussed in Section 4.1, Functional Grammar has identified a number of Focus marking strategies such as prosodic prominence, special positions within a clause, special particles, special Focus marking constructions (e.g. Dik 1997: 291) and, as proposed in this study, the use of subject pronouns in subject predicative complements (cf. Section 5.2; H3). It was also mentioned in Section 4.1 that several Focus marking devices may not only co-occur but may even cancel each other out so that there can be a “trade-off” between the different focussing devices (cf. Dik 1989: 278). Thus, two pragmatically salient features or Focus marking strategies in a clause or sentence may suffice to mark a certain constituent as being the most important one. On the one hand, this would explain the lesser likelihood of observing first person subject pronouns in it-clefts, as discussed above. On the other hand, this could explain why subject forms are less likely to be used in it-clefts in the spoken mode of discourse, since this mode of discourse not only allows for highlighting a pronoun by means of the focal position of an it-cleft, but also for enduing it with particular prosodic prominence, which is of course not available for focal pronouns in the written mode of discourse. As a consequence, prosodic prominence and the use of it-clefts may be deemed sufficient to mark a constituent as the Focus of a clause in the spoken BNC data. For it BE sentences, however, syntactic highlighting is only partly available for the particularly focussed ones (cf. Section 6.1.2.5). Hence, the use of subject pronoun case forms may be used as a second Focus marking device in addition to or complementing prosodic prominence in the spoken data, which would explain why the factor SPOK tends to promote rather than inhibit the use of subject pronoun case forms in it BE sentences in the spoken data of the BNC.

The applied minimal-adequate regression model itself is statistically very highly significant, which signals a strong correlation between the dependent variable and the retained independent variables ($\chi^2 = 361.19, p<0.001$). It can be inferred from the Nagelkerke’s $R^2$ value that the statistical model can account for 52 per cent of the variation observed in the dependent variable, which is a very good result, especially when compared to that of the it BE sentences (cf. Table 21). In addition, the current model is also evidently superior to the baseline model, improving it by more than ten per cent, since the applied binary logistic regression model correctly predicts the outcome of the dependent variable in 83.35 per cent of all cases, whereas the baseline model only predicts 73.09 per cent of all instances correctly.
In sum, we can state that for this subset of subject predicative complements following *it* and a form of *BE* the variable *NUMBER* is again non-significant. However, all factors retained in the minimal-adequate model influence the dependent variable in the predicted manner. *As_SBJ* strongly promotes the use of subject pronoun forms in *it*-clefts, whereas *FIRST* and *SPoK* markedly inhibit the use of subject pronouns in *it*-clefts. The strongest factor retained in this model is clearly *As_SBJ*, which operationalises the co-reference of the focal pronoun of the main clause with the subject of the dependent clause. This factor emphatically promotes the use of subject pronoun case forms. This is consistent with the trend observed throughout the BNC that factors from the domains of syntax and pragmatics or their interface, such as *CLEFT*, *FOCUS* or *As_SBJ*, not only promote the use of subject pronoun case forms but also turn out to be consistently strong predictors in the analysed contexts.

8.4 The Distribution of Pronoun Case Forms in the British National Corpus: Interim Summary

The analysis of the subject predicative complements in the BNC has yielded very remarkable results. First of all, we observed enormous differences in the distribution of pronoun case forms between the *it BE* sentences and *it*-clefts subsumed under the heading of the more general subject predicative complement category. This justifies a separate examination of each sentence type, which was conducted in Sections 8.2 and 8.3.

Secondly, all three individual analyses executed in this chapter have shown that syntactic and pragmatic factors, such as the difference between *it*-clefts and *it BE* sentences, the difference between particularly focussed and normal *it BE* sentences and the co-reference of the focal pronoun of *it*-clefts with either the subject or an object of the dependent clause, exert an enormous influence on the distribution of pronoun case forms in each of the analysed contexts. Thus, the BNC data show that the distribution of pronoun case forms in the analysed contexts is not solely determined by the position of the personal pronoun in the clause nor by the membership of a pronoun to a weak or strong pronoun class, which severely challenges the predictions of several mainly theoretical accounts (cf. Section 5.2; H1).
Furthermore, the data show that all these factors from the domains of syntax, pragmatics and the interface thereof clearly promote the use of subject pronoun forms in the analysed contexts. This corroborates hypothesis H3 in Section 5.2 according to which subject pronoun case forms are re-functionalised as Focus markers in subject predicative complements. This hypothesis is further supported by the fact that in the BNC both necessary and sufficient conditions for such a conclusion are met by the preceding analyses (cf. Section 6.1.2.5). Subject pronoun forms are much more likely to occur in both it-clefts and particularly focussed it BE sentences than in normal it BE sentences. This eminent influence of Focus on the distribution of pronoun case forms is also demonstrated by the fact that FOCUS is the only really significant factor retained in the regression model used for the it BE sentences.

With regard to the core-grammatical variables NUMBER and PERSON analysed in the BNC data, the results show a differentiated picture. Number distinctions turn out to be non-significant in each dataset. Hence, at least in the BNC, the distinction between singular and plural pronouns does not influence the distribution of pronoun case forms in subject predicative complements. The distinction between first and third person pronouns, operationalised via FIRST, only plays a role in it-clefts, where first person pronouns are significantly less likely to occur in their subject form. This may be due to the fact that first person pronouns are per se cognitively more salient than third person pronouns, since the referents of the former are more immediately involved in the communicative situation. As a consequence, first person pronouns receive no further Focus marking in it-clefts, because the cleft-construction itself, together with the first person pronoun, may be enough to mark the focal pronoun as the relatively most important piece of information in the clause.

The only sociolinguistic variable in the models distinguishing between different modes of discourse also yields also remarkable results: While SPOK clearly inhibits the use of subject forms in it-clefts, the same factor tends to promote the use of subject pronouns in it BE sentences. A possible explanation may be that in spoken discourse the speaker can use another Focus marking device, i.e. prosody, which is, of course, not available in written discourse. Hence, similar to the interaction between FIRST and CLEFT, two pragmatically prominent features – in this case prosody and the it-cleft construction – may be enough to mark a piece of information as Focus of a clause, which could account for why subject forms are significantly less likely to occur in spoken it-clefts.
9 The Distribution of Pronoun Case Forms in the Corpus of Contemporary American English

The following sections report the results of the statistical analyses of the data obtained from the *Corpus of Contemporary American English* (COCA). As in Chapter 8, we begin directly by discussing the results for the general category of subject predicative complements in Section 9.1. Then, Section 9.2 presents the results for the *it BE* sentences before Section 9.3 examines and discusses the distribution of pronoun case forms in *it*-clefts in this corpus. Finally, Section 9.4 concludes this chapter with a brief summary of the most important findings.

9.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the Corpus of Contemporary American English

As shown in Table 24, the processing of the raw data, i.e. the exclusion of doublets and irrelevant and uncertain data points, has yielded a dataset of 3887 subject predicative complements following *it* and a form of *BE* for the COCA, which form the basis for the further multivariate analysis (cf. Chapter 6 and Chapter 7).

<table>
<thead>
<tr>
<th>Subject Predicative Complements (COCA)</th>
<th>Total N</th>
<th>Subject Forms N</th>
<th>Object Forms N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3887 (100%)</td>
<td>1131 (29.10%)</td>
<td>2756 (70.90%)</td>
</tr>
</tbody>
</table>

*Table 24: Subject Predicative Complements in the COCA: The Token Numbers*

When comparing the COCA results depicted in Table 24 with those obtained for the BNC in Section 8.1, we can note that subject predicative complements are obviously even rarer in the COCA than in the BNC. Although we detect a frequency of roughly 13.8 subject predicative complements with a case-sensitive pronoun per million words in the BNC, the rate of occurrence is considerably lower in the COCA, with only approximately 9.8 instances
of subject predicative complements per million words in this corpus.\footnote{Note that the normalised frequencies of \textit{it}-clefts and \textit{it} \textit{BE} sentences per million words are based on the idealised token counts of 100 million words for the BNC and 400 million words for the COCA. The author is of course aware that the actual token counts differ from these idealised numbers. However, since it was not possible to retrieve the exact token count of the 400 million word version of the COCA afterwards (Davies 2011, p.c.), this approach offers at least a crude approximation to the per-million-word frequencies of the analysed constructions.} Furthermore, we observe that the share of subject pronoun case forms in subject predicative complements in the COCA is considerably smaller than in the BNC. In the COCA, subject pronouns account for only roughly 29 per cent of all tokens whereas object forms are used in nearly 71 per cent of all subject predicative complements. In the BNC, however, the share of subject forms amounts to more than 46 per cent whereas object forms account for nearly 54 per cent (cf. Section 8.1). However, there may be a variety of reasons for this difference, such as different compositions of the datasets with regard to the proportions of \textit{it} \textit{BE} sentences and \textit{it}-clefts included in the superordinate subject predicative complement category (cf. Maier 2013). Therefore, we will now turn to the multivariate statistical analysis of the COCA data to uncover which factors exert a significant influence on the distribution of pronoun case forms, and to what extent.

As illustrated in Table 25 below, all five of the independent variables motivated and discussed in Chapter 6 that are applicable to the superordinate category of subject predicative complements are retained in the minimal-adequate regression model. Furthermore, one very highly significant interaction between two of these main effects is also retained in the final model.\footnote{The following interactions were eliminated in the course of the model-building process: \textsc{cleft*spok}, \textsc{cleft*singular}, \textsc{spok*first}, \textsc{spok*singular}, \textsc{spok*focus}, \textsc{first*singular}, \textsc{first*focus}, \textsc{singular*focus}.}
To start with, the factor CLEFT operationalising the CONSTRUCTION variable is not only very highly significant, but also exerts a very strong influence on the distribution of pronoun case forms because it substantially promotes the use of subject pronoun forms in subject predicative complements. This becomes particularly apparent if we translate its odds ratio of 128.77 into a percentage value, since the likelihood of observing a subject pronoun case form in an it-cleft is 12777 per cent higher than in an it BE sentence. Thus, CLEFT is again clearly the strongest factor maintained in the regression model (cf. Section 8.1). As with the observations made for the BNC data, this finding supports those accounts highlighting the differences in the distribution of pronoun case forms between it BE sentences and it-clefts (e.g. Biber et al. 1999: 335-336; Erdmann 1978). At the same time, however, this observation challenges accounts arguing for a similarity of these two sentence types in terms of the distribution of pronoun case forms as well as their underlying theoretical assumptions, i.e. that the distribution of pronoun case forms is either exclusively determined by the position of the pronoun form or by a split in the pronominal paradigm into weak or strong pronoun classes (e.g. Emonds 1986: 96-100; Harris 1981: 19-20). Hence, the discourse-pragmatic function of cleft sentences seems to clearly promote the use of subject pronoun case forms in this construction, which is in line with the observations made for the BNC data (cf. Section 8.1).
The second factor retained in the minimal-adequate model is \textsc{first} operationalising the variable \textsc{person}, i.e. the differences between first and third person pronouns with regard to pronoun case distribution. In contrast to the BNC data, \textsc{first} is significant by itself and not only when interacting with another factor (cf. Section 8.1). In the present dataset, \textsc{first} is statistically very highly significant. With an odds ratio of 0.33, this factor clearly inhibits the use of subject pronoun case forms in subject predicative complements. On the one hand, this finding conflicts with studies reporting a higher share of first person subject pronouns than third person pronouns in \textit{it BE} sentences (e.g. Quinn 2005a: 246). On the other hand, this result is in line with observations made about \textit{it}-clefts according to which third person subject pronouns are more likely to be used than first person subject forms in this sentence type (e.g. Quinn 2009: 42; Wales 1996: 95-96). Similar to the situation described in Section 8.3, the greater probability of observing third person subject forms than first person subject pronoun forms may be explained in terms of the fundamental differences between first and third person pronouns (e.g. Croft 2003: 130; Langacker 1991: 307). As stated, first person pronouns are by definition deictic. Hence, these pronouns, as well as their referents, are automatically prominent in the communicative situation in which they are used and thus are most likely to receive our empathy and attention (cf. Siewierska 2004: 5-8; Schmid 2007: 132-133). Since the typical use of third person pronouns is, however, phoric rather than deictic (e.g. Siewierska 2004: 7), speakers or writers may feel the need to highlight the referents of these pronouns as particularly prominent for the present communicative situation, given they are the most important piece of information, i.e. Focus, of the clause. This possible explanation would indeed account for not only the greater likelihood of third person pronouns to be used in subject predicative complements but would also be in line with hypothesis H3 according to which subject pronoun case forms may be used as a Focus marking device (cf. Section 5.2).

In contrast to the BNC, \textsc{singular} which operationalises the difference between singular and plural pronouns, is also retained as significant in the minimal-adequate model accounting for the distribution of subject predicative complements in the COCA. In the present dataset, this factor exhibits an odds ratio of 0.66 and thus inhibits the use of subject pronoun forms in subject predicative complements. This observation stands in marked contrast to the expectations of the literature which predict that plural subject forms are considerably less likely than singular subject pronouns to occur in \textit{it}-clefts and are even ruled
out in *it BE* sentences (Sobin 1997: 334). The question crops up why plural pronouns are more likely than singular pronouns to occur in their subject forms in subject predicative complements in the COCA. To account for this observed difference in purely structural terms is not easy. However, this difference may in principle be tackled again by means of an implicational hierarchy (cf. Section 6.1.2.3; Siemund 2008). Similar to the more or less explicit implicational hierarchy assumed thus far for the person distinctions (e.g. Sections 6.1.2.2 and 8.3), the referents of plural pronouns may be deemed to need more Focus marking than the referents of singular pronouns in order to be recognisable as the most important discourse entity or the Focus of a clause (cf. Croft 2003: 126-132; Section 6.1.2.3). This need to stress plural forms more explicitly than singular forms may result from the much broader scope of possible referents of the former. This broader scope of referents is particularly true for third person plural pronouns, which can refer to basically every plurality of animate, inanimate, concrete and abstract referents. Hence, their immediate relevance for an ongoing communicative situation may not be self-evident, which could be a reason why plural pronouns receive additional highlighting by means of subject pronoun case forms in order to mark their referents as Focus, i.e. the most salient and important piece of information delivered by the clause (cf. Section 4.1). However, due to the fact that this observation, i.e. the higher likelihood of observing plural pronouns rather than singular pronouns in their subject form, corresponds neither with the assumptions of the literature nor with the observations made for the BNC, this possible explanation must be considered tentative until we can assess the effect of this factor in the other datasets still to come.

As Table 25 shows, SPOK is also identified as being very highly significant and it clearly decreases the chances for observing subject pronoun case forms in the analysed contexts. Hence, as can be inferred from the odds ratio of 0.48, the likelihood of observing a subject pronoun case form diminishes by 52 per cent if a subject predicative complement occurs in the spoken subset of the COCA data. This result is very much in accordance with the expectations of the literature, which more or less unequivocally assumes a general preference for object pronoun case forms and a lower probability of observing subject pronoun forms in more informal or spoken contexts (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Wales 96: 91-108). However, this inhibiting effect of SPOK contrasts with the observations made for the BNC data in Section 8.1, where this factor promotes rather than inhibits the use of subject forms. The difference between these two corpora
with regard to this factor is even more striking when bearing in mind that the spoken data of the COCA is in its totality more formal than the spoken data of the BNC (cf. Sections 7.1.2 and 7.1.3). Hence, this difference between the BNC and the COCA with regard to the influence of the factor SPOK may be cross-varietal. The reason subject pronoun forms are less likely to be used in the spoken data of the COCA can be explained with the help of different approaches. On the one hand, the underlying reason for the lower likelihood of subject pronouns occurring in the spoken COCA data may be that the assumed “prescriptive bias in favour of the subjective form” (Quirk et al. 1985: 338) is indeed less pronounced in the spoken than in the written data of this corpus. This would be in line with the assumptions of much of the relevant literature (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459). On the other hand, this distributional difference between the spoken and the written mode of discourse may also be explained from a more Focus-oriented perspective. As discussed before (e.g. Section 4.1; Dik 1997: 291), languages employ several different Focus marking strategies in order to highlight a certain constituent as particularly prominent. Similar to the account proposed for the *it*-clefts in the BNC (cf. Section 8.3), the availability of other devices indicating Focus may also explain the lower likelihood of observing subject pronouns in the spoken data of the COCA. In the present case, it is possible that Focus is mainly marked by prosodic prominence in the spoken data. Since this Focus marking strategy is not available for the written data, subject pronouns are more likely to occur in this mode of discourse to mark the referent of a pronoun as the “relatively most important or salient information” (Siewierska 1991: 174). This explanation is not only in line with the findings of the multivariate analysis for this dataset, but also with all results obtained thus far in this study suggesting a very important influence of pragmatic factors in general and Focus marking in particular on the distribution of pronoun case forms in subject predicative complements (e.g. Chapter 8).

In view of these considerations, it is therefore not surprising that Focus is also retained in the regression model depicted in Table 25 as a very highly significant factor that substantially promotes the use of subject pronoun case forms. Its odds ratio of 4.87 means that the likelihood of observing a subject pronoun case form increases by 387 per cent in a context that has been identified as particularly focussed (cf. Section 6.1.2.5). Furthermore, this high odds ratio means that Focus turns out to be the second strongest factor retained in the minimal-adequate model. Hence, as is the case for the BNC data (cf. Section 8.1), this
factor not only promotes very strongly the use of subject pronoun case forms, but also firmly corroborates hypothesis H3 according to which subject pronouns forms have been reanalysed as Focus marking devices in subject predicative complements (cf. Section 5.2).

Finally, the regression model also identifies an interaction between two of the main effects as statistically very highly significant, i.e. the interaction between CLEFT and FIRST. These two factors work against each other, which means that the probability of observing an it-cleft with a first person subject pronoun case form as focal pronoun is lower than observing an it-cleft with a third person subject pronoun case form as focal pronoun. The inhibiting effect of this interaction becomes apparent from the odds ratio of 0.33, which clearly indicates its constraining effect on the occurrence of subject pronoun case forms in subject predicative complements. As the interplay of these factors will be discussed again in Section 9.3 when analysing the it-cleths of the COCA in particular, it suffices here to say that this finding is in accordance with both the results of other studies observing a higher probability of third person than of first person pronouns to occur in their subject form as focal pronoun of it-cleths (e.g. Quinn 2005a: 135-136; Wales 1996: 95-96) and with the observations about the BNC data in Sections 8.1 and 8.3.

With regard to the overall significance of the minimal-adequate model used to analyse this dataset, Table 25 indicates a very strong correlation between the independent variables and the dependent variable, since the applied model is statistically very highly significant ($\chi^2 = 2704.16, p<0.001$). Furthermore, the model is able to account for 72 per cent of the variation observed in the dependent variable as can be deduced from the model's Nagelkerke's $R^2$ value of 0.72. Moreover, the applied binary logistic regression model is able to correctly predict the outcome of the dependent variable in 90.40 per cent of all cases. Hence, it improves the baseline model substantially, namely by 19.50 per cent.

As was the case with the corresponding model for the BNC data (cf. Section 8.1), the results obtained for the general category of subject predicative complements in the COCA indicate that discourse-pragmatic features such as CLEFT operationalising the CONSTRUCTION variable (cf. Section 6.1.2.1.1) and FOCUS not only exert a very strong influence on the choice of pronoun case forms, but also substantially promote the use of subject pronoun case forms. This strengthens hypothesis H3 according to which pragmatic factors in general and Focus in particular have a significant influence on the distribution of pronoun case forms in the contexts analysed in this study (cf. Section 5.2). At the same time, this finding also
weakens hypothesis H1 proposing that the position or membership to a certain pronoun class are the major or sole determinants of pronoun case distribution in Present-Day English (e.g. Burridge 2004: 1118; Harris 1981; Section 5.2). The factors FIRST and SINGULAR operationalising the variables PERSON and NUMBER also significantly influence the choice of pronoun case forms since they both significantly inhibit the use of subject forms in the analysed context. While the difference between first and third person pronouns could also be explained by means of pragmatic factors inherent in the nature of these different persons, the motivation for the difference between singular and plural forms is more challenging. Although the observed difference could be explained, in principle, with the help of an implicational hierarchy similar to that assumed for the differences between first and third person, the validity of such an assumption cannot yet be affirmed since the results obtained for the COCA are in contrast to the BNC results and the expectations of the literature. The analysis of the different modes of discourse shows that subject forms are more likely to be used in written than in spoken data in the COCA, which is in line with the expectations of the literature and is also compatible with an approach accounting for this difference in terms of Focus marking. However, this finding is not consistent with the observations made for the BNC (cf. Section 8.1). Finally, the interaction included in the model again suggests that it BE sentences and it-clefts may be differently influenced by the individual factors, which supports accounts pointing out the distinctiveness of these two sentence types in terms of pronoun case distribution (e.g. Erdmann 1978).

9.2 The Distribution of Pronoun Case Forms in it BE Sentences in the Corpus of Contemporary American English

In the COCA, 2595 tokens of the 3887 subject predicative complements following it and a form of BE fall into the category of it BE sentences, which corresponds to a share of 66.76 per cent. Table 26 demonstrates that there are again pronounced differences in the distribution of pronoun case forms between it BE sentences on the one hand and the more general category of subject predicative complements on the other (cf. Table 24). Whereas the share of subject pronoun case forms in the latter category still amounts to 29.10 per cent, the share of subject pronoun case forms in the subclass of it BE sentences only
accounts for 4.47 per cent of all tokens. Thus, 95.53 per cent of all *it BE* sentences in the COCA are used with an object pronoun case forms. Furthermore, the share of subject pronoun case forms attested for the *it BE* sentences in the COCA (cf. Table 26) is also much smaller than that observed for the BNC, which accounts for roughly eight per cent of all *it BE* sentences (cf. Section 8.2).

<table>
<thead>
<tr>
<th>IT BE SENTENCES (COCA)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2595 (100%)</td>
<td>116 (4.47%)</td>
<td>2479 (95.53%)</td>
</tr>
</tbody>
</table>

*Table 26: it BE Sentences in the COCA: The Token Numbers*

Turning to the discussion of the multivariate analysis, Table 27 below shows that only two factors are retained in the minimal-adequate regression model applied to the data. At the same time, this implies that some variables were eliminated in the course of the model-building process. The first factor omitted from the final model is SINGULAR operationalising the variable NUMBER, i.e. the influence the distinction between singular and plural pronoun forms might have on the distribution of pronoun case forms. Contrary to the expectations of the literature (Sobin 1997), this factor does not significantly influence the choice of pronoun case forms although the use of plural subject pronoun forms in *it BE* sentences has been ruled out (Sobin 1997: 334). However, this finding matches the observations made for the BNC data where this factor is also not significant (cf. Section 8.2).

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRST</td>
<td>-1.05</td>
<td>***</td>
<td>0.35</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.55</td>
<td>***</td>
<td>4.72</td>
</tr>
</tbody>
</table>

MODEL $\chi^2$ 52.23 (***)

R$^2$ 0.07

% CORRECTLY PREDICTED 95.53

% BASELINE 95.53

N 2595

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

*Table 27: it BE Sentences in the COCA: Logistic Regression Results*
Another factor excluded from the minimal-adequate model is *Spok*. That the difference between spoken and written data does not play a role in the distribution of pronoun case forms in *it BE* sentences is rather unexpected, since much of the current literature assumes that subject forms are less likely to occur in spoken or informal data (e.g. Biber et al. 1999: 335-336; Harris 1981: 18). Furthermore, *Spok* has been identified as significantly inhibiting the use of subject pronoun forms in the superordinate category of subject predicative complements in the COCA (cf. Section 9.1). In view of these discrepancies, the separate analysis of each sentence type, i.e. *it BE* sentences and *it*-clefs, is justified once more.42

Turning now to the factors that are retained in the minimal-adequate model, we note in Table 27 that the binary logistic regression model identifies *First* as a very highly significant factor which decreases the likelihood of observing a subject pronoun case form in an *it BE* sentence by 65 per cent, as can be inferred from its odds ratio of 0.35. This finding contrasts with assumptions found in the literature according to which subject forms of first person pronouns should occur more frequently than those of third person pronouns – particularly singular ones (e.g. Quinn 2005a: 246). As has been discussed before (e.g. Sections 8.3 and 9.1), this attested preference for third person subject forms may be attributed to the underlying differences between prototypically deictic first person pronouns and prototypically phoric third person pronouns (e.g. Croft 2003: 130; Siewierska 2004: 5-8). As a result of the prototypically phoric use of third person pronouns (e.g. Siewierska 2004: 7), the use of special focussing strategies, such as the use of subject pronoun forms, may be deemed necessary to highlight the referents of third person pronouns in order to indicate their prominence and relevance for the ongoing communicative situation, if they are the Focus of the clause (cf. Section 8.3). Referents of first person pronouns, in contrast, do not need necessarily additional highlighting since they are automatically relevant and also highlighted in the communicative situation because their referents are always a part of it. This explanation would nicely tie in with hypothesis H3 (cf. Section 5.2) assuming that subject pronoun case forms may have been reanalysed as a Focus marking device. Furthermore, the observations made in this section are also in line with similar observations made, for example, in Sections 8.3 and 9.1.

---

42 In addition to the factors discussed in the text, the following interactions were also eliminated in the course of the model-building process: *Spok* *First*, *Spok* *Singular*, *Spok* *Focus*, *First* *Singular*, *First* *Focus*, *Singular* *Focus*.
Whereas the discourse-pragmatic relevance of the difference between first and third person pronouns is rather indirect, the immediate pragmatic relevance of Focus is more straightforward (cf. Section 4.1). As Table 27 shows, Focus is not only very highly significant, but also by far the strongest factor in the regression model exhibiting an odds ratio of 4.72. Thus, the likelihood of observing a subject pronoun case form in an *it BE* sentence in the COCA increases by 372 per cent, if it occurs in an *it BE* sentence that has been identified as particularly focussed (cf. Section 6.1.2.5). This strong and fostering impact of Focus is not only in accordance with the findings made for the *it BE* sentences in the BNC (cf. Section 8.2), but also with one of these study’s central assumptions, i.e. that subject pronoun case forms may be used as Focus marking devices in subject predicative complements (cf. Chapter 4).

Table 27 also demonstrates that the minimal-adequate model as a whole is statistically very highly significant ($\chi^2 = 52.23$, $p<0.001$). As was the case with the corresponding model analysing the *it BE* sentences in the BNC (cf. Section 8.2), the amount of variance explained by the model is, however, rather small. The current model is only capable of explaining roughly seven per cent of the variation observed in the dependent variable (Nagelkerke’s $R^2 = 0.07$). Nevertheless, this is again sufficient to render the model substantially significant (Szmrecsanyi 2006: 55). Considering the small share of variance that can be explained with the help of the applied regression model, it is perhaps also not surprising that the applied minimal-adequate model does not substantially improve the baseline model, since both of which are able to predict the outcome of the dependent variable in 95.53 per cent correctly.

In sum, the statistical analysis of the *it BE* sentences in the COCA has again produced very interesting results. The factor *SPOK*, which is deemed to be very influential according to the body of literature, turned out to be non-significant in the present dataset. Similarly, the factor *SINGULAR* was also eliminated since the number distinction it operationalises does also not significantly influence the choice of pronoun case forms in the *it BE* sentences of the COCA. The other grammatical factor included in the maximal model, i.e. *FIRST*, is, however, retained in the minimal-adequate model as very highly significant. Although it is difficult to account for its impact in structural terms, the influence of this factor could be explained, however, with the pragmatic differences inherent in the nature of first and third person pronouns. Since first person pronouns are per se more salient in a communicative discourse
than third person pronouns, they may simply need less additional highlighting to mark their referents as extremely important discourse entities. Speaking of salience, highlighting, and discourse-pragmatic variables, the present analysis also shows that Focus is the strongest factor in the minimal-adequate model, clearly promoting the use of subject pronoun case forms in *it BE* sentences. Since this factor also plays an outstanding role in the *it BE* sentences of the BNC, we can assume that Focus, which has so far been largely ignored in the discussion of pronoun case assignment in the analysed context, does indeed play an important role in the distribution of pronoun forms in the varieties or databases discussed so far. This, in turn, strongly corroborates the hypotheses of this study arguing for functional factors and particularly pragmatic ones to influence the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H2 and H3).

### 9.3 The Distribution of Pronoun Case Forms in *it*-Clefts in the Corpus of Contemporary American English

As Table 28 shows, 1292 tokens of the 3887 subject predicative complements observed in the COCA are *it*-clefts. As was the case with the BNC data, the distribution of pronoun case forms depicted in Table 28 demonstrates marked differences particularly when comparing it to the distribution of subject and object pronoun case forms observed for the *it BE* sentences. We noted in the preceding section that the share of subject pronouns only amounts to 4.47 per cent in the *it BE* sentences, while the remaining 95.53 per cent of all tokens exhibit an object pronoun form (cf. Table 26). For the construction at hand, however, 78.56 per cent of all *it*-clefts with a case-sensitive focal pronoun are used with a subject form, whereas 21.44 per cent of all *it*-clefts in the COCA exhibit an object pronoun as focal pronoun (cf. Table 28). This shows again how differently *it*-clefts and *it BE* sentences behave in terms of pronoun case distribution.

<table>
<thead>
<tr>
<th></th>
<th><strong>Total N</strong></th>
<th><strong>Subject Forms N</strong></th>
<th><strong>Object Forms N</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>it</em>-CLEFTS (COCA)</td>
<td>1292 (100%)</td>
<td>1015 (78.56%)</td>
<td>277 (21.44%)</td>
</tr>
</tbody>
</table>

*Table 28: *it*-CLEFTs in the COCA: The Token Numbers*
What Table 28 also indicates is the fact that the share of subject forms in the COCA *it*-clefs is higher than that in the BNC, where subject forms are used only in 73.09 per cent of all analysed *it*-clefs (cf. Section 8.3). The reason why the share of subject pronoun forms in the superordinate category of subject predicative complements accounting for only 29.10 per cent of all tokens in the COCA is still smaller than that in the BNC, which amounts to 46.27 per cent, is the fact that *it*-clefs with a case-sensitive focal pronoun are considerably less frequent in the COCA than in the BNC (cf. also Sections 8.3 and 9.2). While we can observe a frequency of approximately 8.1 *it*-clefs with either a first or third person pronoun as clefted element per million words in the BNC, the frequency for the COCA data only amounts to roughly 3.2 *it*-clefs per million words. The frequencies of *it BE* sentences in the two corpora are, in contrast, more similar, since we observe approximately 5.7 *it BE* sentences per million words in the BNC compared to a frequency of 6.5 *it BE* sentences per million words in the COCA. Hence, the lower frequency of *it*-clefs with a case-sensitive focal pronoun in the COCA data seems to constitute a difference between British and American English, since these differences in terms of frequency are simply too big to be solely attributable to differences in the corpus composition (cf. Maier 2013). Nevertheless, the subject pronoun shares of 78.56 per cent in the *it*-clef data and 29.10 per cent in the superordinate subject predicative complement category observed in the COCA are still noteworthy taking the expectations of some accounts into consideration which assume much smaller shares for these contexts (e.g. Emonds 1986; Harris 1981; Quinn 2005a).

Table 29 below presents the results for the statistical analysis of the *it*-clef data obtained by means of binary logistic regression modelling. As we can see, all main effects included in the maximal model are retained in the minimal-adequate model depicted in Table 29.44

---

43 Again, the normalised frequencies of *it*-clefs and *it BE* sentences per million words are based on the idealised token counts of 100 million words for the BNC and 400 million words for the COCA. The author is still aware that the actual token counts differ from these idealised numbers. However, as it was not possible to retrieve the exact token count of the 400 million word COCA afterward (Davies 2011, p.c.), this approach offers at least an – admittedly – coarse-grained approximation to the per-million-word frequencies of the analysed constructions. Furthermore, even if the numbers were slightly different, they would probably not diminish the vast differences in the token frequencies of *it*-clefs between the BNC and the COCA.

44 Although all main effects are retained in the final model, there are several interactions that were eliminated in the model-building process: Spok*First, Spok*Singular, Spok*As_Subj, First*Singular, First*As_Subj, Singular*As_Subj.
The factor with the largest effect size in the final regression model is \( \text{AS}_\text{SUBJ} \) operationalising the CO-REFERENCE variable, i.e. the question of whether or not the co-reference of the focal pronoun with either the subject or an object of the following dependent clause influences the distribution of pronoun case forms in \textit{it}-clefts (cf. Section 6.1.2.1.2). As Table 29 demonstrates, \( \text{AS}_\text{SUBJ} \) is very highly significant and its odds ratio of 27.26 indicates that the likelihood of observing a subject pronoun case form in an \textit{it}-cleft is 2626 per cent higher in tokens in which the clefted pronoun is co-referential with the subject of the following clause than in \textit{it}-clefts in which the focal pronoun is co-referential with an object of the dependent clause. This result is in line with the observations made for the BNC, where \( \text{AS}_\text{SUBJ} \) is also clearly the strongest factor in the \textit{it}-cleft data (cf. Section 8.3). Moreover, this finding also accords with the results and expectations of earlier studies assuming and observing a higher likelihood of subject forms to occur in \textit{it}-clefts in which the focal pronoun is co-referential with the subject of the dependent clause (e.g. Erdmann 1978: 76-77; Huddleston and Pullum 2002: 459; Quinn 2005a: 133). Thus, this hinge function of the clefted pronoun as central constituent of both the main clause and the dependent clause in an \textit{it}-cleft has indeed a very strong influence on the distribution of pronoun case forms (cf. Quirk et al. 1985: 338; Section 6.1.2.5). Furthermore, the strong influence of \( \text{AS}_\text{SUBJ} \) corresponds to the observations made so far in both the COCA and the BNC that virtually all factors with syntactic, pragmatic and/or discourse-pragmatic functions play an important role in the distribution of pronoun case forms in subject predicative complements.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{AS}_\text{SUBJ} )</td>
<td>3.31</td>
<td>***</td>
<td>27.26</td>
</tr>
<tr>
<td><strong>FIRST</strong></td>
<td>-2.50</td>
<td>***</td>
<td>0.08</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>-0.49</td>
<td>*</td>
<td>0.61</td>
</tr>
<tr>
<td>SPOK</td>
<td>-1.45</td>
<td>***</td>
<td>0.24</td>
</tr>
<tr>
<td>MODEL ( \chi^2 )</td>
<td>458.87 (***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R(^2)</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td>85.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% BASELINE</td>
<td>78.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1292</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 29: \textit{it}-Clefts in the COCA: Logistic Regression Results
The second factor retained in the minimal-adequate model is FIRST, operationalising the PERSON variable (cf. Section 6.1.2.2). That this factor has a very highly significant influence on the distribution of pronoun case forms in it-clefts has already been indicated by the respective interaction in the statistical model analysing the more general subject predicative complement category (cf. Section 9.1). As can be seen from Table 29, FIRST considerably restrains the occurrence of subject pronoun forms in it-clefts in the COCA. The odds ratio of 0.08 shows that the likelihood of observing a subject pronoun case form in an it-cleft decreases by 92 per cent if the clefted pronoun is a first person pronoun. The influence and the direction of the impact FIRST exerts is not surprising, since the body of literature expects this finding as we observe it (e.g. Quinn 2005a: 135; Wales 1996: 95-96). Moreover, we have observed a similar influence of FIRST on the distribution of pronoun case forms before in both the BNC (cf. Section 8.3) and the COCA (cf. Sections 9.1 and 9.2). As discussed above (e.g. Sections 6.1.2.2 and 8.3), a possible explanation for why first person pronouns are less likely to be used in their subject forms are the inherent pragmatic differences between first and third person pronouns. Since the former are deictic and typically refer to one or more of the discourse participants, referents of first person pronouns are thus automatically salient and also most likely to receive the empathy and attention of speakers and hearers in an ongoing communicative situation (cf. Croft 2003: 130; Langacker 1991: 307; Siewierska 2004: 5-8). Third person pronouns are, however, prototypically phoric (cf. Croft 2003: 130; Siewierska 2004: 5-8). Hence, the speakers or writers may feel the need to highlight the importance of the referents of third person pronouns by the use of subject pronoun case forms in case they are the Focus (cf. e.g. Sections 8.3, 9.1 and 9.3). Therefore, if FIRST and CLEFT coincide, the combination of two such pragmatically salient factors may suffice to mark a constituent as the Focus (cf. Section 4.1).

A further factor that has a significant impact on the distribution of pronoun case forms in it-clefts is SINGULAR (cf. Section 6.1.2.3). With an odds ratio of 0.61, SINGULAR evidently inhibits the use of subject pronoun case forms in it-clefts in which the clefted element is a case-sensitive singular pronoun form. This finding is not only in contrast to the findings obtained from the BNC, where this variable was eliminated as being non-significant, but also to the expectations of some other studies which consider plural subject pronoun forms in it-clefts to be unexpected (Sobin 1997: 334) This difference could be again explained in terms of an implicational hierarchy. In analogy to the differences between first
and third person (cf. Section 6.1.2.2), plural pronouns may need additional Focus marking in order to be identifiable as particularly salient and important discourse entities or even as the Focus of a clause (cf. Croft 2003: 126-132; Sections 6.1.2.3 and 9.1). The necessity to stress plural forms more than singular pronouns may be due to the scope of potential referents. Particularly third person plural pronouns can be used – in contrast to all other case-sensitive pronoun forms – to refer to virtually every plurality of animate, inanimate, concrete and abstract referents as long as they are countable, which means that their scope of potential referents is extremely broad. Hence, the immediate relevance of the entities they refer to may not be easily recognisable for the participants, particularly the hearer(s) or reader(s), of an ongoing communicative situation (cf. Sections 4.1 and 9.1). This may account for the fact that plural pronouns may receive additional highlighting by means of subject pronoun case forms in order to mark their referents as Focus. Similar to the observed distributional differences between first and third person pronouns and their assumed interaction with the sentence type in which these forms are used, it is possible that if SINGULAR and CLEFT coincide, the combination of these factors may be sufficient to highlight a singular pronoun form as Focus, which would explain why plural pronoun forms in it-clefts in the COCA are more likely to be used in their subject form than singular pronouns. However, this possible explanation has to remain tentative or even speculative at this point of time, since this factor does not play a role in the BNC and these findings even contradict the assumptions of the literature (cf. Sobin 1997: 337). Hence, the observed differences in terms of pronoun case distribution between singular and plural forms may also be due to other reasons, such as cross-varietal differences or idiosyncrasies of the COCA. Thus, the subsequent analyses of the other datasets will shed further light on this issue and will help to assess whether there is any general pattern observable in the different distributional profiles between singular and plural pronoun case forms or whether this is a variety-specific phenomenon.

Finally, Table 29 also identifies SPOK as a very highly significant factor in the distribution of pronoun case forms in the it-cleft data of the COCA. With an odds ratio of 0.24, this factor inhibits the use of subject pronoun case forms. In terms of percentages, we can say that the likelihood of observing a subject pronoun case form as focal pronoun is 76 per cent lower in the spoken data than in the written data of the COCA. This finding corresponds well not only to the observations made for the BNC (cf. Section 8.3), but also to the predictions of the relevant literature (e.g. Biber et al. 1999: 335-336; Harris 1981: 19). As
discussed above (cf. Section 9.1), the reason for the lower share of subject forms in spoken
it-clefts could be either due to the generally lower degree of formality in spoken data or to
the fact that speakers also possess other Focus marking devices, such as prosodic
prominence, which in combination with the cleft construction may be deemed sufficient to
mark a pronominal referent as the Focus of the sentence (cf. Sections 8.3 and 9.1 above).
The latter explanation is also indirectly supported by the fact that we do not observe a
significant influence of SPOK in the it BE sentences of the COCA (cf. Section 9.2). If this
difference in the distribution of pronoun case forms between spoken and written data were
indeed solely due to aspects of formality, we would expect to observe this formal–informal
split also for the it BE sentences of this corpus, which is, however, not the case.

As far as the significance of the minimal-adequate model is concerned, Table 29
shows that the model itself is very highly statistically significant, which indicates a very
strong correlation of the dependent variable with the retained independent ones ($\chi^2 =
458.87$, $p<0.001$). Furthermore, the applied model can account for 46 per cent of the
variation observed in the dependent variable, which is indicated by the model’s Nagelkerke’s
$R^2$ value of 0.46. Thus, this model is evidently substantially significant (cf. Szmrecsanyi 2006:
55). In addition, the minimal-adequate model predicts the outcome of the dependent
variable in 85.29 per cent of all instances correctly and therefore clearly improves the
baseline model (cf. Table 29).

For the it-clefts in the COCA, we can conclude that the findings for the variable
NUMBER, operationalised with the help of SINGULAR, are neither in line with the corresponding
observations made for the BNC nor with the expectations of the relevant literature (e.g.
Sobin 1997: 334). However, possible explanations for this deviation, i.e. variety-specific
feature or highlighting strategy based on typological considerations, are discussed above. All
other factors retained in the regression model affect the dependent variable as predicted by
the body of literature and as observed in the BNC. Whereas AS_SUBJ strongly promotes the
use of subject pronoun case forms, FIRST and SPOK notably inhibit the use of subject forms in
it-clefts in the COCA. As Table 29 demonstrates, AS_SUBJ exerts a very strong influence on
the dependent variable, which conforms very well to the trend observed so far for both the
BNC and the COCA that those factors that can be identified as having syntactic, pragmatic
and/or discourse-pragmatic functions seem to influence consistently and strongly the
distribution of pronoun case forms.
9.4 The Distribution of Pronoun Case Forms in the Corpus of Contemporary American English: Interim Summary

To conclude the discussion of the COCA data, we can state that the analysis of the subject predicative complements and the two subcategories of it BE sentences and it-clefs has again produced very remarkable findings.

To begin with, we have again observed considerable distributional differences between it-clefs and it BE sentences in terms of pronoun case usage. Hence, the separate analysis of each construction type has once again proved reasonable, since the respective factors affect the two constructions quite differently. This corresponds to what we have observed before for the BNC data.

As a second major finding that is also in line with the observations made for the BNC, we can note that all three analyses conducted in this chapter clearly demonstrate that syntactic and pragmatic factors, i.e. CLEFT, FOCUS and AS_SUBJ, very strongly influence the distribution of pronoun case forms in the respective contexts (cf. Section 5.2; H2 and H3). Hence, the COCA data also indicate that accounts trying to explain and predict the distribution of pronoun case forms in terms of pronoun-class membership or in terms of purely positional considerations are too simplistic, since the distribution of pronoun forms is strongly influenced by several functional factors (cf. Section 5.2; H1).

Moreover, these three syntactic, pragmatic or discourse-pragmatic factors also have in common that each of them evidently promotes the use of subject pronoun forms in the respective contexts. This finding supports one of the central assumptions of this study namely that subject pronoun case forms have been reanalysed as Focus markers in subjective predicative complements (cf. Section 5.2; H3). Furthermore, we have observed that subject pronouns are much more likely to occur in it-clefs than in it BE sentences and we have also noted that subject pronoun forms are much more likely to be used in particularly focussed than in normal it BE sentences in the COCA. Hence, both necessary and sufficient conditions postulated in Section 6.1.2.5 as requirements for allowing the conclusion that a reanalysis from former case forms to Focus markers may have taken place are fulfilled. This finding is in accordance with the results obtained for the BNC (cf. Section 8.4).
The analysis of the grammatical variables PERSON and NUMBER yields a very differentiated picture. With regard to the former, the factor FIRST inhibits the use of subject pronoun forms in each of the datasets analysed in Sections 9.1–9.3. As was the case for the relevant sections in the BNC, this difference could be explained with an implicational hierarchy according to which third person pronouns are due to their prototypically phoric status more likely to receive additional Focus marking by means of subject pronouns than first person pronouns in order to mark their relevance for the current communicative situation. This assumption is corroborated both by the results of the BNC and the COCA as well as the observations of other studies (e.g. Wales 1996: 96; Quinn 2005a: 134-135). The findings obtained for the variable NUMBER in the present chapter are somehow more difficult to deal with. On the one hand, they are not attested by other studies (e.g. Quinn 2005a: 135) or even contradicting their assumptions (Sobin 1997: 337). On the other hand, they also do not match the observations made for the BNC. Nevertheless, the fact that in it-clefs and in the superordinate subject predicative complements category plural pronouns are more likely to be used as subject forms than singular pronouns may also be accounted for with the help of an implicational hierarchy according to which plural pronouns may need more Focus marking than singular ones in order to identify them as most important piece of information. However, since this has not been attested before this might as well be a particularity of this variety or database (cf. Bisang 2004: 26-27).

The factor SPOK turned out to be significantly inhibiting the use of subject pronoun case forms in the general subject predicative complement category as well as in the it-cleft category. In the class of it BE sentences, however, it is not significant, which makes the results for this factor again difficult to interpret. Although informal contexts are generally assumed to disfavour subject forms (e.g. Huddleston and Pullum 2002: 459), we would expect to observe this formal–informal split not only for it-clefs but also for it BE sentences, if this spoken–written distinction is due to formality, which is, however, not the case in the COCA. Thus, in spoken language, it is also possible that the speaker of an it-cleft uses another Focus marking device, i.e. prosodic prominence, and not a subject pronoun case form, which leads to a trade-off between the different Focus markers (cf. Section 4.1). It is therefore possible that two Focus markers, i.e. prosodic prominence and the use of an it-cleft, may be deemed enough by the speaker to highlight a pronoun as the Focus of the clause or sentence.
10 The Distribution of Pronoun Case Forms in the British (.uk) Internet Data

After having analysed the data acquired from the closed corpora in the preceding Chapters 8 and 9, the subsequent Chapters 10–14 will discuss the results obtained from the Web-derived datasets which were compiled and classified prospectively for this study (cf. Section 7.2.4). To begin with, the following sections present the results of the multivariate statistical analyses obtained for the British Internet data (henceforth .uk data). As has become customary by now, we start again with the presentation and discussion of the results obtained for the superordinate category of subject predicative complements in Section 10.1. In a second step, Section 10.2 shows the results for the it BE sentences as observed in the .uk data, before Section 10.3 focuses on the distribution of pronoun case forms in it-clefts in this dataset. Eventually, Section 10.4 completes this chapter by providing an interim summary of the most important observations made for the .uk data.

10.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the British (.uk) Internet Data

Both for the sake of brevity and also in analogy to the two preceding Chapters 8 and 9, a reiteration of the discussion of the tested variables as well as of the mode of data compilation, verification and classification is not considered necessary at this point. Instead, the interested reader is referred to the relevant Chapters 6 and 7 discussing these issues in detail.

As illustrated in Table 30, the processing of the raw data (cf. Sections 7.2.4.1–7.2.4.3) has left us with 1926 subject predicative complements following it and form of BE for the top-level domain of the United kingdom (.uk). When compared to the 1379 tokens obtained from the BNC, one may be tempted to consider the .uk token number as puny considering the vast amount of data the Internet as a whole and even a national top-level domain such as .uk can offer. However, it has to be borne in mind that the token numbers of the Web-derived data collections are limited due to the fact that commercial Web crawlers usually only display the first 1000 hits retrieved for a given search string. Moreover, a considerable
amount of these 1000 hits had to be excluded in the course of the subsequent data processing, i.e. cleaning, verification and classification of the data (cf. Sections 7.2.4.2 and 7.2.4.3). It also has to be taken into consideration that the Web-queries were conducted for only one form of BE, i.e. is, and not also for was and ‘s as was the case for the closed corpora (cf. Sections 7.1 and 7.2.4.1). Although the combination with the form is yielded considerably more hits for potential subject predicative complements in the Web data than combinations with was and ‘s, this observation does not correspond to those made for the corpus data, where quite the opposite can be found. In the BNC, for example, only 201 of the 1379 subject predicative complements, i.e. a share of only 14.58 per cent, have is as their finite form of BE. Thus, taking this rather modest number of 201 tokens observed in a 100 million word corpus into account, the 1926 tokens obtained from the .uk domain appear all of a sudden in a very different light since this number highlights how comprehensive a closed corpora would have to be to yield a number of tokens comparable to that offered by the Web. In addition, this comparison gives a vital impression of how much data is needed to examine the phenomena discussed in this study.

<table>
<thead>
<tr>
<th>Subject Predicative Complements (.uk)</th>
<th>Total N</th>
<th>Subject Forms N</th>
<th>Object Forms N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1926 (100%)</td>
<td>416 (21.60%)</td>
<td>1510 (78.40%)</td>
</tr>
</tbody>
</table>

Table 30: Subject Predicative Complements in the .uk Data: The Token Numbers

Of the 1926 subject predicative complements obtained from the .uk domain which form the basis for the subsequent statistical analysis, a share of 21.60 per cent have a subject pronoun case form as pronominal complement whereas 78.40 per cent of all tokens exhibit an object pronoun case form. If the proportion of subject pronoun forms attested in this dataset is compared to those of the closed corpora (cf. Sections 8.1 and 9.1), we observe that the share of subject forms in the .uk data is considerably lower than that of the COCA and particularly than that of the BNC, the latter of which amounts to 46.27 per cent and is thus more than twice as large. However, this comparatively small share of subject forms in the .uk dataset might be due to several reasons, such as the composition of the superordinate category with regard to the proportions of its subsets, i.e. it-clefs and it BE sentences (cf. Section 9.3). Another factor that may also be responsible for the lower share of subject forms in the .uk data when compared to the BNC is the fact that commercial Web
crawlers ignore punctuation marks (cf. Section 7.2.2). Since this shortcoming particularly affects subject pronoun case forms, which are of course more likely to occur clause-initially than object pronouns, we have already assumed in Section 7.2.4.2 that the Web-based datasets may exhibit a lower share of subject forms when compared to closed corpora. This prediction has so far been confirmed by the data and it will be interesting to see whether or not the subsequent statistical analysis yields further evidence corroborating this prediction.

As Table 31 below demonstrates, five factors and one interaction are retained in the minimal-adequate model as significantly influencing the distribution of pronoun case forms in subject predicative complements in the .uk dataset. The only main effect eliminated from the final model is SPOK. Thus, the difference between spoken and traditional written data does not significantly influence the distribution of pronoun case forms in subject predicative complements in the .uk data. Although this is not in accordance with the predictions and expectations of much of the literature (e.g. Huddleston and Pullum 2002: 459; Wales 1996: 91-108), the fact that SPOK has been omitted from the minimal-adequate model depicted in Table 31 is not too surprising. On the one hand, the non-significance of this factor in the .uk data can be partially attributed to the generally rather formal character of much of the spoken data obtained from the Internet (cf. Section 7.2.4.3.1). On the other hand, the analysis of the BNC data has also shown that the assumption that object forms are generally favoured in informal contexts does not hold – at least not for British data (cf. Sections 8.1–8.3).45

45 In addition, the following interactions were eliminated in the course of the model-building process: SPOK*FIRST, SPOK*SINGULAR, SPOK*CLEFT, SPOK*FOCUS, CMC*FIRST, CMC*SINGULAR, CMC*CLEFT, CMC*FOCUS, FIRST*CLEFT, FIRST*FOCUS, SINGULAR*CLEFT, SINGULAR*FOCUS.
As far as the statistical analysis of the data is concerned, the results indicate that **CLEFT** is very highly significant in this dataset and strongly promotes the use of subject pronoun case forms in subject predicative complements (cf. Table 31). In analogy to both the BNC and the COCA (cf. Sections 8.1 and 9.1), this factor exhibits again the highest odds ratio, which amounts in the present model to a value of 52.41. This means that the probability of observing a subject case form is 5141 per cent higher in an *it*-cleft than in *it BE* sentence in the .uk dataset. This finding is both in line with the results obtained for the BNC and the COCA and also with those accounts arguing for marked differences between *it*-clefts and *it BE* sentences in terms of pronoun case distribution (e.g. Biber et al. 1999: 335-336; Maier 2013). However, this observation challenges again accounts proposing the similarity of *it BE* sentences and *it*-clefts in terms of pronoun case usage as well as their underlying theoretical assumptions (e.g. Harris 1981: 18-19; Quinn 2005a). Once again, the statistical analysis shows that the highlighting or discourse-pragmatic function of *it*-clefts seems to clearly foster the use of subject pronoun forms.

With regard to the variable **PERSON**, we can note that **FIRST** is very highly significant and markedly inhibits the use of subject pronoun forms, which can be deduced from the factor’s odds ratio of 0.36. As was the case before (cf. Section 9.1), this observation is in contrast to assumptions of the literature according to which a higher share of first person

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>3.96</td>
<td>***</td>
<td>52.42</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.02</td>
<td>***</td>
<td>0.36</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>0.65</td>
<td>***</td>
<td>1.92</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.56</td>
<td>***</td>
<td>0.21</td>
</tr>
<tr>
<td>FOCUS</td>
<td>2.20</td>
<td>***</td>
<td>9.04</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>-2.00</td>
<td>***</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Model $X^2$ = 669.26 (***)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R$^2$</td>
<td>0.45</td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td>84.11</td>
</tr>
<tr>
<td>% BASELINE</td>
<td>78.40</td>
</tr>
<tr>
<td>N</td>
<td>1926</td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 31: Subject Predicative Complements in the .uk Data: Logistic Regression Results
than third person subject pronoun forms is expected in *it BE* sentences (e.g. Quinn 2005a: 246). At the same time, however, it accords with findings made in earlier studies for *it*-clefts stating that third person pronouns are more likely than first person pronouns to be used in their subject form in this context (e.g. Quinn 2009: 42; Wales 1996: 95-96). As discussed before (e.g. Sections 8.3 and 9.1), the higher likelihood of observing third person subject forms may be accounted for with the pragmatic differences in the referential nature of first and third person pronouns (e.g. Croft 2003: 130; Siewierska 2004: 5-8). These fundamental differences may cause speakers and writers to additionally highlight the referents of third person pronouns in subject predicative complements by means of subject pronoun forms as particularly prominent in case they are the most important piece of information of the clause. This line of reasoning could explain why third person rather than first person pronouns are more likely to be used in their subject form in subject predicative complements. Furthermore, this explanation would also be fully consistent with hypothesis H3 assuming that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements (cf. Section 5.2). Hence, this possible pragmatic re-functionalisation of this core-grammatical distinction may also help to explain why we have observed so far more or less systematically significant differences in the distribution of pronoun case forms between first and third person pronouns in different datasets, i.e. in the .uk dataset, the COCA data and to some extent also in the BNC (cf. Chapters 8 and 9).

The other core-grammatical variable retained in the statistical model, i.e. the distinction between singular and plural pronouns operationalised with the factor SINGULAR, is not nearly as uniform in its effect across varieties of English as the factor FIRST. In the present dataset, SINGULAR is very highly significant and substantially promotes the use of subject forms in subject predicative complements. As can be seen from its odds ratio of 1.92, subject forms are 92 per cent more likely to occur in subject predicative complements with a singular pronoun than in subject predicative complements with a plural pronoun. While this finding meets the expectations of some theoretical accounts (Sobin 1997: 334), it is in contrast to the effect of SINGULAR observed for the COCA data, where we observe an inhibiting effect of SINGULAR. Furthermore, this result is also not in accordance with that obtained for the BNC data, where SINGULAR is not significant (cf. Sections 8.1 and 9.1). This means that this factor behaves differently in each of the three datasets analysed so far. Hence, the preliminary hypothesis voiced in Section 9.1 assuming that an implicational
hierarchy could be at work according to which plural pronouns may need more Focus marking than singular pronouns in order to be recognisable as the Focus of a clause cannot be confirmed in view of the present and previous results obtained for SINGULAR (cf. Sections 8.1 and 9.1). Indeed, the results for this factor have turned out to be maximally different from each other in the three datasets analysed so far.

The next factor maintained in the minimal-adequate model depicted in Table 31 is CMC assessing the influence computer-mediated communication has on the distribution of pronoun case forms in subject predicative complements (cf. Section 6.1.2.4). This factor is discussed in this chapter for the first time since neither the BNC nor the COCA contains data from this mode of discourse. As Table 21 shows, CMC markedly inhibits the use of subject pronouns in subject predicative complements, which becomes apparent from its odds ratio of 0.21. The inhibiting effect of this factor may be partly attributed to the differences in formality between the traditional written mode of discourse and the computer-mediated one, since CMC shares not only many aspects with traditional written data but also with the spoken mode of discourse (e.g. Crystal 2011: 32; Herring 2010a). This explanation solely in terms of formality would also correspond to expectations of the literature according to which subject pronoun forms in subject predicative complements are more closely associated with formal than informal varieties of English (e.g. Quirk et al. 1985: 337-338; Huddleston and Pullum 2002: 459). However, what is probably as important as formality itself to account for the differences between CMC and the traditional written data are the numerous features of CMC that are very much reminiscent of characteristics of spoken discourse and that can be attributed either to a lack of formality in CMC or – to put it more neutrally – to fewer text conventions, which allow for more flexibility in using alternative expressive options. For example, CMC can employ certain typographical practices imitating prosody and intonation units (e.g. Herring 2010a). This means, that CMC also possesses more options to highlight the Focus of a clause than traditional written data, which may also account for why CMC is less likely to use subject forms in subject predicative complements than traditional written data. Cases in point are sentences like the following taken from the CMC subset of the .uk data in which capitalisation is used rather than subject pronoun case forms to mark certain constituents as the relatively most important ones (cf. (139)):

(139) Capitalisation as Focus Marking Strategy in CMC

246
Why do YOU have to do anything? It is HIM who has to make a choice - do something positive about finding employment or carry on sponging on the social just because work doesn’t fit in with his social life. (.uk/it is him/15.07.2008)

Thus, although the differences between written and CMC data can be accounted for simply in terms of formality, they can also be explained in terms of Focus marking. CMC may not only be on average less formal than prototypical traditional written data, but it is also definitely more flexible in terms of its expressive repertoire. This higher degree of expressive flexibility allows for alternative Focus marking strategies in CMC, such as capitalisation (cf. (139)). Hence, it could also be argued that CMC is less likely to use subject pronoun forms because it can use other or additional means to highlight a certain clausal constituent as the Focus of a clause, which is in line with hypothesis H3 of this study according to which subject pronoun case forms may have been re-functionalised in subject predicative complements as Focus markers.

Speaking of Focus markers, this leads us straight to the next factor retained in the minimal-adequate model in Table 31, i.e. FOCUS. As was the case with both the BNC and the COCA (cf. Sections 8.1 and 9.1), FOCUS is very highly significant and with an odds ratio of 9.04 decidedly promotes the use of subject pronoun case forms in subject predicative complements. Translating this odds ratio into a percentage value, this means that the probability of observing a subject form increases by 804 per cent when the pronoun occurs in an it BE sentence that has been identified as being particularly focussed rather than in a normal it BE sentence (cf. Section 6.1.2.5). As was also the case with the BNC and the COCA (cf. Sections 8.1 and 9.1), FOCUS does not only very strongly promote the use of subject pronoun case forms in the .uk dataset, but its effect also emphatically corroborates hypothesis H3 assuming that subject pronouns case forms have been reanalysed as Focus markers in subject predicative complements (cf. Section 5.2). This seems true particularly for British English, since the influence of FOCUS on the distribution of pronoun case forms has by now been affirmed by two different datasets.

Finally, there is also one very highly significant interaction retained in the minimal-adequate model, namely that between FIRST and SINGULAR. This interaction clearly inhibits the use of subject pronoun case forms in subject predicative complements, which is indicated by its odds ratio of 0.14 (cf. Table 31). Hence, first person singular subject forms are significantly less likely to occur than other subject pronoun case forms in subject predicative complements in this dataset. The reason why the interaction between FIRST and SINGULAR
strongly inhibits the use of subject pronoun case forms can be explained, however, with the limitations of commercial Web crawlers concerning their potential to compile data for corpus linguistic studies. As outlined in Section 7.2.2, commercial Web crawlers ignore punctuation marks, even if the search string is set in quotation marks, which results in a large number of false positives (cf. Keller and Lapata 2003: 468). This means that – compared to the data of both the BNC and of the COCA – a smaller proportion of subject pronoun forms in subject predicative complements is to be expected in the Web-derived datasets, as has been predicted in Section 7.2.4.2. For the .uk dataset, this expectation has already been confirmed by the results described in Table 30, since the share of subject pronouns is indeed much smaller than that observed for the BNC, where the share of subject pronoun case forms is more than twice as high as in the .uk dataset (cf. Section 8.1). Furthermore, as stated in Section 7.2.4.2, the non-consideration of punctuation marks particularly affects the data for the first person singular for which the highest share of false positives can be observed. Thus, this significant interaction between FIRST and SINGULAR can be explained solely with the peculiarities of the mode of data compilation by means of a commercial search engine and it is very likely that this interaction is identified as significantly influencing the distribution of pronoun case forms in other Web-derived datasets as well.

Finally, the minimal-adequate model itself that is used to analyze the subject predicative complements in the .uk dataset is also very highly significant ($\chi^2 = 669.26$, $p<0.001$), as Table 31 shows. The Nagelkerke’s $R^2$ value of 0.45 indicates that the model can account for 45 per cent of the variation observed in the dependent variable. Furthermore, the applied model is able to predict the outcome of the dependent variable correctly in 84.11 per cent of all cases, thereby improving the baseline model, which only accounts for 78.40 per cent. Very interestingly, the predictive accuracy of both models applied to the British datasets is very similar. Whereas the model applied to the BNC data in Section 8.1 correctly predicts 84.41 per cent of all data, the present model can account for 84.11 per cent of all data points.

In sum, the results obtained for the superordinate category of subject predicative complements in the .uk dataset have corroborated the findings of the BNC and the COCA in that the functional, discourse-pragmatic factors CLEFT and FOCUS have a very strong impact on the distribution of pronoun case forms in subject predicative complements, both of which emphatically promote the use of subject pronoun forms. Thus, this finding again
corroborates one of this study’s central assumptions hypothesising that subject pronoun case forms may have been reanalysed as Focus markers (cf. Section 5.2; H3). With regard to the results obtained for the variables PERSON and NUMBER, the data yields again a differentiated picture. As was the case before (cf. Section 9.1), FIRST clearly inhibits the use of subject pronoun case forms and this difference between first and third person pronouns could be explained again by means of the pragmatic properties inherent in the nature of these different persons (cf. Section 6.1.2.2). The factor SINGULAR, however, does not seem to allow for generalisations as easily as the other factors discussed so far. While SINGULAR is not significant in the BNC, it inhibits the use of subject forms in the COCA and promotes it in the .uk dataset (cf. Section 8.1 and 9.1). Thus, the results obtained so far for SINGULAR seem to vary from dataset to dataset. With regard to the MODE OF DISCOURSE variable (cf. Section 6.1.2.4), SPOK does not significantly influence the choice of pronoun case forms in the present dataset, which may partly be due to the rather formal character of the spoken data obtained from the Internet data collection (cf. Section 7.2.4.3.1). The other mode of discourse distinguished for the analysis, i.e. CMC, turns out to be significantly inhibiting the use of subject forms. Whereas this inhibiting effect can be partly attributed to differences in the degree of formality itself, it has also been pointed out that exactly because of the fewer text conventions constraining CMC, it possesses more means to highlight a certain constituent as pragmatically prominent, which may also account for the differences between written and CMC data.

10.2 The Distribution of Pronoun Case Forms in it BE Sentences in the British (.uk) Internet Data

As can be seen from Table 32, 802 tokens of the 1926 subject predicative complements obtained from the Web-derived data collection for the .uk domain fall into the category of it BE sentences. As was already the case for both the BNC and the COCA, Table 32 indicates that the share of subject pronoun forms is considerably lower in the subclass of it BE sentences than in the superordinate category of subject predicative complements (cf. Section 10.1). Only 3.37 per cent of all it BE sentences in the .uk dataset exhibit a subject pronoun form as pronominal complement, while in the general subject predicative
complement category 21.60 per cent of all tokens are used with a subject form. Compared to the shares of subject pronoun forms observed in the corresponding subsets of the BNC and the COCA, the share of subject forms in the .uk data is clearly the smallest, since it is not even half as high as that attested for the BNC (cf. Sections 8.2 and 9.2). This, however, may also be partly attributed to the peculiarities of the mode of data collection used for this dataset (cf. Sections 7.2.4.2 and 10.1).

<table>
<thead>
<tr>
<th></th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT BE SENTENCES (.uk)</td>
<td>802 (100%)</td>
<td>27 (3.37%)</td>
<td>775 (96.63%)</td>
</tr>
</tbody>
</table>

*Table 32: it BE Sentences in the .uk Data: The Token Numbers*

Table 33 below shows the results for the multivariate analysis of the it BE sentences in the .uk dataset. For this dataset, three factors are retained in the minimal-adequate model as significantly influencing the distribution of pronoun case forms in it BE sentences. In the BNC and the COCA, in contrast, only two factors are maintained in the corresponding models (cf. Sections 8.2 and 9.2). Although three predictors are identified as significant, this also means that some predictors have been excluded to arrive at the minimal-adequate model.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGULAR</td>
<td>2.08</td>
<td>*</td>
<td>7.97</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.41</td>
<td>**</td>
<td>0.24</td>
</tr>
<tr>
<td>FOCUS</td>
<td>2.77</td>
<td>***</td>
<td>16.00</td>
</tr>
</tbody>
</table>

MODEL $\chi^2$ 71.56 (***)

<table>
<thead>
<tr>
<th>R$^2$</th>
<th>0.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td>96.63</td>
</tr>
<tr>
<td>% BASELINE</td>
<td>96.63</td>
</tr>
<tr>
<td>N</td>
<td>802</td>
</tr>
</tbody>
</table>

* + significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

*Table 33: it BE Sentences in the .uk Data: Logistic Regression Results*

The first of these omitted non-significant factors is FIRST. The exclusion of this factor may to some extent be unexpected, since some accounts state that “[n]on-1sg nominatives would seem to be extremely rare, and restricted to certain discourse contexts” (Quinn

250
Moreover, the analysis of the COCA data has demonstrated that First is very highly significant and clearly inhibits the use of subject forms in the *it BE* sentences in that corpus (cf. Section 9.2). However, the distinction between first and third person pronouns does also not significantly influence the distribution of pronoun case forms in the *it BE* sentences of the BNC. Hence, the elimination of this factor in the other British dataset is not too surprising.

While the exclusion of First is in line with earlier observations made for the corresponding BNC data, the omission of SPOK is more reminiscent of the COCA data (cf. Sections 8.2 and 9.2), since SPOK also plays no role in the distribution of pronoun case forms in the *it BE* sentences of the COCA. In the BNC, in contrast, this factor in tendency even promotes the use of subject pronoun forms in *it BE* sentences (cf. Section 8.2). However, the spoken data obtained from the Internet is on average more formal than that of the BNC, as outlined in Section 7.2.4.3.1. As a consequence, the non-significance of SPOK may not be surprising given the comparatively modest cline of formality between the written and the spoken data in the .uk dataset, even though much of the current literature discussing the distribution of pronoun case forms in *it BE* sentences assumes that subject forms are less likely to occur in spoken or informal varieties than in written or more formal ones (e.g. Biber et al. 1999: 335-336; Harris 1981: 18).

Turning eventually to the significant factors of the applied regression model, Table 33 indicates that the factor SINGULAR is significant and clearly promotes the use of subject pronoun case forms in *it BE* sentences in the .uk dataset, as its odds ratio of 7.97 demonstrates. Although this finding is expected by Sobin (1997: 334), who even rules out plural subject pronouns in *it BE* sentences altogether, this finding contrasts with the observations made for the BNC, where this factor plays no role in the distribution of pronoun case forms whatsoever (cf. Section 8.1–8.3). Moreover, it is also in contrast with the observations made for parts of the COCA data, where SINGULAR inhibits the use of subject forms, particularly in the *it*-cleft category (cf. Section 9.1 and 9.3).

The second factor identified in Table 33 as significantly influencing the distribution of pronoun case forms in *it BE* sentences in the .uk data is CMC, operationalising the difference between traditional written and computer-mediated written data (cf. Sections 6.1.2.4 and

46 In addition, the following interactions were eliminated in the course of the model-building process: SPOK*First, SPOK*SINGULAR, SPOK*Focus, CMC*First, CMC*SINGULAR, CMC*Focus, First*SINGULAR, First*Focus, SINGULAR*Focus.
7.2.4.3.3). As is illustrated in Table 33, CMC is highly significant and markedly inhibits the use of subject pronoun case forms in *it BE* sentences, which becomes apparent from its odds ratio of 0.24. This finding is in accordance with accounts proposing that subject forms are less likely to be used in informal than in formal varieties (e.g. Biber et al. 1999: 335-336; Harris 1981: 18), since CMC has often been considered as exhibiting many features of spoken communication (e.g. Crystal 2011: 32; Herring 2010a). However, this lower degree of formality may also help to explain the lesser likelihood of observing subject pronoun case forms in CMC from a more Focus-oriented perspective. Exactly because CMC may prototypically be subject to fewer formal restrictions than traditional written data, it has a larger stock of expressive options, including additional Focus marking devices, which are not as readily available for the traditional written mode of discourse (cf. Section 10.1). An example from the .uk data for such an alternative strategy in CMC to highlight a pronominal constituent as particularly prominent is given in (140), where capitalisation is used as emphasis or Focus marker:

(140) Capitalisation as Focus Marking Strategy in *it BE* Sentences in CMC

Oh and to bleeding heart AC, it is not any of us affecting the prisoner’s family life, it is THEM, their direct actions caused the problem, and they ought to deal with it instead of being bailed out. (.uk/it is him/15.07.2008)

Thus, the distributional differences between traditional written data and CMC data with regard to the use of pronoun case forms cannot only be explained with their prototypical degree of formality but also more specifically in terms of the different text conventions arising from these differences in formality. As illustrated in (140), these different text conventions allow CMC, for example, to use more expressive devices, such as capitalisation, to indicate the discourse-pragmatic prominence of a certain clausal constituent. Hence, the inhibiting effect of CMC observed for the .uk dataset is not only compatible with accounts assuming that the differences in the distribution of pronoun case forms between more formal and less formal varieties or modes of discourse are due to the putatively more formal character of subject pronoun case forms (e.g. Harris 1981; Quirk et al. 1985: 337-338). This finding is also in line with the stance taken in this study hypothesising that subject pronoun case forms may be used as Focus markers in subject predicative complements (cf. Chapter 4). However, the exact impact of capitalisation and other means to highlight discourse-pragmatic prominence in CMC still has to be addressed in future studies, since this lies beyond the scope of the present work, which focuses on the distribution of pronoun case
forms. Suffice it here to say that the example in (140) indicates that subject pronoun forms seem to compete with more alternative Focus marking strategies in CMC than in traditional written data, which could also partly account for why subject forms are less often used in CMC than in traditional written data.

The last factor retained in the minimal-adequate regression model used to analyse the *it BE* sentences of the .uk dataset is Focus (cf. Table 33). This factor is not only very highly significant but also emphatically promotes the use of subject pronoun case forms as its odds ratio of 16.00 illustrates. Thus, the likelihood of observing a subject form in an *it BE* sentence increases by 1500 per cent, when the pronoun occurs in a particularly focussed *it BE* sentence rather than a normal *it BE* sentence (cf. Section 6.1.2.5). Hence, this finding corresponds to the results obtained for the BNC and the COCA where Focus is also very highly significant and also strongly promotes the use of subject pronoun case forms in *it BE* sentences. Indeed, in each of the respective datasets analysed so far, Focus has been very highly significant and it has also been the factor with the highest odds ratio evidently promoting the use of subject pronoun forms in *it BE* sentences in the BNC, the COCA and the .uk dataset (cf. Sections 8.2 and 9.2). Thus, this factor seems to play an important role in the distribution of pronoun case forms in *it BE* sentences – at least in the varieties or datasets analysed so far.

Finally, the minimal-adequate regression model used to analyse the present dataset is itself also very highly statistically significant ($\chi^2 = 71.56, p<0.001$). This indicates, of course, a strong correlation between the dependent variable and the independent variables retained in the model. As the Nagelkerke’s $R^2$ value of 0.33 indicates, the statistical model can account for 33 per cent of the variation observed in the dependent variable, which is a fairly good result, especially when compared to the amount of variation explained by the corresponding minimal-adequate models applied to the *it BE* sentences in the BNC and COCA (cf. Sections 8.2 and 9.2). Although the present minimal-adequate model is itself very highly significant and can account for 33 per cent of the variation observed in the dependent variable, it does not, however, substantially improve the baseline model (cf. Table 33), since both models are able to correctly predict the outcome of the dependent variable in 96.63 per cent of all cases.

In a nutshell, the multivariate analysis of the *it BE* sentences in the .uk data has shown that the factor SINGULAR promotes the use of subject pronoun forms in the present
The distribution of pronoun case forms in subject predicative complements is influenced by functional and particularly pragmatic features. This is confirmed by the distribution of pronoun case forms in the .uk Internet Data (Sobin 1997: 334). However, the factor CMC markedly inhibits the use of subject pronoun case forms in the it BE sentences of the .uk data. In view of the fewer formal restrictions of CMC, this finding is in accordance with accounts assuming differences in the distribution of pronoun case forms between more formal and less formal varieties or modes of discourse (e.g. Harris 1981; Quirk et al. 1985: 337-338). However, this finding is also consistent with a more Focus-oriented perspective hypothesising that due to the fewer text conventions, CMC can use a larger stock of expressive means to highlight a constituent as the Focus of a clause, which may also partly explain why subject forms are less likely to occur in CMC data. In analogy to the results obtained for the respective sections in the BNC and the COCA, the most important factor in the domain of it BE sentences is again Focus. It is the only very highly significant factor retained in the regression model and it is also the factor with the highest odds ratio clearly promoting the use of subject forms. Thus, the analysis of the it BE sentences of the .uk data confirms again the assumptions of this study according to which functional and particularly pragmatic features play an important role in the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H2 and H3).

10.3 The Distribution of Pronoun Case Forms in it-Clefts in the British (.uk) Internet Data

Turning now to the second subgroup of subject predicative complements distinguished in this study, Table 34 shows that 1124 tokens of the 1926 subject predicative complements obtained from the Web-derived data collection for the .uk domain are it-clefts. When we compare the shares of subject and object pronoun case forms of the it-clefts given in Table 34 with those of the it BE sentences in the .uk data (cf. Section 10.2), we observe again marked differences between these two sentence types in the distribution of pronoun case forms. Whereas only 3.37 per cent of all it BE sentences in the .uk data are used with a subject pronoun case form, 34.61 per cent of all it-cleft tokens in this dataset exhibit a subject pronoun case form. Hence, the share of subject forms in the it-cleft data is more than ten times higher than that observed for the it BE sentences in the .uk data. This marked
difference not only highlights again how different it-clefs and it BE sentences are with regard to pronoun case distribution, but also firmly corroborates accounts stressing the differences between these two sentence types (e.g. Erdmann 1978; Maier 2013). When comparing the distribution of pronoun case forms in this it-cleft section with those of the closed corpora in Sections 8.3 and 9.3, we also note that the share of subject forms is rather modest in the present dataset. The reason for this comparatively low share of subject forms in the .uk data can be partly attributed, however, to the particularities of the Web-derived data collection (cf. Section 7.2.4.2), which have already been discussed in this chapter (cf. Section 10.1) and which will be briefly readdressed below.

<table>
<thead>
<tr>
<th>IT-CLEFTS (.uk)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1124 (100%)</td>
<td>389 (34.61%)</td>
<td>735 (65.39%)</td>
</tr>
</tbody>
</table>

*Table 34: it-Clefs in the .uk Data: the Token Numbers*

Table 35 summarises the results obtained for the multivariate analysis of the it-cleft data in the .uk dataset. As Table 35 demonstrates, four main effects and one interaction are identified as significantly influencing the distribution of pronoun case forms in the it-clefs of the .uk data. Thus, only one main effect has been eliminated in the model-building process as being non-significant. Although much of the literature expects SPOK to inhibit the use of subject forms in this context (e.g. Huddleston and Pullum 2002: 459; Wales 1996: 91-108), the non-significance of SPOK is not too much of a surprise in view of the rather formal character of much of the spoken data retrieved from the Internet (cf. Section 7.2.4.3.1).47

---

47 In addition to the factor SPOK, the following interactions were eliminated to arrive at the minimal-adequate model depicted in Table 35: SPOK*FIRST, SPOK*SINGULAR, SPOK*AS_SUBJ, CMC*FIRST, CMC*SINGULAR, CMC*AS_SUBJ, FIRST*AS_SUBJ*SINGULAR*AS_SUBJ.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.81</td>
<td>***</td>
<td>6.13</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.17</td>
<td>***</td>
<td>0.31</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>0.70</td>
<td>***</td>
<td>2.01</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.64</td>
<td>***</td>
<td>0.19</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>-2.28</td>
<td>***</td>
<td>0.10</td>
</tr>
</tbody>
</table>

| MODEL $\chi^2$ | 328.96 (***), $R^2$ 0.35, $\%$ CORRECTLY PREDICTED 77.14, $\%$ BASELINE 65.39, N 1124 |

Table 35: it-Clefts in the .uk Data: Logistic Regression Results

As indicated in Table 35, AS_SUBJ is very highly significant and it is also the factor with the highest odds ratio in the minimal-adequate regression model. With an odds ratio of 6.13, the likelihood of observing a subject pronoun case form as focal pronoun is more than 500 per cent higher in it-clefts in which the focal pronoun is co-referential with the subject of the depending clause than in it-clefts where the focal pronoun is co-referential with an object of the dependent clause in the .uk dataset. Hence, this finding is consistent with the observations made for the it-cleft data in both the BNC and the COCA, where AS_SUBJ is also the factor with the highest odds ratio in the it-cleft data (cf. Sections 8.3 and 9.3). Furthermore, this result meets the expectations of the literature according to which this factor promotes the use of subject forms as focal pronouns of it-clefts (e.g. Erdmann 1978: 76-77; Huddleston and Pullum 2002: 459; Quinn 2005a: 133). On a more general level, this finding also corroborates the general trend observed so far that factors with clearly syntactic, pragmatic and/or discourse-pragmatic functions, i.e. CLEFT, AS_SUBJ and FOCUS, have a very strong and promoting effect on the use of subject pronoun forms in the respective contexts.

The second factor identified in Table 35 as being very highly significant is FIRST, which markedly constrains the use of subject pronoun case forms in it-clefts in the .uk data. Its odds ratio of 0.31 signals that first person pronouns are 69 per cent less likely to occur in a subject form than third person pronouns in the it-clefts of the .uk dataset. The inhibiting
effect of FIRST is in line with the observations made for this factor in the BNC and in the COCA, where FIRST has a similar effect (cf. Sections 8.3 and 9.3), and also with the observations of earlier studies (e.g. Quinn 2005a: 135-140; Wales 1996: 95-96). As outlined before (e.g. Sections 8.3 and 9.1), a possible explanation for why third persons are more likely than first person pronouns to be used in their subject forms are the differences concerning their prototypical referential status. Since they are more or less by definition highlighted, the immediate discourse-pragmatic relevance of first person pronouns as well as their referents is obvious to the recipient(s) from the very use of them (cf. Langacker 1991: 307; Siewierska 2004: 5-8). The referents of third person pronouns, however, may need additional highlighting to mark them as the Focus of the clause, which may be done by means of subject pronoun case forms (cf. e.g. Sections 6.1.2.2, 8.3 and 9.3). In analogy to the BNC and COCA data, a possible explanation for why first person subject pronouns are particularly less likely to occur in it-clefts may be the circumstance that if FIRST and CLEFT coincide, the interaction of two pragmatically salient factors may be sufficient to mark a constituent as particularly prominent (cf. Sections 8.3 and 9.3; cf. Dik 1989: 278).

The third factor retained in the regression model in Table 35 as very highly significant is SINGULAR. In this model, SINGULAR exerts a considerable effect on the distribution of pronoun case forms in it-clefts in the .uk data and it clearly promotes the use of subject forms, which becomes evident from its odds ratio of 2.01. While this finding meets the expectations of Sobin (1997: 334), who assumes that subject forms are mainly restricted to the singular, this result is neither in line with the observations of studies which could not corroborate such a hypothesis (e.g. Quinn 2005a: 134-135) nor with the observations made in this study so far. In the BNC, this factor does not play any role in the distribution of pronoun case forms, and in the COCA SINGULAR actually inhibits the use of subject pronoun case forms in it-clefts (cf. Sections 8.3 and 9.3). Hence, whereas the impact of the other grammatical distinction, i.e. FIRST, on the distribution of pronoun case forms seems to be quite uniform across the different datasets examined so far, the results obtained for SINGULAR do not lend themselves as easily to robust generalisations, since the effect and impact of this factor have so far been different from dataset to dataset. As a consequence, the preliminary hypothesis formulated in Section 9.1 suggesting an implicational hierarchy according to which plural pronouns may need more Focus marking than singular pronouns in order to mark the immediate relevance of their referents for the ongoing discourse and to
render them recognisable as the most important discourse entity of a clause cannot be confirmed in view of the very heterogeneous results obtained so far (cf. Sections 8.3 and 9.3).

In addition to the three factors discussed so far, CMC determining the impact computer-mediated communication has on the distribution of pronoun case forms in *it*-clefts in the .uk data is also identified as being very highly significant. In analogy to the observations made for preceding sections (cf. Sections 10.1 and 10.2), this factor markedly lessens the likelihood of observing subject pronoun forms in *it*-clefts, as its odds ratio of 0.19 indicates. Since CMC is often considered to bear many traits of spoken discourse and because it is subject to fewer formalised text conventions (e.g. Crystal 2011: 57-75; Herring 2010a), this observation is in line with accounts assuming a lower likelihood of subject forms to occur in less formal contexts (e.g. Quirk et al. 1985: 335-337). The greater flexibility with regard to formal conventions also enables CMC to employ more options, such as capitalisation, to mark a constituent as the Focus of a clause, which may also account for why subject pronoun case are less likely to occur in CMC than in the traditional written mode of discourse. Since this Focus-oriented analysis of the difference between CMC and written data has already been discussed in detail in Sections 10.1 and 10.2, the interested reader is referred to the preceding sections for examples illustrating possible alternative Focus marking strategies in CMC.

Finally, Table 35 also identifies one interaction as very highly significant, namely that between FIRST and SINGULAR, the result of which is a very pronounced inhibiting effect on the use of subject pronoun case forms in *it*-clefts, which becomes apparent from the interaction’s odds ratio of 0.10. However, as outlined before (cf. Section 10.1), this interaction can be attributed to the mode of data collection and to the particularities of commercial search engines, such as Google. Since commercial search engines ignore punctuation marks, a data collection by means of such a Web crawler generates a huge number of false positives, which generally results in a lower share of subject pronoun case forms when compared to closed corpora (cf. Sections 7.2.2 and 10.1). Moreover, as noted in Section 7.2.4.2, this non-consideration of punctuation marks affects particularly the subject form of the first person singular, which suffers from the highest share of false positives. Hence, the interaction between FIRST and SINGULAR in Table 35 can be more or less exclusively attributed to the shortcomings of commercial search engines for corpus linguistic data.
compilations and we can expect to observe this interaction also in the minimal-adequate regression models of some of the following Web-derived datasets.

As indicated by Table 35, the minimal-adequate regression model as a whole is also very highly statistically significant ($\chi^2 = 328.96$, p<0.001). Its Nagelkerke’s $R^2$ value of 0.35 indicates that the regression model is able to account for 35 per cent of the variation observed in the dependent variable, which renders the model substantially significant (Szmrecsanyi 2006: 55). Furthermore, the minimal-adequate regression model used to analyse the $it$-cleft data is able to predict the outcome of the dependent variable in 77.14 per cent of all cases correctly, which clearly improves the baseline model, which can only correctly account for 65.39 per cent of all instances.

With regard to the $it$-clefts in the .uk dataset, we have observed both similarities and differences to the other datasets analysed so far. In analogy to the BNC and the COCA and as predicted by much of the literature (e.g. Quirk et al. 1985: 337-338), As_subj strongly promotes the use of subject pronoun case forms in $it$-clefts, which is consistent with the trend observed so far according to which functional factors strongly influence the distribution of pronoun case forms. Also similar to the BNC and the COCA and the observations of earlier studies (e.g. Quinn 2005a: 135-140), the factor First clearly inhibits the use of subject pronouns in the analysed context. The factor Singular, in contrast, clearly promotes the use of subject pronoun case forms in the .uk $it$-clefts. Although this is line with the literature (Sobin 1997: 334), this finding does not correspond to the observations made for the BNC and the COCA. As the analysis of the $it$-cleft data also demonstrates, the factor CMC markedly constrains the use of subject pronoun case forms. This inhibiting effect can be attributed not only to matters of formality but also to the fact that CMC possesses more means to highlight certain constituents as pragmatically salient.
10.4 The Distribution of Pronoun Case Forms in the British (.uk) Internet Data: Interim Summary

To sum up the results of the preceding three sections, the analysis of the .uk data has yielded results that are for the most part consistent with the observations made for the BNC and the COCA. Differences concern mainly the significance and impact of factors that are either not applicable to the BNC and the COCA, such as CMC, or that exhibit a considerable extent of variation even in these carefully compiled corpora, such as SINGULAR and SPOK.

To start with the similarities between the .uk data and the closed corpora, the analysis of the subject predicative complements has again demonstrated how different it-clefts and it BE sentences are in terms of pronoun case distribution. Hence, the examination of each sentence type separately has again been justified.

Also in analogy to the results of the BNC and the COCA, the individual analyses conducted in this chapter have consistently demonstrated that functional factors that can be assigned to the domains of syntax and pragmatics strongly influence the choice of pronoun case forms in all analysed contexts (cf. Section 5.2; H2 and H3). This strong influence particularly of CLEFT, FOCUS and AS_SUBJ confirms not only the trends already attested for the closed corpora in Chapters 8 and 9, but also severely challenges once again accounts trying to reduce pronoun case distribution in subject predicative complement contexts simply to matters of position within a clause or a pronoun-class membership (cf. Section 5.2; H1).

Furthermore, the results obtained for CLEFT, FOCUS and AS_SUBJ in the .uk data also corroborate the findings of the BNC and the COCA in that these syntactical and discourse-pragmatic factors also markedly foster the use of subject pronoun case forms in the respective analyses of the .uk dataset. Particularly with regard to the factors CLEFT and FOCUS, the findings of the .uk data strongly corroborate hypotheses H3 of this study assuming that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements. Since these factors emphatically promote the use of subject forms, both necessary and sufficient condition required for such an assumption are met (cf. Section 6.1.2.5).

With regard to the grammatical factors FIRST and SINGULAR, the analyses have produced the following results: FIRST does not play a role in the distribution of pronoun case
forms in the *it BE* sentences but in the category of *it*-clefts and the general subject predicative complements category, where it clearly constrains the use of subject forms in both categories. This is more or less in line with the BNC findings, where the same has been attested for the *it*-clefts and the *it BE* sentences (cf. Section 8.2 and 8.3), and also partly with the COCA results, where this factor even inhibits the use of subject forms in the *it BE* sentences (cf. Section 9.2 and 9.3). A possible explanation for why first person subject pronouns are significantly less likely to occur in *it*-clefts may be the concurrence of *FIRST* and *CLEFT*. If these two pragmatically prominent factors co-occur, this may be sufficient to mark a pronominal constituent as the Focus of a clause and additional marking by means of a subject form may not be deemed necessary (cf. Sections 9.3 and 10.3). More challenging to interpret are, however, the results obtained for *SINGULAR*. This factor markedly promotes the use of subject forms in the .uk dataset. Although *SINGULAR* thus turns out as predicted by parts of the literature (e.g. Sobin 1997: 334), the results obtained for the .uk data do neither correspond to the observations of other studies (e.g. Quinn 2005a: 134-135), nor to the findings for the BNC, nor to those obtained for the COCA. Hence, the impact of this factor seems to be dependent on the respective dataset or variety and does not allow easily for cross-varietal generalisations.

The *MODE OF DISCOURSE* variable has also yielded interesting results. The factor *SPOK* does not play any role in the .uk dataset, which can partly be attributed to the rather formal character of the spoken data in the Web-derived datasets and also to the fact that this factor is also only marginally significant in the BNC. The second *MODE OF DISCOURSE* factor *CMC*, however, strongly inhibits the use of subject pronoun forms in the .uk data. This has been attributed to the differences in formality and also to greater expressive freedoms resulting from the lower degree of formality, which allow, for example, the use of other Focus marking strategies in CMC. Whether these differences between CMC and traditional written data are particular to the .uk data or whether they can also be attested for other Web-derived datasets, will be examined in the subsequent chapters.
11 The Distribution of Pronoun Case Forms in the Australian (.au) Internet Data

In the following sections, this study analyses the distribution of pronoun case forms in the Australian Internet data (henceforth .au data) obtained from the Australian top-level domain .au (cf. Section 7.2.4). In analogy to the preceding three chapters, Section 11.1 examines and discusses the distribution of pronoun case forms in the superordinate subject predicative complement category. In a second step, this study will zoom in on the subgroup of *it BE* sentences in Section 11.2 and analyse which factors determine the distribution of pronoun case forms in this context in the .au data. Then, Section 11.3 will focus on the distribution of pronoun case forms in the Australian *it*-cleft data before this chapter will be concluded with an interim summary of the major results in Section 11.4.

11.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the Australian (.au) Internet Data

Table 36 illustrates that 1777 tokens remain after cleaning, verifying and classifying the raw data obtained from the Australian .au top-level domain, which are again the data on which the subsequent multivariate analyses are based (cf. Chapter 6 and Chapter 7).

<table>
<thead>
<tr>
<th></th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SUBJECT PREDICATIVE COMPLEMENTS (.au)</strong></td>
<td>1177 (100%)</td>
<td>238 (20.22%)</td>
<td>939 (79.78%)</td>
</tr>
</tbody>
</table>

*Table 36: Subject Predicative Complements in the .au Data: The Token Numbers*

With regard to the distribution of subject and object pronoun case forms, Table 36 shows that 20.22 per cent of all subject predicative complements are used with a subject pronoun form, compared to a share of 79.78 per cent of tokens exhibiting an object pronoun form. Thus, the share of subject pronoun case forms attested in the .au data is even slightly smaller than that observed for the .uk data, which accounts for 21.60 per cent of all tokens (cf. Section 10.1). One reason for why the shares of subject forms in the Web-derived data
are in general comparatively small are the limitations of commercial search engines, which have already been discussed (cf. Sections 7.2.4.2 and 10.1). Furthermore, this modest share of subject forms may also be partly due to the composition of the superordinate category in terms of the proportions of it BE sentences and it-clefs included in the data (cf. Section 9.1). However, to really identify the factors determining the distribution of pronoun case forms in subject predicative complements in the .au data, let us now turn to the multivariate analysis of the data.

Table 37 indicates that five main effects and one interaction between two of these factors are retained in the minimal-adequate model as significantly influencing the distribution of pronoun case forms in the .au data. Hence, only one factor was eliminated in the model-building process. This factor is again SPOK, which was also eliminated in the corresponding .uk model. The reason why SPOK may not play a role in the Internet datasets discussed so far, although much of the literature assumes it to inhibit the use of subject forms (e.g. Quirk et al. 1985: 337-338; Wales 1996: 91-108), may be the rather formal character of much of the spoken data obtained from Google, which may level potential differences between the spoken and written mode of discourse (cf. Section 7.2.4.3.1). Furthermore, the analysis of the BNC data has shown that even in closed corpora SPOK may not have a striking impact on the distribution of pronoun case forms, and particularly not as predicted by much of the literature (cf. Sections 8.1 and 8.2).48

---

48 In addition to the factor SPOK, the following interactions were eliminated in the model-building process: SPOK*FIRST, SPOK*SINGULAR, SPOK*CLEFT, SPOK*FOCUS, CMC*FIRST, CMC*SINGULAR, CMC*CLEFT, CMC*FOCUS, FIRST*CLEFT, FIRST*FOCUS, SINGULAR*CLEFT, SINGULAR*FOCUS.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-Value</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>3.84</td>
<td>***</td>
<td>46.49</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.41</td>
<td>***</td>
<td>0.25</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>0.88</td>
<td>**</td>
<td>2.40</td>
</tr>
<tr>
<td>CMC</td>
<td>-2.35</td>
<td>***</td>
<td>0.10</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.16</td>
<td>+</td>
<td>3.18</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>-2.19</td>
<td>***</td>
<td>0.11</td>
</tr>
</tbody>
</table>

| MODEL $\chi^2$   | 551.84 (***)|
| R$^2$             | 0.59        |
| % CORRECTLY       | 88.79       |
| PREDICTED         |             |
| % BASELINE        | 79.78       |
| N                 | 1177        |

+ significant at $p<0.1$, * significant at $p<0.05$, ** significant at $p<0.01$, *** significant at $p<0.001$

Table 37: Subject Predicative Complements in the .au data: logistic regression results

Table 37 further demonstrates that CLEFT is not only very highly statistically significant, but that it is also again the factor with the highest odds ratio retained in the minimal-adequate model. If we translate its odds ratio of 46.69 into a percentage value to make the impact of this factor more transparent, we can note that the likelihood of observing a subject form in a subject predicative complement is 45.69 per cent higher in an it-cleft than in an it BE sentence in the .au data. This observation is very much in line with the findings obtained so far for the other datasets and it further corroborates accounts arguing for pronounced differences in terms of pronoun case distribution between it-clefs and it BE sentences (e.g. Biber et al. 1999: 335-336; Erdmann 1978: 75-78). At the same time, this finding again undermines studies trying to explain the distribution of pronoun case forms in terms of positional constraints or in terms of weak versus strong pronoun classes (e.g. Emonds 1986: 96-100; Harris 1981: 18-19; Quinn 2005a).

The factor FIRST is also again identified as very highly significant in the present regression model. In analogy to the COCA and the .uk data (cf. Sections 9.1 and 10.1), FIRST has a clearly constraining effect on the use of subject pronouns in subject predicative complements in the .au data, which is signalled by the factor’s odds ratio of 0.25. On the one hand, this observation contrasts again with those accounts assuming a higher share of first person than third person subject pronoun forms in it BE sentences (e.g. Quinn 2005a: 246).
On the other hand, this is in line with studies attesting a higher share of third person subject pronoun case forms in it-cLEFTs (e.g. Quinn 2009: 42; Wales 1996: 95-96). Furthermore, the same effect of this variable has also been attested for the COCA and the .uk data and to some extent also for the BNC data. As was the case before (e.g. Section 8.3 or 9.3), this distinction can be explained from a rather pragmatic point of view with the help of the differences between the prototypical referential uses of first and third person pronouns (cf. also Sections 6.1.2.2 and 9.1). In a nutshell, this study has proposed that speakers or writers may deem it more necessary to put additional emphasis on the referents of these third person pronouns by means of a subject pronoun than on the referents of first person pronouns (cf. Sections 8.1–10.4). This possible explanation is not only in line with the data in Table 37 but also with this study’s assumption according to which subject pronoun case forms may be used as Focus markers in subject predicative complements (cf. Section 5.2; H3).

As Table 37 also shows, SINGULAR is highly significant and, as its odds ratio of 2.40 demonstrates, it markedly promotes the use of subject pronoun case forms in subject predicative complements in the .au data. As was the case in Section 10.1, this observation meets the expectations of some accounts postulating a clear preference of subject forms to occur in the singular (Sobin 1997: 334) and it also complies with the findings of the .uk data, where SINGULAR has a very similar effect. It is, however, neither compatible with the findings made for the BNC, where this factor has no significant effect (cf. Section 8.1), nor with the findings of the COCA, where SINGULAR even significantly inhibits the use of subject pronoun case forms in subject predicative complements (cf. Section 9.1). Hence, this factor obviously does not lend itself as easily to ascertain preliminary cross-varietal trends as other factors do, but SINGULAR seems to be more prone to inter-varietal variation.

The next factor identified as very highly significant is CMC. As can be seen from Table 37, CMC clearly inhibits the use of subject pronoun case forms, which becomes apparent from its odds ratio of 0.10. A possible explanation for why this factor constrains the use of subject pronoun case forms may be the differences in formality and the implications arising from them (cf. also Sections 10.1–10.3). Computer-mediated communication does not only exhibit many traits of spoken discourse (e.g. Crystal 2011: 32; Herring 2010a), but also imitates spoken discourse by trying to represent prosody and intonation units with the help of certain typographical practices, which is possible due to fewer text conventions and
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

stylistic restrictions (e.g. Herring 2010a). Furthermore, both the imitation of spoken discourse and the fewer text conventions are also responsible for the fact that CMC can employ more strategies than traditional written data to highlight the Focus of a clause. This, in turn, may also partly explain why subject pronoun forms are less likely to occur in subject predicative complements in CMC than in traditional written data (cf. Section 10.1). These alternative Focus marking strategies can again be illustrated with the help of examples taken from the CMC subset of the .au data, in which capitalisation is used to highlight certain pronouns as particularly prominent (cf. (141)):

(141) Capitalisation as Focus Marking Strategy in CMC in the. au Data
    a. BTW in our household it is ME who does the gardening. (.au/it is me/14.07.2008)
    b. Confuses independent views with individual ones. A rebel? No, to lazy... No admiration for You Alex... So it is HIM. (.au/it is him/14.07.2008)

Hence, the smaller share of subject forms in CMC can also be partly attributed to the fact that this mode of discourse has simply more options to highlight a pragmatically prominent constituent than traditional written data. This possible explanation is not only consistent with the data but also with this study’s assumption that subject pronoun case forms may have been re-functionalised in subject predicative complements as Focus markers (cf. Section 5.2; H3).

The fourth factor retained in the minimal-adequate regression model is FOCUS. In contrast to the previous datasets (cf. sections 8.1, 9.1 and 10.1), FOCUS is only marginally significant in the .au data (cf. Table 37). However, as its odds ratio of 3.18 illustrates, this factor – at least in tendency – clearly promotes the use of subject pronoun case forms in subject predicative complements. This promoting effect or tendency of this factor accords well with both the findings obtained from the other datasets analysed so far as well as with this study’s assumption that Focus marking plays a vital role in the distribution of pronoun case forms in subject predicative complements (cf. Chapter 4). Whether this factor will pass another level of significance in the subsequent analysis of the it BE sentences, or whether it remains marginally significant will be interesting to see in the following section (cf. 11.2).

Last but not least, Table 37 identifies also one interaction as very highly significant. As its odds ratio of 0.11 illustrates, the interaction between FIRST and SINGULAR markedly constrains the use of subject pronoun case forms in subject predicative complements in the .au data. As discussed before (e.g. Sections 10.1 and 10.3), this interaction can, however, be mainly attributed to the shortcomings of Google as far as its potential for the compilation of
Web-derived databases is concerned (cf. Section 7.2.2). The assumption that this interaction is due to the particularities of the data compilation employed for this study is further corroborated by the fact that the same interaction with a similarly strong inhibiting effect has already been attested for the .uk data (cf. Section 10.1). Indeed, due to the systematic nature of this effect, it is even likely that this interaction will also be significant in other Web-derived datasets still to come.

Turning from the individual factors to the overall model, Table 37 indicates that the minimal-adequate model applied to the .au data is very highly significant ($\chi^2 = 551.84$, $p<0.001$). Furthermore, the regression model is able to account for 59 per cent of the variation observed in the dependent variable, as the model’s Nagelkerke’s $R^2$ value of 0.59 indicates. With regard to the model’s predictive power, Table 37 also shows that the applied regression model clearly improves the baseline model. While the minimal-adequate model used here is able to predict the outcome of the dependent variable correctly in 88.79 per cent of all cases, the baseline model can only account for 79.78 per cent of all tokens. Thus, the statistical model used here improves the baseline model by more than nine per cent.

To conclude, the results obtained for the superordinate category of subject predicative complements in the .au dataset confirm some of the major findings made in the other databases analysed so far. As in all other corresponding datasets (cf. Sections 8.1, 9.1 and 10.1), CLEFT clearly promotes the use of subject pronouns and the same is true, at least in tendency, for FOCUS, which is, however, only marginally significant in the present analysis. In view of the impact of these syntactic and/or pragmatic factors, this study’s assumption according to which subject pronoun case forms may have been reanalysed as Focus markers is further corroborated, while accounts trying to account for pronoun case distribution only in terms of position or pronoun class membership are again severely challenged (e.g. Burridge 2004: 1118; Harris 1981: 18-21). As was the case before (e.g. Sections 9.1 and 10.1), the results for the variables PERSON and NUMBER are less uniform. With regard to PERSON, the factor FIRST clearly constrains the use of subject pronoun case forms in subject predicative complements, which is accordance with the findings of the COCA and the .uk data, and partly also with those of the BNC (cf. Sections 8.1, 9.1 and 10.1). Again, this inhibiting effect can be attributed to the differences between first and third person pronouns in terms of their prototypical referential statuses (cf. Section 6.1.2.2). With regard to NUMBER, the factor SINGULAR clearly promotes the use of subject forms in this dataset. This is in line with the
findings of the .uk data and the expectations of some accounts (e.g. Sobin 1997: 334), but neither with the COCA nor with the BNC findings, which makes robust generalisations for SINGULAR very difficult. As far as the MODE OF DISCOURSE variable is concerned, SPOK is again not significant, as in the preceding Web-derived dataset (cf. Section 10.1). Also in accordance with the .uk data is the effect of CMC which clearly inhibits the use of subject pronoun forms in subject predicative complements. As before (cf. Section 10.1), this inhibiting effect can be attributed to the greater expressive repertoire arising from fewer and more relaxed text conventions. Hence, CMC has more options to mark a pronoun as pragmatically prominent, which may also partly account for why CMC is less likely to use subject pronouns in subject predicative complements.

11.2 The Distribution of Pronoun Case Forms in it BE Sentences in the Australian (.au) Internet Data

In the .au dataset, 527 tokens of the 1177 subject predicative complements belong to the subcategory of it BE sentences. Table 38 demonstrates that only 2.09 per cent of all it BE sentences in the .au data are used with a subject pronoun form as pronominal complement, whereas the remaining 97.91 per cent are used with an object pronoun case form. Thus, in analogy to all datasets analysed so far, the share of subject forms in the it BE sentences is considerably lower than in the superordinate category of subject predicative complements, where the share of subject forms accounts for 20.22 per cent of all tokens in the .au data (cf. Section 11.1). This modest share of only 2.09 per cent subject pronoun forms observed for the Australian it BE sentences is indeed the smallest one, when compared to all other datasets analysed so far (cf. Sections 8.2–10.2).

<table>
<thead>
<tr>
<th>IT BE SENTENCES (.au)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td>527 (100%)</td>
<td>11 (2.09%)</td>
<td>516 (97.91%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 38: it BE Sentences in the .au Data: The Token Numbers

Table 39 below illustrates the results obtained for the multivariate analysis of the it BE sentences in the .au data. What is immediately striking is the fact that only one factor is
retained in the minimal-adequate model as significantly influencing the distribution of pronoun case forms in the *it BE* sentences of the .au data (cf. Table 39). This is in contrast to all corresponding *it BE* sentence datasets analysed so far, where a minimum of two factors has been identified as – at least marginally – significant (cf. Sections 8.2–10.2).

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOCUS</td>
<td>2.19</td>
<td>***</td>
<td>8.91</td>
</tr>
<tr>
<td>MODEL $\chi^2$</td>
<td>10.14 (**)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td>97.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% BASELINE</td>
<td>97.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>527</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

*Table 39: it BE Sentences in the .au Data: Logistic Regression Results*

Since only one factor is identified as significant by the applied regression model, all other factors are eliminated in the course of the model-building process.

To begin with, the factors FIRST and SINGULAR operationalising the variables PERSON and NUMBER are excluded from the final regression model. The elimination of FIRST is, on the one hand, in line with the observations made for the BNC and the .uk data but, on the other hand, also is in contrast to the COCA data, where this factor markedly constrains the use of subject forms. Furthermore, the non-significance of this factor is not in accordance with accounts proposing that “[n]on-1sg nominatives would seem to be extremely rare, and restricted to certain discourse contexts” (Quinn 2005a: 246) in *it BE* sentences.

The factor SINGULAR is also eliminated in the model-building process. The exclusion of SINGULAR complies with the findings made for the BNC and the COCA, where this factor is also not significant in the subclass of *it BE* sentences (cf. Sections 8.2 and 9.2). It is, however, neither consistent with the results obtained for the .uk data, where SINGULAR clearly promotes the use of subject pronoun case forms, nor with studies ruling out that plural subject forms can actually occur in *it BE* sentences and thus expect a promoting effect of this factor (Sobin 1997: 334).
In addition to these rather grammatical factors, the socio-linguistic factors SPOK and CMC are likewise eliminated in the model-building process. While the elimination of SPOK is consistent with the results obtained for the COCA and the .uk data, it is in contrast to both the BNC results, where SPOK in tendency even promotes the use of subject pronoun forms in *it* *BE* sentences, and also to the expectations of much of the current literature, assuming an inhibiting effect of this factor on the use of subject pronoun case forms in *it* *BE* sentences (e.g. Quirk et al. 1985: 337-338). In a similar vein, CMC is also identified as being non-significant and consequently eliminated, which is in contrast to the findings made for the .uk data. The non-significance of CMC does also not fully correspond to expectations voiced in the literature assuming that less formal varieties are less likely to exhibit subject forms in *it* *BE* sentences (e.g. Harris 1981: 18-19; Quirk et al. 1985: 337-338), especially not when we take into consideration that CMC exhibits and even imitates many features of spoken discourse (e.g. Herring 2010a).49

As Table 39 indicates, the only factor that is identified by the minimal-adequate regression model as significantly influencing the distribution of pronoun case forms in the .au data is Focus. It is very highly significant and emphatically promotes the use of subject pronoun case forms in the Australian *it* *BE* sentences (cf. Table 39). Its odds ratio of 8.91 indicates that the probability of observing a subject pronoun form is 791 per cent higher in those *it* *BE* sentences that have been identified as particularly focussed than in normal *it* *BE* sentences (cf. Section 6.1.2.5). This finding is consistent with the observations made for the BNC, COCA and .uk data, where Focus is also identified as very highly significant and clearly promoting the use of subject pronoun case forms (cf. Sections 8.2–10.2). Furthermore, the strong effect Focus exerts on the distribution of pronoun case forms in the Australian *it* *BE* sentences evidently corroborates two of this study’s major assumptions, namely that functional factors in general and Focus in particular play an enormous role in the distribution of pronoun case forms (cf. Section 5.2; H2 and H3).

The minimal-adequate model itself is highly significant ($\chi^2 = 10.14$, $p<0.01$). Although only one factor is retained, the regression model applied to the *it* *BE* sentences in the .au data is still able to account for ten per cent of the variation observed in the dependent variable, as the model’s Nagelkerke’s $R^2$ value of 0.10 indicates. Although this share is rather

49 In addition, the following interactions were eliminated to arrive at the minimal-adequate model: $\text{SPOK} \times \text{FIRST}$, $\text{SPOK} \times \text{SINGULAR}$, $\text{SPOK} \times \text{FOCUS}$, $\text{CMC} \times \text{FIRST}$, $\text{CMC} \times \text{SINGULAR}$, $\text{CMC} \times \text{FOCUS}$, $\text{FIRST} \times \text{SINGULAR}$, $\text{FIRST} \times \text{FOCUS}$, $\text{SINGULAR} \times \text{FOCUS}$. 
modest when we compare it to that obtained for the superordinate subject predicative complement category in the preceding section (cf. 11.1), the share of variance explained by the present model is still higher than those observed for the corresponding models of the BNC and the COCA (cf. Sections 8.2 and 9.2). Moreover, the share of variance explained by this model is also enough to render it substantially significant (e.g. Szmrecsanyi 2006: 55). As far as the predictive power of the applied regression model is concerned, Table 39 shows that it is not able to substantially improve the baseline model, since both of which can correctly predict the outcome of the dependent variable in 97.91 per cent of all cases.

To sum up, the multivariate analysis of the *it BE* sentences in the .au data shows that the only significant factor retained in the minimal-adequate regression model is FOCUS, which markedly promotes the use of subject pronoun forms in this context. Although this is in line with the observations made so far for the other databases, where FOCUS also plays an eminent role (cf. Sections 8.2, 9.2 and 10.2), it is still remarkable that the only factor that evidently influences the distribution of pronoun case forms in this dataset has so far been largely neglected by the literature trying to account for the distribution of pronoun case forms in subject predicative complements. Indeed, all variables that are deemed relevant by the literature, such as NUMBER, PERSON, or MODE OF DISCOURSE, do not play a role in the present dataset. What is even more remarkable is the fact that FOCUS is even the only factor that has consistently been identified as significant in each of the *it BE* sentence datasets analysed so far (cf. Sections 8.2, 9.2 and 10.2). This corroborates not only the assumption that this factor plays an important role in the distribution of pronoun case forms, but also this study’s hypothesis that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements (cf. Chapter 4).

**11.3 The Distribution of Pronoun Case Forms in *it*-Clefts in the Australian (.au) Internet Data**

Table 40 shows that 650 tokens of the 1177 subject predicative complements in the .au data are *it*-clefts. As Table 40 also illustrates, 34.92 per cent of these *it*-clefts exhibit a subject form as focal pronoun while the remaining 65.08 per cent are used with an object pronoun form as focal pronoun. Although this share of subject pronoun forms attested in the .au data
may appear modest when compared to those accounted for the BNC and the COCA, both of which are more than twice as high (cf. Sections 8.3 and 9.3), the proportion of subject pronoun forms in the .au data is very similar to that observed in the .uk data in Section 10.3. Indeed, the difference is only very marginal between these two Web-derived datasets. However, as stated before (cf. Section 10.1), a very important reason for why the share of subject pronoun forms in the Web-derived databases are considerably lower than in the closed corpora are the particularities of the data compilation by means of a commercial Web crawler leading to a high number of false positives (cf. Section 7.2.4.2). If we compare the proportions of subject and object forms observed for the it-clefts in this section to that attested for the it BE sentences in the preceding Section 11.3, we note again that there are very marked differences between it-clefts and it BE sentences with regard to the distribution of pronoun case forms (cf. also Sections 8.3, 9.3 and 10.3). The observation of these marked differences is in line with the findings made for the other datasets examined so far and corroborates studies arguing for these marked differences between the two sentence types (e.g. Erdmann 1978; Maier 2013).

<table>
<thead>
<tr>
<th>IT-CLEFTS (.au)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>650 (100%)</td>
<td>227 (34.92%)</td>
<td>423 (65.08%)</td>
</tr>
</tbody>
</table>

Table 40: it-Clefts in the .au Data: The Token Numbers

Table 41 shows the results obtained for the multivariate analysis of the it-clefts in the .au data. As is illustrated by Table 41, four main effects and two interactions are identified in the minimal-adequate regression model as significantly influencing the distribution of pronoun case forms in the Australian it-cleft data. The only factor that is eliminated to arrive at the minimal-adequate model is SPOK. Although many accounts assume a constraining effect of this factor on the likelihood of observing subject pronoun forms (e.g. Huddleston and Pullum 2002: 459; Wales 1996: 91-108), the elimination of SPOK corresponds to the results for the it-cleft data observed in both the BNC and the .uk data, where SPOK is also not significant. Furthermore, the non-significance of SPOK can be partly attributed again to the generally rather formal character of much of the spoken data obtained from the Web, which
may level potential differences between spoken and traditional written data (cf. Sections 8.3 and 10.3).\textsuperscript{50}

\begin{table}
\centering
\begin{tabular}{|l|l|l|l|}
\hline
\textbf{Variable} & \textbf{Coefficient} & \textbf{P-Value} & \textbf{Odds Ratio} \\
\hline
\textit{As\_Subj} & 3.62 & *** & 37.23 \\
\textit{First} & -1.81 & *** & 0.16 \\
\textit{Singular} & 1.25 & *** & 3.49 \\
\textit{CMC} & -2.67 & *** & 0.07 \\
\textit{First}\_*\textit{Singular} & -2.48 & *** & 0.08 \\
\textit{CMC}\_*\textit{First} & 0.85 & + & 2.35 \\
\hline
\textbf{Model} $\chi^2$ & 332.16 (***) & & \\
\textbf{R}$^2$ & 0.55 & & \\
\% Correctly Predicted & 83.85 & & \\
\% Baseline & 65.08 & & \\
\textbf{N} & 650 & & \\
\hline
\end{tabular}
\caption{it-Clefts in the .au Data: Logistic Regression Results}
\end{table}

The first factor retained in the minimal-adequate regression model is \textit{As\_Subj} (cf. Table 41). \textit{As\_Subj} is not only very highly significant but also exerts a very strong effect on the distribution of pronoun case forms in the Australian cleft data, which is clearly indicated by its odds ratio of 37.23. Translating this ratio into a percentage value, we can note that the likelihood of observing a subject pronoun case form is 3623 per cent higher in \textit{it}-clefts in which the focal pronoun is co-referential with the subject of the dependent clause than in \textit{it}-clefts in which the focal pronoun is co-referential with an object of the dependent clause. This observation is in accordance with the relevant literature, which also assumes or even attests a promoting effect of this factor on the use of subject pronoun forms in \textit{it}-clefts (e.g. Erdmann 1978: 75-78; Huddleston and Pullum 2002: 459). Moreover, this finding is fully consistent with the analyses of the \textit{it}-cleft data in the preceding chapters, where \textit{As\_Subj} has also been identified as very highly significant and as clearly promoting the use of subject forms (cf. Sections 8.3, 9.3 and 10.3). Finally, the effect attested for this factor further corroborates the trend observed so far that functional factors, particularly syntactic and

\textsuperscript{50} The following interactions were likewise eliminated to arrive at the minimal-adequate model: \textit{Spok}\_*\textit{First}, \textit{Spok}\_*\textit{Singular}, \textit{Spok}\_*\textit{As\_Subj}, \textit{CMC}\_*\textit{Singular}, \textit{CMC}\_*\textit{As\_Subj}, \textit{First}\_*\textit{As\_Subj}, \textit{Singular}\_*\textit{As\_Subj}.
discourse-pragmatic ones, play a very important role in the distribution of pronoun case forms in subject predicative complements.

The next factor identified in Table 41 as very highly significant is FIRST. With an odds ratio of 0.16, FIRST markedly constrains the use of subject pronoun forms in it-clefts. This result conforms to both the observations of earlier studies (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96) and to the findings made in all it-cleft datasets analysed so far, where FIRST also clearly inhibits the use of subject pronoun case forms (cf. Sections 8.3, 9.3, 10.3). The underlying reason for why first person pronouns may be less likely to surface in their subject form may again be explained in terms of their prototypical referential status. Since first person pronouns can be considered in general as more salient than third person pronouns, third person pronouns may be more likely to receive additional Focus marking by means of a subject form (cf. e.g. Sections 8.3 and 9.1). Moreover, if FIRST and CLEFT coincide, this may be deemed enough to mark a focal pronoun as the relatively most important constituent or Focus of a clause (cf. Sections 4.1 and 9.3). This possible explanation can account for why first person pronouns are particularly less likely to occur in their subject form in it-clefts, a trend which has so far been attested in four different databases (cf. 8.3, 9.3 and 10.3), and it would also perfectly comply with this study’s assumption according to which subject pronoun forms may have been reanalysed as Focus markers (cf. Chapter 4).

Table 41 shows that SINGULAR is also very highly significant in the present regression model. With regard to its actual effect, the odds ratio of 3.49 indicates that SINGULAR evidently promotes the use of subject forms in the it-clefts of the .au data. This finding conforms again with the expectations of some studies (Sobin 1997: 334) as well as with the observations made for the .uk data, where SINGULAR also clearly promotes the use of subject pronoun case forms, particularly in the cleft data (cf. Section 10.3). At the same time, however, this observation conflicts with the findings made for the BNC, where this factor is not significant, and even more so with those obtained for the COCA, where SINGULAR even constrains the use of subject forms in it-clefts. Thus, as noted before, SINGULAR seems to be much more susceptible to cross-varietal variation than other factors analysed in the present study, such as As_SUBJ or FIRST, which seem to behave more uniformly across different datasets.

The fourth very highly significant factor retained in Table 41 is CMC, which, as its odds ratio of 0.07 demonstrates, severely inhibits the use of subject pronoun case forms in
the Australian *it*-cleft data. This is line with the findings obtained for the corresponding analysis of the .uk data in Section 10.3 and also with the observations made for the superordinate category of subject predicative complements in this chapter (cf. Section 11.1). Furthermore, this observation is consistent with those accounts assuming that more formal varieties of English are more likely to exhibit subject pronoun case forms in subject predicative complements (e.g. Quirk et al. 1985: 337-338; Huddleston and Pullum 2002: 459). However, the smaller share of subject pronoun case forms can also be attributed to the greater repertoire of expressive options that arise from the lesser degree of codification of computer-mediated discourse and the fewer text conventions this mode of discourse is bound to. As illustrated and discussed in Sections 10.1 and 10.2, this means in particular that CMC can employ more strategies to highlight the importance of a clausal constituent than traditional written data, such as the deliberate use of capital letters to highlight a pronoun as pragmatically prominent or as the Focus of a clause (cf. (141) in Section 11.1).

In addition, two interactions are retained in the minimal-adequate model depicted in Table 41. The first one is again that between First and Singular, which is very highly significant and markedly inhibits the use of subject pronoun case forms in *it*-clefts in the .au data, which is indicated by its odds ratio of 0.08. This is in line with the corresponding finding in the .uk data, where more or less the same has been observed for the interaction between First and Singular (cf. Section 10.3). The linguistic significance of this interaction remains, however, questionable, since it can be attributed to the systematic data skewing resulting from the use of commercial search engines for the compilation of the Web-derived data (cf. Sections 7.2.2, 10.1 and 11.1). Due to the systematic nature of this skewing effect, it is even possible that this interaction will be identified as being significant in the analysis of the remaining Web-derived datasets as well.

The second interaction retained in the applied regression model is that between CMC and First. This interaction is only marginally significant but in tendency promotes the use of subject pronoun case forms, as its odds ratio of 2.35 indicates. Thus, while both factors by themselves markedly inhibit the occurrence of subject pronoun case forms in *it*-clefts, their co-occurrence seems to reverse the effects of the single variables and – at least in tendency – rather promotes the use of subject pronoun case forms as focal pronouns in *it*-clefts in the .au data.
The regression model is itself very highly statistically significant ($\chi^2 = 332.16$, $p<0.001$), which indicates a strong correlation between the dependent and the independent variables in the minimal-adequate model used to analyse the .au clefts (cf. Table 41). As far as the share of variance explained by it is concerned, the model’s Nagelkerke’s $R^2$ value of 0.55 shows that the applied statistical model is able to account for 55 per cent of the variation observed in the dependent variable, which is more than enough to qualify the model as substantially significant. The applied regression model can also substantially improve the baseline model. While the latter is able to account for 65.08 per cent of all cases, the minimal-adequate model can even correctly predict the outcome of the dependent variable in 83.85 per cent of all cases.

In sum, the analysis of the it-clefts in the .au dataset confirms the results obtained for the other datasets examined so far as well as the expectations of the relevant literature (e.g. Erdmann 1978: 76-78; Quirk et al. 1985: 337-338) in that As_SUBJ markedly promotes the use of subject pronouns. This finding is, of course, also in line with the more general observation made so far that functional factors, particularly syntactic and discourse-pragmatic ones, markedly influence the distribution of pronoun case forms (cf. Section 5.2; H2 and H3). The results obtained for FIRST, i.e. its significantly inhibiting effect, are also in accordance with the results obtained in earlier sections (cf. 9.3 and 10.3) and much of the relevant literature (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96). The promoting effect of SINGULAR, however, is again more difficult to account for. While the observed effect corresponds to the results of the .uk data and the expectations of some accounts (e.g. Sobin 1997: 334), it does not correspond to the findings of the BNC and even contradicts the findings of the COCA, where SINGULAR has an adverse effect (cf. Sections 8.3 and 9.3). In analogy to the .uk data, the analysis also shows a clearly inhibiting effect of CMC in the present dataset. This, however, has been explained by combining formal considerations with the Focus-oriented perspective of this study assuming that the lower degree of formality of CMC also implies more options to mark pragmatically salient entities in a clause.
11.4 The Distribution of Pronoun Case Forms in the Australian (.au) Internet Data: Interim Summary

The analysis of the .au data has confirmed many of the trends that have emerged from the analyses of the preceding databases, i.e. the BNC, the COCA and the .uk data. Once again, the analysis of the .au data has exposed the vast differences between *it*-cLEFTs and *it* *BE* sentences in terms of pronoun case distribution, which does not only further legitimise the analysis of each sentence type separately, but also is difficult to account for by means of positional considerations or a split in the pronominal paradigm (cf. Section 5.2; H1).

These accounts trying to explain pronoun case distribution solely by the pronoun’s position or a certain pronoun class membership are further challenged by the results of this chapter which again show the strong influence of functional factors on the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H2). Particularly the impact of the factors CLEFT, FOCUS and AS_SUBJ on the distribution of pronoun case forms, confirming the findings made for all datasets examined so far, is very difficult to reconcile with the positional approach or the weak vs. strong pronoun class approach (cf. Sections 3.2 and 3.3).

The results for the .au data are also in accordance with the results of the previous three datasets in that CLEFT, FOCUS and AS_SUBJ emphatically promote the use of subject pronoun case forms in subject predicative complements (cf. Chapters 8–10). Particularly with regard to CLEFT and FOCUS, this means that both necessary and sufficient condition for us to assume that Focus has a strong influence on the distribution of pronoun case forms are met (cf. Section 6.1.2.5). The importance of the factor FOCUS becomes further evident by the fact that it is the only factor retained in the minimal-adequate model analysing the distribution of pronoun case forms in the *it* *BE* sentences in the .au data. This strongly corroborates this study’s assumption that subject pronoun case forms may have been reanalysed as Focus markers in the analysed contexts (cf. Chapter 4).

In contrast to CLEFT, FOCUS and AS_SUBJ, the results for the rather grammatical factors FIRST and SINGULAR are less coherent across the different datasets. In the .au data, FIRST is only significant in the *it*-cLEFT and the superordinate subject predicative complement categories, where it markedly inhibits the use of subject forms in both datasets. This observation is largely consistent with the .uk data, with the BNC data as well as with parts of the COCA,
where FIRST even significantly inhibits the use of subject forms in the it BE data (cf. Chapters 8–10). The inhibiting effect of FIRST, particularly in it-clefts, can again be accounted for by means of the Focus-oriented perspective. If the pragmatically salient factors FIRST and CLEFT coincide (cf. Biber et al. 1999: 958-959; Siewierska 2004: 5-8), this may suffice to highlight a pronoun’s referent as the most important piece of information in a clause and Focus marking by means of subject pronouns is not required (cf. Sections 8.3 and 9.3). While the effect of FIRST is more or less coherent across the different datasets, SINGULAR has so far exhibited much more variation. In the .au data, SINGULAR promotes the use of subject forms in it-clefts and the superordinate subject predicative complement category, confirming the expectations of some theoretical accounts (Sobin 1997: 334) and also partly the findings of the .uk data (cf. Section 10.1–10.3). It does, however, conflict with the results of other studies (e.g. Quinn 2005a: 134-135) and the findings obtained for the BNC and the COCA, where SINGULAR is either non-significant or even inhibiting the use of subject forms (cf. Chapters 8 and 9). Thus, the effect of SINGULAR seems to be much more variety-specific than those of the other factors discussed so far.

The factor SPOK does not play a role in the .au data, which is in line with the observations made for the .uk data and which can be explained with the rather formal character of the spoken data obtained from the Web (cf. Section 7.2.4.3.1). CMC again significantly inhibits the use of subject forms, particularly in the Australian it-cleft data. The constraining effect of CMC can be accounted for by combining the concept of formality with the Focus-oriented perspective of this study (cf. Section 10.3). Thus, this finding is also consistent with this study’s hypothesis H3 that subject forms may have been reanalysed as Focus markers in subject predicative complements (cf. Section 5.2).
12 The Distribution of Pronoun Case Forms in the Irish (.ie) Internet Data

In this chapter, the distribution of pronoun case forms in the data obtained from the Irish Internet top-level domain .ie (henceforth .ie data) will be examined. This chapter is structured as the preceding ones. Thus, Section 12.1 will analyse and discuss the distribution of pronoun case forms in the superordinate subject predicative complement category. In a second step, Section 12.2 will focus on the distribution of pronoun case forms in *it* BE sentences before Section 12.3 examines and discusses the *it*-clefs in the .ie data. Finally, Section 12.4 complements this chapter with a brief interim summary of the major results.

12.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the Irish (.ie) Internet Data

Table 42 indicates that 425 tokens remain after cleaning, verifying and classifying the raw data obtained from the Irish .ie top-level domain with the help of Google (cf. Chapter 7). This number of tokens may seem low when compared to the token numbers obtained from the closed corpora or even when compared to the preceding two Web-derived datasets, both of which also contain more than 1000 tokens. And indeed, the .ie dataset is the smallest one examined in this study. Yet, it must be taken into consideration that each Web-derived dataset is subject to severe limitations posed by the commercial search engine, which, for example, only displays the first 1000 hits for a given search string. In addition, many of these maximally 1000 tokens retrieved per query may be junk or repetitions, which subsequently have to be filtered out (cf. Sections 7.2.4.2 and 7.2.4.3). Furthermore, the Web-queries for the present study were only conducted for one form of *BE*, i.e. *is*, since this form yielded the best results in order to obtain as many potential subject predicative complements from the Web data as possible (cf. Section 7.2.4.1). Finally, in view of the fact that even a 100 million word corpus such as the BNC only contains 201 tokens of subject predicative complements with *is* as their finite form of *BE*, the 425 tokens contained in the .ie dataset constitute a considerable amount of data and also roughly suggest how big a closed corpora of Irish
English would have to be to contain a comparable number of tokens (cf. Sections 7.2.3 and 7.4).

<table>
<thead>
<tr>
<th>Subject Predicative Complements (.ie)</th>
<th>TOTAL N</th>
<th>Subject Forms N</th>
<th>Object Forms N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>425 (100%)</td>
<td>299 (70.35%)</td>
<td>126 (29.65%)</td>
</tr>
</tbody>
</table>

*Table 42: Subject Predicative Complements in the .ie Data: The Token Numbers*

However, the .ie dataset is not only remarkable because of its size but also due to its composition in terms of subject and object forms, when compared to all other datasets examined so far. Table 42 illustrates that 70.35 per cent of the 425 subject predicative complements in the .ie data exhibit a subject form as pronominal complement, while the remaining 29.65 per cent are used with an object pronoun form. This means that the .ie dataset is not only the first of the databases analysed so far which contains more subject than object pronoun forms in the superordinate subject predicative complement category, but it also surpasses the share of subject pronoun case forms attested in the BNC, the database with the thitherto highest share of subject pronoun case forms by more than 24 per cent (cf. Section 8.1). The shares of subject forms in the corresponding datasets for the .uk and the .au data even only account for 21.60 and 20.72 per cent respectively (cf. Sections 10.1 and 11.1). This high proportion of subject forms in the .ie data becomes even more remarkable when considering the fact that search engines ignore punctuation marks, which leads to a high number of false positives particularly affecting subject pronoun case forms (cf. Section 7.2.2). As a consequence, a considerably lower share of subject forms in subject predicative complements can be expected for Web-derived datasets when compared to closed corpora, which has so far been more or less confirmed by the respective shares observed for the BNC and the .uk data (cf. Sections 8.1 and 10.1). However, the actual shares of subject and object pronoun case forms are also at least co-determined by the proportions of *it BE* sentences and *it*-clefts contained in the total sample of subject predicative complements (cf. Sections 9.1 and 9.3 for example). Thus, only the subsequent analyses will reveal to what extent the .ie data is really more likely to use subject forms in subject predicative complements than the other datasets analysed so far.

Table 43 shows the results obtained for the multivariate analysis of the subject predicative complements of the .ie data. While four factors are retained, two others have
been eliminated to arrive at the minimal-adequate regression model, the results of which are given in Table 43. The first non-significant factor excluded from the minimal-adequate model is SINGULAR. While this finding matches the observations made in the BNC (cf. Section 8.1), it is not in line with the findings of the COCA, where SINGULAR inhibits the use of subject pronoun forms (cf. Section 9.1), nor with the results of the .uk and .au data, where SINGULAR promotes the use of subject forms (cf. Sections 10.1 and 11.1), nor with accounts assuming such a fostering effect of this factor on the use of subject pronoun case forms (cf. Sobin 1997: 334). However, the non-significance of SINGULAR further corroborates the assumption that the effect of this factor may be strongly variety- or at least data-specific, as the results obtained so far indicate.

The other factor eliminated in the model-building process is SPOK. Considering the fact that SPOK is not significantly influencing the distribution of pronoun case forms in the .uk and .au data, the exclusion of this factor is also not really surprising. Although the non-significance of SPOK in the Web-derived databases can be partly attributed to the on average rather formal nature of the spoken data obtained from the Web, it still has to be borne in mind that this factor – in contrast to the predictions of the literature (e.g. Harris 1981: 18-19; Quirk et al. 1985: 337-338) – only inhibits the use of subject forms in the COCA (cf. Section 9.1). In the BNC, however, this factor is only marginally significant and in tendency even promotes the use of subject forms. Thus, a general trend suggesting that spoken data generally disfavour subject pronoun forms in subject predicative complements cannot be observed so far.\footnote{The following interactions were also eliminated to arrive at the minimal-adequate model: SPOK*FIRST, SPOK*SINGULAR, SPOK*CLEFT, SPOK*FOCUS, CMC*FIRST, CMC*SINGULAR, CMC*CLEFT, CMC*FOCUS, FIRST*SINGULAR, FIRST*CLEFT, FIRST*FOCUS, SINGULAR*CLEFT, SINGULAR*FOCUS.}
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>2.60</td>
<td>***</td>
<td>13.49</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.51</td>
<td>***</td>
<td>0.22</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.84</td>
<td>***</td>
<td>0.16</td>
</tr>
<tr>
<td>FOCUS</td>
<td>0.86</td>
<td>n.s.</td>
<td>2.37</td>
</tr>
<tr>
<td><strong>MODEL</strong></td>
<td><strong>X^2</strong></td>
<td><strong>182.16 (</strong>*)**</td>
<td><strong>R^2</strong></td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 43: Subject Predicative Complements in the .ie Data: Logistic Regression Results

Table 43 indicates that CLEFT is very highly significant and that it is also again the factor with the highest odds ratio, which amounts to 13.49 in the present dataset. This confirms the findings of the previous four chapters, in which we have observed a similar strong effect of CLEFT (cf. 8.1–11.1). Also in line with the previous observations is the direction of this factor’s effect, since the chances for observing a subject form are 1249 percent higher in an it-cleft than in an it BE sentence in the .ie data. While this finding supports accounts arguing for the distinctiveness of it-clefs and it BE sentences in terms of pronoun case distribution (e.g. Erdmann 1978: 75-78; Maier 2013), it challenges again accounts trying to boil down the distribution of pronoun case forms to rather simplistic generalisations in terms of position and/or pronoun class (e.g. Emonds 1986: 96-100; Harris 1981: 18-19; Quinn 2005a).

The second factor retained in the model depicted in Table 43 is FIRST. This factor is very highly significant and, as its odds ratio of 0.22 indicates, markedly inhibits the use of subject pronoun forms in subject predicative complements. This finding is in line with the observations made for the corresponding datasets of the COCA, the .uk and .au data and to some extent even the BNC (cf. Sections 8.1, 9.1, 10.1, and 11.1). In addition, this observation is also in accordance with studies reporting a higher share of third person subject pronoun forms in it-clefs (e.g. Quinn 2005a: 134-135; Wales 1996: 95-96), while it is at the same time in contrast to those accounts expecting a higher share of first person subject pronouns than third person ones in it BE sentences (e.g. Quinn 2005a: 246). In analogy to the foregoing
chapters, this difference in the distribution of pronoun case forms between first and third person pronouns can be explained with the differences regarding their prototypical referential functions. As outlined in the preceding sections (e.g. Sections 9.1, 10.1 and 11.1), speakers and writers may feel the need to additionally emphasise the importance of third person referents by means of a subject form, where this is admissible, thus exploiting the variability in these contexts (e.g. Wales 1996: 95). This Focus-oriented interpretation of the effect of FIRST is not only consistent with the observations made for the .ie dataset but also with one of this study’s central hypotheses assuming that Focus plays a prominent role in the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H3).

In addition to CLEFT and FIRST, CMC is also identified as very highly significant (cf. Table 43). Its odds ratio of 0.16 indicates that CMC markedly constrains the use of subject pronoun case forms in subject predicative complements. A very similar effect of this factor has already been observed for the preceding two Web-derived datasets and may again be explained with the differences between CMC and traditional written discourse in terms of their formal and orthographic conventions (cf. also Sections 10.1 and 10.2). Since CMC is subject to fewer text conventions (e.g. Herring 2010a), it can also employ more strategies than traditional written data to highlight a certain clausal constituent, for example by deliberately capitalising it (cf. Section 11.1). Assuming that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements, it could be argued that CMC uses less subject forms simply because it is less reliant on using them than traditional written data since CMC possesses alternative means to highlight the Focus of a clause (e.g. Section 10.1). While this possible explanation complies well with the hypotheses of this study (cf. Section 5.2), it is at the same time in line with those accounts assuming that subject pronoun forms are less likely to occur in less formal than in more formal varieties of English (e.g. Quirk et al. 1985: 337-338; Huddleston and Pullum 2002: 459). Hence, the reason for this difference between traditional written and CMC data is not so much that one mode of discourse is more formal than the other. Instead, it is perhaps more appropriate to say that CMC is simply more flexible than traditional written data since it has more options to mark pragmatically salient entities (cf. Sections 10.1 and 11.1).

The last factor retained in the regression model depicted in Table 43 is FOCUS. As can be seen from Table 43, FOCUS is not significant, but bootstrap validation tells us to retain this
factor in the minimal-adequate model (cf. Section 6.2). The reason for this result may be that the token number may be simply too small to render this factor significant (cf. Szmrecsanyi 2006: 102). The odds ratio given for this variable in Table 43 indicates the theoretically expected impact of Focus — had it passed a threshold of statistical significance (cf. Szmrecsanyi 2006: 102). Thus, at least in theory, this factor seems to promote the use of subject forms in subject predicative complements in the .ie dataset. Whether or not Focus will turn out significant in the analysis of the it BE sentences of the .ie data will be examined in the subsequent section (cf. Section 12.2).

As far as the applied statistical model as a whole is concerned, the minimal-adequate regression model is very highly statistically significant ($\chi^2 = 182.16$, $p<0.001$). Furthermore, the model can account for 49 per cent of the variation observed in the dependent variable, which is signalled by the model’s Nagelerke’s $R^2$ value of 0.49 in Table 43. The predictive power of the regression model used to analyse the subject predicative complement category is also fairly good. The applied model correctly predicts 83.53 per cent of all instances, which evidently improves the baseline model by 13.18 per cent (cf. Table 43).

The analysis of the .ie data has so far yielded very remarkable results. One of the most interesting observations made so far is that the share of subject pronoun forms in the .ie data is considerably higher than in all other datasets analysed so far. Although this may be partly due to a high share of it-clefts included in the data, it is still remarkable that the share of subject forms in the .ie data is more than thrice as high as in the .uk and .au data (cf. Sections 10.1 and 11.1). Furthermore, this high share of subject pronoun case forms is also very remarkable given the assumptions of positional and weak versus strong pronoun form accounts, which expect a clear preference of object pronoun forms in this context (cf. Sections 3.2 and 3.3). Also challenging for these accounts is again the strong impact of CLEFT, which clearly promotes the use of subject forms. FOCUS, however, is not significant in the present dataset, yet it is retained in the regression model. Again, FIRST markedly inhibits the use of subject pronoun case forms, which is consistent with the observations made for the .au and the .uk data, the COCA and to some extent even with the results of the BNC. This could again be attributed to the differences in the prototypical referential functions of first and third person pronouns and the resulting need to additionally mark third person as Focus by means of a subject form (cf. Sections 6.1.2.2 and 8.3). The factor SINGULAR is not significant, which corresponds to the findings made for the BNC but neither to the findings...
of all other datasets, where this factor either promotes or inhibits the use of subject forms, nor to some accounts expecting a promoting effect of this predictor (cf. Sobin 1997: 334). In analogy to the preceding two Web-derived datasets, CMC markedly inhibits the use of subject forms, which could again be explained by the fact that this mode of discourse possesses more options to highlight clausal constituents as pragmatically salient (cf. Section 11.1).

12.2 The Distribution of Pronoun Case Forms in *it BE* Sentences in the Irish (.ie) Internet Data

Table 44 indicates that only 87 tokens of the 425 subject predicative complements in the .ie dataset belong to the subclass of *it BE* sentences. However, not only this low token number is very remarkable but even more so the distribution of pronoun case forms in this data subset. Table 44 shows that 27.59 per cent of all *it BE* sentences in the .ie data are used with a subject pronoun form, while the remaining 72.41 per cent employ an object form as pronominal complement. This means that the share of subject forms in *it BE* sentences in the .ie data is by far the highest one thitherto observed, being for example more than three times as high as that of the BNC and even more than ten times as high as that of the .au data (cf. Sections 8.2 and 11.2). This is even more remarkable given the skewing effect this database as well as all other Web-derived datasets is subject to, which is caused by the disregard of punctuation marks by commercial search engines, and which results in a lower share of subject forms when compared to closed corpora (cf. Section 7.2.4.2). Thus, in view of this observed distribution, Irish *it BE* sentences seem more likely to exhibit subject pronoun case forms as pronominal complements than those observed in the British, American and Australian datasets discussed so far.

<table>
<thead>
<tr>
<th>IT BE SENTENCES (.ie)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>87 (100%)</td>
<td>24 (27.59%)</td>
<td>63 (72.41%)</td>
</tr>
</tbody>
</table>

*Table 44: *it BE* Sentences in the .ie Data: The Token Numbers*
With regard to the multivariate analysis of the *it BE* sentences, Table 45 indicates that only one factor is retained in the minimal-adequate model applied to this dataset. This conforms to the results obtained for the .au data but is in contrast to the other models analysing *it BE* sentences, in which a minimum of two factors is retained in the respective model-building processes (cf. Sections 8.2–11.2).

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOCUS</td>
<td>1.76</td>
<td>***</td>
<td>5.79</td>
</tr>
<tr>
<td>MODEL $\chi^2$</td>
<td></td>
<td></td>
<td>12.24 (***)</td>
</tr>
<tr>
<td>R$^2$</td>
<td></td>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td></td>
<td></td>
<td>75.86</td>
</tr>
<tr>
<td>% BASELINE</td>
<td></td>
<td></td>
<td>72.41</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>87</td>
</tr>
</tbody>
</table>

Table 45: *it BE* Sentences in the .ie Data: Logistic Regression Results

As in the corresponding .au analysis, the factors FIRST and SINGULAR are eliminated in the course of the model-building process. While the elimination of FIRST is in line with the BNC as well as the .au and .uk findings, it does not comply with the COCA, where this factor significantly inhibits the occurrence of subject forms and also contrasts with some accounts expecting a promoting effect of this factor (Quinn 2005a: 246). Furthermore, SINGULAR is also excluded from the minimal-adequate model, which is in accordance with the results obtained for the BNC, the COCA and the .au database, where this factor is also not significant. However, it is neither in accordance with the .uk data, where SINGULAR promotes the use of subject forms in *it BE* sentences, nor with accounts assuming a fostering effect of this factor on the use of subject forms (cf. Sobin 1997).

Furthermore, the factors SPOK and CMC are also not significant and thus excluded from the minimal-adequate model (cf. Table 45). The exclusion of SPOK is coherent with the findings obtained for all other *it BE* sentence datasets except for that of the BNC, where SPOK in tendency fosters the use of subject forms. The elimination of CMC is in line with the findings of the .au data but not with the results obtained for the .uk data. Moreover, the non-significance of these factors is in contrast to the many accounts assuming that less
formal varieties are less likely to use of subject forms (e.g. Huddleston and Pullum 2002: 459; Quirk et al. 337-338), which could, however, neither be attested for Spok nor for CMC in this dataset.\(^{52}\)

The only significant factor retained in the minimal-adequate model is again Focus, as is the case in the corresponding .au model (cf. Section 11.2). Its odds ratio of 5.79 signals that Focus emphatically promotes the use of subject pronoun case forms in it BE sentences. Converting this ratio into a percentage value, we can note that the chances for observing a subject form are 479 per cent higher in a particularly focussed than in a normal it BE sentence in the .ie data (cf. Section 6.1.2.5). This markedly fostering effect is in accordance with all corresponding datasets analysed so far, where a very similar impact of Focus can be observed (cf. Sections 8.2, 9.2, 10.2 and 11.2). In particular, this observation is in line with the findings obtained for the .au data, where Focus is also the only factor retained in the minimal model and also markedly fosters the use of subject forms. On a more general level, the impact and significance of Focus in the .ie data further corroborates this study’s hypothesis assuming that Focus plays a very important role in the distribution of pronoun case forms in the present context (cf. Chapter 4).

The statistical model as a whole used to analyse the it BE sentences in the .ie data is very highly statistically significant ($\chi^2 = 12.24$, $p<0.001$). With a Nagelkerke’s $R^2$ value of 0.18, the model can account for 18 per cent of the variation observed in the dependent variable. This is the second highest Nagelkerke’s $R^2$ value observed so far for the models particularly analysing the it BE sentences of their respective datasets. Furthermore, this is enough to render the model substantially significant (e.g. Szmrecsanyi 2006: 55). Table 45 further demonstrates that the applied regression model also improves the baseline model. While the latter can account for 72.41 per cent of all tokens, the minimal-adequate regression model even correctly predicts the outcome in 75.86 per cent of all cases.

In sum, the analysis of the it BE sentences in the .ie data has shown that the share of subject forms in the Irish subject predicative complement data in general and in it BE sentences in particular is substantially higher than those in the other databases analysed so far. Furthermore, the multivariate analysis of the data has identified Focus as the only factor significantly influencing the choice of pronoun case forms in it BE sentences in this dataset. This is in line with all other corresponding datasets examined so far where Focus also

\(^{52}\) The following interactions were also eliminated in the course of the model-building process: Spok*First, Spok*Singular, CMC*First, CMC*Singular, CMC*Focus, First*Singular, First*Focus, Singular*Focus.
markedly boosts the use of subject forms in *it BE* sentences. Furthermore, the fact that *Focus* is again identified as the only factor that actually significantly influences the distribution of pronoun case forms in *it BE* sentences further supports this study’s assumption according to which subject forms may have been reanalysed as Focus markers in this context (cf. Section 5.2).

12.3 The Distribution of Pronoun Case Forms in *it*-Clefts in the Irish (.ie) Internet Data

With regard to the second subgroup of subject predicative complements distinguished in this study, Table 46 indicates that 338 tokens of the 425 subject predicative complements obtained from the Web-based data compilation for Irish English are *it*-clefts. Table 46 also demonstrates that the vast majority of *it*-clefts in the .ie data, i.e. a share of 81.36 per cent, are used with a subject form as focal pronoun. Thus, only 18.64 per cent of the .ie *it*-clefts are used with an object form as focal pronoun. When comparing the share of subject forms in this subgroup to the share of subject forms observed for the *it BE* sentences, amounting to 27.59 per cent, the high share of subject forms attested for the *it*-clefts confirms the findings of the other datasets examined so far in that there are marked differences in the distribution of pronoun case forms between *it*-clefts and *it BE* sentences (cf. Chapters 8–11). Once again, this pronounced difference also corroborates accounts reporting marked differences between these two similar yet different sentence types in terms of pronoun case distribution (e.g. Erdmann 1978; Maier 2013). Furthermore, the high share of subject pronoun forms attested for the *it*-clefts in the Irish data is in accordance with the overall higher proportions of subject forms used in this dataset when compared to all other datasets examined so far (cf. Chapters 8–11). Indeed, the present dataset is that with the highest share of subject pronoun forms attested so far. As noted before (e.g. Section 12.1), to observe such a high share of subject pronoun forms in a Web-derived database is rather surprising not only in view of the many accounts expecting a high or even near-exclusive share of object forms in this context (cf. Sections 3.2 and 3.3), but also because of the systematic data skewing in Web-based datasets, which mainly affects subject pronoun forms (cf. Section 7.2.4.2). Thus, we actually expect to observe lower shares of subject forms in
Web-derived datasets than in closed corpora, as in the case of the BNC and the .uk data (cf. Sections 8.3 and 10.3). In view of all these facts, we can indeed conclude that Irish English seems to be on average more likely to use subject pronoun case forms in subject predicative complements than the other databases or varieties analysed so far.

<table>
<thead>
<tr>
<th>IT-CLEFTS (.ie)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>338 (100%)</td>
<td>275 (81.36%)</td>
<td>63 (18.64%)</td>
</tr>
</tbody>
</table>

*Table 46: it-Clefts in the .ie Data: The Token Numbers*

As far as the multivariate analysis of the *it*-clefts in the .ie data is concerned, Table 47 indicates that three factors are retained in the minimal-adequate model. This means that two of the factors motivated and discussed in Chapter 6 are excluded in the course of the model-building process. The first of these eliminated factors is *SINGULAR*, which is in line with the corresponding model of the BNC, where this factor is also not significant (cf. Section 8.3). The exclusion of *SINGULAR* is, however, in contrast to the significantly inhibiting effect of this factor observed in the COCA, and it is also in contrast to the significantly fostering effect of *SINGULAR* attested for the .uk and .au data (cf. Sections 9.3, 10.3 and 11.3). Moreover, it does also not meet the expectations of some accounts assuming a promoting effect of this factor on the use of subject forms in *it*-clefts (Sobin 1997: 334). Thus, as noted before, the impact of *SINGULAR* on the distribution of pronoun case forms seems to strongly depend on the analysed dataset or variety.

Furthermore, *SPOK* is also excluded from the minimal-adequate model (cf. Table 47). The elimination of this factor is in line with the results obtained for the .uk and .au data but does not correspond to the effect of *SPOK* observed in the BNC and COCA data, where it markedly constrains the use of subject pronoun case forms in *it*-clefts. On the one hand, the non-significance of *SPOK* in the present dataset can again be partly attributed to the rather formal character of much of the spoken Web data (cf. Section 7.2.4.3.1). On the other hand, it has also to be borne in mind that we have so far not attested a general trend according to which *SPOK* generally inhibits the use of subject pronoun case forms in subject predicative complements in general and *it*-clefts in particular as is suggested by much of the literature (e.g. Harris 1981: 18-19; Quirk et al. 1985: 337-338). Thus, this factor too, exhibits a certain
degree of cross-varietal variability, when looking at its effect in the different datasets analysed and discussed so far.\textsuperscript{53}

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>As_Subj</td>
<td>2.08</td>
<td>**</td>
<td>7.98</td>
</tr>
<tr>
<td>First</td>
<td>-2.06</td>
<td>***</td>
<td>0.13</td>
</tr>
<tr>
<td>CMC</td>
<td>-2.10</td>
<td>***</td>
<td>0.12</td>
</tr>
<tr>
<td>Model $\chi^2$</td>
<td>101.29</td>
<td>(***)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
<td>0.41</td>
</tr>
<tr>
<td>% Correctly Predicted</td>
<td></td>
<td></td>
<td>86.39</td>
</tr>
<tr>
<td>% Baseline</td>
<td></td>
<td></td>
<td>81.36</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>338</td>
</tr>
</tbody>
</table>

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 47: it-CLEFTS in the ie Data: Logistic Regression Results

Table 47 demonstrates that As_Subj is very highly significant and emphatically promotes the use of subject pronoun case forms in it-clefts, which is indicated by the factor’s odds ratio of 7.98. This odds ratio shows that the probability of observing a subject form as focal pronoun is 698 per cent higher in it-clefts in which the focal pronoun is co-referential with the subject of the following dependent clause than in it-clefts in which the focal pronoun is co-referential with an object. The impact of this factor does not only confirm the results obtained for all other corresponding datasets examined so far, but also the expectations voiced in the literature assuming a markedly fostering effect of As_Subj on the use of subject forms in it-clefts (e.g. Erdmann 1978: 76-78; Huddleston and Pullum 2002: 459). Thus, once again, a very strong impact of a functional factor on the distribution of pronoun case forms in subject predicative complements can be noted, which clearly corroborates the functional perspective of this study.

The second factor identified as being very highly significant is First. In analogy to all it-cleft datasets analysed so far, this factor evidently constrains the use of subject pronoun forms in it-clefts, which is indicated by its odds ratio of 0.13. In addition, this finding is also

\textsuperscript{53} Moreover, the following interactions were eliminated to arrive at the minimal-adequate regression model: Spok*First, Spok*Singular, Spok*As_Subj, CMC*First, CMC*Singular, CMC*As_Subj, First*Singular, First*As_Subj, Singular*As_Subj.
again in line with the assumptions and observations of earlier studies (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96). This difference between first and third person pronouns can be again explained with their prototypical referential functions (e.g. cf. Sections 6.1.2.2 and Sections 9.3, 10.3 and 11.3). This means that first person pronouns are per se more pragmatically salient than third person ones. If first person pronouns are used as focal pronouns in *it*-clefts, which are themselves a focussing device (cf. Section 4.1), this may be sufficient to mark a pronoun or its referent as the Focus of a clause and hence additional highlighting by means of a subject form may not be necessary (cf. Dik 1989: 278). This account is not only in line with the present data, but could also explain why FIRST inhibits the use of subject pronouns in *it*-clefts in all respective datasets observed so far. Moreover, this explanation is fully consistent with one of this study’s central hypotheses assuming a re-functionalisation of subject pronoun forms as Focus markers (cf. Section 5.2; H3).

Finally, CMC is not only very highly significant but also – as its odds ratio of 0.12 demonstrates – evidently inhibits the use of subject pronoun case forms in *it*-clefts in the .ie data. This result confirms the findings of the corresponding analyses made for the .uk and .au data and also those accounts reporting that less formal varieties may be less likely to use subject forms in subject predicative complements (e.g. Quirk et al. 1985: 337-338; Huddleston and Pullum 2002: 459). As before (cf. Sections 10.1 and 11.3), this clearly constraining effect of CMC can be accounted for with the fact that CMC can – as a consequence of its lower degree of formalisation – employ a much wider repertoire of expressive options ranging from the imitation of intonation patterns by means of orthography (cf. Herring 2010a) to the highlighting of pragmatically salient referents by means of capitalisation (e.g. Section 11.1). Thus, simply because CMC can use other means to highlight the Focus of a clause, it may be less likely to use subject pronoun forms in subject predicative complements to do so.

When shifting our attention from the individual factors to the statistical model as a whole, Table 47 shows that the minimal-adequate regression model itself is also very highly statistically significant ($\chi^2 = 101.29, p<0.001$). As its Nagelkerke’s $R^2$ value of 0.41 indicates, the model is not only substantially significant (cf. Szmrecsanyi 2006: 55), but can also account for 41 per cent of the variation observed in the dependent variable (cf. Table 47). In addition, the applied regression model correctly predicts the outcome of the dependent
variable in 86.39 per cent of all instances, which clearly improves the baseline model, accounting only for 81.36 per cent of all tokens, by more than five per cent.

In a nutshell, the strongly fostering effect of As_SUBJ on the use of subject forms in the .ie data strongly corroborates not only the results obtained from the examinations of all other corresponding datasets analysed so far but also those accounts which also assume or even attest a promoting effect of this factor on the use of subject forms in it-clefts (e.g. Erdmann 1978: 75-78; Quirk et al. 1985: 337-338). Moreover, this finding further corroborates the general trend observed so far that functional factors – contrary to the expectations of many accounts (e.g. Burridge 2004: 1118) – exert a remarkable influence on the distribution of pronoun case forms in subject predicative complements (cf. H2 and H3 in Section 5.2). The effect attested for FIRST is also in line with the results obtained for all other it-cleft datasets examined so far in that this factor markedly constrains the use of subject forms. This also corresponds to the observations and assumptions of much of the literature on this subject (e.g. e.g. Quinn 2005a: 133-135; Wales 1996: 95-96) and could again be explained with the help of the different referential properties of first and third person pronouns. Furthermore, the impact of CMC is also in line with the results obtained for the .uk and .au data. In all three datasets containing this mode of discourse, CMC clearly inhibits the use of subject pronoun case forms in it-clefts.

12.4 The Distribution of Pronoun Case Forms in the Irish (.ie)

Internet Data: Interim Summary

The analyses in Sections 12.1–12.3 have demonstrated that subject pronoun forms are considerably more often used in the .ie data than in all other datasets analysed so far. Bearing in mind that the share of subject forms are skewed in the Web-derived datasets due to the particularities of the data compilation, this very high share of subject forms attested for the .ie data is even more remarkable, since it could be hypothesised that it may be even higher in a corresponding closed corpus (cf. Sections 7.2.4.2 and 10.1). Moreover, the high shares of subject pronoun forms consistently observed in this chapter severely question the assumptions of studies expecting a clear preference or even predominance of object pronoun forms in these contexts (e.g. Burridge 2004: 1118; Harris 1981: 18-19).
Even more challenging for these accounts trying to explain the distribution of pronoun case forms merely by the pronoun’s position or by its pronoun class membership is, however, the very strong impact functional factors and particularly syntactic and pragmatic ones exert on the distribution of pronoun case forms in the .ie data. The analyses have again demonstrated that there are marked differences between it-clefts and it BE sentences in the distribution of pronoun case forms. Furthermore, the very strong influence that is exerted by CLEFT, FOCUS and As_SUBJ is very difficult to accommodate with the positional approach or the weak vs. strong pronoun class approach or its derivatives outlined in Chapter 3.

In addition, the findings for CLEFT, FOCUS and As_SUBJ in the .ie data are in line with the observations made for all other datasets examined so far in that these factors unequivocally foster the use of subject pronoun case forms in the respective contexts. At the same time, this also implies that both necessary and sufficient condition for us to assume that Focus plays an important role in the distribution of pronoun case forms are again met (cf. Section 6.1.2.5). The strong impact of FOCUS on the distribution of pronoun case forms is further witnessed by the fact that it is, as in the corresponding .au dataset (cf. Section 11.2), the only factor that significantly influences the distribution of pronoun case forms in the it BE sentences of the .ie data. This strongly corroborates the hypothesis according to which subject pronouns have been reanalysed as Focus markers in subject predicative complements (cf. Section 5.2; H3).

As far as the variables NUMBER and PERSON are concerned, the analysis of the .ie data has demonstrated that SINGULAR plays no role in the present dataset. Although this corroborates the findings obtained for the BNC, it is neither in accordance with the COCA results, where SINGULAR has partly an inhibiting effect, nor with the factor’s fostering effect observed particularly in the it-clefts of the .uk and .au data, nor with the assumptions of some accounts expecting such a fostering effect (cf. Sobin 1997: 334). FIRST, however, again significantly inhibits the use of subject pronoun case forms – particularly in the it-cleft subset. This is in line with the .uk and .au data, with the BNC data and also to some extent with the COCA, where the constraining effect of FIRST is even attested for the it BE sentences. This inhibiting effect can again be attributed to the inherent differences between first and third person pronouns as far as their prototypical referential characteristics as well as the salience of their referents are concerned (cf. also Section 6.1.2.2). In short, the results of this study so far suggest that the referents of third person pronouns are more likely to
receive additional highlighting by means of a subject form to mark their pragmatic prominence within a clause. Again, this Focus-oriented perspective is not only able to explain the differences between first and third person pronouns as attested in the data but is also in line with this study’s hypothesis that pragmatic considerations play an important role in the distribution of pronoun case forms in the analysed contexts (cf. Section 5.2).

With regard to the MODE OF DISCOURSE variable, SPOK is not significant in the .ie data, confirming the results obtained for the previous Web-derived datasets. As already observed for the .uk and .au data, CMC again significantly inhibits the use of subject forms, particularly in the it-cleft data. The reason for this cannot only be attributed to the fact that less formal varieties are simply less likely to use subject pronouns, but also to the fact that exactly because CMC is subject to fewer formal conventions, it can employ strategies to highlight pragmatically salient entities, such as capitalisation, that are available yet not admissible in traditional written varieties.
13 The Distribution of Pronoun Case Forms in the South African (.za) Internet Data

The following chapter presents and discusses the distribution of pronoun case forms as observed in the South African Internet data (henceforth .za data). Although this is by now the fourth Web-derived dataset examined in this study, it is the first database obtained from a region where English is not the dominant L1. Thus, it will be interesting to see whether or not the trends that have been observed so far for the other datasets will also bear out for the .za data. As in the preceding chapters, Section 13.1 begins with the analysis and the discussion of the results of the general subject predicative complement category before Section 13.2 examines the distribution of pronoun case forms in the *it BE* sentences of this dataset. Section 13.3 then analyses and discusses the distribution of pronoun case forms in the *it*-clefts of the .za data. Finally, Section 13.4 concludes this chapter with an interim summary of the obtained results.

13.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the South African (.za) Internet Data

As has become customary, the following sections immediately start with the presentation and discussion of the obtained results, since the motivation and discussion of the tested variables as well as a meticulous description of the data compilation and data processing are provided in Chapters 6 and 7.

<table>
<thead>
<tr>
<th>Subject Predicative Complements (.za)</th>
<th>Total N</th>
<th>Subject Forms N</th>
<th>Object Forms N</th>
</tr>
</thead>
<tbody>
<tr>
<td>651 (100%)</td>
<td>261 (40.09%)</td>
<td>390 (59.91%)</td>
<td></td>
</tr>
</tbody>
</table>

*Table 48: Subject Predicative Complements in the .za Data: The Token Numbers*

Table 48 shows that the processing of the raw data has left us with 651 subject predicative complements forming the empirical foundation of the subsequent statistical analysis. Hence, although the token number obtained for the .za top-level domain is considerably lower than
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

those obtained for .uk and .au datasets (cf. Chapters 10 and 11), it is still markedly higher
than the 425 tokens constituting the .ie dataset (cf. Chapter 11). As far as the distribution
of subject and object pronoun case forms is concerned, Table 48 indicates that 40.09 per cent
of all subject predicative complements in the .za data employ a subject form, which also
implies that the remaining 59.91 per cent are used with an object pronoun form. Comparing
these numbers to the corresponding datasets of the closed corpora, we can say that the
share of subject forms in the .za data is lower than that observed for the BNC, but it still
remarkably higher than that attested in the COCA (cf. Sections 8.1 and 9.1). When
comparing these shares in Table 48 to those of the other Web-derived datasets analysed so
far, we can note that even though the share of subject pronoun case forms in the .za data is
much smaller than that observed for the .ie data, it is still larger than those of the .uk and
the .au data, which only amount to 21.60 and 20.72 per cent respectively (cf. Sections 10.1
and 11.1). However, whether or not subject forms are indeed more likely to be used in South
African English than in the .uk and .au datasets can only be conclusively assessed after
having conducted the separate analyses of the it BE sentences and the it-clefts in terms of
pronoun case distribution (cf. Sections 13.2 and 13.3).

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>2.49</td>
<td>***</td>
<td>12.05</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.45</td>
<td>***</td>
<td>0.24</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>0.78</td>
<td>**</td>
<td>2.19</td>
</tr>
<tr>
<td>SPOK</td>
<td>-0.98</td>
<td>***</td>
<td>0.38</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.54</td>
<td>***</td>
<td>0.22</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>-0.82</td>
<td>+</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Model $\chi^2$ 274.29 (***)

R$^2$ 0.46

% Correctly Predicted 78.65

% Baseline 59.91

N 651

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 49: Subject Predicative Complements in the .za Data: Logistic Regression Results
As can be seen from Table 49, five factors and one interaction are retained in the minimal-adequate model applied to the .za data. Thus, only one of the variables motivated and discussed in Chapter 6 is eliminated, which is – very surprisingly – Focus. The exclusion of Focus is not only in contrast to this study’s assumption that subject pronoun case forms may have been reanalysed as Focus markers in subject predicative complements (cf. Chapter 4), but also with the results obtained so far from the other corresponding datasets. Even in the .ie data, where Focus is also not significant in the superordinate category of subject predicative complements (cf. Section 12.2), this factor is retained in the corresponding minimal-adequate regression model (cf. Table 43). In the present model, however, bootstrap validation tells us to eliminate Focus from the model (cf. Section 6.2).

While the elimination of Focus is so far unique, the results for Cleft in the .za data are consistent with those obtained for the corresponding models of all other datasets examined so far (cf. Sections 8.1–12.1). Again, Cleft is very highly statistically significant and markedly promotes the use of subject pronoun case forms in subject predicative complements, as its odds ratio of 12.05 indicates (cf. Table 49). In terms of percentages, this means that the probability of observing a subject form in subject predicative complement position is 1105 per cent higher in an it-cleft than in an it BE sentence in the .za data. On the one hand, this finding confirms not only the observations made for all other databases with regard to this factor but it also supports the assumptions and results of those accounts highlighting the marked differences between it BE sentences and it-clefs as far as the distribution of pronoun case forms is concerned (e.g. Biber et al. 1999: 335-336; Maier 2013). On the other hand, however, these results do not confirm accounts claiming that positional considerations or the assignment to certain pronoun classes are the decisive criteria determining the distribution of pronoun case forms (e.g. Emonds 1986: 96-100; Harris 1981: 18-19; Quinn 2005a).

Table 49 also shows that First is likewise retained in the minimal-adequate model. This factor is very highly significant and clearly inhibits the use of subject pronoun case forms in subject predicative complements, which is signalled by its odds ratio of 0.24. Hence, this finding is also consistent with this factor’s effect observed for all other corresponding databases examined so far except for that of the BNC, where First is only significant in the it-

54 In addition to Focus, the following interactions were eliminated in the course of the model-building process: Spok*First, Spok*Singular, Spok*Cleft, Spok*Focus, Cmc*First, Cmc*Singular, Cmc*Cleft, Cmc*Focus, First*Cleft, First*Focus, Singular*Cleft, Singular*Focus.
cleft subsection. With regard to the expectations of the literature, the impact of First in the .za data is again not in line with those accounts assuming a higher share of first person subject pronouns than of third person subject pronouns in it BE sentences (e.g. Quinn 2005a: 246). At the same time, however, this finding complies with accounts observing a greater likelihood of third person pronouns than of first person pronouns to be used in their subject form in it-clefs (e.g. Quinn 2005a: 134-135; Wales 1996: 95-96). As discussed before (e.g. Sections 10.1 or 11.1), these more or less systematic differences between first and third person pronouns can be explained by the fundamental differences between these two persons with regard to their prototypical referential characteristics and functions. In a nutshell, it can be argued that the referents of first person pronouns are per se more salient than the referents of third person pronouns, which means that the latter ones may need more additional highlighting by means of a subject pronoun form – where this is admissible – in order to be recognisable as the most important piece of information, i.e. the Focus, of a clause (cf. Sections 6.1.2.2 and 8.3). This possible explanation is not only consistent with the observations made for this factor across most of the corresponding datasets analysed so far, but also with the general stance taken in this study assuming that pragmatic factors play an important role in the distribution of pronoun case forms (cf. Chapter 4).

Furthermore, Table 49 also identifies Singular as highly significant. In the .za data, Singular exerts a markedly fostering effect on the use of subject pronoun case forms in subject predicative complements, which is indicated by its odds ratio of 2.19. This observation is in line with some accounts assuming such a promoting effect of this factor (Sobin 1997: 334) as well as with the effect of Singular as attested in the.uk and the.au data (cf. Sections 10.1 and 11.1). However, this finding does neither correspond to the results obtained for the BNC and the .ie data, where Singular is eliminated from the respective minimal models (cf. Sections 8.1 and 12.1), nor to the COCA results, where even a contrary effect, i.e. an inhibiting one, is attested for this factor (cf. Section 9.1). Indeed, the impact of Singular seems to vary from database to database and does not behave as uniform across different datasets as the effect of other factors (e.g. cf. Section 10.1).

Another factor that is retained in the minimal-adequate model as very highly statistically significant is Spok. As can be seen from its odds ratio of 0.38, Spok evidently constrains the use of subject pronoun case forms in subject predicative complements in the .za data. Although such a constraining effect of Spok is actually expected by much of the
relevant literature (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Quirk et al. 1985: 337-338; Wales 96: 91-108), it has only been observed so far in one other corresponding dataset, i.e. that of the COCA (cf. Section 9.1). In the other Web-derived datasets, SPOK is eliminated from the minimal-adequate models as being non-significant. Even more remarkably, the BNC data yields a very differentiated picture with regard to this factor, since SPOK in tendency even promotes the use of subject forms in the subject predicative complement category and the *it BE* sentences of that corpus (cf. Sections 8.1 and 8.2), whereas a markedly inhibiting effect of this factor is observed for the *it*-cleft section of that corpus (cf. Section 8.3). The effect of SPOK may be accounted for in terms of formality, meaning that spoken data is less likely to be subject to the “prescriptive bias in favour of the subjective form” (Quirk et al. 1985: 338), which is often deemed responsible for the use of subject forms in subject predicative complements (cf. Section 6.1.2.4). However, we can also try to explain this difference again from a more pragmatic point of view. Since languages and language varieties may employ several different devices to mark a constituent as the Focus of a clause (cf. Section 4.1), it has already been proposed that the availability of other Focus marking strategies may also explain why subject forms are less likely to occur in the spoken data than in written data of some databases (cf. Sections 9.1, 10.1 and 11.1). Thus, speakers of this variety of English might rather use, for example, prosodic prominence than subject pronoun forms to highlight the Focus of a clause. Producers of traditional written data, in contrast, cannot employ this highlighting strategy and therefore may rather rely on subject forms to mark the referent of a pronoun as the Focus of a clause (cf. Section 4.1). Hence, the availability of other pragmatically salient features or strategies might actually inhibit the use of subject forms in some contexts (cf. Sections 9.1, 10.1 and 11.1). Such an explanation is not only in line with results obtained for this dataset, but also with the more general perspective of this study assuming that Focus marking plays an eminent role in the distribution of pronoun case forms in subject predicative complements – even though Focus as operationalised in Section 6.1.2.5 is admittedly not significant in the present model. Furthermore, this functional approach does not only describe that there may be differences between modes of discourse, but also attempts to provide an explanation for the differences observable between the different modes of discourse.

In addition to SPOK, CMC is also identified as very highly significant and – with an odds ratio of 0.22 – evidently constrains the use of subject pronoun case forms in subject
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

predicative complements in the .za data. This observation corresponds to the impact of CMC as observed for the previously analysed Web-derived datasets, where CMC also emphatically inhibits the use of subject pronoun case forms in subject predicative complements (cf. Sections 10.1–12.1). Moreover, this finding is also in line with accounts postulating a lower likelihood of subject pronoun forms to occur in less formal varieties (e.g. Huddleston and Pullum 2002: 459; Quirk et al. 1985: 337-338). Similar to the factor SPOK, CMC can, however, also employ more means, such as capitalisation, to mark a clausal constituent as the relatively most important one, which is not as easily possible in traditional written texts, since they are subject to more rigid text and spelling conventions (cf. Sections 10.1 and 11.1). Hence, it could be argued that CMC is less dependent on the use of subject forms in subject predicative complements to mark the pronoun’s referent(s) as the Focus of the clause, as is also indicated by the examples in (142) from the South African CMC data:

(142) Capitalisation as Focus Marking Strategy in CMC in the. za Data
a. If someone takes it as an attack on Blacks, then it is THEM that makes it racial and they have a racial problem. (.za/it is them/15.07.2008)
b. I mean it is THEM isn’t it?. (.za/it is them/15.07.2008)

These examples in (142) illustrate that although Focus as operationalised in Section 6.1.2.5 is not significant in the .za data, the CMC subset of this data sample still exhibits tokens in which the pronouns in subject predicative complement position are particularly highlighted as Focus – though by other means than subject pronoun forms. This observation corroborates the assumption that CMC may use fewer subject forms simply because it possesses other options to mark pragmatically salient entities. Furthermore, this finding suggests that although these contexts can be considered as clearly focussed, different varieties – both within and across geographic continua – may make use of different strategies to mark the Focus of a clause in subject predicative complement positions.

Finally, the interaction between FIRST and SINGULAR is also identified as marginally significant in Table 49. With an odds ratio of 0.44, this interaction tends to inhibit the use of subject pronoun case forms in subject predicative complements in the .za data. However, as already discussed in Sections 10.1 and 11.1, this interaction can be mainly accounted for by the limitations of commercial search engines for corpus linguistic data compilations (cf. Section 7.2.4.2). The resulting skewing effect in the shape of the interaction between FIRST and SINGULAR can be observed in the statistical analyses of three of the four Web-derived

300
datasets analysed so far, namely in the present one and in the corresponding models of the .uk data and the .au data (cf. Sections 10.1 and 11.1).

The minimal-adequate regression model used to analyse the subject predicative complements of this dataset is itself also very highly statistically significant ($\chi^2 = 274.29$, $p<0.001$). With a Nagelkerke’s $R^2$ value of 0.46, the model applied to the .za data is able to account for 46 per cent of the variation observed in the dependent variable. Furthermore, the regression model clearly improves the baseline model by nearly 19 per cent. While the baseline model is able to correctly predict 59.91 per cent of all tokens, the applied regression model can correctly predict the outcome of the dependent variable in 78.65 per cent of all instances.

Although many of the results obtained for the .za data confirm earlier findings, this analysis has also yielded one remarkable difference: In contrast to all other datasets analysed so far, FOCUS is eliminated from the minimal-adequate model, which is unexpected not only considering the results obtained so far for this factor but also in view of the assumptions of this study (cf. Chapter 4). CLEFT, however, is again identified as being very strongly promoting the use of subject pronoun forms in subject predicative complements, which is in line with the results of all other datasets examined so far and also with those accounts attesting marked differences between *it*-clefts and *it BE* sentences in terms of pronoun case distribution (e.g. Erdmann 1978; Maier 2013). This, in turn, weakens again accounts contesting the influence of functional factors on the distribution of pronoun case forms (cf. Chapter 3). The factor FIRST evidently inhibits the use of subject pronoun case forms in subject predicative complements, which is in accordance with the findings obtained for all other datasets except for the BNC, where such an impact of FIRST is only attested for the *it*-cleft data (cf. Sections 8.1 and 8.3). As before, the constraining effect of FIRST could be attributed to the differences between first and third person pronouns with regard to their prototypical referential functions and properties (cf. Sections 6.1.2.2 and 8.3). SINGULAR, in contrast, again evidently fosters the use of subject forms in the .za data, which corresponds to the results obtained for this factor in the .uk data and the .au data as well as with the expectations of some studies (e.g. Sobin 1997: 334). However, it does neither match the results obtained for the COCA, nor those of the BNC and the .ie data. SPOK clearly inhibits the use of subject pronoun case forms. This finding is expected by much of the literature and it is in line with the observations made for the COCA. It is in contrast, however, to the .uk and .au
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English datasets, where this factor plays no role, and also to the BNC data, where this factor in tendency even promotes the use of subject pronoun case forms. Similarly, CMC also inhibits the use of subject pronoun case forms in the .za data, which corresponds to its effect as observed in all other Web-derived datasets so far. The inhibiting effect of Spok and CMC could be explained with the fact that both spoken and CMC discourse cannot only be considered as less formal than written data but are also more flexible and can employ more Focus marking strategies to highlight pragmatically prominent entities, which may also account for why subject pronoun case forms are less likely in these discourse modes.

13.2 The Distribution of Pronoun Case Forms in *it BE* Sentences in the South African (.za) Internet Data

In the dataset obtained from the South African top-level domain .za, 171 tokens of the 651 subject predicative complements are *it BE* sentences. Hence, the number of *it BE* sentences in this dataset is nearly twice as high as that in the .ie data (cf. Section 12.2). Furthermore, Table 50 indicates that a share of 5.85 per cent of these *it BE* sentences are used with a subject pronoun form, while the remaining 94.15 per cent are used with an object pronoun form. Although the share of subject forms attested for this dataset is much smaller than that attested for the *it BE* sentences in the .ie data and also than that observed in the BNC (cf. Sections 8.2 and 12.2), it is still higher than the corresponding shares of subject forms in the .uk data, the .au data and the COCA (cf. Sections 9.2, 10.2 and 11.2). What the shares of subject and object forms depicted in Table 50 also illustrate once again are the marked differences in the distribution of pronoun case forms between the superordinate category of subject predicative complements amounting to 40.09 per cent (cf. Section 13.1) and the *it BE* sentences subsumed under this heading.

<table>
<thead>
<tr>
<th></th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IT BE SENTENCES (.za)</strong></td>
<td>171 (100%)</td>
<td>10 (5.85%)</td>
<td>161 (94.15%)</td>
</tr>
</tbody>
</table>

*Table 50:* *it BE* Sentences in the .za Data: The Token Numbers
As far as the multivariate analysis of the *it BE* sentences in the .za data is concerned, Table 51 below indicates that two factors are retained in the minimal adequate model applied to the data. What is very remarkable, however, is the fact that none of these factors is statistically significant (cf. Table 51).

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGULAR</td>
<td>1.50</td>
<td>n.s.</td>
<td>4.46</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.41</td>
<td>n.s.</td>
<td>4.10</td>
</tr>
<tr>
<td>MODEL x2</td>
<td></td>
<td>5.08 (+)</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>% CORRECTLY PREDICTED</td>
<td></td>
<td></td>
<td>94.15</td>
</tr>
<tr>
<td>% BASELINE</td>
<td></td>
<td></td>
<td>94.15</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td>171</td>
</tr>
</tbody>
</table>

* + significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

*Table 51: it BE Sentences in the .za Data: Logistic Regression Results*

With regard to the factors excluded from the model in Table 51, we note that First is eliminated to arrive at the minimal adequate model. Although the elimination of First complies with its exclusion in the corresponding BNC, .uk, .au and .ie datasets, it is neither in accordance with the significant inhibiting effect attested for this factor in the COCA, nor with accounts expecting a fostering impact of this factor on the use of subject pronoun forms (Quinn 2005a: 246). Likewise, SPOK and CMC are also not significant and therefore also eliminated from the final regression model (cf. Table 51). With the exception of the BNC, where SPOK in tendency fosters the use of subject forms, the elimination of SPOK is in line with the results of all other datasets examined so far. In a similar vein, the non-significance of CMC confirms the results obtained for the .au and .ie data but not for the .uk data, where this factor has an inhibiting effect on the use of subject pronoun forms in *it BE* sentences. Indeed, the elimination of these variables is also not in line with accounts assuming that less formal varieties are less likely to use subject forms in subject predicative complements (e.g. Huddleston and Pullum 2002: 459; Quirk et al. 1985: 337-338).55

---

55 The following interactions were also eliminated to arrive at the minimal adequate model: SPOK*FIRST, SPOK*SINGULAR, SPOK*FOCUS, CMC*FIRST, CMC*SINGULAR, CMC*FOCUS, FIRST*SINGULAR, FIRST*FOCUS, SINGULAR*FOCUS.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

The first factor actually retained in the minimal-adequate model is SINGULAR. Even though it is not significant, bootstrap validation tells us to maintain it in the regression model (cf. Section 6.2). When compared to the corresponding analyses of the other datasets discussed so far, the non-significance of SINGULAR is in accordance with the analyses of the BNC, COCA, .au and .ie datasets. However, the non-significance of SINGULAR in the it BE sentences of the .za data does not confirm its promoting effect observed for the .uk data, nor accounts expecting such a fostering effect (cf. Sobin 1997: 334).

The second factor retained in Table 51 is FOCUS, which is also not significant in the present dataset. This is of course neither in line with one of these study’s hypotheses assuming that FOCUS plays an enormous role in the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H3), nor with the results obtained for the corresponding datasets of all other corpora and databases examined so far, in which FOCUS consistently exhibits a highly significant and promoting effect on the use of subject pronoun forms in it BE sentences (cf. Sections 8.2–12.2).

Having a look at the odds ratios of these retained, yet non-significant, factors in Table 51, we can observe the theoretically expected impact of these factors – had they not missed statistical significance (cf. Szmrecsanyi 2006: 102). Since both of them are clearly above one, at least their theoretically expected effect, particularly that of FOCUS, complies with earlier observations and the central assumptions of this study.

With regard to the overall model, Table 51 shows that the minimal-adequate model is marginally significant ($\chi^2 = 5.08$, $p<0.1$), which indicates that the correlation between the dependent and the independent variables is not too strong. This, however, is not surprising considering the fact that none of the factors retained in the model is actually significant. The model’s Nagelkerke’s $R^2$ value is 0.07, which means that the model could account for seven per cent of the variation attested in the dependent variable if it were significant. As far as its predictive power is concerned, the applied regression model cannot improve the baseline model since both minimal-adequate regression and baseline model correctly predict the outcome of the dependent variable in 94.15 per cent of all cases (cf. Table 51).

To sum up, the analysis of the it BE sentences in the .za dataset has yielded very remarkable results in that none of the variables motivated and discussed in Chapter 6 is significant in the regression model applied to the data. This result is very surprising, since not even FOCUS, which has so far consistently been identified as significantly influencing the
distribution of pronoun case forms in *it BE* sentences is significant in the .za data. Furthermore, the non-significance of Focus as operationalised in Section 6.1.2.5 does also not corroborate our assumption that subject pronoun forms may have been reanalysed as Focus markers in subject predicative complements (cf. Chapter 4). In general, the non-significance of all tested independent variables is remarkable because it cannot be merely attributed to the token number of this dataset. Although there are only 87 tokens in the corresponding .ie dataset, the analysis of this much smaller dataset yields statistically very highly significant results (cf. Section 12.2). Thus, the factors used for the present analysis are obviously not as helpful to account for the distribution of pronoun case forms in the South African data as they are for the other datasets examined so far. This, in turn, suggests marked cross-varietal differences between the .za data and the other datasets and varieties examined so far.

13.3 The Distribution of Pronoun Case Forms in *it*-Clefs in the South African (.za) Internet Data

Of the 651 subject predicative complements constituting the .za dataset, 480 tokens are *it*-clefs with a case-sensitive focal pronoun (cf. Table 52). Table 52 also indicates that 52.39 per cent of these *it*-clefs are used with a subject form as focal pronoun, whereas 47.71 per cent of these tokens exhibit an object form as focal pronoun. This means that the share of subject pronoun forms is nearly ten times higher in the *it*-cleft than in the *it BE* sentence section of the .za dataset (cf. Section 13.2), which once again corroborates accounts pointing out the differences between these two sentence types in terms of pronoun case distribution (e.g. Biber et al. 1999: 334-336; Erdmann 1978). When comparing the shares of subject and object pronoun case forms given in Table 52 to those of the corresponding datasets examined before, we note that the share of subject forms in the .za *it*-cleft data is markedly lower than those in the corresponding BNC, COCA and .ie datasets (cf. Sections 8.3, 9.3 and 12.3). At the same time, however, the share of subject pronoun forms attested for the .za *it*-clefs is still considerably higher than the corresponding shares in the .uk and .au data. Bearing in mind that all Web-derived databases are subject to the same skewing
mechanisms, this suggests again that we observe marked cross-varietal differences between the different datasets.

| TABLE 52: IT-CLEFTS IN THE .ZA DATA: THE TOKEN NUMBERS |

As can be seen from Table 53, all five main effects motivated and discussed in Chapter 6 are retained in the minimal-adequate regression model analysing this dataset. Furthermore, two interactions are also kept in this model. Hence, the present minimal-adequate regression model is – compared to all other corresponding models analysing it-clefts – the one that retains the most factors (cf. Sections 8.3–12.3).56

Table 53 indicates that AS_SUBJ is very highly significant. Furthermore, with an odds ratio of 5.64, AS_SUBJ also markedly fosters the use of subject pronoun forms in it-clefts in

---

56 Although many factors and interactions have been retained in this model, the following interactions were eliminated in the course of the model-building process: SPOK*FIRST, SPOK*SINGULAR, SPOK*AS_SUBJ, CMC*FIRST, CMC*SINGULAR, CMC*AS_SUBJ, FIRST*SINGULAR, FIRST*AS_SUBJ, SINGULAR*AS_SUBJ.
the .za data. Percentagewise, this means that the likelihood of observing a subject pronoun form is 464 per cent higher in *it*-clefts in which the focal pronoun is co-referential with the subject of the following clause than in *it*-clefts in which the focal pronoun is co-referential with an object of the subsequent dependent clause. First of all, this finding is consistent with the results obtained for AS.SUBJ in all other *it*-cleft datasets examined so far, where this factor is also that with the highest odds ratio and also very highly significantly promotes the use of subject pronoun case forms. Moreover, the impact of AS.SUBJ as observed in the .za data is also in line with the considerations of those accounts assuming an emphatically promoting effect of this factor on the use of subject forms in *it*-clefts (e.g. Erdmann 1978: 76–78; Huddleston and Pullum 2002: 459; Quirk et al. 1985: 338). Thus, even though a strong impact of functional factors cannot be observed for the *it BE* sentences of the .za data (cf. Section 13.2), the impact and significance of AS.SUBJ suggests that at least in the *it*-cleft section of this dataset functional factors play an important role in the distribution of pronoun case forms.

Table 53 further identifies FIRST as very highly significant, and its odds ratio of 0.17 signals that FIRST again strongly constrains the use of subject pronoun forms. This observation is consistent with the factor’s impact as observed in all other *it*-cleft datasets examined so far as well as with the findings of previous studies (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96). As in the previous chapters (cf. 8–12), this more or less systematic difference between first and third person pronouns regarding their likelihood of occurring in their subject form can be again accounted for with the higher pragmatic salience of first person pronouns due to their prototypical deictic nature. Third person pronouns, in contrast, are prototypically phoric and thus their referents may need additional marking to be more easily identifiable as the Focus, i.e. the most important piece of information in a clause (cf. Siewierska 2004: 5-8). In case first person pronouns are used as focal pronouns of *it*-clefts, the co-occurrence of two such pragmatically prominent features may suffice to mark the referent or referents of the focal first person pronoun as the Focus of the clause. This can account for the differences between first and third person pronouns observed in this dataset and it can also explain why first person subject pronouns are systematically less likely to occur in *it*-clefts across varieties of English (cf. Sections 8.3–12.3).

While the effect of FIRST is more or less uniform across all datasets analysed so far, the second grammatical distinction operationalised in this study, i.e. SINGULAR, is subject to
much more variation. In the .za it-cleft data, SINGULAR is very highly significant and strongly promotes the use of subject forms in it-clefts, as its odds ratio of 2.79 indicates. Thus, this result confirms the observations made for the. uk and the .au data, where SINGULAR also significantly fosters the use of subject pronoun forms in it-clefts (cf. Sections 10.3 and 11.3). In addition, this finding is in line with studies expecting a fostering effect of this factor on the use of subject pronoun case forms (cf. Sobin 1997). However, the fostering effect observed for SINGULAR in this dataset does neither comply with the findings made for the BNC and the .ie data, where SINGULAR is not significant (cf. Sections 8.3 and 12.3), nor with the COCA, where it actually significantly inhibits the use of subject pronoun case forms in it-clefts (cf. Section 9.3). Thus, this factor is much more prone to variation than FIRST operationalising the other grammatical category, i.e. PERSON.

Table 53 also shows that SPOK is also very highly significant. As its odds ratio of 0.29 indicates, the likelihood of observing subject pronoun forms in the spoken subset of the .za it-cleft data is much lower than in the traditional written data. While such an inhibiting effect of SPOK is expected by the relevant literature (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Quirk et al. 1985: 337-338), it has only been observed so far for the it-cleft datasets of the BNC and the COCA, but not for any of the Web-derived it-cleft datasets. Furthermore, the inhibiting force of SPOK seems to be restricted to the it-cleft data, since it has not been attested for any of the it BE sentence analyses conducted so far (cf. Sections 8.2–13.2). This, in turn, indirectly corroborates the possible explanation already put forth earlier in this study according to which the differences between the different modes of discourse are not merely the result of differences in formality but can also be partly attributed to the greater repertoire of marking pragmatic prominence which spoken and CMC data can employ when compared to written data. As discussed in Section 4.1, prosodic prominence is cross-linguistically one of the most common strategies to mark the Focus of a clause. As this Focus marking strategy is, however, not available in traditional written discourse, this may also partly explain – in addition to matters of formality – why this discourse mode is more likely to use subject forms in subject predicative complements (cf. Sections 9.1 and 13.1). Again, this possible explanation is not only in line with the observations made for the .za cleft data, but also with this study’s assumption that functional and particularly pragmatic considerations play an important role in the distribution of pronoun case forms in subject predicative complements – even though FOCUS
as operationalised in Section 6.1.2.5 has not turned out to be significant in this particular dataset.

In addition to SPOK, CMC is also very highly significant and also severely constrains the use of subject pronoun case forms in \textit{it}-clefts in the .za data (cf. Table 53). This constraining effect of CMC, which is signalled by its odds ratio of 0.20, confirms the observations made for the previous Web-derived datasets, where the same effect is attested for this factor in the respective analyses of the \textit{it}-cleft data (cf. Setions 10.3–12.3). Moreover, this result is also consistent with the expectations of the relevant literature assuming an overall lower probability of subject pronoun case form usage in subject predicative complement position in less formal varieties (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Quirk et al. 1985: 337-338). In analogy to SPOK, the impact of CMC cannot only be explained solely in terms of differences in formality but also by the higher number of strategies to highlight pragmatically salient discourse entities resulting from the fewer textual conventions CMC is supposed to adhere to, when compared to traditional written data. Hence, as outlined before (e.g. Sections 10.1 and 11.1), CMC possesses more options, such as capitalisation, to mark pragmatically salient clausal constituents. This implies that CMC is less dependent on the use of subject pronoun forms to mark the Focus in subject predicative complements.

Furthermore, two interactions are also retained in the minimal-adequate regression model (cf. Table 53). The first of these interactions maintained in this model is that between FIRST and SINGULAR. Although the effect of this interaction is not significantly influencing the distribution of pronoun case forms in \textit{it}-clefts, bootstrap validation tells us to retain it in the final regression model. This interaction can again be attributed to the particularities of the data compilation by means of a commercial search engine, which leads to certain skewing effects, one of which mainly affects the first person singular and in particular its subject form (cf. Sections 7.2.4.2, 10.3 and 11.3).

In addition to the by now familiar interaction between FIRST and SINGULAR, the interaction between FIRST and SPOK is not only retained in the minimal-adequate regression model but also identified as statistically highly significant. While each of the factors involved in this interaction by themselves markedly constrain the use of subject pronoun case forms in \textit{it}-clefts in the .za data, the co-occurrence of these factors fosters the use of subject pronoun case forms in this context, as its odds ratio of 2.12 indicates. Thus, first person subject forms are comparatively more likely to be observed in the spoken than in the written
it-cleft data of the .za dataset. This effect is actually quite difficult to account for, since it does neither correspond to the expectations and observations of the relevant literature (e.g. Harris 1981: 18-19; Huddleston and Pullum 2002: 459; Quinn 2005a: 133-135; Quirk et al. 1985: 337-338; Wales 1996: 95-96), nor to the assumptions derived from earlier observations of this study according to which the co-occurrence of several pragmatically salient factors and/or factors that allow for different or additional Focus marking strategies, such as FIRST and CLEFT or CLEFT and CMC, may actually decrease the likelihood of observing subject pronoun case forms. On the one hand, the effect observed for this interaction could suggest that the observations made so far as well as the Focus-oriented assumptions derived from them cannot be extended to the present dataset. This would also be in line with the observation that the role of Focus marking is not as evident in this dataset as in the previously examined corpora and datasets, which is indicated by the non-significance of FOCUS. On the other hand, the effect of this interaction could, however, also be a variety-specific or data-specific particularity, for which no general or universal explanation is available (cf. Bisang 2004: 25-27). This is at present difficult to assess and further research will be needed to shed more light on this issue.

The regression model itself is also very highly statistically significant ($\chi^2 = 169.92$, $p<0.001$), which once again indicates a strong correlation between the dependent and the independent variables included in the minimal-adequate regression model. Moreover, the model’s Nagelkerke’s $R^2$ value of 0.39 signals that the statistical model can account for 39 per cent of the variation observed in the dependent variable. As far as the predictive power of the model is concerned, Table 53 indicates that the baseline model is improved by more than 20 per cent by the regression model. While the former can only correctly predict the outcome of the dependent variable in 52.29 per cent of all instances, the latter, i.e. the minimal-adequate regression model, correctly predicts the outcome in 72.71 per cent of all cases.

The analysis of the it-clefts in the .za dataset has shown that As_SUBJ emphatically promotes the use of subject pronoun forms, which complies with both the expectations of the relevant literature (e.g. Erdmann 1978: 75-78; Quirk et al. 1985: 337-338) and the observations made for the corresponding analyses of all other datasets examined so far. In addition, this finding also supports this study’s assumption according to which functional factors play a prominent role in the distribution of pronoun case forms (cf. Section 5.2; H2).
Furthermore, the inhibiting effect of FIRST attested for this dataset is also in line with the results of all previously conducted it-cleft analyses in this study as well as with the findings of earlier studies (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96) and could be again accounted for by the different prototypical referential functions of first and third person pronouns and the implications deriving from them. The fostering effect of SINGULAR is in line with some accounts expecting such an impact (Sobin 1997: 334) as well as with the .uk and the .au cleft data, where such an effect can also be observed. However, this finding does not match the results obtained for the BNC and the .ie data where SINGULAR is not significant, and it even contradicts the findings of the COCA where SINGULAR exerts a constraining force on the use of subject forms. Complying both with the results of the BNC and COCA as well as the expectations of the literature, SPOK clearly inhibits the use of subject forms, which can, however, not be attested for the other Web-derived datasets analysed so far, where SPOK is not significant. Similar to SPOK, CMC also clearly inhibits the use of subject pronoun case forms in it-clefts in the .za data. For both SPOK and CMC, the inhibiting effect cannot only be attributed to the lower degree of formality of these modes of discourse when compared to traditional written data, but also to the higher number of highlighting devices, such as prosodic prominence or its imitation by means of capitalisation, spoken and computer-mediated communication can make use of.

13.4 The Distribution of Pronoun Case Forms in the South African (.za) Internet Data: Interim Summary

The results of this chapter again indicate remarkable differences in the distribution of pronoun case forms between the different datasets. From an inter-varietal perspective this means that in the South African data, subject forms are – particularly in it-clefts – considerably more likely to occur than in the .uk and the .au data but are less likely to be used than in the .ie data, the BNC and the COCA. This indicates again cross-varietal differences with regard to the distribution of pronoun case forms in subject predicative complements. From an intra-varietal perspective, the analysis of the .za data reveals again marked differences in the distribution of pronoun case forms between it-clefts and it BE sentences, which are difficult to reconcile with accounts explaining the distribution of
pronoun case merely with the help of the position or the class membership of a given pronoun (cf. Sections 3.2 and 3.3).

Also difficult to accommodate with the positional and the weak vs. strong pronoun approach is the fact that the clearly functional factors CLEFT and AS_SUBJ exert a very strong influence on the distribution of pronoun case forms and are, as in all previously examined datasets, the factors with the highest odds ratio in their respective regression models (cf. Sections 13.1 and 13.3). The impact of CLEFT and AS_SUBJ corroborates, however, this study’s assumption that functional factors are very important in the distribution of pronoun case forms (cf. Section 5.2; H2).

Although the results for CLEFT and AS_SUBJ correspond to the observations made for all other datasets examined so far in that they foster the use of subject pronoun case forms, the same does not hold for FOCUS, which is not significant in the .za data. Thus, although the necessary condition for us to assume that Focus plays a role in the distribution of pronoun case forms is met, the sufficient condition is not achieved (cf. Section 6.1.2.5). Thus, Focus marking seems not to be as important in the distribution of pronoun case forms in the .za data as in all other datasets examined so far – at least not in the way it is operationalised in this study.

The analyses of the superordinate category of subject predicative complements and of the it-cleft subset (cf. Sections 13.1 and 13.3) have demonstrated that FIRST constrains the use of subject forms in these datasets, which is consistent with the results of the .uk data, .au data, .ie data, BNC data and with parts of the COCA, where FIRST even constrains the use of subject forms in the it BE data. Again this inhibiting effect can be accounted for by means of the prototypically different referential functions of first and third person pronouns. These may result in the fact that if the pragmatically more salient first person pronouns coincide with CLEFT (Siewierska 2004: 5-8), this might suffice to mark a pronoun’s referent as the most important piece of information in a clause and further Focus marking by means of subject pronouns may not be necessary. Although FOCUS as such is not significant, this possible explanation can still account for the differences observed in the data and perfectly complies with the general assumptions of this study (cf. Chapter 4). SINGULAR fosters again the use of subject pronoun forms in both it-clefts and the superordinate subject predicative complement category, which is in accordance with the expectations of some studies (Sobin 1997: 334) and also the observations made for the .uk data and .au data. Its promoting
effect is, however, in contrast to the results of other studies (e.g. Quinn 2005a: 134-135) and to those obtained for the BNC and the .ie dataset, where this factor is eliminated, and to those of the COCA, where SINGULAR even constrains the use of subject forms.

With regard to the MODE OF DISCOURSE variable, the analysis of the .za data shows that SPOK inhibits the use of subject pronoun case forms, particularly in the it-cleft and the superordinate subject predicative complement categories. This is in line with the COCA results as well as with accounts assuming a correlation between lower degree of formality and lower likelihood of observing subject forms but it is not in accordance with the observations made for the other Web-derived datasets nor with the BNC, where we observe a very differentiated effect of SPOK depending on the respective sentence type. In a similar vein, CMC also constrains the use of subject pronoun case forms in the it-cleft and the superordinate subject predicative complement categories, which matches the observations made for the other Web-derived datasets. While the inhibiting force of both SPOK and CMC could be explained again by the common assumption that less formal varieties are less likely to exhibit subject forms in subject predicative complements (e.g. Harris 1981), this view can be complemented by the stance taken here that spoken discourse and CMC also possess more Focus marking strategies than written discourse, which may also partly account for the differences observed. Although Focus as operationalised in Section 6.1.2.5 is not significant, the examples in 13.1 still suggest that South African English may particularly highlight focussed pronouns, for example by means of capitalisation, which still corroborates a Focus-oriented approach to explain this mode of discourse difference.
14 The Distribution of Pronoun Case Forms in the Indian (.in) Internet Data

Finally, Chapter 14 provides an analysis and discussion of the distribution of pronoun case forms in the data gathered from the Indian top-level domain .in (henceforth .in data). Following the structure of the preceding six chapters, Section 14.1 presents the result obtained for the superordinate subject predicative complement category. Secondly, Section 14.2 focuses on the *it BE* sentences in the .in data, before Section 14.3 examines the distribution of pronoun case forms as well as the factors influencing it in the *it*-clefts of this dataset. Finally, Section 0 concludes this chapter with a summary of the major findings.

14.1 The Distribution of Pronoun Case Forms in Subject Predicative Complements in the Indian (.in) Internet Data

As Table 54 indicates, 492 subject predicative complement tokens constitute the .in dataset and thus are the input for the following multivariate analysis. This means that this dataset is the second smallest examined in the present study with only the Irish dataset discussed in Chapter 12 consisting of fewer tokens.

<table>
<thead>
<tr>
<th>SUBJECT PREDICATIVE COMPLEMENTS (.IN)</th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>492 (100%)</td>
<td>426 (86.59%)</td>
<td>66 (13.41%)</td>
</tr>
</tbody>
</table>

*Table 54: Subject Predicative Complements in the .in Data: The Token Numbers*

As far as the distribution of pronoun case forms is concerned, Table 54 demonstrates that 86.59 per cent of all subject predicative complements in the .in data are used with a subject pronoun. This also implies, of course, that only 13.41 per cent of all subject predicative complements are used with an object pronoun. When comparing these shares to those observed for the other datasets examined so far, the share of subject forms observed for this dataset is by far the highest one, since it surpasses the Irish one by more than 15 per cent (cf. Section 12.1) and it is even more than four times as high as the shares of subject
forms attested for the .uk data and .au data (cf. Sections 10.1 and 11.1). This very high share of subject pronoun forms is also very remarkable in view of the many theoretical accounts assuming very high shares of object pronoun case forms in subject predicative complements (e.g. Harris 1981: 17-20; Quinn 2005a: 138, 242-248). Although the share of subject forms in the subject predicative complement category is co-determined by the composition of this category in terms of the proportions of it BE sentences and it-clefts subsumed under this heading, the results in Table 54 superficially suggest a more extensive use of subject forms in the .in data than in all other datasets examined so far. However, whether or not this preliminary observation holds can only be confirmed after the separate analysis of both it BE sentences and it-clefts in Sections 14.2 and 14.3.

Table 55 indicates that four main effects are retained in the minimal-adequate regression model, while two other factors are eliminated in the course of the model-building process. The first of these non-significant factors is SINGULAR. The elimination of SINGULAR is in line with the results of the BNC as well as with those of the .ie data. It is, however, neither in accordance with the inhibiting effect attested for SINGULAR in the COCA, nor with the promoting effect observed for this factor in the .uk data, .au data and .za data, and it does also not correspond to the expectations of some accounts predicting a fostering effect of SINGULAR on the use of subject forms (Sobin 1997: 334).

The second factor excluded from the minimal-adequate regression model is Focus (cf. Table 55). Although the exclusion of Focus is in line with the elimination of this factor in the .za data, it is neither consistent with its strong impact observed in all other datasets nor with the hypothesis of this study assuming a reanalysis of subject pronoun case forms as Focus markers in subject predicative complements (cf. Section 5.2; H3).57

---

57 In addition, the following interactions were also eliminated from the minimal-adequate model: Spok*First, Spok*Singular, Spok*Cleft, Spok*Focus, CMC*First, CMC*Singular, CMC*Cleft, CMC*Focus, First*Singular, First*Cleft, First*Focus, Singular*Cleft, Singular*Focus.
In the .in data, CLEFT is again very highly significant and emphatically promotes the use of subject pronoun case forms, which is signalled by its odds ratio of 84.75. Translating this odds ratio into a percentage value, this means that the probability of observing a subject form is 8375 per cent higher in an it-cleft than in an it BE sentence in the .in data. This fostering effect is in accordance with the results obtained for this factor in all other corresponding datasets and with studies arguing for clear differences between it-clefts and it BE sentences as far as their use of pronoun case forms is concerned (e.g. Erdmann 1978; Maier 2013).

Secondly, Table 55 identifies FIRST as very highly significant and as clearly constraining the use of subject forms in subject predicative complements, as its odds ratio of 0.22 indicates. This effect complies with the results observed for five of the six preceding subject predicative complement analyses. Only in the BNC, a significantly constraining effect of FIRST is restricted to the it-cleft section (cf. Chapter 8). Although some accounts expect a promoting effect of FIRST particularly in the subclass of it BE sentences (e.g. Quinn 2005a: 246), which can, however, not be confirmed for the .in data, other accounts expect an inhibiting force of this factor in the domain of it-clefts (e.g. Wales 1996: 95-96), which is in accordance with the present observation. However, the actual effect of FIRST on the individual subtypes of subject predicative complements can only be exhaustively determined when examining each sentence type separately. As far as a possible explanation for the difference between first and third person pronouns in terms of their likelihood of using a

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>4.44</td>
<td>***</td>
<td>84.75</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.51</td>
<td>***</td>
<td>0.22</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.21</td>
<td>*</td>
<td>3.35</td>
</tr>
<tr>
<td>CMC</td>
<td>-1.72</td>
<td>***</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Model $X^2$ 159.34 (***)

$R^2$ 0.50

% Correctly Predicted 90.65

% Baseline 86.59

N 492

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

Table 55: Subject Predicative Complements in the .in Data: Logistic Regression Results
subject form is concerned, the inhibiting effect of FIRST can again be partly attributed to the fundamental differences between these two persons (cf. Section 6.1.2.2). In analogy to the previous chapters, we can also argue for the .in data that the referents of first person pronouns are automatically more salient. Third person pronouns, however, are often used phorically, which means their referents are not as easily identifiable and thus may need more highlighting, for example by means of a subject pronoun form, than the referents of first person pronouns (e.g. Section 8.3 and 9.1). Although FOCUS as operationalised in Section 6.1.2.5 is not significant in the present regression model, this possible explanation is in line with both the present data as well as the hypotheses that functional factors in general and Focus marking in particular are very important in the distribution of pronoun case forms (cf. Section 5.2; H2 and H3).

The third significant factor in Table 55 is SPOK. With an odds ratio of 3.35, SPOK clearly promotes the use of subject pronoun case forms. This finding is rather surprising since it is neither in line with much of the relevant literature (e.g. Biber et al. 1999: 335-336; Harris 1981: 18-19; Huddleston and Pullum 2002: 459), nor with most of the databases analysed so far. Whereas SPOK is not significant in the corresponding regression models analysing the .uk, .au and .ie datasets (cf. Sections 10.1, 11.1 and 12.1), it significantly constrains the use of subject forms in the COCA and .za data (cf. Section 9.1 and 13.1). Only in the BNC, SPOK is also marginally significant and in tendency fosters the use of subject pronoun case forms in subject predicative complements (cf. Section 8.1). Thus, these highly heterogeneous results obtained for SPOK across different datasets suggest that this factor is subject to cross-varietal variation and that assumptions according to which subject forms are generally less likely to occur in spoken discourse are too simplistic (e.g. Harris 1981).

Finally CMC is also retained in the minimal-adequate regression model (cf. Table 55). As its odds ratio of 0.18 indicates, CMC clearly inhibits the use of subject forms in the .in data. This observation matches the results for this factor as observed in all other corresponding Web-derived datasets. Furthermore, this finding also complies – in contrast to that for SPOK – with the expectations of the literature assuming that subject forms in subject predicative complements are more likely to be attested in rather formal varieties (e.g. Biber et al. 1999: 335-336; Harris 1981: 18-19; Huddleston and Pullum 2002: 459). However, while these accounts explain this difference merely in terms of formality, the present study additionally takes the greater repertoire of Focus marking devices into
consideration, which CMC possesses when compared to traditional written data (e.g. Sections 11.1 and 13.1). This point of view is not only reconcilable with the analysis of the present data but also with the Focus-oriented perspective of this study, even though Focus as operationalised in this study is eliminated from the minimal-adequate model applied to this dataset (cf. Table 55).

Table 55 also indicates that there is a strong correlation between the dependent and the independent variables retained in the minimal-adequate regression model, since the model itself is very highly statistically significant ($\chi^2 = 159.34$, $p<0.001$). Moreover, the regression model can account for 50 per cent of the variation observed in the dependent variable, which is more than enough to render it substantially significant (cf. Szmrecsanyi 2006: 55). With regard to its predictive power, the regression model correctly predicts the outcome of the dependent variable in 90.65 per cent of all cases, which is actually the highest share observed in this study for a regression model analysing the distribution of pronoun case forms in the superordinate subject predicative complement category (cf. Chapters 8–13). Hence, the applied regression model can even improve the baseline model, which is already able to account for 86.59 per cent, by more than four per cent.

To conclude, the analysis of the subject predicative complements in the .in data demonstrates that the non-significance of Focus in the .za is not a particularity but can also be accounted for other datasets. Still, the elimination of this factor is unexpected since it runs counter not only to the expectations of this study (cf. Chapter 4), but also to the results of the other five datasets. In contrast to Focus, Cleft is again significantly and very strongly fostering the use of subject pronoun case forms in the .in data. This is in accordance with all previous analyses of the superordinate subject predicative complement categories as well as with studies arguing for marked differences between it-clefts and it BE sentences (e.g. Erdmann 1978). Furthermore, the strong impact of clearly functionally motivated factors seriously undermines accounts boiling down the distribution of pronoun case distribution in subject predicative complements solely to a matter of the position or class-membership of a pronoun (cf. Sections 3.2 and 3.3). Also in analogy to most other datasets, First markedly inhibits the use of subject forms. This constraining effect can again be accounted for by the differences inherent in the referential nature of first and third person pronouns. Rather surprisingly, however, SPOK clearly fosters the use of subject pronoun case forms in the .in data. This is unexpected given the results obtained for SPOK in most other datasets as well as
the expectations of the literature assuming an inhibiting effect of this factor. Very much in line with all other corresponding Web-derived datasets is again the effect of CMC, which constrains the use of subject pronoun case forms in the subject predicative complements in the .in data. As before, the impact of this predictor can be explained with the wider variety of Focus marking strategies CMC can employ in order to mark pragmatically salient constituents (cf. 10.1).

14.2 The Distribution of Pronoun Case Forms in it BE Sentences in the Indian (.in) Internet Data

As Table 56 indicates, the .in dataset only comprises 29 instances of it BE sentences. This means that only 6.76 per cent of all subject predicative complements in the .in data belong to this subcategory (cf. Section 14.1). This share is considerably lower than those of all other datasets examined so far. Even in the .ie data, which exhibits the second-smallest share of it BE sentences, the proportion of it BE sentences amounts to 20.47 per cent of all subject predicative complements. In the COCA, even 66.76 per cent of all subject predicative complements are it BE sentences (cf. Section 9.2). Thus, this sentence type seems to be considerably less frequent in the Indian English data than in all other datasets or varieties examined so far. With regard to the distribution of subject and object pronoun case forms, Table 56 demonstrates that 10.34 per cent of all it BE sentences are used with a subject pronoun form, while the remaining 89.66 per cent are used with an object pronoun case form. Although the absolute numbers on which these percentages are based are very low, which makes it difficult to assess the robustness of these proportions, the share of subject pronoun case forms attested in the .in data is comparatively high, since only the .ie data exhibits a larger share of subject forms in the in the it BE sentence subset (cf. Section 12.2). In addition, Table 56 indicates again the evident differences in terms of pronoun case distribution between the superordinate subject predicative complements category, where more than 86 per cent of all forms are used with a subject form (cf. Section 14.1), and its subset of it BE sentences, where only 10.34 per cent of all it BE sentences exhibit a subject form (cf. Table 56).
However, the low number of *it BE* sentences observed in the .in dataset poses not only difficulties for the comparison of these results to the corresponding analyses of all other datasets, but even more so for the statistical analysis of the *it BE* sentences in the .in data. In the relevant literature, a minimum of 50 tokens is considered to be necessary to be able to obtain meaningful results from a binary logistic regression model (e.g. Rese 2000: 107). Thus, according to these model requirements, it is not possible to analyse the *it BE* sentences in the .in data with the help of a binary logistic regression. Nevertheless, it was still tried to do so, but no factor could be identified as significantly influencing the distribution of pronoun case forms in the *it BE* sentences in the .in data. Moreover, bootstrap validation also did not retain any factor. Thus, it is not possible for the present dataset to arrive at any conclusions as far as the distribution of pronoun case forms is concerned by means of a multivariate statistical analysis.

If we take, however, the time to look at the three instances in which a subject form is actually used in an *it BE* sentence in the .in dataset, we make a very remarkable observation:

(143) The Three *it BE* Sentences with Subject Pronoun Forms in the .in Data
a. Satgurunathan Oduvar has a rare gift for this unique music. I could see him becoming popular like a Bollywood star, if only his genius could be exposed to a larger audience. If anyone has the ability to revive and re-inspire this fading tradition, it is he. (.in/it is he/14.07.2008)

b. He brought me to Mumbai and taught me everything. And if there is someone who I am missing today, it is he. (.in/it is he/14.07.2008)

c. [...] if somebody touches and asks who is this, it is I. (.in/it is I/14.07.2008)

As can be seen in (143a)–(143c), all three *it BE* sentences in which a subject form is used as pronominal complement are preceded by an *if*-clause in the .in data. Furthermore, the subject or subject complement of these *if*-clauses is either a non-assertive pronoun in (143a) or an assertive pronoun in sentences (143b)–(143c). This means that sentence (143a) is a clear example of the first type of particularly focussed *it BE* sentences identified in Section 6.1.2.5. In addition, the sentences in (143b) and (143c) can be considered as slight variations of this class (cf. Section 6.1.2.5). Although the sentences in (143b) and (143c) are used with an assertive instead of a non-assertive pronoun in the dependent *if*-clause, the pronouns in

<table>
<thead>
<tr>
<th></th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IT BE SENTENCES (.in)</strong></td>
<td>29 (100%)</td>
<td>3 (10.34%)</td>
<td>26 (89.66%)</td>
</tr>
</tbody>
</table>

*Table 56: it BE Sentences in the .in Data: The Token Numbers*
the *it BE* sentences can still be considered as particularly focussed, since all other characteristics that set the pronouns in the first type of particularly focussed *it BE* sentences apart from their unmarked counterparts apply to the pronouns in the *it BE* sentences in (143b) and (143c) (cf. Section 6.1.2.5). Thus, all three *it BE* sentences in which a subject form is used in the .in dataset can be considered as particularly focussed. Hence, although the statistical significance of Focus cannot be proven with certainty due to a lack of data, the examples in (143) suggest that that this factor may play an important role in the distribution of pronoun case forms if only enough data points were available. The real extent to which Focus may influence the distribution of pronoun case forms in Indian English can, however, only be resolved in future studies by compiling even bigger databases in order to obtain enough data allowing for robust multivariate analyses.

In sum, a multivariate analysis of the *it BE* sentences in the .in data is not possible, since this study does not possess enough data points to do so. This, however, is not only due to the overall size of the Indian subject predicative complement dataset, but it is rather due to its composition. *It BE* sentences are considerably less often used in the .in dataset than in all other datasets analysed in this study, which in itself constitutes a particularity of this dataset or variety. Furthermore, although a statistical analysis is not possible, the few *it BE* sentences that exhibit a subject form as subject predicative complement in the .in data can all be considered as particularly focussed. This corroborates, even though it cannot be backed by robust quantitative evidence, one of this study’s central hypotheses according to which subject pronouns may have been reanalysed as Focus markers in subject predicative complements (cf. Section 5.2; H3).

14.3 The Distribution of Pronoun Case Forms in *it*-Clefts in the Indian (.in) Internet Data

While the analysis of the *it BE* sentences in the .in data suffers from a lack of data, this should not be the case for the *it*-clefts in the .in data since 463 tokens belong to this subcategory, as Table 57 indicates. As Table 57 also shows, 91.36 per cent of the *it*-clefts in the .in data are used with a subject form as focal pronoun, whereas only 8.64 per cent of all *it*-clefts are used with an object form. Although the number of *it BE* sentences in the .in data
is very small (cf. Section 14.2), the comparison of *it BE* sentences and *it*-clefts suggests that the distribution of pronoun case forms is nearly complementary in this dataset. While roughly ten per cent of all *it BE* sentences exhibit a subject form as predicative complement, more than 90 per cent of *it*-clefts are used with a subject pronoun case form as focal pronoun in the .in data (cf. Section 14.2). Comparing the share of subject forms of this *it*-cleft dataset to those of the other datasets analysed so far, the share of subject forms attested for the .in data is clearly the highest one, since it even surpasses the ratio of subject focal pronouns observed in the .ie dataset by ten per cent (cf. Section 12.3). Indeed, the use of subject pronouns in *it*-clefts seems to be nearly categorical in the .in data. This is particularly true if we take into consideration that the share of subject forms in a Web-derived dataset is expected to be lower than in a traditional closed corpora due to the particularities of the data compilation (cf. Section 7.2.4.2). This near-categorical use of subject forms in *it*-clefts in Indian English is also attested by an analysis of all cleft sentences in the Indian component of the ICE, which shows that all *it*-clefts with a case-sensitive focal pronoun exhibit a subject form (Beal in prog.). Moreover, this very marked preference of subject pronoun forms in the *it*-clefts of the .in dataset does not only support studies arguing for marked differences between *it BE* sentences and *it*-clefts in the distribution of pronoun case forms (e.g. Erdmann 1978), but it also seriously undermines accounts expecting mainly or only object pronouns in the focal pronoun position of *it*-clefts (e.g. Harris 1981; Quinn 2005a).

<table>
<thead>
<tr>
<th></th>
<th>TOTAL N</th>
<th>SUBJECT FORMS N</th>
<th>OBJECT FORMS N</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT-CLEFTS (.IN)</td>
<td>463 (100%)</td>
<td>423 (91.36%)</td>
<td>40 (8.64%)</td>
</tr>
</tbody>
</table>

Table 57: *it*-Clefs in the .in Data: The Token Numbers

Table 58 below presents the results obtained for the multivariate analysis of the *it*-cleft data. This analysis identifies three factors as significantly influencing the distribution of pronoun case forms. Among the excluded factors is again SINGULAR, the non-significance of which is in line with the BNC and the .ie data, but neither with the constraining impact of SINGULAR observed in the COCA nor with the fostering force of this factor attested in the .uk, .au and .za databases, nor with accounts assuming a promoting effect of this factor (Sobin 1997: 334). Secondly, CMC is also excluded from the minimal-adequate regression model.
This is in contrast to all other Web-derived datasets analysed so far, in which CMC consistently inhibits the use of subject pronoun case forms in *it*-clefts (cf. 10.3–13.3).58

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>P-VALUE</th>
<th>ODDS RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>3.44</td>
<td>***</td>
<td>31.33</td>
</tr>
<tr>
<td>FIRST</td>
<td>-1.98</td>
<td>***</td>
<td>0.14</td>
</tr>
<tr>
<td>SPOK</td>
<td>2.55</td>
<td>**</td>
<td>12.82</td>
</tr>
</tbody>
</table>

**Model \( \chi^2 \)** 59.56 (***)

\(R^2\) 0.27

\% **Correctly Predicted** 91.79

\% **Baseline** 91.36

\(N\) 463

+ significant at p<0.1, * significant at p<0.05, ** significant at p<0.01, *** significant at p<0.001

*Table 58: *it*-Clefts in the .in Data: Logistic Regression Results*

In analogy to all other *it*-cleft analyses conducted so far, Table 58 indicates that AS_SUBJ is not only very highly significant, but also again the factor with the highest odds ratio retained in the minimal-adequate regression model applied to the dataset. Its odds ratio of 31.33 indicates that the probability of observing a subject form as focal pronoun is more than 3000 per cent higher in *it*-clefts in which the focal pronoun is in co-reference with the subject of the following dependent clause. This effect is in line with the observations made for AS_SUBJ in all other *it*-cleft datasets examined so far as well as with accounts asserting such a promoting influence of this factor (e.g. Erdmann 1978: 76-78; Huddleston and Pullum 2002: 459). Furthermore, the fact that a clearly functional factor has been identified as consistently significantly influencing the distribution of pronoun case forms supports also the general view of this study assuming that functional factors are very important in the distribution of pronoun case forms in varieties of English (cf. Section 5.2; H2).

Table 58 also identifies FIRST as very highly significant. With an odds ratio of 0.14, this factor severely constrains the use of subject pronoun forms in *it*-clefts in the .in data. This

---

58 The following interactions were also eliminated in the model-building process: SPOK*FIRST, SPOK*SINGULAR, SPOK*AS_SUBJ, CMC*FIRST, CMC*SINGULAR, CMC*AS_SUBJ, FIRST*SINGULAR, FIRST*AS_SUBJ, SINGULAR*AS_SUBJ.
finding is not only consistent with this factor’s effect as observed for all other corresponding
datasets analysed so far, but this inhibiting force also confirms the findings and assumptions
of earlier studies with regard to the impact of FIRST (e.g. Quinn 2005a: 133-135; Wales 1996:
95-96). As has become customary by now, this difference between first and third person
pronoun forms with regard to their likelihood of being used in their subject form as focal
pronouns of it-clefts can be again explained with their prototypical referential statuses and
the differences arising from them. As discussed above (e.g. Section 8.3), this study assumes
that third person pronouns may be more likely to receive additional highlighting by means of
a subject pronoun case form to indicate their status as the Focus of a clause (cf. Dik 1989:
278; Siewierska 2004: 5-8). While this mode of explanation again does not only nicely fit the
data, it is also be fully consistent with one of the central hypotheses of this study assuming a
re-functionalisation of subject pronoun case forms in subject predicative complements in
general and it-clefts in particular (cf. Section 5.2; H3).

Thirdly, the minimal-adequate regression model identifies SPOK as highly significant.
As its odds ratio of 12.82 indicates, SPOK emphatically fosters the use of subject pronoun
case forms in it-clefts in the .in data (cf. Table 58). This finding is neither in line with the
observations made for the .uk, .au and .ie datasets, where SPOK is not significant, nor with
the results obtained for the BNC, the COCA and the .za dataset, where SPOK markedly
constrains the use of subject pronoun case forms in it-clefts. Moreover, this finding is also
not consistent with the expectations and findings of accounts assuming a clearly inhibiting
effect of this factor on the use of subject pronoun case forms (e.g. Biber et al. 1999: 335-
336; Harris 1981: 19). In some of the preceding sections (e.g. Sections 9.3 and 13.3), it has
been assumed that subject pronoun case forms may be less likely to occur in spoken it-cleft
data because spoken discourse can also use prosodic prominence as Focus marking strategy,
which may be sufficient to mark a pronoun or its referent as the Focus of a clause. In the .in
data, however, this possible co-occurrence of two highlighting strategies, i.e. it-cleft and
prosodic prominence, seems not to inhibit a third one, i.e. the use of subject pronoun forms,
but even to foster it, which seems to be a peculiarity of this variety or dataset.

As far as the statistical model itself is concerned, Table 58 indicates that the minimal-
adequate regression model is very highly significant ($\chi^2 = 59.56, p<0.001$). Furthermore, the
model’s Nagelkerke’s $R^2$ value of 0.27 illustrates that the model can account for 27 per cent
of the variation observed in the dependent variable, which clearly renders the model
substantially significant (Szmrecsanyi 2006: 55). With regard to its predictive power, the applied regression model can only slightly improve the baseline model. While the latter can already account for 91.36 per cent of all cases, the minimal-adequate regression model correctly predicts the outcome of the dependent variable in 91.79 per cent of all cases.

In sum, the markedly promoting effect of the predictor As_SUBJ is in line with both the results obtained for all other datasets and with the accounts and studies predicting or attesting such an effect on the distribution of pronoun case forms in it-clefs (e.g. Erdmann 1978: 75-78; Quirk et al. 1985: 337-338). This observation also ties in with the general perspective of this study assuming a strong influence of functional factors on the distribution of pronoun case forms in subject predicative complements (cf. Section 5.2; H2). The results obtained for FIRST also correspond to those obtained for all other it-cleft analyses conducted so far, since FIRST again emphatically constrains the use of subject forms in it-clefs, which can be accounted for again with the help of the fundamental differences between first and third person pronouns as far as their prototypical referential functions are concerned. In addition, the inhibiting effect of FIRST is also in accordance with the findings of previous studies (e.g. Quinn 2005a: 133-135; Wales 1996: 95-96). Moreover, the analysis has revealed an emphatically promoting effect of SPOK in the present dataset, which is neither in line with the results obtained from the previous it-cleft analyses, nor with the expectations voiced in the literature (e.g. Biber et al. 1999: 335-336; Quirk et al. 1985: 337-338). Indeed, the direction of this effect as observed in the present dataset may be a variety-specific particularity.

14.4 The Distribution of Pronoun Case Forms in the Indian (.in)

Internet Data: Interim Summary

The analysis of the Indian data has once again demonstrated the marked differences that can be observed between the individual datasets in terms of pronoun case distribution. The first very noteworthy particularity of the .in data is the scarcity of it BE sentences attested for this dataset, which even impedes a multivariate analysis of this sentence type (cf. Section 14.2). This low frequency of it BE sentences cannot, however, be explained by the overall size of the dataset alone, since not only the absolute number of tokens but also the relative
share of it BE sentences included in the superordinate subject predicative complements category is markedly lower than those of all other datasets examined in this study. The second very remarkable finding in the .in dataset is the near-categorical use of subject pronoun forms in it-clefs. Although this dataset suffers from the same skewing effects in favour of the object pronouns as all other Web-derived datasets, more than 91 per cent of all it-clefs in the .in data are used with a subject pronoun form as focal pronoun.

This high share of subject forms is again very difficult to accommodate with approaches focussing particularly on the position of the pronoun in a clause or the putative membership to a certain pronoun class to account for the distribution of pronoun case forms (cf. Sections 3.2 and 3.3). Similarly, while these accounts also struggle to account for the strong impact the functional factors CLEFT and AS_SUBJ have on the distribution of pronoun case forms in their respective models, the effect of these factors is, however, in accordance with this study’s perspective assuming a very important role of functional aspects in the distribution of pronoun case forms (cf. Section 5.2; H2).

As far as the rather syntactic and discourse-pragmatic factors are concerned, the observations made for CLEFT and AS_SUBJ, which again markedly foster the use of subject pronoun forms, are also in line with results obtained for these factors in all other datasets analysed for this study. However, as is the case with the .za dataset, FOCUS is again not significant in the .in data. Thus, although the necessary condition is met for us to conclusively assume that Focus marking plays an eminent role in the distribution of pronoun case forms, the sufficient condition is not fulfilled (cf. Section 6.1.2.5). The non-significance of FOCUS in this dataset can be largely attributed, however, to the low frequency of it BE sentences in this dataset. It may be even assumed that FOCUS turns out significant given a larger database, since those few it BE sentences that actually exhibit a subject form in the .in data can all be considered as particularly focussed (cf. Sections 6.1.2.5 and 14.2).

With regard to the variables PERSON and NUMBER, FIRST clearly constrains the use of subject pronoun forms in the subject predicative complement and it-cleft categories of the .in data. This is line with the results obtained for the corresponding subject predicative complement and it-cleft categories of the BNC and all other Web-derived datasets. It is also partly in line with the COCA, where FIRST has even an inhibiting effect on the it BE sentences (cf. Section 9.2). In the present study, this more or less systematic difference between first and third person pronouns has been explained with the prototypically different referential
functions of these persons. Since the referents of third person pronouns may not be automatically as salient as the referents of first person pronouns, the referents of third person pronouns may be more likely to receive additional highlighting by means of a subject form to become more easily accessible as the Focus of a clause. Particularly, if a pronoun is used in focal position of an it-cleft, this may be sufficient to mark the referent or referents of a first person pronoun as Focus, which would account for why this difference is particularly discernible in this sentence type (e.g. cf. 8.3). Thus, although the relevance of Focus cannot be proven statistically due to a lack of data, this line of reasoning is not only consistent with the data but also again with the central hypotheses of this study (cf. Section 5.2). SINGULAR is not significant in this dataset, which is in accordance with the BNC and the .ie dataset as well as with the findings of some previous studies (e.g. Quinn 2005a: 134-135). However, the non-significance of SINGULAR in the .in data is neither in line with the COCA, where it has partly an inhibiting effect, nor with the .uk, .au and .za datasets, where SINGULAR significantly promotes the use of subject forms, which is also expected by some accounts (Sobin 1997: 334).

As far as the MODE OF DISCOURSE variable is concerned, the .in dataset is also very particular: In contrast to all other Web-derived datasets, where this factor has an emphatically constraining effect in both the superordinate and the it-cleft category, CMC is only significant in the subject predicative complement category, where it does, however, inhibit the use of subject forms. Moreover, SPOK is clearly fostering the use of subject pronoun forms in the .in data. While this complies partly with this factor’s effect as observed in the BNC, it is not in accordance with the inhibiting effect attested for SPOK in the COCA and the .za data, nor with the non-significance of SPOK in the .uk, .au and .ie datasets. Additionally, the fostering effect of SPOK and the non-significance of CMC in the it-cleft subset does also not match the assumptions of the literature according to which less formal varieties are less likely to use subject pronoun case forms in subject predicative complements (e.g. Harris 1981; Quirk et al. 1985: 337-338). Thus, such generalisations cannot be confirmed in view of the present results.
15 General Discussion

In the preceding seven chapters, this study has examined the distribution of pronoun case forms in subject predicative complements in seven corpora and Web-derived datasets from six different regional varieties of English, i.e. British, American, Australian, Irish, South African and Indian English. By means of multivariate statistical modelling, this study has already provided unprecedented insights into the distribution of pronoun case forms in English in general and the varieties under examination in particular.

However, while the preceding Chapters 8–14 have presented and thoroughly discussed the results obtained for each of these datasets – even in relation to the results of the other datasets – a more general or bird’s eye perspective on the results, particularly with regard to the research questions and hypotheses outlined in Chapter 5, still has to be provided. Hence, this chapter reassesses and discusses the major findings obtained from the analysis conducted in the previous seven chapters with regard to the Focus-oriented approach introduced in Chapter 4 and the central hypotheses of this study outlined in Section 5.2. Thus, each subsection of this chapter particularly attends to one of the central hypotheses of this study following the order of these hypotheses as given in Section 5.2.

15.1 Position and Pronoun Class as Determinants of the Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

As outlined in Sections 3.2 and 3.3, much of the current linguistic theory expects mainly or only object pronoun forms to occur in subject predicative complement position. To begin with, there are positional accounts that assume that the distribution of pronoun case forms is more or less exclusively determined by the position of the pronoun relative to the finite verb (cf. Section 3.2). In particular, these accounts propose that subject pronoun forms are restricted to the noun phrase slot in declarative sentences which immediately precedes the finite verb. Thus, subject pronoun case forms in subject predicative complements are unexpected. As discussed in Section 3.2, variation in the use of pronoun case forms in that
account is attributed mainly to differences in formality (e.g. Burridge 2004: 1118; Jespersen 1933: 136; Quirk et al. 1985: 337-339).

Furthermore, there are weak vs. strong pronoun accounts that focus on the distinction between different pronoun classes rather than on the identification of certain areas in which specific pronoun forms are supposed to occur (cf. Section 3.3), even though they make the same predictions as the positional account with regard to the distribution of pronoun case forms in subject predicative complements. In analogy to languages such as French, this weak vs. strong pronoun approach assumes that the former subject pronoun case forms have become or are in the process of becoming clitics and thus are restricted to the noun phrase slot immediately preceding the finite verb. The former object pronoun case forms, in contrast, have become or are in the process of becoming strong pronouns, which occur in all other positions and contexts (e.g. Harris 1981; Sweet 1875: 495). In this approach, variation in the use of pronoun case forms is explained mainly in terms of a dynamic transition from a formerly case-governed pronoun system to a system of clitic and non-clitic pronoun forms (cf. Section 3.3). Hence, both approaches assume not only that functional factors do not play an important role in Present-Day English but also that we should observe only or mainly object pronoun case forms in subject predicative complements, since this noun phrase slot is either clearly in “object territory” or requires a “strong pronoun” (cf. Sections 3.2 and 3.3).

However, the results obtained from the analyses of the data in Chapters 8–14 are difficult to reconcile with the predictions of these two approaches and actually seriously challenge the positional and the weak vs. strong pronoun accounts. In particular, we have observed very pronounced differences between it-clefts and it BE sentences in the distribution of pronoun case forms in each and every one of the seven datasets. Specifically, it-clefts are much more likely to be used with a subject pronoun form than it BE sentences. In some datasets such as the BNC, the COCA and the .ie dataset, subject pronoun case forms are used in the majority of it-clefts (cf. Sections 8.3, 9.3 and 12.3). In Indian English, the use of subject forms as focal pronouns in it-clefts seems nearly categorical (cf. Section 14.3). These findings have also been confirmed by the results obtained from statistical analyses of the datasets. The multivariate analyses of the superordinate subject predicative complement categories have identified the variable CONSTRUCTION, which has been operationalised via CLEFT (cf. Section 6.1.2.1.1), as consistently the strongest factor in each of
the examined datasets (cf. Sections 8.1–14.1). Thus, the difference between *it*-clefts and *it BE* sentences plays a very important role in the distribution of pronoun case forms in subject predicative complements in and across varieties of English. However, this finding is in marked contrast to the assumptions of the positional approach and the weak vs. strong pronoun approach, which both assume object pronoun forms occur in more or less equal measures in *it*-clefts and *it BE* sentences (cf. Sections 3.2 and 3.3).

The variable Co-referencE is also a very important factor in the distribution of pronoun case forms in *it*-clefts in varieties of English. In the respective *it*-cleft subsets of the seven datasets examined for this study, this variable has turned out to be the most important factor promoting the use of subject pronoun case forms, if the focal pronoun is co-referential with the subject of the subsequent dependent clause (cf. Sections 8.3–14.3). Such a strong influence of this functional variable is also unexpected according to the positional and the weak vs. strong pronoun approaches to the distribution of pronoun case forms.

Furthermore, with regard to *it BE* sentences, Focus significantly influences the distribution of pronoun case forms in five of the seven analysed datasets. More specifically, Focus, as it is operationalised in Section 6.1.2.5, markedly fosters the use of subject pronoun case forms in *it BE* sentences in the BNC, COCA, .uk data, .au data, and .ie data (cf. Sections 8.2–12.2). This, too, is in marked contrast with the expectations and assumptions of both the positional and the weak vs. strong pronoun accounts, where functional and pragmatic factors are not deemed to play a significant role (cf. Sections 3.2 and 3.3). Thus, at each level of analysis considered for this study, i.e. the superordinate subject predicative complement category, *it BE* sentences and *it*-clefts, we obtained findings that clearly contradict the predictions and assumptions of both the positional and the weak vs. strong pronoun approaches.

Except for the impact of these functional factors, there is also other more indirect evidence that severely undermines the assumptions of both the weak vs. strong and positional accounts. With regard to the former, variation in the use of pronoun case forms in subject predicative complements is often attributed to an assumed dynamic transition from a case-governed pronominal paradigm to a system of weak and strong pronoun classes (e.g. Harris 1981). However, there are no quantitative analyses corroborating such an assumption. For instance, a comparison of Erdmann’s (1978) subject predicative
complement data, representing literary written British English from 1930 to 1970, with the written data of the BNC, which includes data from 1960 to 1994 (cf. Section 7.1.2), yields no statistically significant difference (cf. Maier 2013). If there were indeed a dynamic transition from a case-governed paradigm to a system of weak vs. strong pronoun case forms, we would expect to observe a significant difference between these two datasets; this, however, is not the case (Maier 2013).

With regard to the positional account, the results obtained from the analysis of the seven datasets in Chapters 8–14 also challenge this account’s explanation of the variation observable in the use of pronoun case forms in subject predicative complements. In a nutshell, positional accounts mainly assume that the more formal a variety or context, the more likely it is that one will observe subject pronouns in subject predicative complements (e.g. Quirk et al. 1985: 337-338). Although the various levels of the MODE OF DISCOURSE variable are significant in many of the analysed datasets, the results are not always those expected by the literature. In particular, the inhibiting influence of CMC attested for many Web-derived datasets cannot be attributed only to the lower degree of formality, but also to the larger stock of Focus marking strategies that can be employed by CMC. The larger stock of Focus marking strategies can also account for why subject pronoun case forms are less likely to be used in subject predicative complements in computer-mediated communication (e.g. Chapters 10 and 13). More importantly, however, spoken data does rather promote than inhibit the use of subject pronoun case forms in some datasets, which is unexpected according to the literature (e.g. Quirk et al. 1985: 336-337). Cases in point are the BNC, where SPOK tends to promote the use of subject forms particularly in it BE sentences, and the .in dataset where SPOK emphatically promotes the use of subject forms, especially in it-clefts (cf. Chapters 8 and 14).

In view of all this evidence, hypothesis H1 stating that pronoun case forms are mainly or solely determined by the pronoun’s position in relation to the finite verb or by their assumed membership to a specific pronoun class cannot be confirmed and actually should be abandoned (cf. Section 5.2). This is due to the fact that we can observe neither a dynamic transition from a case-governed pronominal paradigm to a system of weak and strong pronoun classes nor an effect of formality as predicted by these accounts. Most importantly, however, this is due to the very pronounced impact of clearly functional factors on the distribution of pronoun case forms attested for all datasets examined in this study. Indeed,
the strong effect of the clearly functional factors is not only unexpected, but rather irreconcilable with the assumptions and predictions of both the weak vs. strong pronoun and the positional accounts.

Thus, although the importance of the pronoun’s position has been explicitly acknowledged in this study (cf. Section 4.3), position alone is not sufficient to account for the distribution of pronoun case forms in contexts allowing for variability in general and in subject predicative complements in particular, as the marked differences between *it*-clefts and *it BE* sentences observed in the preceding chapters have demonstrated. In fact, several other factors have been identified as significantly influencing the distribution of pronoun case forms in subject predicative complements.

15.2 The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English: The Determinants and their Impacts

In general, the analysis of the datasets in Chapters 8–14 has demonstrated that the independent variables motivated and discussed in Chapter 6 can be categorised broadly into three different groups as far as their impact on the distribution of pronoun case forms in subject predicative complements is concerned: factors that foster the use of subject pronoun case forms, factors that, when significant, mainly inhibit the use of subject forms in the analysed contexts, and factors that are susceptible to a considerable extent of variation across the different datasets or varieties examined in this study.

To begin with, the analysis of the different corpora and datasets in Chapters 8–14 has demonstrated that particularly factors from the domains of syntax and/or pragmatics as well as their interface exert not only a very strong impact on the distribution of pronoun case forms but also, when statistically significant, tend to promote the use of subject forms quite strongly. In particular, the analysis of the superordinate subject predicative complement categories has identified CLEFT, by means of which the CONSTRUCTION variable is operationalised, as the strongest factor in each of the examined corpora and datasets (cf. Sections 8.1–14.1). Furthermore, CLEFT also emphatically promotes the use of subject pronoun case forms in subject predicative complements in each of these datasets. This
finding is in line with the assumptions and results of several other functional studies that observe considerably higher shares of subject pronoun forms in *it*-clefts than in *it BE* sentences (Biber et al. 1999: 335-336; Erdmann 1978: 75-78; Maier 2013; Wales 1996: 95-96).

With regard to the *it*-cleft subsets of each corpus and dataset, the Co-reference variable operationalised via AS_SUBJ is consistently the factor with the highest odds ratio. More specifically, the significance and strong impact of AS_SUBJ attested for each of the *it*-cleft datasets indicates that subject pronoun case forms are much more likely to occur in *it*-clefts in which the focal pronoun is co-referential with the subject in the following dependent clause than in *it*-clefts in which the focal pronoun is co-referential with an object of the dependent clause (cf. Sections 8.3–14.3). Again, this also complies with the assumptions and findings of several – mainly but not exclusively – functional studies that predict or even attest such a fostering impact of AS_SUBJ on the use of subject pronoun forms in *it*-clefts (e.g. Erdmann 1978: 75-78; Huddleston and Pullum 2002: 459; Quinn 2005a: 133-142; Wales 1996: 96).

Moreover, Focus has also turned out to be very highly significant and clearly promotes the use of subject pronoun case forms in *it BE* sentences in five of the seven subsets examined in this study, i.e. the *it BE* sentences of the BNC, COCA, .uk data, .au data and the .ie data (cf. Sections 8.2–12.2). This complies with the Focus-oriented approach adopted in this study (cf. Chapter 4) as well as with its corresponding hypothesis H3 (cf. Section 5.2).

On a more general level, however, the strong impact of CLEFT, AS_SUBJ and Focus also clearly confirms hypothesis H2, assuming that functional factors exert a very strong influence on the distribution of pronoun case forms in subject predicative complements in Present-Day English (cf. Section 5.2).

While CLEFT, AS_SUBJ and Focus are clearly functional and when significant consistently promote the use of subject forms in subject predicative complements, this study has also identified factors that tend to consistently inhibit the use of subject pronoun case forms in the analysed contexts. One of these constraining factors is First operationalising the independent variable PERSON (cf. Section 6.1.2.2). Whenever it is identified as statistically significant, First markedly inhibits the use of subject pronoun case forms in the respective contexts. However, the inhibiting force of First seems to some extent
to be sensitive to the sentence type in which the subject predicative complement occurs. While *FIRST* significantly inhibits the use of subject forms only in one *it* BE sentence dataset, i.e. that of the COCA (cf. Section 9.2), it significantly constrains the use of subject pronouns in each and every one of the seven *it*-clef subsets examined for this study (cf. Sections 8.3–14.3). As far as this factor’s impact on *it* BE sentences is concerned, this observation conflicts with accounts which expect a higher share of first person subject pronouns than third person pronouns in this sentence type (e.g. Quinn 2005a: 246). With regard to *it*-clefts, however, this result is in line with the observations of other studies according to which third person subject pronouns are more likely to be used than first person subject forms in this sentence type (e.g. Biber et al. 1999: 335-336; Wales 1996: 96). While acknowledging the higher likelihood of third person pronouns than of first person pronouns to occur as subject form in *it*-clefts, many accounts fail to offer a thorough explanation for this co-variance pattern between *CLEFT* and *FIRST* (Quinn 2009: 42; Wales 1996: 95-96). However, the current study accounts for this nearly systematically occurring pattern by means of Focus marking and the Focus-oriented perspective, as will be discussed again in more detail in the subsequent section. Suffice it to say that this study at least partly attributes the inhibiting effect of *FIRST* to the inherent differences between first and third person pronouns in their prototypical referential characteristics and to the salience of their referents in the ongoing communication (cf. Sections 6.1.2.2 and 8.3 for example).

In addition to *FIRST*, CMC also consistently inhibits the use of subject forms in those Web-derived datasets where this factor has been identified as statistically significant. While this constraining effect complies with accounts which assume that less formal varieties are less likely to use subject forms in subject predicative complements (e.g. Harris 1981; Quirk et al. 1985: 337-338), the effect of CMC on the use of subject pronoun case forms can also be considered from a more Focus-oriented perspective. Thus, it has been proposed that because computer-mediated communication may be, on average, subject to fewer formal conventions than traditional written data, it possesses a larger stock of expressive options such as capitalisation as Focus marking device, which are not available to the same extent for the traditional written mode of discourse (e.g. Sections 10.1 and 11.1). However, CMC is not only similar to *FIRST* in that it tends to inhibit the use of subject forms, which can also be explained with the help of a Focus-oriented perspective, but also in that CMC exhibits a similar co-variance pattern to *FIRST*. CMC, too, mainly inhibits the use of subject forms in *it*
clefts, namely in the *it*-cleft subsets of the .uk, .au, .ie and .za datasets (cf. Sections 10.3–13.3). In contrast, CMC only inhibits the use of subject forms in one *it BE* sentence dataset, i.e. that of the .uk data (cf. Section 10.2). Such a co-variance pattern is difficult to account for only in terms of formality, since we would expect a similar effect of CMC on both construction types if its influence were solely due to differences in formality. The Focus-oriented perspective adopted in this study, however, can explain this phenomenon by means of Focus marking or, to be more exact, by means of a trade-off between Focus marking strategies according to which not all available focussing devices such as *it*-clefts, subject forms and capitalisation may be used to highlight an entity as the Focus of a clause at the same time (cf. Dik 1989: 278; Section 4.1).

Finally, the analysis of the seven datasets examined in this study has also demonstrated that some of the independent variables exhibit a considerable extent of variation across the different datasets. The effect of the variable NUMBER operationalised via SINGULAR is a case in point. Although linguistic theory assumes that SINGULAR clearly fosters the use of subject forms in both *it BE* sentences and *it*-clefts (cf. Sobin 1997: 334), reality is much more complex. While SINGULAR, for example, inhibits the use of subject forms in the superordinate subject predicative complement category of the COCA (cf. Section 9.1), it significantly promotes the use of subject forms in the corresponding datasets of the .uk, .au and .za datasets (cf. Sections 10.1, 11.1 and 13.1). In the BNC, .ie and .in datasets, however, SINGULAR is eliminated from the respective minimal-adequate as non-significant (cf. Sections 8.1, 12.1 and 14.1). Thus, the effect of this variable seems to be prone to cross-varietal variation or, at least, to variation across different datasets. In view of the observed variation, it is extremely difficult to arrive at robust generalisations as far as the impact of SINGULAR on the distribution of pronoun case forms is concerned.

In a similar vein, SPOK, operationalising the difference between spoken and traditional written data, also exhibits a considerable degree of variation across the different datasets. While SPOK significantly inhibits the use of subject forms in the superordinate subject predicative complement category of the COCA and the .za data (cf. Sections 9.1 and 13.1), SPOK in tendency promotes the use of subject forms in the corresponding dataset of the BNC and even significantly fosters the use of subject pronoun case forms in the subject predicative complement category of the .in dataset (cf. Sections 8.1 and 14.1). In the corresponding datasets of the .uk, .au, .ie datasets, however, SPOK does not exert a
significant influence on the distribution of pronoun case forms. Moreover, a considerable extent of variation with regard to the impact of SPOK can be observed even within single datasets. Whereas SPOK significantly inhibits the use of subject pronoun forms in the it-cleft data of the BNC, it fosters in tendency the use of subject forms in the it BE sentences of this corpus (cf. Sections 8.2 and 8.3). Hence, the effect of SPOK on the distribution of pronoun case forms is susceptible to variation not only across but even within regional varieties of English. This finding is, however, not consistent with accounts assuming that less formal or spoken varieties of English are generally less likely to use subject pronoun case forms in subject predicative complements (e.g. Biber et al. 1999: 335-336; Hopkins 1975: 28; Quirk et al. 1985: 336-338; Wales 1996: 101). Moreover, the impact of SPOK as attested in the different datasets also challenges the assumption underlying many of these accounts, i.e. that the differences between modes of discourse in the distribution of pronoun case forms are attributable to differences in formality (e.g. Biber et al. 1999: 335-336; Hopkins 1975: 28; Quirk et al. 1985: 336-338). If the impact of SPOK were solely due to matters of formality, we would expect more coherent results for this factor across the different datasets, particularly for comparatively similar ones such as the closed corpora. More specifically, we would not expect to observe intra-varietal differences like in the BNC, where SPOK may have a totally different effect depending on the construction type (cf. Chapter 8). Although formality may also matter, it is definitely not sufficient to account for the impact of SPOK on the distribution of pronoun case forms in subject predicative complements in the different datasets examined in this study. Thus, there may be other factors responsible for the impact of SPOK as observed in this study. Due to the fact that SPOK particularly inhibits the use of subject forms in the it-cleft subsets of some varieties, it is proposed that this may also be a trade-off effect between different focussing devices. Since prosodic prominence and it-clefs are focussing strategies available in spoken discourse, subject forms to further highlight the focal pronouns in it-clefs may not be as necessary in spoken as in written discourse (cf. Sections 8.3, 9.3 and 13.3). While such an interpretation may not only explain the inhibiting impact of SPOK as observed in the it-clefs of the COCA, BNC and .za datasets, it is also in line with the fact that these datasets do not exhibit a similar effect of SPOK on their it BE sentences, which is difficult to account for solely in terms of formality. Furthermore, such an interpretation complies with the Focus-oriented approach to the distribution of pronoun case forms proposed and adopted in this study (cf. Chapter 4). However, in the case of the
in clefts, $\text{SPOK}$ seems to reinforce the use of subject forms rather than diminish it (cf. Section 14.3). Although this is, in principle, still reconcilable with a Focus-oriented approach since different Focus marking strategies may co-occur (cf. Section 4.1), the impact of $\text{SPOK}$ on the distribution of pronoun case forms is probably best considered a variety-specific phenomenon that does not easily allow for robust generalisations (cf. Sections 13.3 and 14.3; Bisang 2004: 26-27).

Finally, the impact of the different variables attested in this study not only corroborates hypothesis H2, which assumes that functional factors are important in the distribution of pronoun case forms, but also confirms our approach of refraining from applying a rigid a priori classification to the tested variables (cf. Section 6.1.3). The results of the data analyses in Chapters 8–14 have shown that it may be more useful to classify variables according to their impact on the outcome of the dependent variable than on a pre-defined level of linguistic analysis. $\text{SPOK}$ and CMC, for example, are in principle not only on the same level of linguistic analysis, but are even levels of the same variable, i.e. MODE OF DISCOURSE (cf. Section 6.1.2.4). However, while CMC is much more similar in its impact to the grammatical PERSON variable, operationalised via FIRST, $\text{SPOK}$ is more similar to SINGULAR, operationalising another grammatical variable, i.e. NUMBER. Indeed, basically every variable can also be considered from a pragmatic point of view in this study. Thus, it is more sensible to group variables according to the kind of impact they exert on the outcome of the dependent variable than according to some more or less abstract level of linguistic classification in order to account for the independent variables’ effects on the use of subject pronoun case forms in subject predicative complements.

15.3 The Reanalysis of Subject Pronoun Case Forms as Focus Markers in Subject Predicative Complements

As introduced in Chapter 4, this study assumes that the more focussed a particular context is, the more likely it is to observe a subject pronoun case form in subject predicative complement position. However, as discussed in Sections 4.1 and 6.1.2.5, Focus may be difficult to identify and thus many quantitative studies have so far refrained from testing this factor. In order to overcome the many problems associated with the concept of Focus, this
study employs an operationalisation of the variable Focus that consists of two steps. Firstly, we should observe a higher share of subject forms in it-clefts than in it BE sentences, if the assumption is correct that the more focussed a particular context, the more likely it is to observe a subject pronoun case form (cf. Section 6.1.2.5). Although this condition is not sufficient on its own to ascertain the impact of Focus, it constitutes the necessary condition for our hypothesis H3 (cf. Sections 5.2 and 6.1.2.5). Secondly, we should also observe significant differences between particularly focussed and prototypical it BE sentences, if subject pronouns are indeed more likely to occur in particularly focussed contexts. This assumption forms the sufficient condition for our hypothesis assuming a re-functionalisation of subject pronouns as Focus markers in subject predicative complements (cf. Section 6.1.2.5).

As stated above (e.g. Section 15.1), the analysis of the data in Chapters 8–14 has demonstrated that there are very pronounced differences between it-clefts and it BE sentences regarding the distribution of pronoun case forms in these two constructions. In each of the datasets examined for this study, it-clefts exhibit a considerably higher share of subject pronoun forms in subject predicative complement position than it BE sentences (cf. Chapters 8–14). This observation is confirmed by the multivariate statistical analysis of the data. Cleft has been identified not only as the factor with the highest odds ratio but also as a factor that substantially promotes the use of subject pronoun case forms in each of the superordinate subject predicative complement categories examined in Chapters 8–14. Thus, the necessary condition allowing us to assume that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements is met by all datasets analysed for this study, i.e. the BNC, COCA, .uk data, .au data, .ie data, .za data and .in data.

Furthermore, the analysis of the it BE sentences of the different datasets has shown that Focus, as operationalised in Section 6.1.2.5, significantly influences the distribution of pronoun case forms in the it BE sentences of the BNC, COCA, .uk data, .au data and .ie data. Moreover, in each of these datasets, Focus clearly promotes the use of subject pronoun case forms in it BE sentences and is consistently the factor with the highest odds ratio (cf. Sections 8.2–12.2). In the .au and .ie data, Focus has even been identified as the only factor that significantly influences the distribution of pronoun case forms in it BE sentences (cf. Sections 11.2 and 12.2). Thus, for these five datasets, i.e. the BNC, COCA, .uk data, .au data and .ie data, the sufficient condition that allows us to assume that subject pronoun forms
have been reanalysed as Focus markers in subject predicative complements is also met. As a consequence, hypothesis H3, assuming that we should observe higher shares of subject pronoun forms in more focussed contexts than in less focussed ones (cf. Section 5.2), can be clearly confirmed for these five datasets (cf. Chapters 8–12). Hence, we can conclude that subject pronoun case forms are indeed used as Focus markers in subject predicative complements in British, American, Australian and Irish English.

However, even in the .za and .in datasets where Focus has missed statistical significance (cf. Sections 13.2 and 14.2), there are still hints that suggest that Focus as operationalised in Section 6.1.2.5 may play an important role in the distribution of pronoun case forms in it BE sentences and therefore also in subject predicative complements given large enough datasets. In the it BE sentences of the .za dataset, no factor could be identified as significantly influencing the distribution of pronoun case forms. However, bootstrap validation has retained two factors, i.e. SINGULAR and Focus, in the minimal-adequate model, which still turns out marginally significant, even though it contains no significant factors (cf. Section 13.2). Although Focus has missed statistical significance, probably because of a low token number, its theoretically expected effect corresponds to the observations of the BNC, COCA, .uk data, .au data and .ie data in that Focus, at least in theory, fosters the use of subject pronoun case forms in the .za data (cf. Szmrecsanyi 2006: 102). Thus, future studies relying on larger datasets of South African English might observe a statistically significant fostering impact of Focus on the distribution of pronoun case forms in it BE sentences for this variety as well. A similar observation has been made for the .in data. Although a statistical analysis of the it BE sentences in the Indian dataset has completely failed due to the very small number of tokens, the three it BE sentences included in the .in data which exhibit a subject pronoun form can clearly be considered as particularly focussed (cf. Sections 6.1.2.5 and 14.2). Hence, given enough data, future studies may also identify Focus as significantly promoting the distribution of pronoun case forms in it BE sentences in Indian English. Even though the analysis of the .in and .za data cannot back hypothesis H3 with solid statistical evidence, the results obtained for these two datasets in tendency seem to confirm the results obtained for the other datasets and the central assumption of this study according to which subject forms have been reanalysed as Focus markers in subject predicative complements (cf. Chapter 4).
The assumption that pragmatic factors in general and Focus marking in particular play an eminent role in the distribution of pronoun case forms in subject predicative complements has clearly been confirmed by the results obtained for CLEFT and FOCUS for five of the seven datasets. This hypothesis and the underlying approach are further strengthened by the results obtained for some of the other variables tested in this study, such as PERSON and MODE OF DISCOURSE.

With regard to the variable PERSON, we observe that FIRST markedly inhibits the use of subject forms in all it-cleft datasets analysed in this study but only in one it BE sentence dataset. In general, the differences between first and third person pronouns regarding their likelihood of occurrence in their subject form in subject predicative complements has been accounted for by the fundamental differences inherent in the nature and pragmatic status of both first and third person pronouns (cf. Sections 8.3 and 9.1 for example). In a nutshell, first person pronouns always denote referents, i.e. the speaker or speakers, who are actively involved in a communicative situation. Thus, they are prototypically deictic because their referents are co-referential with persons actively involved in a communicative situation. Moreover, they are those discourse entities that are most likely to receive our attention and our empathy (e.g. Langacker 1991: 307; Siewierska 2004: 5-8). As a consequence, it has been proposed that the relevance of first person pronouns and their referents for an ongoing communicative situation is more or less automatically given simply by using these pronouns, since first person pronouns identify their referents as salient information for both the speaker and the hearer. Although third person pronouns can also be used deictically, they are prototypically used as phoric expressions, i.e. referring backward or even forward to referents mentioned in the previous or subsequent discourse (e.g. Siewierska 2004: 7). Hence, in subject predicative complements speakers or writers may feel the need to additionally highlight the referents of third person pronouns by means of subject pronoun forms in order to highlight their relevance for the immediate communicative situation and to clearly identify the referent of the third person pronoun as the Focus of the clause. In principle, this may account for why first person pronouns are less likely to occur in their subject forms in the examined datasets. Moreover, if FIRST coincides with CLEFT, it has been proposed that the combination of two such pragmatically salient factors, i.e. a first person pronoun in the focal position of an it-cleft, may be sufficient to identify the referent or referents of the first person pronoun as the most salient, highlighted or important piece of
information in a clause (e.g. Dik 1978: 130). Such a “trade-off” (Dik 1989: 278) between different focussing strategies could explain why third person pronouns are more likely to receive additional Focus marking by means of subject pronoun case forms, particularly in *it*-clefts (cf. Section 4.1). In addition, this possible explanation corresponds not only to the *it*-cleft data discussed here but also to the Focus-oriented approach outlined in Chapter 4. In any case, the fact that FIRST has a strongly inhibiting effect in *it*-clefts, but hardly ever in *it BE* sentences, i.e. only in the COCA, suggests that this difference is not a purely structural one but instead may be a pragmatic difference.

In a similar fashion, the results obtained for the MODE OF DISCOURSE variable can also be partly explained from a Focus-oriented perspective. The consistently inhibiting effect of CMC in those datasets where it is both applicable and statistically significant cannot be attributed merely to the fact that less formal varieties are less likely to exhibit subject pronouns in subject predicative complements (e.g. Harris 1981; Quirk et al. 1985: 337-338). This inhibiting effect may also partly be due to the fact that CMC can use more strategies to highlight pragmatically salient entities, including the Focus of a clause, simply because it is subject to fewer formal conventions than traditional written data (e.g. Sections 10.1, 11.1 and 13.1). Hence, CMC can also use capitalisation, for example, to highlight the most important entities in a clause (e.g. Section 13.1).

In analogy to the trade-off effect proposed for the combination of FIRST and CLEFT, a similar impact has been suspected for SPOK and CLEFT (e.g. Section 9.3). In particular, it has been suggested that in some varieties, subject forms may not be as necessary in spoken as in written discourse to highlight the focal pronouns in *it*-clefts as the Focus of a clause because speakers can use prosodic prominence, which is another common Focus marker (cf. Sections 8.3, 9.3 and 13.3). Although this could explain the constraining impact of SPOK as attested for the *it*-clefts of the COCA, BNC and the .za dataset and would also be in line with the fact that these datasets do not exhibit a similar effect of SPOK on their *it BE* sentences, such a trade-off explanation is challenged by the fact that SPOK promotes the use of subject forms in the *it*-clefts of the .in data and in tendency also in the *it BE* sentences of the BNC (cf. Sections 8.2 and 14.3). While the reinforcing effect of SPOK on the .in clefts and the BNC *it BE* sentences may still be reconciled with the Focus-oriented approach to pronoun case distribution (cf. Section 4.1), it seems best to treat the impact of SPOK as a variety-specific phenomenon due to the vast scope of variation attested for this factor (cf. Section 15.2; Bisang 2004: 26-27).
Nevertheless, as discussed in Section 15.2, the impact of SPOK is even more challenging for accounts attributing its effect merely to matters as formality, since the direction of SPOK’s impact is subject not only to inter- but also to intra-varietal variation (cf. Chapter 8). As a consequence, the impact of SPOK will be discussed again in the following section (cf. Section 15.4).

In sum, we can conclude that hypothesis H3, assuming that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements, has been confirmed for the British, American, Australian, and Irish datasets examined in this study. Moreover, we have found leads suggesting that Focus may also turn out to significantly promote the use of subject forms in subject predicative complements in South African and Indian English, given large enough datasets. Eventually, a Focus-oriented perspective to pronoun case variation has been very helpful in accounting for the impact of other variables observed in the data. For example, the inhibiting effect of CMC can be partly explained by the fact that this mode of discourse possesses more Focus marking strategies than traditional written English. In a similar vein, the lesser likelihood of first person pronouns to occur in their subject forms, particularly in it-clefts, can be explained by the fact that first person pronouns are per se pragmatically more salient than third person pronouns and that their occurrence in a cleft may suffice to mark the focal pronoun as Focus of a clause.

15.4 The Distribution of Pronoun Case Forms in Subject Predicative Complements across Varieties of English

As outlined in Section 5.1, this study has also tried to find out whether there are pronounced cross-varietal differences in the distribution of pronoun case forms in subject predicative complements. Although Chapter 2 demonstrated that we observe an impressive amount of variation in the use of pronoun case forms across varieties of English, some accounts still assume there are “global trends” in the distribution of pronoun case forms (e.g. Burridge 2004: 1118). In fact, while some studies cannot find evidence for cross-varietal differences in the distribution of pronoun case forms in it-clefts (Quinn 2009: 42), other bivariate analyses relying on larger databases have demonstrated that there are significant differences in the distribution of pronoun case forms in it-clefts between British and American English (Maier
2013). We will therefore reconsider the results obtained from the analysis of the different datasets in Chapters 8–14 in this section with regard to which differences or global trends can be observed in the distribution of pronoun case (cf. Section 5.2; H4).

Starting with the similarities, the analysis of the data in Chapters 8–14 has shown that in the superordinate subject predicative complement categories, CLEFT has consistently been identified as the factor with the highest odds ratio and also as emphatically promoting the use of subject pronoun forms in each of the seven datasets (cf. Sections 8.1–14.1). Thus, the very strong impact of CLEFT on the distribution of pronoun case forms in subject predicative complements may indeed be considered a global trend in view of the results obtained from the six regional varieties of English examined in this study.

Moreover, CMC exhibits a consistently inhibiting effect on the use of subject pronoun case forms in the subject predicative complement categories of the five Web-derived datasets (cf. Sections 10.1–14.1). In four of the five Web-derived datasets, this strong inhibiting effect of CMC is also attested for the it-cleft subsets (cf. Sections 10.3–13.3).

A further cross-varietal trend observed in the it-cleft data is the impact of FIRST that also clearly inhibits the use of subject pronoun case forms in all it-cleft subsets (cf. Sections 8.3–14.3). As discussed previously, the reason this factor is mainly affecting the it-cleft subsets of the data has been explained with the referential properties inherent in first and third person pronouns and with a possible trade-off effect between different Focus marking strategies (cf. Sections 8.3 and 9.3, for example).

Finally, another very important cross-varietal trend observed in the it-cleft data is the very strong impact of AS.SUBJ. It is consistently the factor with the highest odds ratio and emphatically promotes the use of subject pronoun case forms in the it-cleft subsets of all seven datasets (cf. Sections 8.3–14.3). Thus, the analysis of the different datasets has indeed detected several cross-varietal trends in the distribution of pronoun case forms in subject predicative complements, although some of them, such as the tremendous impact of CLEFT, have been rather unexpected given much of the literature (cf. Sections 3.1–3.3).

Despite these similarities, the analysis of the different datasets has also unearthed many cross-varietal differences that cannot be solely attributed to the heterogeneity of the data or the general difficulty of comparing different corpora and datasets (cf. Sections 6.1.2.6 and 7.5). A first very important difference between the examined corpora and datasets concerns the overall share of subject forms used in a particular construction or
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

sentence type. If we look at the subject pronoun shares in the *it BE* sentences of the .au data and the .ie data, for example, we observe that the share of subject forms in the .au data amounts to only 2.09 per cent of all tokens, whereas 27.59 per cent of all *it BE* sentences in the .ie data take a subject pronoun case form. Bearing in mind not only that these datasets were compiled in the same manner (cf. Section 7.2.4.1), but also that the *it BE* sentences in both datasets are significantly influenced by only one factor, which in both cases is Focus, the vast disparities in the use of subject pronoun forms in *it BE* sentences between these two varieties become even more apparent (cf. Section 11.2 and 12.2). Indeed, the comparison of these two datasets shows very well that varieties differ considerably from each other with regard to the overall likelihood of using subject pronoun forms in subject predicative complements. However, differences in the use of pronoun case forms can be observed not only between Web-derived datasets but also between the two closed corpora. With regard to *it BE* sentences, 8.08 per cent of all BNC tokens are used with a subject pronoun form, while only 4.47 per cent of all *it BE* sentences exhibit a subject form in the COCA (cf. Sections 8.2 and 9.2; Maier 2013).

To be sure, cross-varietal differences are not restricted to the domain of *it BE* sentences; they can also be attested for the *it*-clefts. For example, while only 34.61 per cent of all *it*-clefts in the .uk data use a subject form, 91.36 per cent of the *it*-clefts in the .in data exhibit a subject form as focal pronoun (cf. Sections 10.3 and 14.3). Moreover, significant differences can also be observed in the use of pronoun case forms between the *it*-clefts of the COCA and the BNC. While 73.09 per cent of all BNC *it*-clefts are used with a subject form, 78.56 per cent of all COCA clefts exhibit subject pronouns (cf. Sections 8.3 and 9.3; Maier 2013).

Furthermore, the corpora and databases examined in this study differ from each other not only with regard to the share of subject pronoun case forms attested for the respective *it*-cleft and *it BE* sentence subsets, but also in terms of the overall usage frequency of *it BE* sentences and *it*-clefts. Based on a rough normalisation (cf. Section 9.3), we observe approximately 8.1 *it*-clefts with a case-sensitive focal pronoun per million words in the BNC, whereas the corresponding frequency for the COCA amounts to only 3.2 *it*-clefts per million words. With regard to *it BE* sentences, we observe roughly 6.5 case-sensitive tokens per million words in the COCA, whereas the BNC frequency amounts to only 5.7 *it BE* sentences per million words (cf. Maier 2013; Section 9.3). Although such a normalisation is
not possible for the Web-derived datasets (cf. Section 7.2.4.1), they also exhibit pronounced cross-varietal differences as far as the proportions of *it*-clefs and *it BE* sentences subsumed under the heading of subject predicative complements are concerned. Cases in point are the .ie and .in datasets. Although both of them are similar in size, they differ considerably as far as their proportions of *it*-clefs and *it BE* sentences are concerned. While the share of *it BE* sentences in the .ie dataset amounts to 20.47 per cent of all subject predicative complements, only 5.89 per cent of subject predicative complements in the .in dataset are *it BE* sentences (cf. Chapters 12 and 14). Indeed, *it BE* sentences in the .in dataset are so infrequent that it has not been possible to conduct a multivariate analysis of this dataset (cf. Section 14.2).

Moreover, the multivariate analysis of the data has demonstrated that the seven datasets may also differ considerably from each other as far as the significance, impact and direction of the impact of the independent variables are concerned. Although some variables behave uniformly across the datasets or at least across a particular subset of the datasets, there are also pronounced cross-varietal differences in the multivariate analyses of the different datasets. A brief look at the minimal-adequate models applied to the superordinate subject predicative complement categories suffices to illustrate these differences. These datasets differ from each other considerably in terms of the number and types of variables and interactions retained in the minimal-adequate models, the exact impact of the significant factors and the kind or direction of impact of the significant variables (cf. Sections 8.1–14.1). In fact, only the minimal-adequate models analysing the superordinate subject predicative complement categories of the .uk and the .au datasets retain the same set of variables and interactions. However, even these superficially similar models or datasets exhibit notable differences with regard to the observed significance levels and odds ratios of the retained variables and interactions (cf. Sections 10.1 and 11.1).

With regard to the *it*-clef subsets, the respective models also differ from each other in terms of the maintained factors and interactions, the effect of these variables and the direction of their impact (cf. Sections 8.3–14.3). Only the minimal-adequate models of the BNC and .in *it*-clefs maintain the same set of variables. However, these models exhibit also considerable differences in the odds ratios and significance levels of the retained factors. Moreover, whereas Spox markedly inhibits the use of subject forms in the *it*-clef section of the BNC, it emphatically promotes the use of subject forms in the *it*-clefs of the .in dataset.
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

(cf. Sections 8.3 and 14.3). Hence, we can note pronounced cross-varietal differences in the distribution of pronoun case forms as far as *it*-clefts are concerned.

For the *it BE* sentences, we also observe marked differences. The minimal-adequate models applied to these datasets differ from each other with regard to the number and types of the maintained variables and also in terms of the significance levels and odds ratios of these factors. Interestingly, the only *it BE* sentence regression models that retain the same variable as significantly influencing the distribution of pronoun case forms are those of the *.ie* and *.au* datasets. In each of the models, only *Focus* is maintained and markedly promotes the use of subject forms, though the odds ratio is higher in the *.au* dataset than in the *.ie* dataset (cf. Sections 11.2–12.2). While the regression results are indeed very similar for these datasets, we have to bear in mind that, as noted above, only 2.09 per cent of all *it BE* sentences in the *.au* data use subject pronouns, whereas 27.59 per cent of all *it BE* sentences in the *.ie* data exhibit subject pronoun case forms, which further illustrates the pronounced cross-varietal differences in the distribution of pronoun case forms.

These cross-varietal differences concerning the impact, direction of impact and significance of the tested variables become even more apparent when shifting our attention from the examination of the regression models as a whole to the effects of the individual variables across the different datasets. Cases in point for these vast cross-varietal differences are *Singular* and *SPOK* (cf. Section 15.2).

With regard to *Singular*, it has been assumed that this factor markedly promotes the use of subject pronoun case forms in *it BE* sentences and *it*-clefts (cf. Sobin 1997: 334). However, as the analysis of the data has demonstrated, the significance and impact of *Singular* is subject to considerable cross-varietal variation (cf. Chapters 8–14). While *Singular* significantly constrains the use of subject pronouns in the subject predicative complement category of the COCA (cf. Section 9.1), it significantly fosters the use of subject pronouns in the corresponding subject predicative complement categories of the *.uk*, *.au* and *.za* datasets (cf. Sections 10.1, 11.1 and 13.1). To complicate matters even further, *Singular* is not significant in the minimal-adequate models applied to the subject predicative complement categories of the BNC, *.ie* and *.in* datasets (cf. Sections 8.1, 12.1 and 14.1).

Also susceptible to a considerable extent of cross-varietal variation is *SPOK*, as the analysis of the different datasets has demonstrated (cf. Chapters 8–14). Whereas *SPOK* significantly constrains the use of subject pronouns in the superordinate subject predicative
complement category of both the COCA and the .za data, we observe the inverse effect for the corresponding datasets in the BNC and the .in data. In the BNC, the promoting effect of SPOK is only marginally significant (cf. Section 8.1), whereas it passes even the five per cent level of significance in the subject predicative complement category of the .in dataset (cf. Section 13.1). However, SPOK is not significant in the corresponding datasets of the .uk, .au and .ie data. Moreover, the impact of SPOK is subject to variation even within single varieties or datasets (cf. Chapter 8). In the BNC, for example, SPOK in tendency promotes the use of subject forms in *it* BE sentences while it significantly inhibits the use of subject forms in *it*-clefts (cf. Sections 8.2 and 8.3). Thus, the effect of SINGULAR and SPOK on the different datasets exemplifies the huge extent of variation that can be observed in the distribution of pronoun case forms in subject predicative complements across the analysed datasets and varieties of English.

Finally, the varieties examined in this study also differ from each other with regard to the impact of Focus on the distribution of pronoun case forms. At first sight, the fact that **FOCUS** is only significant in the L1 varieties examined in this study may even suggest that this re-functionalisation process is restricted to L1 varieties. However, such a conclusion may be over-hasty, taking a closer look at the obtained results. Although the Indian data contains too few data points to examine the data by means of multivariate statistical modelling, those few *it* BE sentences that exhibit a subject form are clearly particularly focussed (cf. Section 14.2). Thus, a re-functionalisation can at least be suspected for this variety. In Indian English, it is the scarcity of *it* BE sentences rather than the statistically non-verifiable impact of **FOCUS** that is the really surprising result. Only in the South African data, the impact of **FOCUS** on the distribution of pronoun case forms in *it* BE sentences seems to be less pronounced than in the other datasets examined in this study. However, different varieties may have different preferences in terms of Focus marking (cf. Section 4.1). In fact, that languages and language varieties have different pragmatic norms and conventions and may differ from each other in terms of their pragmatic strategies and inventories is widely acknowledged (e.g. Huang 2007: 119-127; Schneider and Barron 2008: 5-25; Wierzbicka 1991: 2-5). Thus, the results obtained for these varieties are still very much compatible with hypothesis H3 and the Focus-oriented approach to pronoun case distribution.

Now, it would of course also be desirable to account for all the cross-varietal differences that have been discovered in the course of the data analysis (cf. Chapters 8–14)
and that have been discussed in this section. Unfortunately, however, this lies by far beyond the clearly delimited scope of the present study and the possibilities of its data basis as the following examples illustrate (cf. Chapters 5 and 7):

It would, of course, be interesting to ask why Indian and South African English, for example, behave so differently from the other datasets analysed in this study. So far, this difference has been mainly explained either with cross-varietal differences (cf. Section 13.2) or with a lack of data (cf. Section 14.2). With regard to Indian English it is tempting to relate the nearly categorical use of subject forms in the *it*-cleft data also to the often-perceived formal character of Indian English (e.g. Hickey 2004: 542; Sailaja 2009: 6). This explanation would also nicely tie in with the often-postulated “prescriptive bias in favour of the subject forms” of traditional grammar books (e.g. Quirk et al. 1985: 338). However, one should be cautious with such an assumption particularly for two reasons: First, this study has demonstrated that matters of formality have not been as important as assumed by the body of literature and particularly not in the way it has been predicted (cf. Sections 8.1, 13.3, 15.2). Second, this study has also shown that matters of Focus marking play a very prominent role in the distribution of pronoun case forms in several varieties of English. Although Focus has not been identified as significant variable in the *it*-data, all Indian *it BE* sentences with a subject form can be considered to be particularly focussed (cf. Section 14.2). Thus, it would be likewise possible to interpret the use of postverbal subject forms as a kind of generally used Focus marker in subject predicative complements, particularly when bearing in mind that Indian English makes also frequent use of other morphosyntactical Topicalisation and Focus marking strategies, in which it is markedly different from other English varieties (e.g. Bhatt 2004: 1023). Thus, the question whether the nearly categorical use of subject forms in focussed subject predicative complements in Indian English is due to the influence of prescriptivism or whether it is due to its being part of a complex system of Topic- and Focus marking (cf. Bhatt 2004: 1023) has to remain unresolved at this point. In order to answer this question we need both a better understanding of Topic and Focus marking strategies in Indian English and data that allows us to take stylistic effects and social as well sociolinguistic factors more into consideration and which may enable us to account for the actuation, transition and embedding of this feature (e.g. Schreier 2014: 155-156). Then, and only then, we will perhaps be able to conclusively decide whether prescriptivism
or Focus marking is the really decisive factor in accounting for the nearly categorical use of subject forms in subject predicative complements in Indian English.

Likewise it would be very desirable to be able to explain why Focus is not significant in the .za data. To find that out we need again more insights into the Focus marking systems of this variety, better knowledge of South African English itself and the (substrate) languages influencing it as well as its pronoun and Focus marking systems. Moreover, we need also more knowledge about the factors influencing the distribution of pronoun case forms in South African English and in English in general.

Finally, it would also be interesting to account for the vast differences in terms of pronoun case distribution between the Irish and Australian data, especially with regard to the it BE sentences (cf Sections 11.2 and 12.2). In order to be able to account for that, again, we have to find out more about the actuation, transition and embedding of the use of subject pronoun forms as postverbal Focus markers and we also have to obtain data that allows us to tackle these issues (e.g. Schreier 2014: 155-156). Unfortunately, this lies beyond the possibilities of this study and its databases.

Although the limited scope of the present study and the fact that it cannot answer each question that it has evoked in the preceding chapters may leave some readers with a sentiment of dissatisfaction, it must not be forgotten that this study actually had to lay the groundwork in many, many respects – particularly from a variationist perspective. Among others, this study has been able to provide for the first time ever a large-scale quantitative account of the distribution of pronoun case forms in subject predicative complements in British, American, Australian, Irish, South African and Indian English. In addition, this study has also for the first time ever demonstrated that subject pronoun forms have been re-functionalised as Focus markers in subject predicative complements in British, American, Australian and Irish English. On a more general level this study has also identified a mechanism, i.e. “markedness reversal”, that may also explain pronoun case variation in several other contexts (cf. Section 16.2).

With regard to hypothesis H4 of this study assessing whether the distribution of pronoun case forms in subject predicative complements is subject to a global trend or susceptible to cross-varietal variation, we can conclude the following: On the one hand, we have indeed observed cross-varietal trends. These trends are, however, rather unexpected and contrast with the assumptions of much of the relevant literature (cf. Sections 3.2 and
3.3). In particular, the strong impact of functional factors attested in each and every one of the examined varieties is irreconcilable with accounts assuming “a general trend in English towards case selection [being] dictated by position rather than function” (Burridge 2004: 1118). On the other hand, this study has also observed considerable cross-varietal differences. Aspects in which the varieties examined differ from each other comprise not only the observed shares of subject forms in subject predicative complements, but also the occurrence rates of it BE sentences and it-clefts in the examined varieties as well as the significance, strength and direction of impact of the tested variables. Hence, the findings of this study strongly corroborate accounts that emphasise the differences between varieties of English as far as the distribution of pronoun case forms in general and in subject predicative complements in particular is concerned (e.g. Maier 2013; Shorrocks 1992: 432-444).

15.5 The Potential of Google Data for Quantitative Studies of Grammatical Variation

As outlined in Section 7.2.2, the validity and generalisability of results obtained from the analysis of Google data are often questioned because of the unclear size of the underlying corpora as well as the uncertain quality, representativeness, exhaustiveness and origin of the data. Despite these potential problems and uncertainties, Web-derived datasets also offer many opportunities (cf. Section 7.2.3). Most importantly, they enable us to examine morphosyntactic phenomena that are rare in closed corpora, such as it BE sentences and it-clefts with a case-sensitive focal pronoun. As a consequence, five Web-derived datasets were compiled prospectively for this study to counteract the data scarcity that has so far hampered the cross-varietal analysis of the distribution of pronoun case forms in subject predicative complements (cf. Quinn 2009: 46) and to allow for the analysis of the distribution of pronoun case forms in varieties of English for which no freely accessible mega-corpora are available, such as Irish, South African or Indian English. However, since the generalisability of results obtained from Google is still an issue of ongoing controversy (cf. Sections 7.2.2 and 7.5), one aim of this study has been to assess whether or not data obtained from commercial Web crawlers by means of the ‘Web as Corpus’ approach provide
a sound basis for quantitative linguistic studies that yield robust results allowing for stable
generalisations (cf. Section 5.2; H5). Thus, we will reconsider and discuss the results
obtained from the data analysis in Chapters 8–14 from the perspective of whether carefully
refined Google data constitutes reasonable input for quantitative studies (cf. Sections
7.2.4.1–7.2.4.3).

Factors that might argue at first glance against the usefulness of Google data are the
skewing effects the Web-derived datasets compiled for this study are undeniably subject to.
The Web crawler’s disregard of punctuation marks leads to a high proportion of false
positives in the Web-derived datasets, which ultimately results in a lower percentage of
subject forms in subject predicative complements in Web-derived datasets when compared
to closed corpora (cf. Section 7.2.4.2). This assumption has been confirmed by the
proportions of subject forms observed for both British datasets where the share of subject
forms is considerably higher in the BNC than in the .uk data (cf. Sections 8.1 and 10.1).
Furthermore, this non-consideration of punctuation marks particularly affects the results of
the first person singular, for which the highest shares of false positives are attested. This
particular effect has become evident in the interaction between FIRST and SINGULAR, which
significantly influences the distribution of pronoun case forms in the superordinate subject
predicative complement as well as in the it-cleft datasets of the .uk data, .au data and .za
data. However, although they definitely affect the obtained results, these skewing effects
could be predicted in advance of the actual analyses and thus are not surprising (cf. Section
7.2.4.2). Therefore, these predictable skewing effects are no reason to doubt the validity and
generalisability of the results obtained from the Web-derived datasets in general.

More importantly, however, all major trends observed in the closed corpora are also
mirrored or replicated by the results obtained from the analysis of the Web-derived
datasets. This finding very much corroborates the assumption that the analysis of Google
data may provide robust results and may be a viable option to evade the data scarcity that
has so far constrained the quantitative analysis of subject predicative complements.

With regard to the general subject predicative complement categories, the
promoting impact of CLEFT and the inhibiting effect of FIRST are observed not only in the BNC
and the COCA but also in the corresponding analyses of the five Web-derived datasets (cf.
Sections 10.1–14.1). Moreover, the analyses of the it-cleft subsets of the BNC and the COCA
have identified AS_SUBJ as markedly fostering the use of subject pronoun forms, whereas
**FIRST** severely constrains the use of subject pronoun forms in these two datasets (cf. Sections 8.3 and 9.3). The same impact of **AS_SUBJ** and **FIRST** has also been attested for the *it*-cleft subsets of the five Web-derived datasets (cf. Sections 10.1–14.1). As far as the results for the *it BE* sentences are concerned, the analysis of the respective subsets of the BNC and the COCA has shown that they are similar in that **FOCUS** is clearly the strongest factor retained in the minimal-adequate regression models, markedly fostering the use of subject pronoun forms (cf. Section 8.2–9.2). The strong promoting effect of **FOCUS** has also been affirmed by the results obtained from the .uk data, .ie data and .au data (cf. Sections 10.2–12.2). In addition, even in the *it BE* sentences of the .za data and .in data we have found leads that suggest that larger databases might also render **FOCUS** statistically significant in these datasets (cf. Sections 13.2 and 14.2).

Thus, the trends observed in the closed corpora for **CLEFT**, **FIRST**, **AS_SUBJ** and **FOCUS**, particularly their significance and direction of impact (cf. Chapters 8 and 9), are largely mirrored and confirmed by the results of the analysis of the Web-derived datasets in Chapters 10–14, which is line with the results of earlier studies also attesting a correlation between the results obtained from the closed corpora and those from the Web-derived datasets (cf. Keller and Lapata 2003; Mair 2007; Rohdenburg 2007).

Indeed, a remarkable extent of variation between the closed corpora and the Web-derived datasets can only be observed for three factors, i.e. **CMC**, **SINGULAR** and **SPOK**. As far as CMC is concerned, this is due to the fact that neither the BNC nor the COCA contains any computer-mediated communication data. Thus, it is self-evident that the Web-derived datasets differ from the closed corpora with regard to the impact of CMC on the distribution of pronoun case forms. However, the Web-derived datasets exhibit consistent results with regard to this factor (cf. Chapter 10–14). Moreover, the fact that we have not obtained consistent trends or results for **SPOK** and **SINGULAR** across the examined datasets should be attributed to the general susceptibility to variation of these two factors rather than to the possible shortcomings of the Web-derived datasets (cf. Section 15.2). As the analysis of the BNC and the COCA has shown, these two corpora themselves do not exhibit consistent results or patterns for **SPOK** and **SINGULAR**. While **SINGULAR** significantly inhibits the use of subject forms in the superordinate subject predicative complement category and the *it*-cleft subset of the COCA, it is eliminated from the minimal-adequate models of the BNC as not significant. Moreover, **SPOK** exhibits variation even within closed corpora. The analysis of the
BNC has shown that SPOK inhibits the use of subject forms in *it*-clefts but tends to foster it in the *it BE* sentences of this corpus (cf. Sections 8.2 and 8.3). Thus, the differences between the closed corpora and the Web-derived datasets with regard to SINGULAR and SPOK are not sufficient to cast any doubt on the general validity and explanatory power of the results obtained from the five Web-derived datasets compiled for this study. As the previous sections have shown, SINGULAR and SPOK are susceptible to variation even within and across carefully crafted closed corpora.

Moreover, the diagnostic values obtained from the statistical analyses of the five Web-derived datasets, some of which are reported in Appendix A and B, also do not provide evidence suggesting that the results may be problematic or unfit to allow for robust generalisations. Hence, there is no reason to question the validity and explanatory power of the results obtained from the Web-derived datasets for British, Australian, Irish, South African and Indian English specifically compiled for this study.

In sum, this study has not only provided a fine-grained analysis of the distribution of pronoun case forms in subject predicative complements in six varieties of English, but it has also demonstrated that not only Web-derived corpora, but also carefully refined data obtained from the ‘Web as Corpus’ approach can yield robust results that allow for stable generalisations. Thus, hypothesis H5, according to which data from Google can be used as input for quantitative studies, can be confirmed based on the obtained results (cf. Chapters 10–14). This finding is indeed very promising in view of the data scarcity that poses severe problems for the quantitative analysis of many linguistic phenomena in and across varieties of English. In fact, this finding may encourage future studies to analyse low-frequency phenomena with the help of Web-derived datasets. Thus, despite all difficulties associated with the use of Web data (cf. Section 7.2.2), the results obtained from this study clearly confirm accounts assuming that “the web is and probably will be one of our best sources of information” (Hundt, Nesselhauf and Biewer 2007b: 3).
15.6 General Discussion: Interim Summary

In sum, the preceding sections have addressed, reassessed and discussed the findings obtained from the analysis of the data in Chapters 8–14 against the background of the major research questions and central hypotheses of this study as outlined in Chapter 5.

To begin with, Section 15.1 has demonstrated that the results obtained from the multivariate analyses of the different datasets seriously challenge the assumptions of positional and weak vs. strong pronoun accounts. The systematic differences between it-clefs and it BE sentences in their use of pronoun forms and the consistently strong impact of functional factors, which has been attested for each and every dataset, are irreconcilable with accounts trying to account for the distribution of pronoun case forms merely in terms of weak or strong pronoun classes or in terms of the pronouns’ position relative to the finite verb. As a consequence, hypothesis H1, assuming that the distribution of pronoun case forms is solely or mainly determined by the pronoun’s position or pronoun class membership, has to be abandoned. Although position certainly matters, the distribution of pronoun case forms is also influenced by several other factors.

Thus, Section 15.2 has systematised these factors according to the direction of impact which they exert on the dependent variable, i.e. whether they inhibit or promote the use of subject pronoun forms in subject predicative complement position. With regard to the fostering factors, the analyses of the data in Chapters 8–14 have identified CLEFT, FOCUS and AS_SUBJ as consistently promoting the use of subject pronoun case forms in those datasets in which they are both significant and applicable. As far as the constraining factors are concerned, the analyses have identified FIRST and CMC, if significant and applicable, as consistently inhibiting the use of subject pronoun case forms in subject predicative complements. Finally, SINGULAR has been identified to be prone to cross-varietal variation, and SPOK is even susceptible to intra-varietal variation. Hence, the effect of these factors may be dependent both on the dataset and the construction in which they are used. In any case, the strong impact of the functional factors CLEFT, FOCUS and AS_SUBJ confirms hypothesis H2 proposing that the distribution of pronoun case forms is still influenced by functional factors.

Section 15.3 then related the results of this study to its central assumption, i.e. that subject pronoun case forms may be used as Focus markers in subject predicative
complements. In view of the strong effect of CLEFT observed for all datasets and the significant impact of FOCUS attested in the British, American, Australian and Irish datasets, both of which emphatically promote the use of subject pronoun case forms, hypothesis H3 assuming that subject forms have been analysed as Focus markers has been confirmed for the BNC, COCA, .uk data, .au data and .ie data. For, the .za data and .in data, this hypothesis has not yet been confirmed, though there are leads suggesting that FOCUS may also play a role in the distribution of pronoun case forms in these varieties, given enough data.

In Section 15.4, this study has addressed hypothesis H4 which has been proposed in order to test whether the distribution of pronoun case forms in subject predicative complements is subject to global trends or whether there are pronounced cross-varietal differences. While the analysis of the data in Chapters 8–14 has indeed detected general trends, some of them, such as the strong impact of functional factors, are rather unexpected based on much of the current literature. Moreover, this study has also observed many cross-varietal differences that cannot be attributed merely to the heterogeneity of the datasets. The different usage rates of subject forms in subject predicative complements, the different occurrence rates of it-clefts and it BE sentences in the respective datasets and the differences in the significance, strength and direction of impact of the independent variables as observed across the different datasets indicate marked cross-varietal differences. Thus, the results of this study strongly corroborate accounts that emphasise the differences between varieties of English in the distribution of pronoun case forms in subject predicative complements in general and in it-clefts and it BE sentences in particular.

Finally, complying with hypothesis H5, Section 15.5 addressed the question whether data obtained from Google may be used as a database for quantitative studies. In a nutshell, all major trends that were observed in the BNC and the COCA, such as the strongly promoting effect of CLEFT, FOCUS and AS_SUBJ and the inhibiting effect of FIRST, have also been attested for the five Web-derived datasets. Notable differences have been observed only for factors that also exhibit variation across or within the examined closed corpora, such as SINGULAR and SPOK. In sum, this study has shown that the careful compilation and subsequent analysis of Web-derived datasets may help to overcome the data scarcity that has thus far constrained the analysis of many morphosyntactic features in and across varieties of English.
16 Conclusion and Outlook

16.1 Conclusion

Having examined the distribution of pronoun case forms in subject predicative complements in six regional varieties of English, this study has shown that – contrary to widespread assumptions – functional factors exert a tremendous influence on the distribution of pronoun case forms in Present-Day English. More precisely, this study has demonstrated that subject pronoun forms have been re-functionalised as Focus markers in subject predicative complements in British, American, Australian and Irish English. Although we have found leads suggesting a similar re-functionalisation process for South African and Indian English, the reanalysis of subject pronoun forms as postverbal Focus markers cannot yet be confirmed for these varieties by means of significant statistical results.

In particular, this study has shown that subject forms are much more likely to be used in it-clefts than in it BE sentences in all of the datasets examined in the previous chapters. Furthermore, subject forms are also much more likely to be used in particularly focussed than in prototypical it BE sentences, at least in the British, American, Australian and Irish datasets. For these varieties, the central assumption of this study stating that the more focussed a subject predicative complement context is, the more likely it is to observe a subject pronoun in a subject predicative complement context has been clearly confirmed.

Moreover, the analyses and discussions in the preceding chapters have shown that the distribution of pronoun case forms in subject predicative complements differs from variety to variety and is subject to several factors. Thus, monocausal explanations are not sufficient to account for complex variation phenomena, even in well-defined and narrowly delimited contexts such as subject predicative complements.

The very strong impact of functional and pragmatic factors, such as CLEFT, AS_SUBJ and FOCUS, has particularly severe repercussions on existing theoretical models trying to account for the distribution of pronoun case forms. To begin with, the assumptions and predictions of the weak vs. strong pronoun approach are irreconcilable with the findings of this study. In fact, in view of the results obtained, this approach can hardly be maintained. Moreover, the strong impact of functional and pragmatic factors uncovered in the present study also
challenges the assumption that position alone determines the distribution of pronoun case forms. Although position definitely matters, it is not sufficient to account for the distribution of pronoun case forms in Present-Day English.

Instead, the results of the statistical analyses clearly corroborate and for five of the seven analysed datasets even statistically validate the Focus-oriented perspective adopted in this study arguing for a reanalysis of subject pronouns as Focus markers in subject predicative complements. The significant and strong impact of CLEFT attested in all varieties and the significant and strong effect of FOCUS attested for five datasets allow us to conclude that subject pronouns have indeed been re-functionalised as Focus markers in British, American, Australian and Irish English. Although such a reanalysis contrasts with the assumptions of much of the current linguistic literature which discusses the distribution of pronoun case forms in subject predicative complements (e.g. Harris 1981; Quinn 2005a; Quirk et al. 1985), it corresponds very well to the results of the cross-varietal survey indicating a very strong influence of pragmatic factors on the distribution of pronoun case forms in many English varieties (cf. Chapter 2). Furthermore, the Focus-oriented approach proposed in this study can explain and also correctly predict the systematic differences between it-clefs and it BE sentences in their use of pronoun case forms and can also explain why certain subtypes of it BE sentences, i.e. particularly focussed ones, are much more likely to elicit subject forms than more prototypical it BE sentences (cf. Section 6.1.2.5). In fact, the Focus-oriented approach fares much better at predicting the distribution of pronoun case forms in subject predicative complements than the positional or weak vs. strong pronoun approaches do. Even more importantly, however, the Focus-oriented approach adopted in this study also provides a functional explanation for the considerable extent of variation in the use of pronoun case forms that can still be observed in subject predicative complements in Present-Day English. In view of the results obtained both from the cross-varietal surveys and the statistical analyses, this explanation is more plausible than previous approaches attempting to account for the distribution and variation of pronoun case forms in it-clefs and it BE sentences in terms of position and formality or in terms of a telic development from a once case-governed paradigm towards a system of weak or strong pronoun classes. Last but not least, the Focus-oriented approach can also accommodate the cross-varietal differences observed in the course of this study, which are often not addressed or not even suspected in earlier accounts (e.g. Burridge 2004; Quinn 2005a).
The discussion of the Focus-oriented approach leads us back to the theoretical underpinnings on which it rests and thus also to the general theoretical orientation of this study. In sum, the combination of Functional Grammar, Functional Typology and functional morphological theory has proven to be extremely beneficial not only for the explanation of the results, but also for the detection of important variables. The systematic investigation of the distribution of pronoun case forms across varieties of English from the perspective of Functional Typology has helped us to identify pragmatic factors, especially Focus, as important variables for the distribution of pronoun case forms in English varieties. Functional Grammar, in turn, has not only enabled us to arrive at a definition of Focus but has also allowed us to implement an operationalisation of this rather elusive variable. Moreover, combining the insights gained from Functional Typology and Functional Grammar has also made it possible to account for important co-variance patterns such as the lower likelihood of first person pronouns to occur in subject form in it-clefts. Finally, functional morphological theory has equipped us with an explanation as to why subject pronoun forms in particular and not other available pronominal options have been reanalysed as Focus markers in subject predicative complements. Indeed, the responsible mechanism, i.e. markedness reversal (Mayerthaler 1981: 48) or deviation from default expectation (Haspelmath 2006: 62), may be relevant not only for subject predicative complements but also for other contexts allowing for variability in the use of pronoun case forms, as will be discussed below (cf. Section 16.2).

As far as the methodological approach of this study is concerned, this study has clearly demonstrated that thorough qualitative and fine-grained quantitative analyses have to complement each other in order to achieve the best possible outcome. It should be borne in mind that only the comprehensive, qualitative, cross-varietal survey of the use of pronoun case forms has shown that the distribution of pronoun forms is influenced by pragmatic factors in many varieties of English, which has so far been largely neglected in the discussion of contexts allowing for variability in the use of pronoun case forms. However, in order to test and evaluate whether or not Focus indeed significantly influences the distribution of pronoun case forms in the subject predicative complements of the examined varieties, the application of multivariate statistical modelling has been indispensable. This study has demonstrated that qualitative and quantitative analyses are not alternative approaches, but
actually have to work in tandem to obtain the best possible results allowing for robust generalisations.

With regard to the mode of data retrieval, the results of this study show that data obtained from commercial Web crawlers by means of the ‘Web as Corpus’ approach, if carefully refined, can indeed yield robust results and thus can form a suitable basis for the examination of a wide range of linguistic phenomena. Bearing in mind that large corpora are available only for very few varieties of English, this is indeed a very promising finding.

16.2 Outlook

Due to its restricted focus on the distribution of pronoun case forms in subject predicative complements in varieties of English, this study could not do full justice to all questions that have emerged in the previous chapters. Thus, some particularly interesting open issues are briefly addressed in this section since they may serve as a starting point for future studies.

One particularly promising loose end that has not received as much attention as it actually deserves is the principle with which the reanalysis of subject forms as Focus markers has been explained, i.e. the phenomenon of markedness reversal (Mayerthaler 1981: 48) or deviation from default expectation (Haspelmath 2006: 62). According to this principle, subject pronoun case forms can be used as Focus markers in subject predicative complements simply because they are usually restricted to the noun phrase slots immediately preceding the finite verb. Hence, they are very unexpected in postverbal subject predicative complement position, which makes their actual occurrence in these contexts even more salient (cf. Section 4.3).

Interestingly, the principle of markedness reversal or deviation from default expectation can be used not only to account for the variation observed in subject predicative complements, but its range of application might also be extended to cover other contexts which allow for variability in the use of pronoun case forms in English.

Markedness reversal could, mutatis mutandis, also account for the use of subject pronoun case forms in coordinated object noun phrases. In analogy to the use of subject pronouns in subject predicative complements, it is conceivable that subject pronoun forms are used in coordinated object noun phrases in order to explicitly mark the referents of
these pronouns as the Focus or as the most important information of the clause (cf. (144a) and (144b)):

(144) Subject Pronouns in Coordinated Object Noun Phrases (Huddleston and Pullum 2002: 463)
   a. The present was supposed to represent Helen and I, that was the problem.
   b. There’s a tendency for he and I to clash.

In the relevant literature, uses of subject pronouns in coordinated object noun phrases are very often merely considered as hypercorrections resulting from the avoidance of using object pronoun forms in coordinated subjects (e.g. Huddleston and Pullum 2002: 459; Quirk et al. 1985: 338). However, the principle of markedness reversal can offer a functional motivation for the use of subject pronouns in coordinated objects. Moreover, this functional explanation can also explain why the use of coordinated noun phrases in object position, and among them particularly constructions with final subject I, are “so common in speech and used by so broad a range of speakers that it has to be recognised as a variety of Standard English” (Huddleston and Pullum 2002: 463). If this use of subject pronoun forms in coordinated object noun phrases were due solely to hypercorrection, it is at least questionable whether it would occur so frequently as to be considered a form of Standard English. Moreover, matters of formality have also been deemed to be of crucial importance for the distribution of pronoun case forms in subject predicative complements (cf. Section 6.1.2.4). However, as the results of this study have demonstrated, other functional factors impact the distribution of pronoun case forms far more than the degree of formality. As a consequence, it is plausible and likely that functional and particularly pragmatic factors such as FOCUS also play a role in the distribution of pronoun case forms in coordinated object noun phrases (cf. Angermeyer and Singler 2003: 197). Thus, future research may test whether or not markedness reversal and a possible re-functionalisation of subject pronoun forms as Focus markers may also be partly accountable for the use of subject forms in coordinated object noun phrases and even perhaps for the use of subject pronoun forms in postverbal contexts in general.

Moreover, the principle of markedness reversal or deviation from default expectation can be used not only to account for the use of subject pronoun forms in postverbal noun phrase slots, but also for the opposite scenario, i.e. the use of object pronoun case forms in preverbal “subject territory” (Quirk et al. 1985: 337). The preceding chapters have shown that subject pronoun forms are used as Focus markers in subject predicative complements,
at least in British, American, Australian and Irish English. Conversely, we could also hypothesise, however, that object pronoun forms have been re-functionalised in subject territory to mark the Topic of the sentence, i.e. “what the utterance is primarily about” (Siewierska 1991: 149). Cases in point corroborating such an assumption are, for example, left dislocations, i.e. contexts in which a noun phrase occurs in sentence-initial position and an anaphoric pronoun is used as its substitute in the corresponding argument position of the following sentence (cf. Ward and Birner 2004: 162; Quirk et al. 1985: 1310; Section 2.3). Although this context has traditionally been considered as allowing for variation in the use of pronoun case forms (e.g. Erdmann 1978: 68-69; Jespersen and Haislund 1949: 223-225; Quinn 2005a: 214-228), the variability of this context has been questioned since the object form is deemed to be not the preferred, but actually the sole option in left dislocations (Huddleston and Pullum 2002: 462):

(145) Object Pronoun in Left Dislocation (Huddleston and Pullum 2002: 462)
Me, I wouldn’t trust him further than I could throw him.

Indeed, it could be argued that this clear preference of object pronouns in left dislocations is also due to the principle of markedness reversal whereby an unexpected object pronoun form is used in subject territory to mark a pragmatically salient constituent, i.e. in this case the Topic of the sentence. Such an assumption is supported by accounts stating that the functions of left dislocations include both “emphasis” (Biber et al. 1999: 956) as well as establishing a topic (Biber et al. 1999: 957). Furthermore, this preliminary hypothesis is also corroborated by the fact that subject pronouns are very often repeated sentence-initially in order to buy processing time (Biber et al. 1999: 334-335):

(146) Repetition of Subject Pronouns in Sentence-Initial Position
b. And he he got, eh got my hand for five seconds and then I I got – I got his for ten something like that. (CONV) (Biber et al. 1999: 335)

As a consequence, object pronouns are much more salient in sentence-initial position and thus may be better suited to mark the referent of a pronoun explicitly as the Topic of the following clause not only because they are unexpected per se in subject territory, but also because repeated, sentence-initial subject pronoun forms may be rather perceived as processing time buyers than as pragmatically salient Topic markers. Thus, the principle of markedness reversal as well as the reanalysis of former case exponents to pragmatic
markers can account even for the use of object pronouns in subject territory. Hence, this principle as well as the pragmatic reanalysis of former case forms which has already been attested for subject predicative complements in this study can reconcile accounts on the distribution of pronoun case forms that seem at first glance diametrically opposed to each other. Whereas some studies observe a “general tendency in colloquial English to use subject forms in non-subject functions, e.g. after prepositions (between you and I) [...]” (Wagner 2004: 159), other accounts note a “general tendency [...] for the accusative forms to spread into contexts traditionally associated with the nominative case” (Biber et al. 1999: 335). Although these statements may seem to exclude each other, these two opposing views or observations are highly compatible with the principle of markedness reversal and the reanalysis of former case forms to pragmatic markers.

Therefore, the principle of markedness reversal as well as the pragmatic re-functionalisation of the remnants of the former case system are promising starting points for future studies examining the distribution of pronoun case forms in a wide range of contexts allowing for variability.

Moreover, future research may not focus only on pragmatic factors but should also try to assess the impact of sociolinguistic factors other than REGIONAL VARIETY and MODE OF DISCOURSE on the use and distribution of pronoun case forms in subject predicative complements. While the speaker’s age, sex and level of education seem to influence the distribution of pronoun case forms in other variable contexts such as coordinated noun phrases (cf. Angermeyer and Singler 2003: 184-193), the impact of these factors on the distribution of pronoun case forms in subject predicative complements unfortunately could not be assessed in this study because the databases examined for this study do not allow for fine-grained sociolinguistic analysis. Corpora that do contain enough metadata to allow for thorough sociolinguistic studies, such as the components of the ICE, do not, however, contain enough subject predicative complements to conduct thorough quantitative analyses of these variables. Thus, future research may try to use experimental techniques to elicit enough data from a sufficient number of experimental subjects in order to enable an analysis of a wider range of sociolinguistic variables than has been so far possible.

Another open issue pertains to the historical dimension of the uncovered re-functionalisation process. The results of this study clearly demonstrate that former case forms have been reanalysed as pragmatic markers in several varieties of English.
Consequently, the question arises when this re-functionalisation process began. Did a systematic reanalysis begin immediately after the decline of the former case system or was there an in-between stage in which the former case forms were simply retained as non-functional remnants (cf. Section 4.2; Lass 1990)? This, too, is a very interesting issue worth addressing in future research, although it may not be easy to deal with in view of the low frequency of subject predicative complements, which is further exacerbated by the overall modest size of most historical corpora and the scarcity of data for Old and Middle English.

Moreover, it is worth asking where this re-functionalisation of former case forms to markers of pragmatic prominence has its origins. Does it have its roots in those regional varieties where today, it is most pronounced and most concrete in the phenomenon of pronoun exchange, and has it then somehow diffused back in an attenuated form into the more Standard-like varieties? Or is it even the other way round? This, in turn, may help to account for the vast cross-varietal differences that have been observed in the present study (cf. Section 15.4). These are also intriguing questions that deserve further consideration in future research, even though they may not be easy to answer.

Finally, this study has shown that computer-mediated communication can use additional or alternative strategies to highlight the Focus of a clause when compared to traditional written data (cf. Herring 2010a). Due to this study’s interest in the distribution of pronoun case forms, the quantification of these alternative focussing strategies was unfortunately beyond the scope of this thesis. Thus, future studies investigating the particularities of CMC should further analyse the whole set of available focussing devices in CMC since the use of different focussing devices might provide insights into those contexts in which CMC is more similar to either spoken or traditional written language.

Thus, much work lies ahead in order to fully understand the distribution and exact functions of pronoun case forms in Present-Day English. Nevertheless, I hope to have contributed through this thesis to a better understanding of the distribution of pronoun case forms in English in general and in subject predicative complements in particular.
Appendix A: Multicollinearity Statistics

As noted in Section 6.2, independent variables must not strongly correlate with each other in regression modelling, since the resulting multicollinearity may lead to unreliable regression coefficients and unstable results (e.g. Backhaus et al. 2008: 87-89; Szmrecsanyi 2006: 54). Thus, the following sections report the multicollinearity statistics for each regression model presented and discussed in this study (cf. Chapters 8–14). In particular, Variance Inflation Factors are reported for each regression model, because they are probably one of the most common multicollinearity measures for binary logistic regression models (e.g. Peukert 2012; Szmrecsanyi 2006: 54, 215-218). As a rule of thumb, Variance Inflation Factors beyond ten clearly indicate multicollinearity (e.g. Schlittgen 2004: 267; Szmrecsanyi 2006: 215). However, there is no consensus as to what value constitutes a critical value the exceedance of which may require additional caution. Some accounts consider values below four as unproblematic (e.g. Wollschläger 2010: 188). Other accounts state that already values above 2.5 may be problematic (Szmrecsanyi 2006: 215), while again other authors correctly state that in principle any value larger than one may potentially be a problem (Field 2009: 224). In view of this uncertainty, this study adopts Szmrecsanyi’s (2006: 215) point of view and consider Variance Inflation Factors with values below 2.5 as unproblematic. If a model exhibits, however, a variable with a Variance Inflation Factor exceeding 2.5, this study also provides the Condition Index for the respective regression model, which is another multicollinearity measure. With regard to Condition Indices, values above 20 may indicate again multicollinearity (e.g. Schlittgen 2004: 267; Wollschläger 2010: 189). Hence, any Condition Index below 20 is considered acceptable in this study. For further information on Variance Inflation Factors, Condition Indices and regression diagnostics in general, the interested reader is referred to the references quoted in this paragraph.
Multicollinearity Statistics for the British National Corpus

Multicollinearity Statistics for the Subject Predicative Complements in the British National Corpus

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>2.515</td>
</tr>
<tr>
<td>FIRST</td>
<td>3.791</td>
</tr>
<tr>
<td>SPOK</td>
<td>2.369</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.195</td>
</tr>
<tr>
<td>SPOK*CLEFT</td>
<td>2.156</td>
</tr>
<tr>
<td>CLEFT*FIRST</td>
<td>4.590</td>
</tr>
</tbody>
</table>

Table: Subject Predicative Complements in the BNC: Variance Inflation Factors

<table>
<thead>
<tr>
<th>CONDITION INDEX</th>
<th>VARIANCE DECOMPOSITION PROPORTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INTERCEPT</td>
</tr>
<tr>
<td>1.000</td>
<td>0.034</td>
</tr>
<tr>
<td>1.549</td>
<td>0.000</td>
</tr>
<tr>
<td>1.657</td>
<td>0.009</td>
</tr>
<tr>
<td>2.029</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table: Subject Predicative Complements in the BNC: Condition Index

Multicollinearity Statistics for the *it BE* Sentences in the British National Corpus

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPOK</td>
<td>1.066</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.066</td>
</tr>
</tbody>
</table>

Table: *it BE* Sentences in the BNC: Variance Inflation Factors
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Multicollinearity Statistics for the *it*-Clefts in the British National Corpus

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.142</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.176</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.075</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the BNC: Variance Inflation Factors

Multicollinearity Statistics for the Corpus of Contemporary American English

Multicollinearity Statistics for the Subject Predicative Complements in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>2.013</td>
</tr>
<tr>
<td>FIRST</td>
<td>2.779</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.022</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.022</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.079</td>
</tr>
<tr>
<td>CLEFT*FIRST</td>
<td>4.117</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the COCA: Variance Inflation Factors

<table>
<thead>
<tr>
<th>CONDITION INDEX</th>
<th>VARIANCE DECOMPOSITION PROPORTIONS</th>
<th>VARIANCE DECOMPOSITION PROPORTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INTERCEPT</td>
<td>CLEFT</td>
</tr>
<tr>
<td>1.000</td>
<td>0.008</td>
<td>0.023</td>
</tr>
<tr>
<td>1.756</td>
<td>0.000</td>
<td>0.071</td>
</tr>
<tr>
<td>1.920</td>
<td>0.000</td>
<td>0.137</td>
</tr>
<tr>
<td>2.095</td>
<td>0.000</td>
<td>0.335</td>
</tr>
<tr>
<td>3.367</td>
<td>0.029</td>
<td>0.287</td>
</tr>
<tr>
<td>8.000</td>
<td>0.962</td>
<td>0.147</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the COCA: Condition Index*
Multicollinearity Statistics for the *it BE* Sentences in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIRST</td>
<td>1.004</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.004</td>
</tr>
</tbody>
</table>

*Table: it BE Sentences in the COCA: Variance Inflation Factors*

Multicollinearity Statistics for the *it*-Clefts in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.163</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.175</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.042</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.093</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the COCA: Variance Inflation Factors*
Multicollinearity Statistics for the British (.uk) Internet Data

Multicollinearity Statistics for the Subject Predicative Complements in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>2.649</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.349</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.448</td>
</tr>
<tr>
<td>CMC</td>
<td>1.033</td>
</tr>
<tr>
<td>FOCUS</td>
<td>2.510</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.377</td>
</tr>
</tbody>
</table>

Table: Subject Predicative Complements in the .uk Data: Variance Inflation Factors

<table>
<thead>
<tr>
<th>CONDITION INDEX</th>
<th>VARIANCE DECOMPOSITION PROPORTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INTERCEPT</td>
</tr>
<tr>
<td>1.000</td>
<td>0.010</td>
</tr>
<tr>
<td>1.778</td>
<td>0.000</td>
</tr>
<tr>
<td>2.388</td>
<td>0.002</td>
</tr>
<tr>
<td>2.530</td>
<td>0.002</td>
</tr>
<tr>
<td>2.862</td>
<td>0.006</td>
</tr>
<tr>
<td>6.245</td>
<td>0.980</td>
</tr>
</tbody>
</table>

Table: Subject Predicative Complements in the .uk Data: Condition Index

Multicollinearity Statistics for the it BE Sentences in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGULAR</td>
<td>1.002</td>
</tr>
<tr>
<td>CMC</td>
<td>1.011</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.012</td>
</tr>
</tbody>
</table>

Table: it BE Sentences in the .uk Data: Variance Inflation Factors
Multicollinearity Statistics for the *it*-Clefts in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.042</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.296</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.278</td>
</tr>
<tr>
<td>CMC</td>
<td>1.050</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.305</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the .uk Data: Variance Inflation Factors*

Multicollinearity Statistics for the Australian (.au) Internet Data

Multicollinearity Statistics for the Subject Predicative Complements in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>1.699</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.381</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.586</td>
</tr>
<tr>
<td>CMC</td>
<td>1.044</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.562</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.557</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .au Data: Variance Inflation Factors*

Multicollinearity Statistics for the *it BE* Sentences in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOCUS</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table: it BE Sentences in the .au Data: Variance Inflation Factor*
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Multicollinearity Statistics for the *it*-Clefts in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.095</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.726</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.659</td>
</tr>
<tr>
<td>CMC</td>
<td>1.723</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.747</td>
</tr>
<tr>
<td>CMC*FIRST</td>
<td>1.941</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the .au Data: Variance Inflation Factors*

Multicollinearity Statistics for the Irish (.ie) Internet Data

Multicollinearity Statistics for the Subject Predicative Complements in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>1.412</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.050</td>
</tr>
<tr>
<td>CMC</td>
<td>1.022</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.461</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .ie Data: Variance Inflation Factors*

Multicollinearity Statistics for the *it* BE Sentences in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOCUS</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table: it BE Sentences in the .ie Data: Variance Inflation Factor*
Multicollinearity Statistics for the *it*-Clefts in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.052</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.061</td>
</tr>
<tr>
<td>CMC</td>
<td>1.009</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the .ie Data: Variance Inflation Factors*

Multicollinearity Statistics for the South African (.za) Internet Data

Multicollinearity Statistics for the Subject Predicative Complements in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>1.120</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.463</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.719</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.189</td>
</tr>
<tr>
<td>CMC</td>
<td>1.180</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.535</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .za Data: Variance Inflation Factors*

Multicollinearity Statistics for the *it BE* Sentences in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGULAR</td>
<td>1.061</td>
</tr>
<tr>
<td>FOCUS</td>
<td>1.061</td>
</tr>
</tbody>
</table>

*Table: *it BE Sentences in the .za Data: Variance Inflation Factors*
Multicollinearity Statistics for the *it*-Clefts in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.079</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.888</td>
</tr>
<tr>
<td>SINGULAR</td>
<td>1.444</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.683</td>
</tr>
<tr>
<td>CMC</td>
<td>1.167</td>
</tr>
<tr>
<td>FIRST*SINGULAR</td>
<td>1.488</td>
</tr>
<tr>
<td>FIRST*SPOK</td>
<td>1.863</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the .za Data: Variance Inflation Factors*

Multicollinearity Statistics for the Indian (.in) Internet Data

Multicollinearity Statistics for the Subject Predicative Complements in the Indian (.in) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEFT</td>
<td>1.063</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.016</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.116</td>
</tr>
<tr>
<td>CMC</td>
<td>1.091</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .in Data: Variance Inflation Factors*

Multicollinearity Statistics for the *it*-Clefts in the Indian (.in) Internet Data

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>VARIANCE INFLATION FACTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS_SUBJ</td>
<td>1.050</td>
</tr>
<tr>
<td>FIRST</td>
<td>1.053</td>
</tr>
<tr>
<td>SPOK</td>
<td>1.005</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the .za Data: Variance Inflation Factors*
Appendix B: Goodness of Fit Statistics

As discussed in Section 6.2, this study also reports values indicating the classificatory accuracy of the applied regression models. Thus, this study states the C statistics and Somer’s $D_{xy}$ for each minimal-adequate regression model discussed in this study (cf. Chapters 8–14). For more information on the interpretation of the reported values, the interested reader is either referred to Baayen (2008: 204) or to Section 6.2.

Goodness of Fit Statistics for the British National Corpus

Goodness of Fit Statistics for the Subject Predicative Complements in the British National Corpus

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.905</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.811</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the BNC: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it BE* Sentences in the British National Corpus

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.64</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.28</td>
</tr>
</tbody>
</table>

*Table: it BE Sentences in the BNC: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the British National Corpus

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.865</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.731</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the BNC: Goodness of Fit Statistics*
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Goodness of Fit Statistics for the Corpus of Contemporary American English

Goodness of Fit Statistics for the Subject Predicative Complements in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.938</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.877</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the COCA: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it BE* Sentences in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.657</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.314</td>
</tr>
</tbody>
</table>

*Table: *it BE* Sentences in the COCA: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the Corpus of Contemporary American English

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.854</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.709</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the COCA: Goodness of Fit Statistics*
Goodness of Fit Statistics for the British (.uk) Internet Data

Goodness of Fit Statistics for the Subject Predicative Complements in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.873</td>
</tr>
<tr>
<td>SOMER’S ( D_{xy} )</td>
<td>0.745</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .uk Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it BE* Sentences in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.905</td>
</tr>
<tr>
<td>SOMER’S ( D_{xy} )</td>
<td>0.81</td>
</tr>
</tbody>
</table>

*Table: *it BE* Sentences in the .uk Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the it-Clefts in the British (.uk) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.805</td>
</tr>
<tr>
<td>SOMER’S ( D_{xy} )</td>
<td>0.61</td>
</tr>
</tbody>
</table>

*Table: it-Clefts in the .uk Data: Goodness of Fit Statistics*
Goodness of Fit Statistics for the Australian (.au) Internet Data

Goodness of Fit Statistics for the Subject Predicative Complements in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.923</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.846</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .au Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it* BE Sentences in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.685</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.369</td>
</tr>
</tbody>
</table>

*Table: *it* BE Sentences in the .au Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the Australian (.au) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.884</td>
</tr>
<tr>
<td>Somer’s $D_{xy}$</td>
<td>0.768</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the .au Data: Goodness of Fit Statistics*
**Goodness of Fit Statistics for the Irish (.ie) Internet Data**

Goodness of Fit Statistics for the Subject Predicative Complements in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.874</td>
</tr>
<tr>
<td>Somer's $D_{xy}$</td>
<td>0.748</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .ie Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it* $BE$ Sentences in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.691</td>
</tr>
<tr>
<td>Somer's $D_{xy}$</td>
<td>0.383</td>
</tr>
</tbody>
</table>

*Table: *it* $BE$ Sentences in the .ie Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the Irish (.ie) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.848</td>
</tr>
<tr>
<td>Somer's $D_{xy}$</td>
<td>0.697</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the .ie Data: Goodness of Fit Statistics*
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

**Goodness of Fit Statistics for the South African (.za) Internet Data**

Goodness of Fit Statistics for the Subject Predicative Complements in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.86</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.72</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .za Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it BE* Sentences in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.668</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.335</td>
</tr>
</tbody>
</table>

*Table: *it BE* Sentences in the .za Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the South African (.za) Internet Data

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.822</td>
</tr>
<tr>
<td>SOMER’S $D_{xy}$</td>
<td>0.643</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the .za Data: Goodness of Fit Statistics*
Goodness of Fit Statistics for the Indian (.in) Internet Data

Goodness of Fit Statistics for the Subject Predicative Complements in the Indian (.in) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.89</td>
</tr>
<tr>
<td>SOMER’S Dₜₛₜ</td>
<td>0.781</td>
</tr>
</tbody>
</table>

*Table: Subject Predicative Complements in the .in Data: Goodness of Fit Statistics*

Goodness of Fit Statistics for the *it*-Clefts in the Indian (.in) Internet Data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.812</td>
</tr>
<tr>
<td>SOMER’S Dₜₛₜ</td>
<td>0.623</td>
</tr>
</tbody>
</table>

*Table: *it*-Clefts in the .za Data: Goodness of Fit Statistics*
References

Primary Sources

Davies, Mark

2004– BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at: <http://corpus.byu.edu/bnc>, July 22, 2010 [date of data compilation].


Secondary Sources

Aarts, Bas and April McMahon (eds.)


Anderwald, Lieselotte


Angermeyer, Philipp S. and John Victor Singler


Ariel, Mira


Aston, Guy and Lou Burnard

Ayafor, Miriam

Baayen, Rolf Harald

Backhaus, Klaus, Bernd Erichson, Wulff Plinke and Rolf Weiber

Bailey, Beryl Loftman

Bailey, Robert and Otto Santa Ana

Baroni, Marco and Silvia Bernardini (eds.)

Bauer, Laurie

Beal, Joan


Beal, Kalynda

Bernardini, Silvia, Marco Baroni and Stefan Evert
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English


<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Title</th>
<th>Publisher/Location</th>
<th>Pages/Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burridge, Kate</td>
<td>2004</td>
<td>Synopsis: morphological and syntactic variation in the Pacific and</td>
<td>Berlin/New York: Mouton de Gruyter</td>
<td>1116–1131</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Australasia. In: Kortmann et al. (eds.), A Handbook of Varieties of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>English, Volume 2: Morphosyntax.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>In: Reuland, Eric (ed.), Arguments and Case: Explaining Burzio’s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Butt, Miriam</td>
<td>2006</td>
<td>Theories of Case.</td>
<td>Cambridge: Cambridge University</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Press.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bybee, Joan</td>
<td>2010</td>
<td>Language, Usage and Cognition.</td>
<td>Cambridge: Cambridge University</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Press.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chomsky, Noam</td>
<td>1981</td>
<td>Lectures on Government and Binding.</td>
<td>Dordrecht: Foris</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>J. Keyser (eds.), The view from Building 20: Essays in Linguistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>in Honor of Sylvain Bromberger [Current Studies in Linguistics 24].</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Press.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collins, Peter</td>
<td>1989</td>
<td>Divided and debatable Usage in Australian English. In: Collins, Peter</td>
<td>St Lucia: University of Queensland</td>
<td>138–149</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Routledge.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Collins, Peter and Pam Peters
2004

Colomb, Gregory G. and Joyce A. Simutis
1996

Cooper, Christopher
1685 [1968]

Cotterill, Janet
2010

Crawley, Michael J.
2007
*The R Book*. Chichester: Wiley.

Croft, William
2003

Crowley, Terry
2004

Crystal, David
2011

Denison, David
1993

Diessel, Holger
2004
*The Acquisition of Complex Sentences*. Cambridge: Cambridge University Press.

Dik, Simon C.
1978


Domsch, Sebastian  

Edwards, Viv  

Eid, Michael, Mario Gollwitzer and Manfred Schmitt  

Elworthy, Frederic T.  

Emonds, Joseph  


Erdmann, Peter  

Escure, Geneviève  

Faráclas, Nicholas  

The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English


Green, Lisa J.  

Greenbaum, Sidney  


Greenbaum, Sidney (ed.)  

Gries, Stefan, Thomas  


Haegeman, Liliane  

Haegeman, Liliane and Jacqueline Guéron  

Hancock, I.  


Hilpert, Martin

Hoffmann, Sebastian

Hollmann, Willem B.

Hopkins, Edwin A.

Hosali, Priya

Huang, Yan

Huber, Magnus


Huddleston, Rodney and Geoffrey Pullum

Hudson, Richard

Hundt, Marianne, Jennifer Hay and Elizabeth Gordon
Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (eds.)
2007a  *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi.

Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer

Ihalainen, Ossi


James, Winford and Valerie Youssef

Jespersen, Otto

Jespersen, Otto and Niels Haislund

Johnson, Daniel Ezra

Johnston, Paul A.

Joseph, Brian

Jourdan, Christine
Kachru, Braj, Yamuna Kachru and Cecil L. Nelson (eds.)  
2006  

Kautzsch, Alexander  
2004  

Keller, Frank and Mirella Lapata  
2003  

Kelly, Sr. Francis  
1975  
The English spoken colloquially by a group of adolescents in Suva. Fiji English Teachers Journal 11: 19–43.

Kerswill, Paul  
2007  

Kilgariff, Adam  
2006a  

2006b  

Kilgariff, Adam and Gregory Grafenstette  
2003  

Kiparsky, Paul  
1997  

Kjellmer, Göran  
1986  

Klima, Edward S.  
1964  

Kortmann, Bernd  
2004a  
### The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Title</th>
</tr>
</thead>
</table>


Leech, Geoffrey 2007 New resources, or just better old ones?. In: Hundt, Marianne, Nadja Nesselhauf and Carolin Biewer (eds.), *Corpus Linguistics and the Web.* Amsterdam/New York: Rodopi, 133–149.

Leech, Geoffrey, Paul Rayson and Andrew Wilson 2001 *Word Frequencies in Written and Spoken English – Based on the British National Corpus.* Harlow: Longman.


Maier, Georg 2013 As the Case May Be: A Corpus-Based Approach to Pronoun Case Variation in Subject-Predicative Complements in British and American English. In: Mukherjee, Joybrato and Magnus Huber (eds.), *Proceedings of ICAME 31.* Varieng: Studies in Variation, Contacts and Change in English.
Mair, Christian

Malcolm, Ian G.

Manning, Christopher D. and Hinrich Schütze

Mason, C. P.

Mayerthaler, Willi

Maynor, Natalie

McCormick, Kay

McEnery, Tony, Richard Xiao and Yukio Tono

Melchers, Gunnel
Menuet, Christian  

Meyer, Charles F.  

Meyer, Charles F., Roger Grabowski, Hung-Yul Han, Konstantin Mantzouranis and Stephanie Moses  

Miller, Jim  


Milroy, James  

Milroy, James and Lesley Milroy (eds.)  

Montgomery, Michael B.  


Mufwene, Salikoko S.  
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Mugler, France and Jan Tent  
2004  

Mühlhäusler, Peter  
1978  
Pidginization and Simplification of Language. Canberra: Australian National University.

1983  

1997  

Mukherjee, Joybrato  
2009  

Murray, Thomas E. and Beth Lee Simon  
2004  

Mustanoja, T. F.  
1960  

Nelson, Gerald  
1996  

Newhook, Amanda  
2002  

O’ Brien, Robert M.  
2007  

O’Keeffe, Anne and Michael McCarthy (eds.)  
2010  
Odlin, Terence  
1997  

OED Online  
2012  

Paddock, Harold  
1991  

1994  

Palmer, Harold E.  
1929  

Pampel, Fred C.  
2000  

Parker, Frank, Kathryn Riley and Charles Meyer  
1988  

Parrott, Jeffrey K.  
2007  

Patrick, Peter L.  
2004  

Pawley, Andrew  
2004  
Peitsara, Kirsti

Peng, Chao-Ying Joanne, Kuk Lida Lee and Gary M. Ingersoll

Penhallurick, Robert


Peters, Pam, Peter Collins and Adam Smith (eds.)

Peukert, Hagen
2012 Hidden Structures in English Corpora. In: Mukherjee, Joybrato and Magnus Huber (eds.), *Corpus Linguistics and Variation in English: Theory and Description* [Language and Computers 75]. Amsterdam: Rodopi, 131–141.

Phythian, B. A.

Pietsch, Lukas

Poussa, Patricia

Quattlebaum, Judith A.

Quinn, Heidi
2005a The Distribution of Pronoun Case Forms in English. Amsterdam: Benjamins.


Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik


Radford, Andrew


The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Ramisch, Heinrich  

Reaser, Jeffrey and Benjamin Torbert  

Reese, Mario  

Rickford, John, R.  

Rogers, Norman  

Rohdenburg, Günter  

Rosenbach, Anette  

Sailaja, Pingali  

Sakoda, Kent and Jeff Siegel  

Sand, Andrea  


Schilk, Marco 2011 *Structural Nativization in Indian English Lexicogrammar.* Amsterdam/Philadelphia: John Benjamins.


The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English


Shnukal, Anna 1988  *Broken: An Introduction to the Creole Language of Torres*. Canberra: Research School of Pacific and Asian Studies, Australian National University.


2004  *Person*. Cambridge: Cambridge University Press.

Singler, John Victor  

Sistren, with Honor Ford-Smith  

Smith, Geoff  

Sobin, Nicholas  

Story, G., W. J. Kirwin and J. D. A. Widdowson  

Sutcliffe, David  

Swan, Michael  

Sweet, Henry  


Szmrecsanyi, Benedikt  

Tagliamonte, Sali A.  
The Distribution of Pronoun Case Forms in Subject Predicative Complements in Varieties of English

Thomas, Larry
1991 3 plays: Outcast; Yours dearly; Men, Women and Insanity. Suva: University of the South Pacific.

Treble, Henry Arthur and George Henry Vallins

Trudgill, Peter

Trudgill, Peter (ed.)

Trudgill, Peter and J. K. Chambers (eds.)

Trudgill, Peter and J. K. Chambers

Trudgill, Peter and Jean Hannah

Ueyama, Motoko

Váradi, Tamás

Volk, Martin
Wagner, Susanne  


Wakelin, Martyn  

Wales, Katie  

Ward, Gregory and Betty Birner  

Weinert, Regina and Jim Miller  

Weinrich, Harald  

Wierzbicka, Anna  

Wilson, Sheila and Rajend Mesthrie  

Winford, Donald and Bettina Migge  

Wolfram, Walt  


**Web Sources**


“IE Domain Registry”. Ireland’s Domain Registry. 


“Registrations statistics”. Nominet. 


Liste der aus dieser Dissertation hervorgegangenen Vorveröffentlichungen

Maier, Georg
2013

As the Case May Be: A Corpus-Based Approach to Pronoun Case Variation in Subject-Predicative Complements in British and American English. In: Mukherjee, Joybrato and Magnus Huber (eds.), Proceedings of ICAME 31. VariEng: Studies in Variation, Contacts and Change in English.

Maier, Georg
2014


Maier, Georg
2009

Summary of the Major Findings/Zusammenfassung der Ergebnisse der Dissertation

A.) English Summary: The Distribution of Pronoun Case Forms in Subject Predicative Complements: A Corpus- and Web-Based Study of Pronoun Case Variation

As early as the 1930s, Jespersen (1933) noted that the “right use” of pronominal case forms “is one of the knottiest points in English grammar” (Jespersen 1933: 132). The reason for this knottiness is the fact that the distribution of subject and object pronoun case forms in English is not as “straightforward” as is often assumed (Biber et al. 1999: 335). Although subject pronoun forms are indeed mainly used as subjects of finite clauses and object pronoun forms largely as objects of transitive verbs and as complements of prepositions (cf. Biber et al. 1999: 335), Present-Day English also exhibits a considerable number of contexts that permit a choice between the use of either subject or object pronoun forms.59 As summarised in Table 1, prominent examples of these contexts that allow for variability in the use of pronoun case forms include the focal position of it-clefts, subject predicative complements following it and a form of BE, i.e. it BE sentences (cf. Quinn 2005a), coordinated noun phrases in subject and object position, pronoun–noun phrase constructions and than-comparatives to mention but a few:

<table>
<thead>
<tr>
<th>Contexts Exhibiting Variability in the Use of Pronoun Case Forms</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>It-clefts</td>
<td>It is they/them who are to blame</td>
</tr>
<tr>
<td><em>It BE</em> Sentences</td>
<td>[Who is it? –] It is I/me</td>
</tr>
<tr>
<td>Coordinated NPs in Subject Position</td>
<td>Rita and I/me will get divorced</td>
</tr>
<tr>
<td>Coordinated NPs in Object Position</td>
<td>This issue has to stay between you and I/me</td>
</tr>
<tr>
<td>Pronoun–Noun Phrase Constructions</td>
<td>We/Us Irish are a happy people</td>
</tr>
<tr>
<td>Than-Comparatives</td>
<td>Susan is younger than he/him</td>
</tr>
</tbody>
</table>

Table 1: Variable Contexts of Pronoun Case Distribution in Present-Day English

---

59 This study uses the terms subject pronoun and object pronoun instead of the also common labels subjective pronoun and objective pronoun or nominative and accusative pronoun because they are probably the most neutral ones simply referring to the prototypical function of these pronouns in a clause.
Although much scholarly attention has been devoted to these contexts, so far only a few studies have analysed these phenomena quantitatively (e.g. Angermeyer and Singler 2003; Biber et al. 1999; Erdmann 1978; Maier 2013; Quinn 2009). This scarcity of quantitative studies is particularly true for the first two contexts which allow for variability in the use of subject pronoun case forms listed in Table 1, i.e. *it*-clefs and *it BE* sentences. These are subsumed under the heading of subject predicative complements (cf. Biber et al. 1999: 335-336; Maier 2013) and are the main focus of this study.

Indeed, there are not many quantitative studies so far – let alone multivariate ones – that try to examine and account for the distribution of pronoun case forms in *it BE* sentences and *it*-clefs subsumed in the general category of subject predicative complements. However, there are several good reasons why a quantitative analysis of particularly these contexts can provide very valuable insights, not only regarding the contexts under consideration, but also for other contexts allowing for variability, as well as for the distribution of pronoun case forms in Present-Day English in general.

To begin with, the distribution of pronoun case forms in subject predicative complements has been the subject of linguistic reasoning for more than a century (e.g. Sweet 1875: 495). However, even though the variability in the use of pronoun case forms in both focal pronoun position of *it*-clefs and *it BE* sentences is still discussed today in most comprehensive grammar books of English (e.g. Biber et al. 1999; Greenbaum 1996a; Jespersen 1933; Quirk et al. 1985; Huddleston and Pullum 2002), there are hardly any quantitative studies examining the distribution of pronoun case forms in these two contexts – neither within nor across varieties of English (cf. Maier 2013; Quinn 2009). Thus, by analysing two mega-corpora as well as five Web-derived datasets that were specifically compiled for the present purpose, this study sheds more light on the use and distribution of pronoun case forms in *it*-clefs and *it BE* sentences in the analysed varieties, i.e. British, American, Australian, Irish, South African and Indian English.

Moreover, much of the current linguistic theory discussing contexts allowing for variability in the use of pronoun case forms assumes that the distribution of pronoun case forms in *it*-clefs and *it BE* sentences should be the same (e.g. Burridge 2004; Emonds 1985; Harris 1981; Quinn 2005a). Thus, by analysing these two constructions, this study is able to test and evaluate the predictions and validity of some well-established theoretical approaches accounting for the distribution of pronoun case forms in Present-Day English.
Finally, the focus of this study on the distribution of pronoun case forms in subject predicative complements also results from this study’s central assumption, i.e. that subject pronoun case forms have been reanalysed as Focus markers in subject predicative complements. Indeed, this study argues that the more focussed a subject predicative complement context is, the more likely it is to observe a subject pronoun case form in a subject predicative complement context. Admittedly, this assumption may appear initially rather unorthodox given the expectations and approaches of much of the current linguistic theory as well as of accounts associating subject pronouns with preverbal, unstressed topical occurrences (e.g. Pietsch 2007: 167, 2009: 146). However, based on a comprehensive survey of the distribution and use of pronoun case forms in varieties of English and on a thorough survey of previous research, this study is able to show that functional and particularly pragmatic factors are very important in determining the distribution of pronoun case forms both in varieties of English and in other contexts allowing for variability in the use of pronoun case forms (e.g. Angermeyer and Singler 2003; Huber 1999; Paddock 1994). As a consequence, this study proposes a Focus-oriented approach to the distribution of pronoun case forms in subject predicative complements that can account for such an assumed re-functionalisation process and helps to explain the considerable extent of variation that can be observed in the distribution of pronoun case forms in it BE sentences and it-clefs in Present-Day English (cf. Erdmann 1978; Maier 2013).

On a very general level, the main objectives of this study are to analyse the distribution of pronoun case forms in subject predicative complements in six varieties of English and thus to shed further light on the factors and variables influencing and determining the distribution of pronoun case forms in English in general.

Moreover, this study sets out to demonstrate that functional and particularly pragmatic factors play a prominent role in the distribution of pronoun case forms in Present-Day English. In order to achieve this, this study analyses two mega-corpora of British and American English, i.e. the British National Corpus (Davies 2004–) and the Corpus of Contemporary American English (Davies 2008–), as well as five prospectively compiled, Web-derived datasets of British, Australian, Irish, South African and Indian English. By providing a

---

Whenever this study uses the concept of Focus as defined in the framework of Functional Grammar, it also adopts the convention of spelling Focus with a capital letter. In Functional Grammar, Focus is defined as the relatively most important, salient or highlighted information in a clause (cf. Dik 1978: 130; Siewierska 2004: 159; Weinert and Miller 1996: 179).
meticulous quantitative analysis of the distributions of pronoun case forms in six regional varieties of English, this study is not only able to assess which factors influence the distribution of pronoun case forms within varieties of English, and to what extent, but it can also determine whether there are any cross-varietal differences or trends observable in the distribution of pronoun case forms in subject predicative complements. In addition, this study empirically tests the Focus-oriented approach to pronoun case distribution, particularly its central assumption proposing that subject pronouns have been reanalysed as Focus markers in subject predicative complement position. The final question this study addresses pertains to the adopted methodology. This study assesses whether or not it is possible to arrive at robust conclusions and generalisations with the help of data obtained from Google.

Since the main objective of this study is to account for the distribution of pronoun case forms in subject predicative complements in six varieties of English, this study is quite naturally embedded in the field of variationist research. In order to analyse the data, this study uses multivariate statistical methods to conduct the analyses of the different corpora and datasets. This is due to the fact that it is has become increasingly accepted in variationist research that a certain variation phenomenon is most often subject to several factors that influence its shape or – to use the terminology of statistics – that affect the outcome of the dependent variable (e.g. Gries 2003; Lohmann 2011; Schweinberger 2011; Szmrecsanyi 2006). Moreover, the application of multifactorial statistical models also allows to determine the exact magnitude and direction of each significant factor (cf. Szmrecsanyi 2006: 53). Hence, this study is a corpus- and Web-based, quantitative account of morphosyntactic variation in and across varieties of English.

However, observing differences and similarities is one thing, accounting for them another. Thus, in order to explain and interpret the results obtained from the multifactorial analyses, this study mainly draws on concepts and insights of traditional Functional Grammar (e.g. Dik 1989; 1997; Siewierska 1991), Functional Typology (e.g. Bisang 2004; Kortmann 2004b; Siemund 2008) and also functional morphological theory (Hauspelmth 2006; Mayerthaler 1981). In principle, however, the analyses, results and assumptions of this study are compatible with a wide range of current functional approaches to the study of language.
As far as the major results of this thesis are concerned, this study demonstrates that – contrary to widespread assumptions – functional factors exert a tremendous influence on the distribution of pronoun case forms in Present-Day English. More precisely, this study shows that subject pronoun forms have been re-functionalised as Focus markers in subject predicative complements in British, American, Australian and Irish English. Although this study has found leads that suggest such a re-functionalisation process is also at work in South African and Indian English, the reanalysis of subject pronoun forms as postverbal Focus markers cannot be confirmed yet for these varieties by means of significant statistical results.

In more detail, this study shows that subject forms are much more likely to be used in *it*-clefts than in *it BE* sentences in all of the datasets examined in this study. Furthermore, subject forms are also much more likely to be used in particularly focussed than in prototypical *it BE* sentences – at least in the British, American, Australian and Irish datasets. Hence, the statistical results of this study confirm for these varieties the central assumption of this study which claims that the more focussed a subject predicative complement context is, the more likely it is to observe a subject pronoun form in a subject predicative complement context.

Moreover, the analyses and discussions of the different datasets examined for this study show that the distribution of pronoun case forms in subject predicative complements differs from variety to variety and is subject to several factors simultaneously. Thus, monocausal explanations do not suffice to account for complex variation phenomena – not even in narrowly delimited contexts, such as subject predicative complements.

In fact, the analysis of the data shows that the different variables can be systematised according to the direction of impact which they exert on the dependent variable, i.e. the likelihood of observing a subject form in subject predicative complement position. On the one hand, the analyses of the data identify a group of factors that consistently and strongly promote the use of subject forms in subject predicative complements in those datasets where they are both significant and applicable. These factors are the pronoun’s occurrence in an *it*-cleft rather than in an *it BE* sentence, the pronoun’s occurrence in a particularly focussed *it BE* sentence rather than in a normal *it BE* sentence and the co-reference of a focal pronoun in an *it*-cleft with the subject rather than with an object of the following dependent clause. These findings comply very well with the Focus-
oriented approach adopted in this study, since each of these factors is not only of functional, but also of discourse-pragmatic relevance.

On the other hand, the analyses of the different datasets identify factors that, if significant and applicable, consistently inhibit the use of subject pronoun case forms in subject predicative complements. These factors are first person pronouns and computer-mediated communication. This means that first person pronouns are less likely to occur as subject forms in subject predicative complements than third person pronouns. While it is difficult to account for this difference in structural terms, the higher likelihood of third person pronouns to occur as subject forms can be accounted for with the Focus-oriented approach adopted in this study. Indeed, these more or less systematic differences between first and third person pronouns can be explained by the fundamental differences between these two persons with regard to their prototypical referential characteristics. In a nutshell, it can be argued that the referents of first person pronouns are per se more salient than the referents of third person pronouns, which means that the latter ones may need more additional highlighting by means of a subject pronoun form – where this is admissible – to be recognisable as the most important piece of information, i.e. the Focus, of a clause (e.g. Croft 2003: 130; Langacker 1991: 307; Siewierska 2004: 5-8).

Moreover, subject pronoun forms are also less likely to be used in computer-mediated communication than in traditional written data. While this difference can be explained to some extent with differences in formality, this inhibiting effect can also be partly accounted for in terms of different focussing strategies. Since computer-mediated communication is subject to fewer formal conventions than traditional written data, it can employ more focussing strategies, such as the imitation of prosody by means of capitalisation. This, too, can explain why computer-mediated communication is generally less likely to use subject pronoun case forms in subject predicative complements.

Interestingly, the inhibiting effect of the factors first person and computer-mediated communication is for the most part restricted to the it-cleft subsets of the respective corpora and Web-derived datasets. While this co-variance pattern is difficult to account for in structural terms or in terms of formality, it can be explained in terms of Focus-marking. This study shows that several Focus-marking devices may not only co-occur but also may cancel each other out so that there may be a “trade-off” between the different focussing devices (cf. Dik 1989: 278). As a consequence, the combination of two pragmatically salient
factors, i.e. first persons pronoun occurring in the focal position of an *it*-cleft, may be considered to be highlighted enough to identify the referent or referents of the first person pronoun as the most salient or important piece of information in a clause. Thus, two pragmatically salient features or Focus marking strategies in a clause or sentence may suffice to mark a certain constituent as being the most important one and additional Focus marking by means of a subject pronoun case form may not be deemed necessary. This “trade-off” effect can explain the lower likelihood of observing first person subject pronouns in *it*-clefts, as discussed above. Moreover, this could also explain why subject forms are less likely to be used in *it*-clefts in computer-mediated communication, since this mode of discourse not only allows for highlighting a pronoun by means of the focal position of an *it*-cleft or by using a subject form, but also for capitalising a pronoun in order to imitate prosodic prominence, which is of course not available for focal pronouns in the traditional written mode of discourse (cf. Herring 2010a).

Finally, this study identifies factors that are prone not only to inter-, but even to intra-varietal variation as far as their impact on the distribution of pronoun case forms is concerned. Cases in point are the differences between singular and plural pronouns and the difference between the spoken and the traditional written mode of discourse. Indeed, the effect of these factors on the distribution of pronoun case forms seems to depend both on the respective corpus or dataset and on the construction, i.e. *it BE* sentence vs. *it*-cleft.

However, particularly the very strong impact of functional and pragmatic factors, such as the differences between *it*-clefts and *it BE* sentences or between particularly focussed and normal *it BE* sentences in the use of pronoun forms, has severe repercussions on existing theoretical models accounting for the distribution of pronoun case forms. To begin with, the assumptions and predictions of accounts proposing a split in the pronominal paradigm of Present-Day English into clitic and non-clitic pronouns as in French are virtually irreconcilable with the findings of this study (e.g. Harris 1981; Sweet 1875; Quinn 2005a). Indeed, in view of the obtained results, this approach can hardly be maintained. Moreover, the strong impact of functional and pragmatic factors uncovered in the present study also challenges the assumption that position alone determines the distribution of pronoun case forms (e.g. Emonds 1985; Quirk et al. 1985; Burridge 2004). Although position definitely matters, it is not sufficient to fully account for the distribution of pronoun case forms in Present-Day English.
Instead, the results of the statistical analyses clearly corroborate and for five of the seven analysed datasets even statistically validate the Focus-oriented perspective adopted in this study, arguing for a reanalysis of subject pronouns as Focus markers in subject predicative complements. The significantly higher likelihood of *it*-clefts to use subject pronouns attested in all varieties and the significantly higher likelihood of observing subject pronouns forms in particularly focussed *it BE* sentences attested for five datasets allows the conclusion that subject pronouns have been re-functionalised as Focus markers in British, American, Australian and Irish English. Although such a reanalysis contrasts with the assumptions of much of the current linguistic literature discussing the distribution of pronoun case forms in subject predicative complements (e.g. Harris 1981; Quinn 2005a; Quirk et al. 1985), this finding corresponds very well to the results of the cross-varietal survey which attests a very strong influence of pragmatic factors on the distribution of pronoun case forms in many English varieties (e.g. Huber 1999; Paddock 1994). Furthermore, the Focus-oriented approach proposed in this study can explain and also correctly predict the systematic differences between *it*-clefts and *it BE* sentences in their use of pronoun case forms and can also explain why certain subtypes of *it BE* sentences, i.e. particularly focussed ones, are much more likely to exhibit subject forms than more prototypical *it BE* sentences. In fact, the Focus-oriented approach fares much better at predicting the distribution of pronoun case forms in subject predicative complements than approaches that account for the distribution of pronoun case forms by means of the pronoun’s position relative to the finite verb or by distinguishing between clitic and non-clitic pronoun classes (e.g. Burridge 2004; Emonds 1985; Harris 1981; Quinn 2005a). Even more importantly, however, the Focus-oriented approach adopted in this study also provides a functional explanation for the considerable extent of variation in the use of pronoun case forms that can still be observed in subject predicative complements in Present-Day English. In view of the obtained results both from the cross-varietal surveys and the statistical analyses, this explanation is more plausible than previous approaches attempting to account for the distribution and variation of pronoun case forms in *it*-clefts and *it BE* sentences in terms of position and formality or in terms of a purportedly telic development from a once case-governed paradigm towards a system of clitic and non-clitic pronoun classes. Last but not least, the Focus-oriented approach can also accommodate the
cross-varietal differences observed in the course of this study, which are often not addressed or not even suspected in earlier accounts (e.g. Burridge 2004; Quinn 2005a).

With regard to the general theoretical orientation of this study, the combination of Functional Grammar, Functional Typology and functional morphological theory proves to be extremely beneficial not only for the explanation of the results, but also for the detection of important variables. The systematic investigation of the distribution of pronoun case forms across varieties of English from the perspective of Functional Typology helps to identify pragmatic factors and particularly Focus as important factor for the distribution of pronoun case forms in English varieties. Functional Grammar, in turn, not only enables this study to arrive at a definition of Focus but also to implement an operationalisation of this rather elusive variable. Furthermore, combining the insights gained from Functional Typology and Functional Grammar allows this study to account for important co-variance patterns such as the lesser likelihood of first person pronouns to occur as subject forms in it-clefts. Finally, functional morphological theory has equipped this thesis and its Focus-oriented approach with an explanation why particularly subject pronoun forms have been reanalysed as Focus markers in subject predicative complements and not other available pronominal options. Indeed, the responsible mechanism, i.e. “markedness reversal” (Mayerthaler 1981: 48) or deviation from “default expectation” (Haspelmath 2006: 62), may not only be relevant for subject predicative complements but also for many other contexts allowing for variability in the use of pronoun case forms, such as left dislocations or coordinated noun phrases and can even reconcile accounts on the distribution of pronoun case forms that seem at first glance mutually exclusive (Biber et al. 1999: 335; Wagner 2004: 159).

As far as the methodological approach of this study is concerned, this study clearly demonstrates that thorough qualitative and fine-grained quantitative analyses have to necessarily complement each other in order to achieve the best possible outcome. Only the comprehensive, qualitative, cross-varietal survey of the use of pronoun case forms shows that the distribution of pronoun forms is influenced by pragmatic factors in many varieties of English, which has so far been largely neglected in the discussion of contexts allowing for variability in the use of pronoun case forms. However, in order to test and evaluate whether Focus indeed significantly influences the distribution of pronoun case forms in the subject predicative complements of the examined varieties, the application of multivariate statistical modelling is absolutely indispensable. Thus, this study demonstrates that qualitative and
quantitative analyses are not alternative approaches, but actually have to work in tandem to obtain the best possible results allowing for robust generalisations.

With regard to the mode of data retrieval, the results of this study show that data obtained from commercial Web crawlers, if carefully refined, can indeed yield robust results and thus can form a suitable basis for the examination of a wide range of linguistic studies. Bearing in mind that large corpora are available only for very few varieties of English, this is indeed a very promising finding.

Thus, I hope to have contributed through this thesis to a better understanding of the distribution of pronoun case forms in English in general and in subject predicative complements in particular.

Bereits in den 1930er Jahren hielt Jespersen (1933) fest, dass die "richtige Verwendung" pronominaler Kasusformen einen der schwierigsten Aspekte der englischen Grammatik darstellt (Jespersen 1933: 132). Der Grund, weshalb die Wahl pronominaler Kasusformen im Englischen als schwierig empfunden wird, ist die Tatsache, dass die Verteilung von Subjekt- und Objektpronomen im Englischen nicht so unkompliziert ist, wie gemeinhin angenommen wird (Biber et al. 1999: 335). Obwohl Subjektpronomen tatsächlich vor allem als Subjekte finiter Sätze und Objektpronomen hauptsächlich als Objekte transitiver Verben und als präpositionale Komplemente verwendet werden (vgl. Biber et al. 1999: 335), weist das heutige Englisch eine beträchtliche Anzahl von Kontexten auf, die eine Wahl zwischen der Verwendung von Subjekt- (I, she, he, we, they) und Objektpronomen (me, her, him, us, them) zulassen. Als bekannte Beispiele für solche Kontexte, die Variabilität in der Verwendung pronominaler Kasusformen erlauben, sind unter anderem die Fokusposition von Spaltsätzen (it-clefts), Subjektkomplemente nach it und einer Form der Kopula BE (it BE sentences vgl. Quinn 2005a), koordinierte Nominalphrasen in Subjekt- und Objektposition (coordinated NPs), durch Pronomen modifizierte Nominalphrasen (pronoun–noun phrase constructions) und Komparative (than-comparatives) zu nennen, wie auch in Tabelle 1 dargestellt wird:

<table>
<thead>
<tr>
<th>CONTEXTS EXHIBITING VARIABILITY IN THE USE OF PRONOUN CASE FORMS</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>If</em>-clefts</td>
<td>It is they/them who are to blame</td>
</tr>
<tr>
<td><em>It BE</em> sentences</td>
<td>[Who is it? –] It is I/me</td>
</tr>
<tr>
<td>Coordinated NPs in subject position</td>
<td>Rita and I/me will get divorced</td>
</tr>
<tr>
<td>Coordinated NPs in object position</td>
<td>This issue has to stay between you and I/me</td>
</tr>
<tr>
<td>Pronoun-Noun Phrase Constructions</td>
<td>We/Us Irish are a happy people</td>
</tr>
<tr>
<td>Than-Comparatives</td>
<td>Susan is younger than he/him</td>
</tr>
</tbody>
</table>

Table 1: Variable Kontexte Pronominaler Kasuswahl im Heutigen Englisch

Tatsächlich gibt es bislang kaum quantitative Studien, geschweige denn multifaktorielle, die versuchen die Verteilung pronominaler Kasusformen in *it*-clefs und *it BE* sentences statistisch zu untersuchen. Dabei gibt es mehrere gute Gründe, warum gerade eine quantitative Analyse von *it*-clefs und *it BE* sentences sehr wertvolle neue Einblicke nicht nur in die pronominale Kasusverteilung in Subjektkomplementen und anderen Kontexten mit variabler pronominaler Kasusverwendung ermöglichen kann, sondern auch neue Erkenntnisse über die Verteilung pronominaler Kasusformen im Englischen im Allgemeinen liefern kann.

Darüber hinaus ist die pronominale Kasusverteilung in *it*-clefts und *it BE* sentences auch deshalb äußerst interessant, weil eine Vielzahl aktueller linguistischer Studien und Ansätze, in denen die Verteilung pronominaler Kasusformen in variablen Kontexten erörtert wird, die Meinung vertritt, dass die Verteilung pronominaler Kasusformen in *it*-clefts und *it BE* sentences gleich oder zumindest sehr ähnlich sein soll (z.B. Burridge 2004; Emonds 1985; Harris 1981; Quinn 2005a). So ist die hier vorliegende Studie in der Lage, durch die Analyse dieser beiden Konstruktionen die Validität und Vorhersagekraft einiger äußerst etablierter theoretischer Erklärungsmodelle zur Verteilung pronominaler Kasusformen im heutigen Englisch zu testen und zu evaluieren.


---

Kasusformen zu pragmatischen Fokusmarkern erklären kann, sondern auch das beträchtliche Ausmaß an Variation erklären kann, das in der pronominalen Kasuswahl in *it*-clefs und *it BE* sentences im heutigen Englischen beobachtet werden kann (vgl. Erdmann 1978; Maier 2013).

Ganz allgemein verfolgt diese Studie das Ziel, die Verteilung pronominaler Kasusformen in Subjektkomplementposition nach *it* und einer Form von *BE* in sechs Varietäten des Englischen zu analysieren, um dadurch ein vertieftes Verständnis und weitere Erkenntnisse über die Faktoren zu erlangen, die die Verteilung pronominaler Kasusformen nicht nur in *it*-clefs und *it BE* sentences, sondern in der englischen Sprache im Allgemeinen bestimmen.

wurden, verlässliche Resultate liefern, die robuste Schlussfolgerungen und Generalisierungen erlauben.


Die wichtigsten Ergebnisse dieser Studie lassen sich wie folgt zusammenfassen. Diese Studie zeigt entgegen weit verbreiteter Annahmen, dass funktionale Faktoren einen sehr starken Einfluss auf die Verteilung der pronominalen Kasusformen im heutigen Englisch ausüben. Mit Hilfe signifikanter statistischer Ergebnisse kann gezeigt werden, dass Subjektpronomen in Subjektkomplementposition nach it und einer Form von BE im britischen, amerikanischen, australischen und irischen Englisch als Fokusmarker
refunktionalisiert wurden. Obwohl diese Studie ferner Indizien ermittelt hat, die solch einen Refunktionalisierungsprozess auch für südafrikanisches und indisches Englisch nahelegen, konnte die Reanalyse von Subjektpronomen zu postverbalen Fokusmarkern für diese Varietäten vorläufig nicht mittels statistisch signifikanter Ergebnisse abgesichert werden.


Darüber hinaus ergibt die Analyse und Diskussion der statistischen Ergebnisse der verschiedenen Datensätze, dass die Verteilung pronominaler Kasusformen in Subjektkomplementposition sehr stark variiert zwischen den untersuchten Varietäten und dass die Verteilung der verschiedenen Kasusformen meist von mehreren Faktoren gleichzeitig determiniert wird. Daher kann festgehalten werden, dass monokausale Modelle nicht ausreichen, um komplexe Variationsphänomene hinlänglich zu erklären – nicht einmal in eng umrissten Kontexten wie Subjektkomplementen nach it und einer Form von BE.

Ferner zeigt die Analyse der Daten, dass die verschiedenen unabhängigen Variablen bezüglich der Richtung des Einflusses, den sie auf die abhängige Variable (das Auftreten von Subjektformen in postverbalen Subjektkomplementkontexten) ausüben, systematisiert werden können.

Zum einen identifiziert die statistische Analyse der verschiedenen Korpora und Datensets eine Reihe von Faktoren, die konsequent und mit Nachdruck das Auftreten von Subjektformen in Subjektkomplementkontexten begünstigen, sofern sie für das jeweilige Modell anwendbar sind bzw. signifikant sind. Zunächst ist hier die Verwendung eines Pronomens in einem it-cleft anstatt in einem it BE sentence zu nennen, das die Wahrscheinlichkeit erheblich steigert, dass eine Subjektform in Komplementposition verwendet wird. Als zweiter sehr stark begünstigender Faktors ist der Grad der Fokussierung in it BE sentences anzuführen, da Subjektpronomen häufiger in speziell fokussierten als in

424


Ferner ergibt die Analyse der Daten, dass Subjektpronomens auch weniger häufig in webbasierten Kommunikation als in traditioneller geschriebener Sprache verwendet werden.

Fokussierungsstrategien kann erklären, warum es besonders in *it*-clefts weniger wahrscheinlich ist, Subjektformen der ersten als solche der dritten Person vorzufinden.

Schließlich werden durch die Datenanalyse dieser Studie auch Faktoren identifiziert, deren Effekt auf die Verteilung pronominaler Kasusformen in den untersuchten Subjektkomplementkontexten nicht nur über Varietäten hinweg, sondern sogar innerhalb der einzelnen Datensets variiert. Beispiele hierfür sind die Unterschiede zwischen Singular- und Pluralpronominen und der Unterschied zwischen gesprochener und traditioneller geschriebener Sprache. Die Effekte, die diese Faktoren auf die Verteilung pronominaler Kasusformen ausüben, sind den Ergebnissen der Datenanalyse zufolge sowohl vom jeweiligen Korpus bzw. Datensatz als auch von dem jeweiligen Konstruktionstyp, d.h. *it*-cleft oder *it BE* sentence, abhängig.


Im Gegensatz dazu wird der in dieser Arbeit vorgeschlagene Fokus-orientierte Erklärungsansatz, der eine Reanalyse der Subjektpronomien zu Fokusmarkern in

das Prinzip der Markiertheitsumkehrung sogar Studien und deren Beobachtungen in Einklang bringen, die sich ansonsten gegenseitig auszuschließen scheinen (Biber et al. 1999: 335; Wagner 2004: 159).

Was die methodische Vorgehensweise dieser Dissertation anbelangt, so zeigt diese Studie, dass gründliche qualitative und eingehende quantitative Analysen sich notwendigerweise gegenseitig ergänzen müssen, um bestmögliche Ergebnisse zu erzielen. Tatsächlich hat nur der umfassende, qualitative Varietätenüberblick gezeigt, dass die Verwendung und Verteilung pronominaler Kasusformen in vielen Varietäten des Englischen maßgeblich von pragmatischen Faktoren beeinflusst wird. Dieser Sachverhalt war bislang weitgehend unbeachtet in der Literatur, die sich mit Kontexten beschäftigt, in denen pronominale Kasusvariation beobachtet wird. Um jedoch zu exakt zu bestimmen, inwiefern Fokus tatsächlich die Verteilung pronominaler Kasusformen in *it*-clefts und *it BE* sentences beeinflusst, war die Anwendung multifaktorieller statistischer Analyseverfahren unerlässlich. Somit zeigt diese Arbeit, dass möglichst genaue qualitative und quantitative Analysen nicht alternative Ansätze darstellen, sondern sich ergänzen müssen, um bestmögliche Ergebnisse zu erzielen, die dann zu robusten Generalisierungen führen können.

Im Hinblick auf die in dieser Studie durchgeführte Datenakquise zeigen die Ergebnisse dieser Arbeit, dass Daten, die mit Hilfe kommerzieller Internetsuchmaschinen erhoben wurden und die entsprechend aufbereitet wurden, durchaus robuste Ergebnisse liefern können und somit eine adäquate Grundlage für eine Vielzahl sprachwissenschaftlicher Studien bilden können. Angesichts der Tatsache, dass große Sprachkorpora nur für sehr wenige Varietäten des Englischen bislang vorliegen, ist dies ein äußerst vielversprechendes Ergebnis.
Eidesstattliche Erklärung

Hiermit versichere ich an Eides Statt, dass ich die Arbeit selbständig angefertigt, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die aus den herangezogenen Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Hamburg, den 8. November 2012, ______________________

Georg Maier