Improving Recombination in a Linear EBMT System by Use of Constraints

Dissertation

zur Erlangung des akademischen Grades
Dr. rer. nat
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
der Universität Hamburg

vorgelegt von

Monica Roxana Gavrilă

aus Bukarest (Rumänien)

Hamburg, Juli 2011

Genehmigt von der Fakultät für Mathematik, Informatik und Naturwissenschaften, Fachbereich Informatik der Universität Hamburg auf Antrag von

Erstgutachter:	Prof. Dr. Walther von Hahn (Betreuer) Fachbereich Informatik Universität Hamburg
Zweitgutachter	Prof. Dr.-Ing. Wolfgang Menzel Fachbereich Informatik Universität Hamburg
Externer Gutachter	Dr. David Farwell Llenguatges i Sistemes Informàtics Universitat Politècnica de Catalunya
Vorsitzender	Prof. Dr.-Ing. Dietmar P.F. Möller Fachbereich Informatik Universität Hamburg

Hamburg, den 01. Februar 2012 (Tag der Disputation)

Dedicated to

my loving parents
and to J.

Abstract

(Automatic) machine translation (MT) is one of the most challenging domains in Natural Language Processing (NLP) and plays an important role in ensuring global communication, especially in a multilingual world with access to large amounts of Internet resources. As rule-based MT approaches need manually developed resources, new MT directions have been developed over the last twenty years, such as corpus-based machine translation (CBMT): statistical MT (SMT) and example-based machine translation (EBMT). These new directions are based mainly on the existence of a parallel aligned corpus and, therefore, can be easily employed for lower-resourced languages.

In this dissertation we showed how EBMT systems behave when a lower-resourced inflecting language (i.e. Romanian) is involved in the translation process. For this purpose we built an EBMT baseline system based only on surface forms (the *Lin-EBMT* system). One of our main goal was to investigate the impact of word-order constraints on the translation results: we integrated constraints extracted from generalized examples (i.e. templates) in *Lin-EBMT* and built an extended system: $Lin - EBMT^{REC+}$. Although constraints represent a well-known method which is employed quite often in NLP, the use of word-order constraints in an EBMT system is an innovative approach which can open new paths in the domain of example-based MT. We run our experiments for two language-pairs in both directions of translation: Romanian-German and Romanian-English. This aspect raises interesting questions, as Romanian and German present language specific characteristics, which make the translation process even more challenging. Both EBMT systems developed are easily adaptable for other language-pairs. They are platform and language-pair independent, provided that a parallel aligned corpus for the language-pair exists and that the tools used for obtaining the needed intermediate information (e.g. word-alignment) are available. As a side question, we studied how EBMT reacts in comparison to SMT. We compared the EBMT results obtained to results provided by a Moses[1]-based SMT system and the Google Translate on-line system.

To provide a complete view on CBMT, the performance of each MT system was assessed in several experimental settings, using different corpora (type and size), various system settings and additional part-of-speech (POS) information. We evaluated the translation results by means of three automatic evaluation metrics: BLEU, NIST and TER. A subset of the results was manually analyzed for a better overview on the translation quality.

Our experiments showed that constraints improve translation results, although a clear

[1] `www.statmt.org/moses` - last accessed on June 27th, 2011.

decision which constraint-combination works best could not be taken. Although the SMT system outperformed the EBMT system in all experiments, the manual analysis provided cases in which EBMT offered more accurate results. The behavior of the systems while changing the experimental settings confirmed that (training and test) data have a substantial impact on both MT approaches. The difference between the results of the two MT approaches decreased when a more restricted corpus was used. As expected, both CBMT approaches worked better for shorter sentences.

Zusammenfassung

Die automatische maschinelle Übersetzung (MÜ) ist einer der kompliziertesten Bereiche in der Sprachverarbeitung. Die MÜ spielt eine wichtige Rolle bei der Gewährleistung der globalen Kommunikation in der mehrsprachigen Welt, die vor allem von Internetressourcen gestützt wird. Da regelbasierte MÜ-Ansätze manuell entwickelte Ressourcen benötigen, wurden neue MÜ-Richtungen entwickelt, wie zum Beispiel die korpusbasierte maschinelle Übersetzung (KMÜ): die statistische MÜ (SMÜ) und die beispielbasierte maschinelle Übersetzung (BMÜ). Der Vorteil dieser neuen MÜ-Richtungen ist, dass sie auch für Sprachen eingesetzt werden können, für die weniger Ressourcen zur Verfügung stehen.

In dieser Dissertation zeigen wir wie BMÜ-Systeme reagieren, wenn eine flektierende Sprache mit weniger Ressourcen (d.h. Rumänisch) in die Übersetzung einbezogen wird. Zu diesem Zweck erstellen wir ein BMÜ-Grundsystem, das nur auf der Oberflächenform der Wörter basiert (das *Lin-EBMT* System). Darüber hinaus untersuchen wir den Einfluss der Wortstellungsbeschränkungen (Constraints) auf die Übersetzungsergebnisse. Wir extrahieren diese Constraints aus allgemeinen Beispielen (d.h. Templates) und integrieren sie in *Lin-EBMT*: das $Lin - EBMT^{REC+}$ System. Obwohl die Verwendung von Constraints eine bekannte Methode in der Sprachverarbeitung ist, ist die Verwendung der Wortstellungsconstraints in einem BMÜ-System ein innovatives Konzept, das neue Wege in dem BMÜ-Bereich öffnen könnte. Wir führen unsere Experimente für zwei Sprachpaare in beide Richtungen der Übersetzung durch: Rumänisch-Deutsch und Rumänisch-Englisch. Dieser Aspekt beinhaltet interessante Fragen, weil Rumänisch und Deutsch spezifische Spracheigenschaften haben, die den Übersetzungsprozess noch komplizierter machen können. Die beiden entwickelten BMÜ-Systeme lassen sich sehr einfach an andere Sprachpaare anpassen. Die Systeme sind plattform- und sprachpaarunabhängig, vorausgesetzt ein Textkorpus von zweisprachigen Texten existiert und die Werkzeuge für die Beschaffung der erforderlichen Informationen (zB Wort-Alignment) vorhanden sind. Als Nebenfrage untersuchen wir, wie BMÜ im Vergleich zu SMÜ reagiert. Daher vergleichen wir die BMÜ-Ergebnisse mit denen eines Moses[2]-basierten SMÜ-Systems und denen des Google Translate Online-Systems.

Die Leistung jedes MÜ-Systems wird in mehreren experimentellen Einstellungen untersucht. Wir verwenden verschiedene Korpora (sowohl Typ, als auch Größe), verschiedene Systemeinstellungen sowie zusätzliche Wortartinformationen. Wir evaluieren die Übersetzungsergebnisse automatisch mit BLEU, NIST und TER. Ein Teil der Ergebnisse wird

[2] `www.statmt.org/moses`.

manuell analysiert, um einen besseren Überblick über die Qualität der Übersetzung zu erhalten.

Unsere Experimente zeigen, dass Constraints die Übersetzungsergebnisse verbessern können, obwohl eine klare Entscheidung darüber, welche der Constraint-Kombinationen am besten funktioniert, nicht getroffen werden kann. Obwohl das SMÜ-System in allen Versuchen besser als das BMÜ-System ist, entdecken wir in der manuellen Analyse Fälle, in denen BMÜ-Systeme besser als das SMÜ-System funktionieren. Das Verhalten der Systeme bestätigt während des Wechsels der experimentellen Einstellungen, dass (Training- und Test-) Daten einen hohen Einfluss auf beide MÜ-Ansätze haben. Der Unterschied zwischen den Ergebnissen der beiden MÜ-Ansätze verringert sich, wenn ein eingeschränktes Korpus verwendet wird. Wie schon erwartet, sind beide KMÜ-Ansätze besser, wenn kürzere Sätze übersetzt werden.

Acknowledgments

I would like to express my gratitude to a number of persons who supported me in various ways while preparing and writing my dissertation.

Above all, I would like to thank my supervisor, Professor Dr. **Walther von Hahn**, for his support, advice and encouragement over the past years: "*Thank you for all your help, patience and guidance throughout the whole time.*"

I am especially grateful to Professor Dr. **Wolfgang Menzel** and Professor Dr. **David Farwell**, for reading and commenting on this thesis.

Various people in the Department of Informatics of the University of Hamburg deserve a special acknowledgment. In the first place I would like to mention Dr. **Cristina Vertan**: "*Thank you for your support, constructive comments, friendship and time.*" Many thanks to all my colleagues with whom I have worked or socialized – especially to **Dr. Kerstin Fischer**, **Karin Jarck**, **Hildegard Westermann**, **Reinhard Zierke** –: "*Thank you for your support and advice over the past years.*". Special thanks to **Natalia Eliţa** for sharing the work for the compilation of the RoGER corpus, listening to me and being a wonderful friend.

It is a pleasure to express my highly appreciation to all my **PIASTA** colleagues and friends, especially to **Alexandra**, **Ulrike**, **Kristina**, **Lea**, **Canan** and the whole **team for PhD students**, who motivated me and endured all my complaints during this time. Many thanks also to **Gabrielle Warnke** and **Bärbel Launer**.

I am grateful to Ms. **Frauke Narjes** for listening and helping me take the right decision in the right moment.

I owe my deepest gratitude to my **parents**, especially my **mother**, whose infinite patience and encouragement has contributed immeasurably to this success:"*Thank you for supporting and believing in me throughout my life*". Warm thanks go also to my younger brother, **Adrian**, for bearing my stay abroad.

I would like to express my love and deepest gratitude to my boyfriend, **Jordi**, who supported me unconditionally during the whole time: "*Thank you for your love, care, patience and advice. Without your understanding and support would have been impossible to finish this thesis.*"

My dearest friends must be thanked for their immense loyalty, love and support – **Ioana**, **Natalia**, **Martha**, **Daniela** and **Raluca**. "*Thank you for being such steadfast sources of encouragement during the writing of this thesis.*"

Many thanks to all the people I have not mentioned, but were next to me throughout the whole process: "*Without you I would not have managed.*"

——* * *——

Thank you ***God*** *for giving me the wisdom, strength and determination to complete this work!*

Monica Roxana Gavrila

Hamburg, July 2011

Contents

List of Figures

LIST OF FIGURES

List of Tables

List of Abbreviations

AC: Acquis Communautaire

ACC., acc.: Accusative

Adj / adj: Adjective

AI: Artificial intelligence

ALPAC: Automatic Language Processing Advisory Committee

ARPA: Advanced Research Projects Agency

BLEU: **Bi**lingual **E**valuation **U**nderstudy

C: Constraint in $Lin - EBMT^{REC+}$

CL: Computational linguistics

CBMT: Corpus-based machine translation

CBR: Case-based reasoning

CPU: Central processing unit

CT: Common tokens

DAT., dat.: Dative

de: German (in the corpus)

DEU: German ([Lewis, 2009])

doc: Documents

DP: Dynamic programming

EBMT: Example-based machine translation

en: English (in the corpus)

ENG: English ([Lewis, 2009])

EU: European Union

FAMT: Fully automatic machine translation

GEN., gen.: Genitive

GPL: GNU Public License

HAMT: Human-aided machine translation

HMM: Hidden Markov Models

HPA: Hierarchical Phrase Alignment

HTML: Hypertext Markup Language

KBMT: Knowledge-based machine translation

LCS: Longest Common Subsequence

LCSS: Longest Common Subsequence Similarity

LD: Levenshtein Distance

LDC: Linguistic Data Consortium

LGPL: GNU Lesser General Public License

LM: Language model

MAHT: Machine-aided human translation

MSD: Morpho-Syntactic Descriptor

MT: Machine translation

no / No.: Number

NOM., nom.: Nominative

NER: Named Entity Recognizer

NIST: **N**ational **I**nstitute for **S**tandards and **T**echnology

NLP: Natural Language Processing

NN: Noun

NP: Noun phrase

OOV-words: Out-of-vocabulary words

PL / pl: Plural

POS: Part of speech

PP: Prepositional phrase

prep: Preposition

List of Abbreviations

RBMT: Rule-based machine translation

ro: Romanian (in the corpus)

RON: Romanian ([Lewis, 2009])

RTF: Rich Text Format

ru: Russian (in the corpus)

SER: Sentence Error Rate

SG / sg: Singular

SL: Source language

SMT: Statistical machine translation

SP: Source pattern

SRILM: The SRI Language Modeling Toolkit

TER: **T**ranslation **E**rror **R**ate

TF / tf: Text fragment in a translation template

TL: Target language

TM: Translation model

TP: Target pattern

TSC: Tree-string correspondence

VAR: Variable in a translation template

VOC. / voc.: Vocative

WER: Word Error Rate

WSD: Word Sense Disambiguation

WWW: World Wide Web

XML: Extensible Markup Language

Chapter 1

Introduction

Machine translation (**MT**), one of the most challenging domains in Natural Language Processing (**NLP**), plays an important role in ensuring global communication. This happens in a multilingual world, where people have access to a large amount of digital data in a multitude of languages, especially over the Internet. Documents in various domains need to be translated in a large number of language-pairs[1]. Furthermore, it is often very hard to find human translators having both domain and bilingual knowledge. In these cases MT offers, at least theoretically, the frame to overcome this gap. For several languages (e.g English) the implementation of MT-systems based on rules or corpora has a long tradition. For these languages, research can now concentrate on improving the translation results through combination of linguistic and statistical methods (see [Uszkoreit, 2009]).

Unfortunately, MT systems with a target or source language different than e.g. English are still less widespread and researched. Lesser researched languages have to overcome a major gap in language resources and tools, training data and reference systems for evaluation, which all ensure the development of a good MT-system. Some of these under-resourced languages are highly inflected, with a more complex grammar and are often presenting linguistic phenomena which have not been encountered in previous language combinations. As they are not present in the languages researched, these (complex) linguistic phenomena are often forgotten or not regarded as translation challenges. On the other side, exactly for these languages, human translators are few or missing, so MT-systems are in high demand. The problem is not only Europe-specific, as, with the spread of information technology, other multilingual communities[2] face similar problems.

[1]For example, in the context of multilingual Europe, the publication of a critical amount of EU-documents is needed in all 23 official languages. This adds to 253 language pairs in both directions of translation (506 combinations).

[2]A list of multilingual countries and regions can be found on `http://en.wikipedia.org/wiki/List_of_multilingual_countries_and_regions` - last accessed on June 27th, 2011.

1.1 Motivation and Statement of the Problem

Based mainly on the existence of parallel corpora, corpus-based machine translation (**CBMT**), together with its two main approaches – statistical MT (**SMT**) and example-based MT (**EBMT**) –, seem to be a solution for under-resourced[3] languages. SMT, based on statistical data extracted from the corpus, offers a flexible framework to develop automatic translation systems in a relatively short time. These systems deliver acceptable results at least for in-domain test data. But for this MT approach tests have been usually performed for languages for which linguistic tools and corpora have been developed to a certain level – see [Callison-Burch et al., 2010]. The other CBMT approach, EBMT, is essentially translation by analogy. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again. Most of the EBMT systems presented in the literature have usually involved a smaller-size corpora in contrast to SMT systems.

The research in this dissertation investigates two language pairs: Romanian (**RON**)-English (**ENG**) and Romanian-German (**DEU**), in both directions of translation[4]. We consider Romanian the under-resourced language in this thesis. While the interest for translation from or into German or English appeared in an early stage of MT, the demand for translation from and into Romanian became more pressing after Romania joined the European Union in 2007[5]. Moreover, for Romanian, not enough linguistic resources were publicly available at the beginning of this research or, when available, comparing with other languages (e.g. English), they were under-developed or not sufficiently tested. There was also no real possibility of choosing among several resources, as only one resource has usually been available. The use of these language pairs raises interesting questions, as most of the example-based translation systems have English as target language (**TL**), which has a 'simpler' syntax and morphology. Romanian and German present language specific (morphological and syntactic) characteristics, which make the process of translation even more challenging (see **Chapter 4** and **Appendix B**).

The aim of the dissertation is to explore how EBMT can be used when translating into or from an inflected under-resourced language. Since over the last few years, the research community has mainly been concerned with the SMT[6] approach, the EBMT results are compared with SMT ones. As we use an under-resourced language, we keep the systems as resource-free as possible. The algorithms are based mainly on surface forms and on corpus statistics.

Another important goal of this thesis is to investigate how word-order constraints influence the translation results for these language pairs.

[3] Lower-resourced.

[4] As part of the results were manually analyzed, only languages which the author knows are analyzed.

[5] On the `www.mt-archive.org` website, some first MT papers which consider German or English are from the 1950s. The first paper on Romanian appeared in the late 1990s.

[6] See the yearly Workshop on SMT, which has started in 2006.

1.2 Contribution of the Work

In this dissertation we show how an EBMT system based on surface forms (a linear EBMT system) behaves when using a lower-resourced inflected language (i.e. Romanian) as source or target language. We test the behavior of the same EBMT system when both the source and target languages are inflected and one language is lower-resourced: the Romanian - German language-pair.

The other main goal of the thesis is to investigate and identify the influence of word-order constraints on the translation results. The constraints are integrated in the last step of the baseline EBMT system we implemented (i.e. the recombination step). The word-order constraints are extracted using information from the template-based EBMT approach[7]. We also test for the impact that part-of-speech (**POS**) information has on the translation quality for the language pair English-Romanian.

SMT has been the main empirical translation approach used in the research community over the last few years. Therefore, we compare the EBMT results with SMT ones. For a better understanding of the SMT approach we also provide several experimental settings. We consider three parallel aligned corpora, of different sizes: a larger corpus (closer to the SMT specifications) and two smaller corpora (of the same size), which are more suitable for the EBMT environment.

In order to achieve these goals, the following tasks have been performed:

- Creation of a parallel domain-restricted corpus, in four languages: Romanian, German, English and Russian: RoGER [Gavrila and Elita, 2006][8]. The work was motivated by the fact that, at the beginning of this research[9], no parallel domain-restricted corpus was available for all the language pairs analyzed in this thesis. The size of the corpus (i.e. 2333 sentences) better fits into the EBMT framework. After 2007, when Romania joined the European Union (**EU**), more linguistic resources appeared or existing resources were extended for Romanian, e.g. JRC-Acquis or OPUS. We will also use JRC-Acquis for our experiments.

- Development and implementation of a linear EBMT baseline system (*Lin-EBMT*), which uses no other linguistic resources but the parallel aligned corpus. The motivation for developing an EBMT system is given by the fact that, to our knowledge, no open source EBMT system was available until the end of 2009[10]. Although the initial ideas of *Lin-EBMT* also appear in previous works (e.g. word-based similarity measures or recombination based on information extracted from a language model), the steps of the EBMT system as implemented during this research (such as the "*recombination-matrix*" and the Longest Common Subsequence Similarity (**LCSS**) metric for matching – see **Chapter 6**) have not been discussed in other papers. A

[7] **Chapter 3** will provide a description of the linear and template-based EBMT approaches.
[8] The corpus compilation was carried out in collaboration with my colleague Natalia Elița.
[9] The research began in the second half of 2005.
[10] This research considers existing resources until the first half of 2009.

direct comparison with other EBMT systems is difficult to make, as the other system components are usually only briefly described and details are missing.

- Implementation of an extended EBMT system ($Lin - EBMT^{REC+}$), which uses word-order constraints in the recombination step, with information extracted from data derived from the template-based EBMT approach. This way, a second EBMT system, an extension of *Lin-EBMT*, has been developed, which combines the linear approach with the template-based EBMT approach. Although constraints represent a well-known method which is used quite often in NLP, the integration of word-order constraints in a linear EBMT system is an innovative approach, which can open new paths in the domain. Both EBMT systems developed during this research are easily adaptable for other language-pairs. They are platform and language-pair independent, provided that a parallel aligned corpus for the language-pair exists and that the tools used for obtaining the needed intermediate information (e.g. word-alignment) are available.

- Experimental settings: during the experiments presented in this thesis several comparisons of corpus-based MT approaches have been investigated, while changing various parameters, such as the MT system and CBMT approach, the language pair, the corpus (type and size) and the test data type (in-domain vs. out-of-domain test data). The corpus-based MT approaches (SMT and EBMT) have been directly compared using the same training and test data. When no other linguistic information was used, the obtained results were also examined in contrast to those given by the Google on-line MT system, Google Translate[11]. The comparison with Google was done using the same test data. In the experiments two frameworks were analyzed: one with a larger corpus closer to the SMT settings and a second with a smaller corpus, which better fits the EBMT framework. The comparison can be considered one-to-one, as the training and test data are identical. Usually in the literature, EBMT and SMT are directly compared in an SMT framework, with a large parallel aligned corpus. The experiments in this thesis were run in a realistic scenario, with no human interference on the data[12]. For example, when users need to translate a text, they do not check before how the text fits into the MT system, but rather just use the MT system. Therefore, we randomly chose the test sentences, without verifying, for example, if these sentences have been included in the training data. We consider the influence of additional linguistic information, i.e. POS, for some experiments. The SMT system setting is the one recommended for the baseline system at the Sixth Workshop on SMT (2011)[13]. Among the results we have obtained, it could be noticed that:

[11]The translation was obtained using the state-of-the-art of the Google Translate system in the second half of 2008. With regard to the Google training data, clear information about its size and type, to the best of my knowledge, is not publicly available.

[12]For abstraction from certain particularities of the corpus itself: see **Chapter 4**.

[13]In the Workshop two baseline systems were provided: "*baseline system*" and "*baseline system 2*". We used only the first configuration of the baseline system.

  - As expected, the degree of inflection of the languages has a direct influence on the translation results;
  - Word-order constraints in the recombination step affects positively the evaluation results;
  - SMT usually outperforms EBMT, although there are cases when the EBMT translations are more correct than the SMT ones.

1.3 Organization of the Thesis

The introduction of this thesis has presented a general view on the need for machine translation, the translation scenario and the contributions of this study. The rest of this work is organized as follows:

- After a very short description of MT and its paradigms in **Chapter 2**, **Chapter 3** will provide a general view on EBMT, its definition and state-of-the-art.
- **Chapter 4** will give a brief overview on the languages used in this thesis and will introduce the parallel corpora employed in the experiments.
- Before explaining in **Chapter 6** the architecture of $Lin - EBMT$, the baseline EBMT system developed, the (open-source) software used in the experiments will be described in **Chapter 5**.
- The constrained version of $Lin - EBMT$, the $Lin - EBMT^{REC+}$ system, will be presented in **Chapter 7**.
- **Chapters 8**, **9** and **10** will present the experimental settings, the automatic evaluation results and a manual analysis and interpretation of the results.
- **Chapter 11** will conclude the thesis and discuss the central results of this approach, showing also the main limitations. Possible future directions of research will also be outlined.
- Further information will be given in the appendices, as follows:
  - **Appendix A** will present, in a tabular form, previous reported EBMT systems and some of their features.
  - A selective analysis of the languages used, extending the information from **Chapter 4**, will be presented in **Appendix B**.
  - **Appendix C** will shortly describe other corpora for the same language-pairs.
  - **Appendixes D** and **E** will give excerpts from the corpora and from the translation examples, respectively.
  - Additional technical detail about the experiments will be presented in **Appendix F**.

– **Appendix G** will show ranking results for the three MT systems trained and developed during the research – the Moses-based SMT system (**Mb_SMT**), *Lin-EBMT* and $Lin-EBMT^{REC+}$. This appendix is an extension of **Section 10.2.1** (**Chapter 10**).

Chapter 2

Machine Translation (MT)

> "*Machine translation was a matter of serious speculation long before there were computers to apply to it; [...] and it has been a subject of lively, sometimes acrimonious debate ever since. [...] (It) has claimed the attention from some of the keenest minds in linguistics, philosophy, computer science, and mathematics. [...] it has always attracted the lunatic fringe, and continues to do so today.*" [Kay, 1992].

Before describing the data used and the main results obtained during this research, in this chapter (**Chapter 2**) and the next (**Chapter 3**), we will present the state-of-the-art of MT in general and EBMT in particular.

2.1 Definition and Classification

Martin Kay's words in the foreword of "*An Introduction to Machine Translation*" describe the topic of machine translation, its history and complexity perfectly [Kay, 1992].

Machine Translation (**MT**) is defined as the branch of computational linguistics (**CL**) that investigates the use of computers in translating text or speech from one natural language, called the source language (**SL**), into another natural language, the target language (**TL**).

As a highly complex area, MT draws ideas and methods from linguistics, computer science, artificial intelligence (**AI**), mathematics (e.g. algebra, statistics) and translation theory. Dealing with at least two[1] natural languages, MT is a multifaceted subject and one of the most challenging domains in computational linguistics (CL). Its difficulty derives from the complexity of the natural language, which is characterized by *ambiguity* and *expressiveness*[2]: a message can be expressed in different ways and can often render several possible interpretations. Given a certain input, several human translators could produce

[1] The translation process might use three natural languages, when a pivot language is considered.

[2] Not all characteristics of natural language are enumerated, as this is not our scope.

different valid translations. Among the challenges for MT there are ambiguity on analysis and selecting among paraphrastic options on generation.[3]

After more than 60 years and major progress in the development of computers and of Natural Language Processing (**NLP**) applications, no fully-automatic MT system has been developed, which can translate any type of input correctly. While it was initially considered a solution, word-for-word translation can only render acceptable results for the translation of some "*very simple*" sentences[4] and for specific language-pairs[5]. An MT system faces several challenges in order to obtain good translation results. These challenges may differ depending on the language-pair used: for example, while it is difficult to find word boundaries in languages like Chinese or Japanese, in European languages the word boundary is clearly represented by the '*space*' character. Researchers split these problems into two categories: "*linguistic*" and "*operational*" challenges. The main *linguistic challenges* are ambiguity (lexical, structural, semantic etc.), text generation (lexical selection, tense generation etc.) and the mappings between the SL and TL representations (divergences: thematic, head-switching, structural etc.): [Dorr et al., 1999] and [Somers, 2000b]. More details and examples of linguistic challenges are also presented in [Eynde, 1993], [Schwarzl, 2001] and [Hutchins and Somers, 1992]. An overview of the linguistic challenges encountered in the data used for the experiments described in this thesis will be presented in **Chapter 4**. A non-exhaustive list of the *operational challenges* includes system maintenance, system integration with other programs and system extendibility to other domains and language pairs.

Although the origin of MT is sometimes thought to date back to the publication of Petr Smirnov-Troyanskij's and Georges Artsrouni's '*mechanical dictionaries*' (1930s), the '*programmatic*' start of machine translation is considered Waren Weaver's work, which emerged in the late 1940s [Hutchins, 2004]. Ever since, different MT approaches have been used and several generations of systems have been developed. A view on the history of MT is described in John Hutchins' work "*Machine Translation: a concise history*" [Hutchins, 2007].

The classification of MT systems has been done according to several criteria, such as:

1. **Degree of automation** – The degree of automation is given by the amount of the user's involvement during the translation process, in this case the involvement of the human translator. Less user involvement means more system automation. Considering the degree of the user's involvement in a descending way, MT systems can be classified into three groups: Machine-aided human translation (**MAHT**), Human-aided MT (**HAMT**) and Fully automatic MT (**FAMT**).
2. **Type of the core technology (the paradigm)** – Regarding the core technology, the MT systems can be divided into two classes: *rule-based* and *corpus-based* (*empirical*). The first are often (linguistic-)theory-driven, the latter do not address either

[3]For further explanations on '*analysis*' and '*generation*', please see Figure 2.1.
[4]Considering their syntax and semantics.
[5]Such language-pairs include similar, low inflected languages.

linguistic or cognitive issues. The following two MT approaches are included in the corpus-based class: statistical machine translation (**SMT**) and example-based machine translation (**EBMT**). Over the last few years, hybrid technologies have been used more frequently. The MT paradigms will be further described in **Section 2.2**.

3. **Input type** – Usually an MT system has as input a text which is expected to be syntactically and semantically correct. In the last few years, systems with speech input have been developed, such as Verbmobil [Wahlster, 2000] and EuTrans-I [Amengual et al., 2000]. The translation task becomes even more complicated for speech input, as the system needs to deal also with ill-formed input. The incorrect input appears due to speech recognition errors, ungrammatical utterances etc. Incorrect text input can also evolve when, for instance, translating the output of another automatic NLP application[6].
4. **Level of analysis (the architecture)** – The current rule-based MT (**RBMT**) architectures can be organized into three classes according to the level of analysis: direct, transfer and interlingua (see Figure 2.1[7]). The first supposes a word-for-word

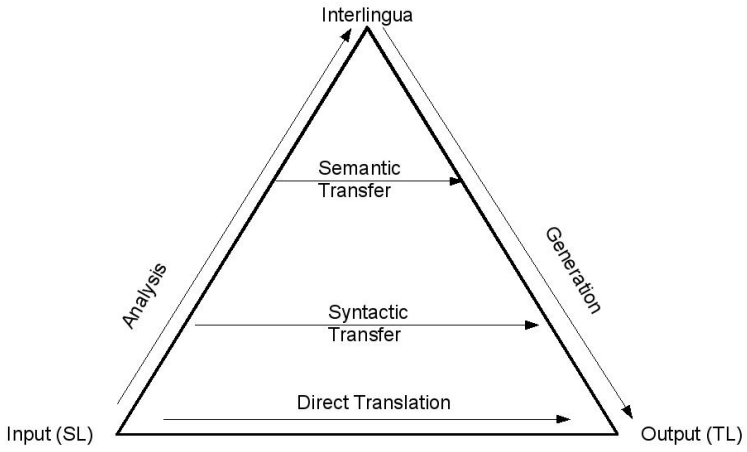

Figure 2.1: Different levels of analysis in an (RB)MT system (the '**Vauquois Triangle**').

translation from the SL to the TL, with no deeper analysis of the input than the one of the word surface forms, and with no other linguistic resources, with the exception of a bilingual dictionary. The second involves a deeper (syntactic and/or semantic) analysis and transfer rules between the SL and TL. The topmost architecture performs the translation using an intermediate (human-created) representation, which is called **interlingua**. Interlingua is a less ambiguous conceptual representation. Systems which use interlingua are also known as knowledge-based MT (**KBMT**) systems. They suppose a complete semantic representation of the input.

5. **Output quality** – The goal of MT has an impact on the expectations for translation quality. The output needs to be of high quality in MT for dissemination

[6]For example, translating the output of a text summarization program.
[7]The 'Vauquois Triangle' is encountered in the literature also under the name of 'Vauquois Pyramid'.

purposes. A comprehensible raw translation might suffice in MT for assimilation. This translation can later be edited by a human translator. A tree diagram of the MT classification according to the output quality is shown in [Carbonell et al., 1992]. A higher quality output is usually obtained when the translation domain is restricted, as in the METEO system [Chandioux, 1976].

The above classification might not be complete, as an exhaustive MT classification is beyond the scope of this dissertation. An extended discussion on MT classification can be found in [Och, 2002] and [Schwarzl, 2001].

According to these classification criteria, the system(s) we developed in this thesis are fully automatic corpus-based MT systems[8], which have as input (syntactically and semantically) well-formed text data. The system(s) can be used in MT for assimilation.

2.2 MT Paradigms

In this section we will discuss the different MT paradigms. The focus is on corpus-based MT (**CBMT**) approaches (SMT and EBMT), as they represent the main part of this thesis: The results of the EBMT systems developed in this dissertation will be compared with the ones of an SMT system.

As mentioned before, MT systems can be classified according to their core technology into two classes: rule-based and corpus-based (empirical). For rule-based systems, human experts have to specify a set of rules, which describe the translation process. This is usually comparatively costly work considering factors such as time, money and man-power. The corpus-based approaches are usually based on translation examples. The MT system analyses automatically the existing examples for translating new sentences. From this perspective, one of the main advantages of this approach is its ability to adapt the MT system to new language pairs and (or) new domains more easily and faster, given that sufficient (training) data is available.

2.2.1 Rule-Based Machine Translation (RBMT)

To have a more complete overview of MT in general, although not directly connected with the main topic of this dissertation, we will present the rule-based MT (**RBMT**) approach in this section.

In the rule-based MT approach, the translation process is based on linguistic rules and consists of three main steps:

- The analysis of the SL text morphologically, syntactically and/or semantically.
- The rule-based transfer from SL to TL.
- The TL text generation using structural conversions.

[8] More exactly, EBMT systems.

These steps need a bilingual dictionary and a grammar which are provided by linguists. It is difficult to manually produce transfer rules to cover a wide variety of input. Moreover, there is always the risk of rule conflicts, which can produce unexpected side effects. Therefore, building rule-based or knowledge-based MT systems such as the PaTrans system [Orsnes et al., 1996] or KANT[9] [Nyberg and Mitamura, 1992], "*is a lengthy, complicated and error-prone process*" [McTait, 2003]. On the other hand, (good) RBMT systems are usually consistent, robust and provide good translations results.

2.2.2 Corpus-Based Machine Translation (CBMT)

In this subsection we will give the definition of a parallel aligned corpus before describing the main CBMT approaches.

A **parallel text (corpus)** is a text together with its translation(s). **Parallel text alignment** is the identification of the corresponding texts in both parts (source and target) of the parallel text. A **parallel aligned corpus** is a collection of parallel aligned texts, which do not necessarily need to be coherent. The corpus contains "examples". An **example** can be a simple or complex sentence or a sub-sentential phrase and it can be stored under different forms, such as strings, parse-trees or templates. In this dissertation 'a corpus' is a (bilingual) parallel aligned corpus, if not mentioned differently.

Statistical Machine Translation (SMT)

In a memorandum written to the Rockerfeller Foundation (1949) Warren Weaver considered that

> "*all [...] (one needs) to do is strip off the code in order to retrieve the information contained in the text.*" (quoted in [Arnold et al., 1994, p.12]).

Waver's work – [Weaver, 1955] – is considered a starting point for statistical methods. This subsection will briefly characterize the SMT approach.

The SMT initial idea was abandoned until the 1990s, when it was reactivated by the work carried out in the TJ Watson Research Center, the IBM Research Division, where Brown et al. [1993] developed an SMT system for French and English.

The SMT approach has contributed to the significant resurgence in interest in MT over the last two decades. At present, there are several SMT approaches (such as word-based or phrase-based SMT) and it is by far the most widely studied MT method.

In SMT the translation process is performed by using two models: a translation model and a language model. SMT treats translation as a machine-learning problem. Formally, SMT can be defined as finding the most likely TL sentence $\tilde{tl}$ for some SL sentence sl:

$$\tilde{tl} = argmax_{tl} P(sl|tl)P(tl), \tag{2.1}$$

[9] `http://www.lti.cs.cmu.edu/Research/Kant/` - last accessed on March 23rd, 2010.

where tl is a target language sentence.

An SMT system has three major components (see Osborne [2010]):

1. A translation model (**TM**), $P(sl|tl)$, which specifies the set of possible translations for a source sentence and assigns probabilities to these translations. The process of extracting the TM uses a bilingual parallel aligned corpus.
2. A language model (**LM**), $P(tl)$, which models the proposed target sentence. In order to obtain an LM, a monolingual corpus for the target language is needed. LMs are usually smoothed n-gram models. Usually the probability of the current word is predicted by conditioning it on two (or more) previous words.
3. A search process (the $argmax$ operator), which is navigating through the space of possible TL translations. This process is called *decoding*. As this process is NP-hard[10] for SMT, most approaches use a beam-search algorithm[11].

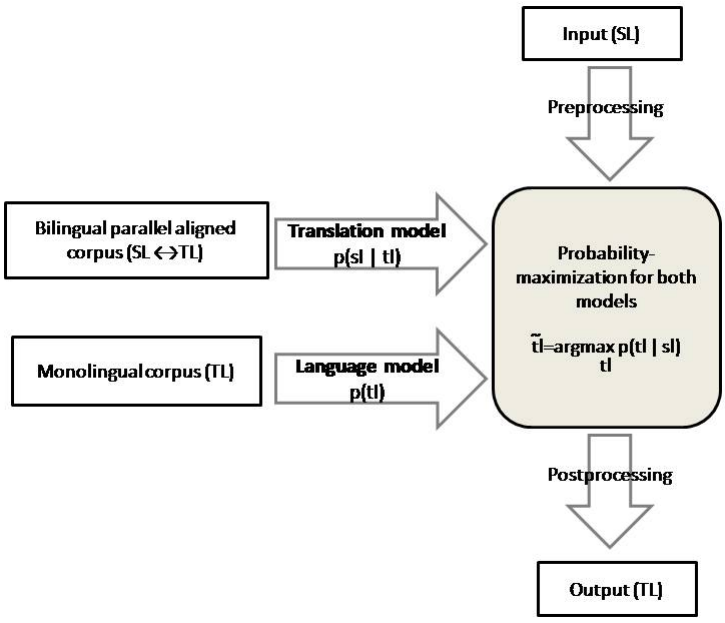

Figure 2.2: The SMT processes – source: [Koehn and Callison-Burch, 2005].

The SMT work-flow is shown in Figure 2.2. An optimal translation is obtained by maximizing the probabilities from the two models. According to the SMT approach used (such as word-based translation like the initial IBM models [Brown et al., 1990], [Brown et al., 1993] or phrase-based translation as in [Koehn et al., 2003], [Och and Ney, 2003]) the complexity level of the models change. A survey on SMT approaches and models is presented in [Lopez, 2008].

SMT systems can be built fast and fully automatically, provided that the needed parallel aligned corpus exists. Open-source projects, such as the phrase-based SMT system Moses

[10]Non-deterministic polynomial-time hard (in computational complexity theory).

[11]Beam-search for SMT is described in [Tillmann and Ney, 2003].

(`http://www.statmt.org/moses/`), and the Workshop on statistical machine translation, which has been organized annually since 2006[12], have stimulated the development of this approach.

Example-Based Machine Translation (EBMT)

The idea of example-based machine translation (**EBMT**)[13] was first put forward in Makoto Nagao's work "*A Framework of a Mechanical Translation between Japanese and English by Analogy principle*" in the early 1980s [Nagao, 1984]. Since then, there has been an enormous interest in approaches which use a bilingual collection of examples (bilingual parallel aligned corpus) as the main bilingual knowledge source. This subsection will shortly present the EBMT approach and its main steps.

Example-based machine translation, called also "*machine translation by example-guided inference*", or "*machine translation by the analogy principle*", follows two main rules, which were first described in Nagao's work:

> **1.** "*Man does not translate a simple sentence by doing deep linguistic analysis*"

and

> **2.** "*Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases [...], then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the* ***analogy translation principle*** *with proper examples as its reference [...].*"

In EBMT a set of phrases in the SL and their corresponding translations in the TL are given: the example database[14]. The MT system uses these examples to translate new similar SL phrases into the TL. The basic premise is that, if a previously translated phrase occurs again, the same translation is likely to be correct again. The way in which an EBMT system determines if an example is equivalent or at least similar enough to the text to be translated varies according to the approach taken by the system in creating the example database: strings, (annotated) tree structures, generalized examples (templates) etc.

After building a database of aligned examples, the '*traditional*' EBMT system follows three steps:

1. **Matching** the SL input against the example database,

[12] For each workshop training and test data were provided.

[13] Here in a narrow sense, not in the general one presented at the beginning of Somers' review, [Somers, 2000a].

[14] Usually the example database is represented by the (preprocessed) parallel aligned corpus, in which the examples are saved, for example, as strings or syntactic structures (parse trees, logical forms etc.). Sometimes it is called an "**example set**".

2. Selecting the corresponding fragments in the TL (**alignment** or **adaptation**), and
3. Recombining the TL fragments to form a correct text (**recombination**). This step sometimes appears in the literature as "*target sentence generation*" [Kit et al., 2002] or as "*synthesis*" [Hutchins, 2005a].

These steps, defined for the first time in Nagao's second rule, are also mentioned in [Somers, 1999]. The representation of the '*Vauquois triangle*' adapted for the EBMT approach as shown in [Somers, 1999] is presented in Figure 2.3.

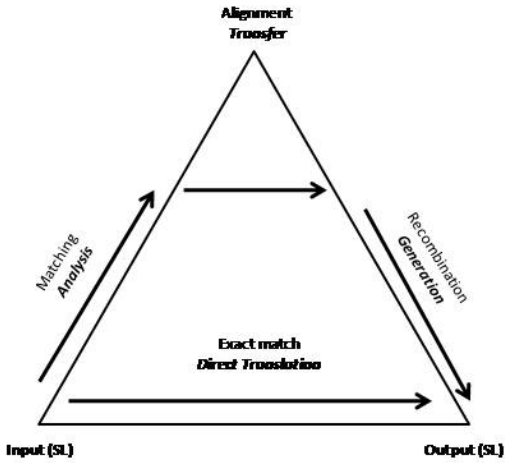

Figure 2.3: The '*Vauquois Triangle*' adapted for EBMT, according to [Somers, 1999]. *(The triangle should not be considered from the point of view of the correctness of the translation results.)*

In some approaches the number of steps differs: for example in [McTait, 2001] the matching and recombination steps are merged into one step, also called "*recombination*". Depending on the EBMT approach (e.g. linear, structure-based, template-based)[15], these steps give rise to different challenges.

As in **Chapter 3** we will describe EBMT in more details, we present below only a brief overview of the main EBMT steps.

Matching

The **matching** step is defined as finding the SL sentences from the corpus similar "enough" to the input sentence. Depending on the approach and data structure chosen for the EBMT system, several algorithms are reported in the literature for this step, such as string-based (e.g. [Matsumoto et al., 1993], [Cranias et al., 1994]) or syntax- or tree-based (e.g. [Watanabe, 1992], [Mandreoli et al., 2002]) algorithms. In the template-based EBMT system presented in [McTait, 2001, p. 76] some kind of rules are implemented for matching.

[15]Details on the classification of the EBMT approaches can be found in **Section 3.2**, **Chapter 3**.

The string-based methods have as basis the edit-distance and can operate at word or character level. These methods sometimes also use semantic information in the form of a thesaurus (e.g. [Nagao, 1984], [Sumita, 2001] etc.) or WordNet[16] ([Nirenburg et al., 1993]). The EBMT systems which perform similarity matching based on syntactic structures (parse trees or logical forms) require the input to be parsed to obtain the needed structure. These example databases sometimes have high maintenance costs as the database sometimes needs to be corrected by humans in order to avoid the propagation of (possible) errors. Another similarity metric based on encoding the sentence in vectors is presented in [Cranias et al., 1994]. An overview on matching metrics can be found in [Cohen et al., 2003] and [Lee, 2001]. Results of some experiments on EBMT matching are presented in [Vertan and Martin, 2005].

Alignment and Recombination

The **alignment** step determines the TL fragments equivalent to the matched fragments in the SL sentence. Once an SL input has been segmented and the TL equivalents of these segments have been determined, the **recombination** task is to recombine the TL fragments appropriately to form "*a legal target text, just as the generation stage of conventional MT puts the finishing touches to the output*" [Somers, 1999].

Usually, recombination techniques are specific to the EBMT approach considered and they are designed according to previous steps[17] and the manner in which the example-database is organized (e.g. parse-trees, templates etc.). Recombination has to adapt itself according to the previous two steps, i.e. if the matching and alignment steps leave unsolved problems or introduce errors (e.g. because of a parser, the alignment or the data), the complexity of the recombination algorithm might increase in order to find a solution to these problems and correct the errors.

In the literature **recombination** is usually described only tangentially. It is also considered "*the most difficult step in EBMT process*" – [Somers, 2003], [Kit et al., 2002] – and it is an "*area that has received little attention*" – [McTait, 2001], [Somers, 1998].

Challenges in the translation process may appear in the "adaptation" and "recombination" steps depending on the underlying approach. One of the most frequent challenges in recombination is represented by disfluencies at the boundaries of the sub-sentential phrases which form the translation (i.e. **boundary friction**). We encounter this phenomenon especially for inflected languages. An example where boundary friction could be a problem for English and German is presented below:

(1) ENG: *The handsome boy entered the room.* / DEU: `Translation needs to be obtained.`

[16] `http://wordnet.princeton.edu/` - last accessed on June 28th, 2011.

[17] The recombination step depends on the sentences extracted by the matching. If the matching method excludes most of the sentences that might help building the output, the recombination step has to provide the translation, with less information. Also, if the word-alignment is wrong, incorrect data is forwarded to the recombination step.

(2) ENG: *The handsome boy ate his breakfast.* / DEU: *Der schöne Junge aßsein Frühstück.*

(3) ENG: *I saw the handsome boy.* / DEU: *Ich sah den schönen Jungen.*

Given the examples in **(2)** and **(3)**, the question is 'which form should be chosen for correctly translating the English text presented in Example **(1)** into German: *der schöne Junge* (nom.) or *den schönen Jungen* (acc.)?'

Possible solutions for this challenge are described in [Somers, 1999]: a grammar of the TL (in a hybrid system) or the consideration of the left and right contexts (which Somers calls "*hooks*"). There are EBMT systems which leave this problem unsolved, such as the Gaijin system [Veale and Way, 1997]. Besides boundary friction, recombination can suffer from loss of information about the relationships between fragments extracted from input sentences. This happens quite often when a language model is chosen as a solution in the recombination step. In this thesis we do not address the boundary friction problem directly, but we include in the recombination step information about the order of the TL fragments to be recombined (see **Chapter 7**).

In [Hutchins, 2005a] recombination "*adapts the extracted TL fragments and combines them into TL (output) sentences*". The author mentions that "*... it can be argued that the operations of synthesis ('recombination'), perhaps the most difficult and complex in EBMT systems, are a consequence of the nature of the output from the matching/extraction process*". On the other hand, in [Hutchins, 2005b], the recombination step is not considered as a part of the core EBMT process, "*since it is a monolingual process, and its nature is determined by the form in which TL fragments are extracted*".

2.2.3 RBMT vs. CBMT Approaches

The difficulties of the RBMT approach (see **Subsection 2.2.1**), known in the literature as "*the knowledge acquisition problem*", motivated the idea of corpus-based (empirical) MT in the 1980s. As stated earlier, the corpus-based approaches make use of a set of previously translated sentences as opposed to the construction of hand-crafted monolingual grammars and transfer rules. These approaches are also found in the literature under the name of data-driven MT. According to [Somers, 2000a], all such approaches (SMT, EBMT[18]) can be grouped together under the generic term of example-based machine translation, as new translations are computed on the basis of previous examples of translated text.

Rule-based MT usually has a consistent and predictable quality. As it knows grammatical rules, it can also provide a fairly good quality for translations from general domains. Conversely, corpus-based MT does not 'know' grammar. Therefore, it has an unpredictable translation quality and delivers poor results for out-of-domain translations.

The advantage of rule-based MT is high performance and consistency between versions.

[18] EBMT in the narrow sense.

On the other hand, empirical MT is inconsistent with respect to the versions. As disadvantage it also has high CPU[19] and memory requirements.

Compared with rule-based approaches, empirical MT can catch exceptions to rules and has a rapid and cost-effective development, provided that the required parallel aligned corpus exists. If sufficient data is available, CBMT can adapt to new language-pairs faster and more easily.

A tabular overview of the advantages and disadvantages for each of the MT approaches is presented in Table 2.1. More information on this topic can be found in [Eisele, 2008].

MT approach	Advantages	Disadvantages
RBMT	• Easy to build an initial system • Based on linguistic theories • Effective for core phenomenon • Consistent and predictable quality • Good general translations	• Knowledge acquisition problem • Difficult to maintain and extend • Experts needed • Ineffective for marginal phenomena
Empirical MT	• Reduces the cost of human work • Extracts knowledge from corpus • Easy adaptation to new language-pairs • Can catch exception to rules	• Unpredictable translation quality • Poor results for out-of-domain translations • High CPU and memory requirements • No grammar known

Table 2.1: RBMT vs. empirical MT.

In [Thurmair, 2004] and in [Labaka et al., 2007] a comparison between rule-based and data-driven MT approaches is presented. Thurmair [2004] evaluates the results of a rule-based and a statistical MT system by classifying the translations as grammatical, understandable and wrong. While the SMT approach has more results in the middle range, the rule-based system provides more grammatical translations. Overall there were better results in the case of the rule-based system (close to 80%) than the statistical MT one (close to 70%). In [Labaka et al., 2007] the evaluation results differ, according to the method (automatic vs. human evaluation) used. While the automatic metrics indicate that the data-driven system outperforms the rule-based system, the human evaluation indicates exactly the contrary.

2.3 Hybrid Approaches

As we have already mentioned, each MT approach has its advantages and disadvantages. In order to gain an advantage by combining the positive sides of each of the approaches, hybrid systems have been developed.

Hybrid MT systems which include EBMT are presented in [Schaeler et al., 2003], where

[19] Central processing unit.

an EBMT system is included in a multi-engine environment. Another hybrid system is shown in [Sumita et al., 2004]. This hybrid system contains two EBMT systems and one SMT selector. The first EBMT system in [Sumita et al., 2004] is based on a Dynamic Programming (**DP**) algorithm[20] and it uses an edit-distance based on words and a semantic distance calculated by means of a thesaurus. It employs thesauri for both source and target languages, as well as a bilingual dictionary. The second EBMT system in [Sumita et al., 2004] uses Hierarchical Phrase Alignment (**HPA**)[21], transfer patterns and a conventional generation. A phrase-based HMM[22] translation model is also used. Paraphrasing is solved using DP-matching.

In an approach to merge statistical and example-based MT, [Watanabe and Sumita, 2003] present a decoder for SMT which takes advantage of the EBMT framework. SMT and EBMT are also combined in [Groves and Way, 2005] and [Smith and Clark, 2009]. An open source platform for data-driven machine translation that puts together the SMT and EBMT approaches is Cunei[23], which is described in [Phillips and Brown, 2009].

In [Carl et al., 1998] an example-based component is included in two rule-based systems. The evaluations and the fine tuning of the components are not presented in the paper, although it can be expected that the information provided by the example-based component is used to improve the results of the rule-based systems. A hybrid rule-based - example-based MT is also presented in [Sanchez-Martinez et al., 2009], where bilingual chunks obtained from parallel corpora are integrated into an MT system built on Apertium, which is an open-source platform for developing rule-based machine translation systems.[24].

Two prototypical architectures in which rule-based systems are combined with SMT systems are presented in [Eisele et al., 2008]. In a first combination multiple MT results from different systems are merged via an SMT decoder. In the second architecture, SMT phrases are fed into a rule-based MT system.

2.4 Chapter Summary

In more than 60 years of MT history several approaches, with a different degree of implication of resources and human experts, have been developed. Each of the approaches has its advantages and disadvantages, and produces results of different quality for various corpora and language-pairs. This is one of the reasons why in the last few years hybrid approaches have been developed (such as the ones presented in **Section 2.3**).

Although in the last few years hybrid MT has been considered a solution for MT, we

[20] Dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems.

[21] [Imamura, 2001].

[22] Hidden Markov Model, [Rabiner, 1990].

[23] `www.cunei.org` – last accessed on June 20th, 2011.

[24] Apertium can be downloaded from `http://sourceforge.net/projects/apertium/` and tested on `http://www.apertium.org/` – both last accessed on June 20th, 2011.

could not have aimed our research at hybrid approaches, as for the language-pairs analyzed (almost) no previous systems had been available. Moreover, additional resources (e.g. corpora, linguistic tools etc.) were quite limited. This is why we have been concentrated our work on EBMT, more exactly on the recombination step of an EBMT system.

In this chapter a short overview of the MT approaches has been presented. For the rest of the thesis corpus-based MT approaches will be analyzed. The SMT system used in the experiments presented in this thesis is based on Moses, an open source SMT system that allows the user to automatically train translation models for any language pair. More information on Moses will be presented in **Chapter 5**. As the SMT system was run as a black box using mainly the parameters recommended for the Sixth Workshop on SMT[25], no further description of the SMT approach will be presented in this thesis. The emphasis is put on the EBMT approach, which will be described in more detail in the next chapter.

[25] `http://www.statmt.org/wmt11/baseline.html` - last accessed on June 20th, 2011.

Chapter 3

Example-Based Machine Translation (EBMT)

In order to place the approach implemented in this dissertation among the existing EBMT systems, in this chapter the EBMT paradigm will be defined and the main EBMT directions will be presented. The main focus will be on the linear and template-based EBMT approaches, as they directly influence the implementation of the EBMT system(s) developed in this thesis.

3.1 Definition

Defining EBMT has proven to be a difficult topic and, as of today, no clear formal definition for EBMT exists. Since Nagao's definition as "*machine translation by the analogy principle*"[Nagao, 1984], researchers have been trying to give different descriptions for the EBMT and to distinguish it from the SMT approach. In this subsection we will provide an overview of these works.

The main characteristics of an EBMT system as presented in [Somers, 2003] are: the use of a bilingual corpus, the examples as "*the main knowledge-base*" and the run-time use of the examples. The classification criterion found in [Turcato and Popowich, 2001] considers the type of linguistic knowledge used by the system as the most important feature, leaving the source and format of the knowledge as secondary. The authors conclude that "*the original idea of translation by analogy stands out as truly example based*".

Hutchins [2005b] views the bilingual process as the essence of the EBMT. This process consists of "*the matching of SL fragments (from the input text) against SL fragments, and the retrieval of equivalent TL fragments (as potential partial translations).*" All other aspects (e.g. recombination, 'run-time' aspect) are judged as auxiliary. The characteristic feature of EBMT is taken to be "*the assumption (or hypothesis) that translation involves the finding of 'analogues' (similar in meaning and form) of SL sentences in existing TL texts.*" [Hutchins, 2005a].

In [Kit et al., 2002] EBMT appears as having four stages: example acquisition, example base management, example application and target sentence synthesis. The first two stages deal only with the example database: how examples are acquired, stored and maintained. The third stage comprises the **matching** and **alignment** steps. The last stage represents the **recombination**. The sentence synthesis is defined as the way to: "*... compose a target sentence by putting the converted examples into a smoothly readable order, aiming at enhancing the readability of the target sentence after conversion*".

In [Carl, 2000], corpus based machine translation (**CBMT**) is seen from the perspective of the **theory of meaning**. Carl discusses the dichotomies of theories of meaning (rich vs. austere, molecular vs. holistic, coarse-grained vs. fine-grained) and classifies nine CBMT systems: four EBMT systems [1], three SMT systems and two translation memories. At the end of the work a model of competence for CBMT – coverage and quality – is given. It was observed that the better the quality, the lower the coverage. The EBMT systems find their place in the middle range for both features.

Further aspects in the (EB)MT definition have been analyzed in [Wu, 2005], [Carl, 2005a] and [Jones, 1992][2]. Differences between phrase-based SMT and EBMT are presented in [Carl, 2005b], [Hovy, 2005] and [Simard, 2005].

This thesis follows the criteria presented in [Somers, 2003] for the EBMT definition underlying the system implemented in this work: the use of a bilingual corpus, the examples as the main knowledge-base and the run-time use of the examples. The additional required information, such as the word alignment and the language model, are extracted prior to the translation process itself.

3.2 Overview of EBMT Systems

As far as the criterion of storing the database of examples is concerned, the EBMT systems are usually separated in the literature – e.g. [McTait, 2001] – into three categories:

1. Linear systems, which are based on raw examples (surface forms),
2. Template-based systems, based on generalized examples (templates) and
3. Structure-based systems, which are built on (parse-) tree structures.

There are systems that use other approaches, such as systems based on proportional analogies (see **Subsection 3.2.3**).

[1] The four EBMT systems are decribed in [Carl, 1999], [Collins, 1998], [Cicekli and Guvenir, 1998] and [Sato and Nagao, 1990]).

[2] In [Wu, 2005] a 3-dimensional MT model space is created. A system-theoretical view of EBMT can be found in [Carl, 2005a]. In an attempt to distinguish EBMT from SMT, Carl considers the following aspects: run-time vs. preprocessing, the structure of the translation unit, rules vs. statistics. Jones [1992] distinguishes between rationalist and empiricist approaches to MT. He compares hybrid example-based systems vs. pure example-based systems.

Sometimes EBMT systems combine several approaches, such as the one presented in [Liu et al., 2006] which is based on tree-string correspondences (**TSCs**) and statistical generation. It views the translation example as a triple consisting of a parse tree in SL, a string in TL and the correspondences between the leaf nodes and the sub-strings in the TL sentence.

An EBMT system can work as a stand-alone application or it can alternatively be integrated into a hybrid MT environment. Independent of the approach, the EBMT systems can include or make use of linguistic resources, such as thesauri, morphological analyzer and generator or POS tagger, to a different degree. The three EBMT steps – matching, alignment and recombination – are implemented and adapted for each of the system types. This chapter will describe the linear and the template-based approaches, which have a direct influence on the EBMT system implemented. Some of the other approaches will be only briefly presented.

3.2.1 Linear EBMT Systems

Linear or non-structural EBMT systems normally employ raw examples rather than abstract representations of them. The systems usually use very little or no preprocessing of the corpus, while the bilingual relationships are computed at run-time. First the input is matched to the examples by overlapping exact matches of the SL input dynamically. The TL equivalents are than extracted by aligning the matched SL fragments. In the final step, the extracted TL fragments are recombined in an appropriate manner to produce the output.

Systems which produce a set of TL fragments with no knowledge about the order of the fragments are problematic. One method to solve this problem is creating a statistical model of the TL, where recombination is expressed statistically. The main idea here is the calculation of the probability of n-gram sequences, as shown in [Brown et al., 1993] or [Brown, 1996], for the PanEBMT system. Since in the recombination step only the LM information extracted from the database of examples is used, possible clues found in the TL sentences provided by the matching and the alignment steps are lost. Grefenstette [1999] verifies alternative translations of ambiguous compound nouns in German and Spanish (when translating into English) by using them as search-terms on the world wide web (**WWW**). The translation with the highest frequency is assumed to be the best.

In linear EBMT, several algorithms have been implemented for the three steps, and an overview of these algorithms will be discussed in this subsection. Some of the systems use extra linguistic information, such as markers – [Gough and Way, 2004] – or semantic information provided by a thesaurus – [Nagao, 1984].

Linear EBMT systems can be categorized according to the type of examples in the bilingual parallel aligned corpus. There are linear systems that consider sub-sentential phrases as examples, i.e. noun phrases (**NPs**). For example, Sumita and Iida [1991] present a linear EBMT system, which translates Japanese NPs of the type "N_1 no N_2" into English

NPs. From similar examples retrieved, the system generates the most likely translation with a reliability factor based on distance and frequency. If there is no similar example within the given threshold, the system provides the user an error message. Matching is done using a distance metric and a thesaurus. Indexing and parallel computing technologies are applied to enhance translation speed.

Another approach which discusses NP translation for Japanese and English in both directions is described in [Sato, 1993]. Here, the NPs represent technical terms. The initial database of examples is transformed into an "*internal translation database*". A record in the internal translation database is formed by an SL and a TL part, where each part consists of three sub-parts: "*focus, previous and next*", where "*previous*" and "*next*" represent the contexts of a "*focus*" to the left and right, respectively. Matching involves a thesaurus and the matching score is the sum of a "*focus matching score*" S_F and a "*context matching score*" S_C. The system generates candidate (complete or partial) translations. The translations are recursively found in order to build the output.

However, most of the linear EBMT systems work with longer phrases than the subsentential ones. A linear translation-aid system (CTM) which uses sentences as examples is presented in [Sato, 1992]. CTM uses a character-based match retrieval method for Japanese and involves an acceleration method using a character index. A character index was necessary due to specific language characteristics for Japanese, such as no spaces between the words or containing more than 7000 characters.

Linear EBMT systems with different types of matching techniques are described in [Doi et al., 2005a] and [Mandreoli et al., 2002]: Doi et al. [2005a] use a search space division, word-graphs and an A^* search algorithm; Mandreoli et al. [2002] utilize SQL mapping together with filtering techniques.

In [Kit et al., 2002] matching is done by decomposing the input sentence into a sequence of seen fragments (examples), using probabilities. In the alignment step the main problem consists in deciding between multiple possible translations. The recombination – sentence synthesis and smoothing – is carried out using a 3-gram LM, which supports word insertion by using a set of "*smoothing*" words. Word insertion is needed when additional words (e.g. function words) improve the readability. Before running the translation algorithm, the example database is created and the examples are aligned. The alignment of the examples is based on a "similarity matrix", which contains the values for the similarity measure for all example-combinations in the database. For choosing the aligned examples, the maximum values on each row and each column are chosen and the union between these two sets is derived.

EBMT systems usually employ non-overlapping fragments. Brown et al. [2003] and Hutchinson et al. [2003] present a new method for improving phrasal translation, i.e "*maximal left-overlap compositional EBMT*"[3]. The method combines overlapping $n-$word

[3]Shortly named "*maximal overlap EBMT*".

fragments, whose translations are consistent. Overlap brings improvement, as it allows a system to use long translations which are not normally considered by standard EBMT.

A problem of CBMT in general and EBMT in particular is data sparseness. In [Gough and Way, 2003] and [Gough and Way, 2004] marker-based segmentation is used for reducing data sparseness. Six sets of markers markers, among which $< PREP >$ (preposition), $< DET >$ (determiner), $< PRON >$ (pronoun), are applied to segment SL and corresponding TL sentences. Marker-lexicons and marker-templates are generated and used. In addition to markers, the EBMT system presented in [Way and Gough, 2003] uses validation and correction via the WWW.

EBMT systems are not always stand-alone applications. Brown [1996] presents the Pangloss EBMT system (**PanEBMT**) as part of a multi-engine MT system. After the matching and alignments steps in the EBMT system, the obtained partial translations are combined with the results of other MT systems to form the final Pangloss translation. All obtained fragment translations are combined into a chart. For determining the best path LM information is used. The EBMT system has several knowledge sources: the sententially-aligned parallel corpus, a bilingual dictionary, a TL root/synonym list and a list of word-classes. Some of the word classes represent language specific information, such as weekdays, countries or measuring units, and some can be considered language independent, such as numbers[4]. In later works, Brown [2001] uses transfer-rule induction[5] followed by word-level clustering to find not only single words, but also transfer rules.

3.2.2 Template-based EBMT Systems

The template-based (or pattern-based) EBMT systems do not only process the surface forms of the examples, but also have the data organized in templates. A template is usually considered as a generalized translation example, where different components can be replaced with variables in both SL and TL sentences, thus establishing bindings (alignments) between these elements. For a better understanding of what a template is, we present below an example for English–Romanian.

Notation. *We use the notation $A \leftrightarrow B$ for an SL sequence A which is aligned to a TL sequence B, where a sequence is represented by one or more tokens[6]. The notation will be used throughout the thesis for any kind of alignment, such as word alignment, sentence alignment or alignments in a template.*

(1) Given the following sentences:
I go to *school* ***by*** *bus.* ↔ ***(Eu) merg la*** *şcoală* ***cu*** *autobuzul.*
I go to *the mountains* ***by*** *train.* ↔ ***(Eu) merg la*** *munte* ***cu*** *trenul.*
the following template can be extracted:

[4]There is not 100% entirely correct, as there are number systems that use comma (,) or point (.) for decimal separation.

[5]Approach also found in [Cicekli and Guvenir, 2001].

[6]A token is a lexical item, a number, a punctuation sign etc.

"*I go to* X_1^{SL} *by* X_2^{SL}.$\leftrightarrow$ *(Eu) merg la* Y_1^{TL} *cu* Y_2^{TL}."

The template from Example (1) has the following (aligned) elements:

- Text fragments:"*I go to* $\leftrightarrow$ *(Eu) merg la* ", "*by*$\leftrightarrow$ *cu*", ".$\leftrightarrow$.";
- Variables: $X_i^{SL} \leftrightarrow Y_i^{TL}$ with $1 \leq i \leq 2$.

Before describing some EBMT systems which are template-based, terminological aspects need to be clarified. In several papers – [Kaji et al., 1992], [Kinoshita et al., 1994], [Carl, 1999] and [Cicekli and Guvenir, 2003] – the generalized examples are called **templates**. The SL and TL parts of a template can be found under different names, such as *pseudosentence* ([Kaji et al., 1992]), *SL/TL expression* ([Carl, 1999]) or *pattern* ([Kinoshita et al., 1994]). In [McTait, 2003] the generalized example is called **pattern**, and the SL and TL parts are named *SL and TL sides*, respectively. We will use mainly the terms "*template*" and "*SL and TL sides*".

The EBMT system based on templates works in two steps: firstly, the translation templates are extracted using a learning algorithm and, secondly, these templates are used in the actual translation process.

The translation template learning algorithm is normally based on heuristics which assumes the following: When given two translation examples

$$SLsentence_i \leftrightarrow TLsentence_i,\ 1 \leq i \leq 2,$$

if the SL sentences ($SLsentence_i$, $1 \leq i \leq 2$) contain similar fragments, then the corresponding parts in the TL sentences ($TLsentence_i$, $1 \leq i \leq 2$) should be similar and should represent the corresponding translation. Moreover, the remaining different fragments in the SL should match different fragments in the TL. These fragments are represented by means of variables.

The systems based on this approach differ in the way the templates are extracted and stored. Depending on the languages involved and on how rich in linguistic resources the system is, the templates might be formed only on surface forms or might include additional linguistic information, such as morphological information (see [Cicekli and Guvenir, 2001] and [McTait, 2001]).

The template-based EBMT systems simplify the recombination step, due to the fact that the correspondences (alignments) between the SL and TL text fragments and variables in translation templates have been computed during the template extraction or alignment phases. This way these correspondences are explicitly labeled and there is direct information about the order of lexical items in the TL.

For template-based EBMT, recombination is based on the output of the matching step. The matching step operates on "*template matching*" or on finding the template that provides the (an) "*optimal cover*" of the SL input.

In the rest of the section we will describe several template-based EBMT systems.

[Kaji et al., 1992] is one of the first works in which translation templates are defined as "*a bilingual pair of pseudo sentences*" which include variables. The variables might have syntactic or semantic information attached to them. Words or phrases which satisfy these conditions can replace a variable. The two pseudo sentences in a template contain the same number of variables and these variables are aligned. By replacing the variables with text fragments, a pair of real sentences, which are translations of each other, is obtained. The system also consists of the two main steps mentioned before: learning the translation templates and the translation process itself. It uses syntactic information and a bilingual dictionary. The translation process contains the three steps of an EBMT system: SL template matching, translation of the matched words and phrases and TL generation. The same central steps also appear in [Kinoshita et al., 1994]. Here, the translation template contains at least a pair of "*patterns*" (source and target patterns), each of them consisting of constants and variables. A source pattern (**SP**) is compared to the source sentence, while the target pattern (**TP**) is used to generate the target sentence. The templates contain syntactic information and are constrained on the source and target part, with source and target conditions respectively.

In [Cicekli and Guvenir, 1998], [Oz and Cicekli, 1998] and [Cicekli and Guvenir, 2001] one of the languages used in the experiments is Turkish, an agglutinative language. Not to limit the template extraction under these conditions, a word is represented in its "*lexical level representation*", decomposed into its stem and its morphemes. It is assumed that "*the generation of surface level representation of words from their lexical level representation is unproblematic*". The system also generates "*atomic translation templates*", which do not contain any variables. The acquisition of translation rules (i.e. translation templates) is a machine learning problem, where the algorithm is applied iteratively, until no additional templates are learned.

In [McTait, 2001] templates with or without morphological information are extracted and their influence on the translation results is compared. The recombination step presented includes both matching and recombination from the three-step algorithm, which originally appears in [Nagao, 1984]. In order to get the best translation a translation confidence score is used in case of several possible translations.

In [McTait, 2001], a translation template[7] is formally defined as a 4-tuple $\{S, T, A_f, A_v\}$, where S represents a sequence of SL text fragments separated by SL variables, T is a sequence of TL text fragments separated by TL variables, and A_f and A_v are the alignments between S and T of the text fragments and variables respectively. In a template, a variable also keeps the place for a text fragment. A text fragment is a continuous series of one or more lexical items. Examples (2) and (3) present templates for the language-pair English–French with and without morphological information.

(2) Given the following sentences:

[7]The author uses the term "*pattern*" for a template.

The commission ***gave*** *the plan* ***up*** ↔ *La commision* ***abandonna*** *le plan*
Our government ***gave*** *all laws* ***up*** ↔ *Notre governement* ***abandonna*** *toutes les lois*

With common text fragments, the template

"*(...) gave (...)up ↔ (...) abandonna(...)*"

can be formed. When using the variables as text fragments, we obtain the "*complement*" templates:

"*The commission (...) the plan (...) ↔ La commision (...) le plan*"
"*Our government (...) all laws (...) ↔ Notre governement (...) toutes les lois*"

(3) Given the corpus examples:
The telephone*s worked* ↔ ***Les téléphone****s functionnaient*
The telephone *failed* ↔ ***Le téléphon****e échouait*,

lemmatisation produces the following results:

"***The telephone****+s work+ed* ↔ ***Le****+s* ***téléphone****+s functionner+aient*"
"***The telephone*** *fail+ed* ↔ ***Le téléphone*** *échouer+ait*"

From these examples, the template

"*The telephone (...) ↔ Le téléphone (..)*"

can be extracted. The morphological templates are formed by replacing lemmas in the translation template above with the correspondent suffixes computed during the morphological analysis of the corpus:

"*[] +s(...) ↔ +s +s (..)*"
"*[] [](...) ↔ [] [] (..)*"

Other systems based on templates are described in [Malavazos and Piperidis, 1999] and [Malavazos et al., 2000], in which additional matching and recombination algorithms are implemented for the cases for which no translation templates have been found.

The **Gaijin** MT system, described in [Veale and Way, 1997], employs template-matching, statistical methods, string-matching and Case-based Reasoning (**CBR**) in order to provide a linguistic-lite EBMT solution. The only linguistic information used by Gaijin is the "*Marker Hypothesis*" [Green, 1979], which is used in the creation of the templates. The markers employed are *DET* (determiner), *QUANT* (quantifier) and *PREP* (preposition). After aligning the bilingual corpus and automatically constructing the lexica, the system uses corpus-based statistics to infer translation templates, which encode a mapping between an SL and TL grammatically-marked sentences. The translation itself is processed in two steps: example (template) retrieval and translation adaptation. After the translation is completed and shown to the user, the new example formed is incorporated into

the database of examples. The original example phrase is adapted, if the new source phrase differs only by a few words, especially if those words represent merely paradigmatic changes, such as singular-plural variations. The system offers no real solution for boundary friction or for the violations of agreement conditions.

The EDGAR system, described in [Carl, 1999], integrates morphological knowledge, simple syntactic rules for analysis and generation and a component which induces translation templates from the translation examples. The source language generalization is expended in the TL by specifying internal constraints (e.g. indexes of the matching example) and external constraints (e.g. morpho-syntactic constraints, such as case and POS) and successively refining the retrieved TL examples. An example of a translation template is given below.

(4) $(X_{dp,nom}\ love_{fin}\ Y_{dp,acc})_s <-> (X_{dp,nom}\ lieben_{fin}\ Y_{dp,acc})_s$

Another system in this category is presented in [Echizen-ya et al., 2000]. It formulates a translation rule that represents the structure of the whole sentence, by automatically forming a Translation Transition Network. The translation rules can be seen as templates and have the following format:

(5) *(He likes @0.; Kare wa @0 ga suki desu.)*

The system in [Sumita, 2001] is a combination of a linear system (the matching step) and a template-based system (the alignment and the recombination). The linear matching retrieves the most similar example by carrying out DP-matching of the input sentence and example sentences while measuring the semantic distance of the words by use of a thesaurus. The approach adjusts the gap between the input and the most similar example by using a bilingual dictionary. The translation process consists of the following steps: retrieval of the most similar translation pair, translation patterns generation, selection of the best translation pattern and generation of the output. The system employs only the best translation pattern and recombination is achieved in a way similar to the approach found in ("*pure*") template-based systems.

3.2.3 Other EBMT Approaches

As already mentioned, next to the linear and template-based approaches, there are several other types of EBMT systems. As these approaches are not incorporated in the development of the EBMT system(s) in this thesis, the systems below provide only a more complete picture of the multitude of EBMT approaches.

Proportional Analogies

EBMT systems based on proportional analogies are found in the work of Yves Lapage and his colleagues and have been developed at the ATR Research Laboratories, Japan. The "*purest EBMT system*" described in [Lepage and Denoual, 2005] is based on previous

work of the same research group: [Lepage, 1998], [Lepage, 2000] and [Lepage and Peralta, 2004]. These works provide the basis of the theory of analogies on words and sentences. The system does not need any preprocessing and uses no variables, templates, training or transfer methodology, as it is based on proportional analogy and analogical equations. A proportional analogy is noted as $A : B :: C : D$ in its general form and reads '*A is to B as C is to D*'. It is a logical predicate that necessarily takes four arguments. An analogical equation has the form $A : B :: x : D$, where x is unknown. Solving the analogical equation means finding a sentence C which can be used in the place of x so that $A : B :: C : D$.

The process of building proportional analogies is based on the algorithm presented in [Lepage and Peralta, 2004], where '*paradigm tables*' are used. A paradigm is created on a number of series of commutations among sentences, which can appear both at the front and at the end of sentences with a certain degree of freedom. This way new (short) sentences are generated, which are similar to existing sentences in a linguistic resource.

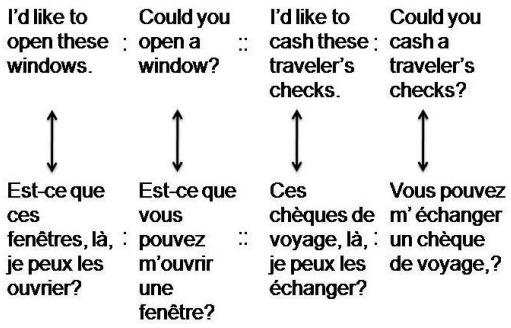

Figure 3.1: Proportional analogies in two different languages that correspond [Lepage and Denoual, 2005].

Figure 3.1 presents proportional analogies in English and French that correspond: each sentence in the lower part of the figure is a possible translation of the sentence above it in the upper part of the figure. The translation of a source sentence D^{SL} is obtained by using the proportional analogies in the TL which correspond to the proportional analogies of the SL that involve D^{SL}. The translation process follows the steps:

- Form all analogical equations which contain the input D^{SL}

$$A_i^{SL} : B_i^{SL} :: x^{SL} : D^{SL} \tag{3.1}$$

- For those sentences $x^{SL} = C_{i,j}^{SL}$, solutions of the analogical equations in 3.1, form all corresponding analogical equations for the target language

$$A_i^{TL} : B_i^{TL} :: C_{i,j}^{TL} : y^{TL} \tag{3.2}$$

- The solutions $y^{TL} = D_{i,j}^{TL}$ of the equations in 3.2 are possible translation for the input D^{SL}

This approach can capture lexical and syntactical variations without explicitly decomposing sentences into fragments [Lepage and Denoual, 2005]. If the number of examples increases, the approach has several drawbacks, such as the increase of run-time and solution space. These drawbacks as well as a compromise solution are reported in [Somers et al., 2009].

Structure-Based EBMT

Structure-based EBMT systems use additional linguistic information and tools, as they have the examples usually stored as tree-structures.

The parsed SL sentence is matched against the SL tree structures until a structure is found that covers the input most accurately. Matching against a set of tree structures is more complex than matching on strings and it requires linguistic resources, such as parsers. This affects the portability, but the translation results should improve as these resources add additional linguistic information. The alignments are established at both lexical and structural level. Since examples are stored as annotated tree-structures and the correspondences between fragments are explicitly labeled, recombination in this approach seems trivial. Structure-based EBMT systems are presented in:

- [Sato, 1995], where the recombination is seen as tree unification,
- [Watanabe, 1992] and [Watanabe, 1995], where graph unification ("gluing"[8] from Graph Grammars) is employed, and
- [Al-Adhaileh and Kong, 1999], in which a process similar to top-down parsing is implemented.

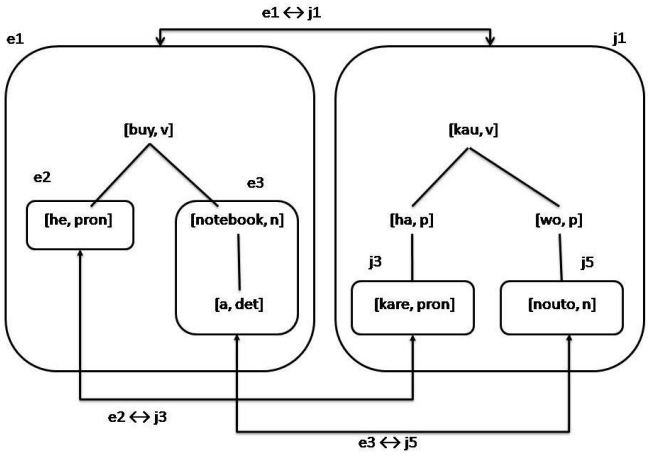

Figure 3.2: SL/TL word dependencies trees and the correspondent alignments – Example from [Sato, 1995].

[8] "Gluing is the process that, given two graph morphisms [...], produces an LDG [Labeled Directed Graph] in which nodes with the same colors (labels), mapped by" the two morphisms, "are glued". For more details see [Watanabe, 1995].

Sato [1995] describes MBT2, a bidirectional English-Japanese system. Here, the examples are represented as pairs of SL/TL word-dependency trees with explicit links between sub-trees (see Figure 3.2). The translation database is a collection of translation examples, where an example consists of three parts: an English word-dependency tree, a Japanese word-dependency tree and a correspondence list.

Similar representations can also be found in [Watanabe, 1992], [Watanabe, 1995], [Matsumoto et al., 1993], [Aramaki and Kurohashi, 2004] and [Vandeghinste and Martens, 2009]. Differences appear in the complexity of the data structures used to represent dependency structures and the correspondences between them.

3.3 Comparison EBMT - SMT

There is no doubt that over the last few years CBMT approaches have been the focus of the MT research and development. As we have already mentioned, among them, the SMT approach has been by far the most dominant MT direction. This is shown also by the available open-source software and the yearly Workshops on statistical machine translation[9]. The Workshop on EBMT in November 2009[10] established a revived interest in EBMT and the intention to develop open-source resources for this MT approach as well.

There has always been a "competition" between these two MT approaches. Similar and unprecise definitions and the mixture of ideas make them difficult to distinguish. The differences became even more difficult to notice since phrase-based SMT systems have appeared. In order to show the advantages of one method over another, several comparisons between SMT and EBMT (or hybrid) systems have been published in the literature. The results, depending on the data type and the MT system, uncovered positive aspects for both approaches.

The marker-based EBMT system described in [Way and Gough, 2005] "*appears to outperform SMT by a factor of two to one*". The evaluation metrics considered are BLEU[11], Precision and Recall [Turian et al., 2003], Word Error Rate (**WER**) and Sentence Error Rate (**SER**). The SMT system is based on Giza++[12], the CMU-Cambridge statistical toolkit[13] and the ISI ReWrite Decoder[14]. The systems use a restricted[15] corpus (a translation memory from Sun Microsystems) for French and English, in both directions. The size of the training corpus varies from 50K to 200K. The test set contains 4K sentences.

[9]The First Workshop on SMT took place in 2006. However, the first steps were done in 2005 with the Workshop on "**Building and using parallel texts: data-driven machine translation and beyond**".

[10]`http://computing.dcu.ie/~mforcada/ebmt3/` - last accessed on January 12th, 2010.

[11]**BiL**ingual **E**valuation **U**nderstudy. More information can be found in Papineni et al. [2002] and **Section 8.2.1**.

[12]`http://code.google.com/p/giza-pp/` - last accessed on June 15th, 2011.

[13]`http://www.speech.cs.cmu.edu/SLM_info.html` - last accessed on June 15th, 2011.

[14]`http://www.isi.edu/licensed-sw/rewrite-decoder/` - last accessed on June 15th, 2011.

[15]By restricted corpus it is meant restricted from the point of view of the domain, syntax, etc.

For French-English the SMT system is better in 9 out of 15 cases; for English-French only in 1 out of 15 cases.

In [Smith and Clark, 2009], the hybrid EBMT-SMT system is outperformed by a Moses-based SMT approach. This approach uses the Europarl corpus in the English-to-French direction of translation. Evaluation is performed using BLEU. The hybrid system consists of an EBMT system which translates the parts of the sentences for which it is confident, followed by an SMT system which fills in the gaps and produces the entire translation. The SMT component is based on Moses[16]. The authors use different matching components in their work, such as string-based and syntax-based.

3.4 Previously Reported Results

This section will present an overview of previously reported results for the language-pairs we used in this dissertation.

For English-Romanian, results for both SMT and EBMT systems have been reported in the literature.

SMT systems, with BLEU results of 0.5464 and 0.3208, are presented in [Cristea, 2009] and [Ignat, 2009] respectively. These systems use parts of the JRC-Acquis corpus (**Chapter 4**) as training and test data. While the architecture described in [Cristea, 2009] involves the use of additional linguistic resources, Ignat [2009] uses pivot languages[17]. Although trained on almost the same type of data (parts of JRC-Acquis), as long as comparisons are not made on identical training and test data-sets, it is difficult to estimate if, overall, the inclusion of linguistic tools available for the moment for Romanian increases the performance significantly.

The SMT results for Romanian-English, German-Romanian and Romanian-German reported in [Ignat, 2009] are 0.3840, 0.2373 and 0.2415, respectively. For Romanian-English the BLEU score reported in [Cristea, 2009] is 0.4604. Koehn et al. [2009] present further results for the same language-pairs and corpus (JRC-Acquis).

Results of an EBMT system that employs linguistic resources (such as morphological analyzers and generators) and the JRC-Acquis corpus are reported in [Irimia, 2009]. Here, the initial hypothesis is that generalization of the data is beneficial for CBMT (as it provides more information, reduces data sparseness and increases the linguistic coverage). In this case lemmatization is used for Romanian and English. As the morphological analyzers and generators are not error-free, the translations results might be negatively influenced. The translation results depend on the degree of inflection of the language. The results are better for the direction English-Romanian, when word forms are considered for the matching, and for Romanian-English, when lemmas are used in the matching step.

[16] See **Chapter 5**.

[17] A pivot language an intermediary language for translation between many different languages. For example, to translate between any pair of languages A and B, one translates A to the pivot language P, then from P to B.

Irimia [2009] reports as maximum BLEU scores 0.3088 for English-Romanian and 0.3689 for Romanian-English. These scores are below the scores of the SMT systems presented in [Cristea, 2009] and [Ignat, 2009]. Although they use the same language pairs and corpus, as the training and test data are not identical, a one-to-one comparison between these three MT systems is not feasible. To the best of our knowledge, a one-to-one comparison between EBMT and SMT using JRC-Acquis has not been attempted so far.

3.5 Chapter Summary

In this chapter, EBMT has been defined and the main EBMT system-types have been described. Those approaches to EBMT that are relevant for this thesis have been presented and analyzed with respect to their advantages and disadvantages. The systems based on surface forms (the linear systems) can be applied to a larger range of languages, but they lack any other linguistic information (e.g syntax) and the end results might be not as accurate as those of the other approaches. The structure-based systems have the advantage of using linguistic information (i.e. syntactic information), which might help in the recombination step. Templates have the advantage of considering longer sequences of a sentence and reducing the error probability. In addition, the word alignment information contained in the template helps in the recombination step. There are also cases where several EBMT approaches are combined. The translation results also depend on factors such as the corpus quality, the number of sentences or the richness of the vocabulary.

Section 3.3 presented a comparison between the results obtained with reported EBMT, SMT and (or) hybrid systems. A tabular overview of EBMT systems, in which the language-pairs and the corpus (type, training ans test data size) have been analyzed, will be presented in **Appendix A**. The last section in this chapter (**Section 3.4**) has shown previous results obtained for the language-pairs used in this thesis and corpus-based MT approaches.

Two EBMT systems have been developed during this research: the former is a linear EBMT system, the latter is a hybrid one, which combines ideas from linear and template-based approaches. The systems will be presented in **Chapters 6** and **7**.

In the next chapter the corpora used in the experiments will be described.

Chapter 4

Corpora Description

The main resource for CBMT in general and for EBMT in particular is the (bilingual) parallel aligned corpus. The decision which corpus to use in a CBMT system depends on the system prerequisites, the (linguistic) resources available and the language-pair in question. Furthermore, this decision directly influences the translation results.

Two domain-restricted corpora are used for the experiments in this dissertation: the JRC-Acquis corpus and the RoGER corpus. JRC-Acquis contains legal texts of the European Community and RoGER is a technical manual of an electronic device. While JRC-Acquis is sufficiently large to train an SMT system on it, RoGER is a small size corpus which better fits the setting of EBMT (which usually uses 'narrow' domains).

A small sub-part of JRC-Acquis is used as a third corpus in some of our experiments: JRC-Acquis$_{SMALL}$. As it is part of JRC-Acquis we will not describe JRC-Acquis$_{SMALL}$ in this chapter, but provide information about the training and test data in **Chapter 8**.

This chapter will motivate the choice of the corpora and describe the data used. Furthermore, the translation challenges (such as mismatches and diverges between the languages) that were found while manually analyzing part of the data will be presented. Other corpora that might be possible candidates for similar experiments, using the same language-pairs, will briefly be described in **Appendix C**.

This chapter will also give a brief overview of the language characteristics for Romanian (**RON**) and German (**DEU**). More information about this topic can be found in **Appendix B**.

4.1 Introduction

EBMT systems are normally used as stand-alone applications for domain-specific cases. They might also be integrated into existing MT architectures (hybrid machine translation) as in [Carl et al., 1998], they can improve coverage in a system that integrates several knowledge sources (see the Pangloss System [Frederking et al., 1994]) or they assist in translating compositional compounds – [Carl, 1999].

Irrespective of how they are used, the main resource for EBMT is the bilingual parallel aligned corpus. The decision which corpus to use depends on the system prerequisites and the language-pair in question. For some language-pairs the choice is restricted by the lack of resources. Even provided the resources and a good MT algorithm, it is not guaranteed that the system provides a good translation. The choice of the (training and test) data and the quality of the corpus are factors which directly influence the MT results.

While in the SMT approach the size of the corpus is considered to be very important – the larger the corpus, the better the results –, in the EBMT approach the size and the type of the corpora used differs: from 32 sentences [McLean, 1992] to the size of the WWW [Way and Gough, 2003], as it can also be seen in Table A.1 (**Appendix A**).

As EBMT got best results for domain restricted corpora, most of the researchers chose such data: Collins [1998] in her thesis uses a Corel Draw manual, Way and Gough [2005] a Sun Microsystem translation memory and McTait [2001] a ScanWorX User Manual (Xerox) and the WHO AFI news titles corpus. The system in [Doi et al., 2005a] employs the Basic Travel Expression Corpus for Japanese and English.

This thesis makes use of two domain restricted corpora, which differ in size and method of compilation: JRC-Acquis and RoGER. As one of the reason for choosing these corpora is represented by the language-pairs under-consideration, we first present a brief overview of the characteristics of the languages used, before describing the data itself. Since English is one of the most frequently used languages for Natural Language Processing (**NLP**) applications, language characteristics are briefly described only for Romanian and German.

4.2 Romanian and German - A Brief Overview

In this section we briefly present some language characteristics for Romanian and German. Both are inflecting languages and have particularities that are absent from English, which can make the translation process even more challenging (e.g. noun inflection, compound words, particle verbs, word order). For example, consider the German sentence in Example (1):

(1) *Den Fisch isst das Mädchen.* (ENG: *The girl eats the fish*)

If the translation systems does not have sufficient information, the following Romanian and English translations could be obtained:[1]

(2) *RON: *Peştele mănâncă fata.*
* ENG: *The fish eats the girl.*

In this case the original semantics is reversed.

[1] We use the character '*' for marking direct, word-for-word translation. The translation is not necessary correct.

An example of a human translation encountered in the RoGER corpus is presented below[2]:

(3) DEU: *andere in diesem handbuch erwaehnte produkt - und firmennamen koennen marken oder handelsnamen ihrer jeweiligen eigentuemer sein .*
RON: *alte nume de produse si de firme mentionate aici pot fi nume comerciale sau marci comerciale apartinind proprietarilor respectivi .*
(ENG: *other product and company names mentioned herein may be trademarks or tradenames of their respective owners .*)

This example contains several changes in the translation which were made by human translators for fluency or appeared due to the languages involved. All these make the process of automatic translation more challenging:

- inversions: "*... erwaehnte produkt - und firmennamen*" ↔ "*nume de produse si de firme mentionate ...*" (*names of products and companies mentioned ...*), "*jeweiligen eigentuemer*" ↔ "*proprietarilor respectivi*" (*respective owners*), etc.
- compounds: "*produkt - und firmennamen*" ↔ "*nume de produse si de firme*" (*names of products and companies*)
- verb position "*... koennen ... sein .*" ↔ "*... pot fi ...*" (*... can be ...*)

Also, no word-for-word translations [e.g. "*in diesem handbuch*" (*in this manual*) ↔ "*aici*" (*here*)] and different ways of expressing possession [e.g. "*ihrer ...*"(*their ...*) ↔ "*apartinind ...*" (*belonging to ...*)] make the automatic MT process more challenging, especially in the word alignment step. Another alignment challenge for the MT system is the case of the word "*owners*": "*eigentuemer*"↔ "*proprietarilor*" (gen., pl., definite article), as "*eigentuemer*" can also be translated with different forms, such as "*proprietari*"(nom./acc., pl., no article), "*proprietarii*" (nom./acc., pl., definite article) or "*proprietar*"(nom./acc., sg., no article).[3]

Romanian

Romanian is an Eastern Romance language [Lewis, 2009], whose grammar and basic vocabulary are closely related with those of its relatives (e.g. Italian, Spanish, French). It has been influenced by several other languages, such as the Slavic languages, Hungarian and Turkish. This influence is encountered especially at lexical level.

Romanian has a rich morphology. Among the language-specific characteristics induced by its Latin origin are the following: a 3-gender system, double negation and pronoun-elliptic sentences. Also, as in all Romance languages, Romanian verbs are highly inflected according to, for example, person, number, tense and mood. Another Latin element that

[2]All examples are presented as they appear in the (training and test) data, after they are pre-processed. The data is tokenized and lowercased (see the Moses-based SMT system - **Chapter 5**.

[3]A perfect automatic word alignment is still not possible. In Example (3) the correct alignment would be "*ihrer ... eigentuemer*" (*their... owners*) ↔ "*apartinind proprietarilor*" (*belonging to the owners*).

has survived in Romanian while having disappeared from other Romance languages is the morphological case differentiation in nouns, albeit reduced from the original seven[4] to only three forms (nominative/accusative, genitive/dative and vocative).

It is the only Romance language where definite articles are attached to the end of the noun or the adjective as enclitics, depending on the position of the adjective before or after the noun. This phenomenon is encountered in some Slavic languages (Bulgarian, Macedonian), in Scandinavian languages (e.g. Danish) and in Albanian[5]. A better overview on this matter can be found in [Himmelmann, 2001]. Within a sentence there is no predefined position for the verb and adjectives can be placed before or after the noun. With respect to the available linguistic resources for NLP, Romanian is under-resourced when compared to other languages, such as German and English.

German

German is a Germanic language that is also inflected. Like Romanian, it also has a 3-gender system and well defined inflectional classes. A one-to-one mapping between Romanian and German inflection is not possible, e.g. the word "*sun*" is feminine in German ("*die Sonne*"), but masculine in Romanian ("*un soare*").

A special feature in German is the verb particle. These particles can be separated from the verb inside the sentence and the particle can also be, in different contexts, preposition or adverb. Depending on (the type of) the particle, the verb changes its meaning. Word order is generally less rigid than in English. However, there are two important rules which establish the position of the verb in the main and subordinate clause (see Example (4)). Another characteristic of the language is that it can contain embedded relative clauses.

(4) Example from RoGER:
DEU: *wir , die nameprod corporation ,* ***erklaeren*** *voll verantwortlich ,* ***dass*** *das produkt npl - num den bestimmungen der folgenden direktive des rats der europaeischen union* ***entspricht*** *: num .*
(ENG: *we , nameprod corporation* ***declare*** *under our sole responsibility that the product npl - num* ***is in conformity*** *with the provisions of the following council directive : num .*)

German forms noun compounds where the first noun modifies the second. These compounds are almost always represented as one orthographic word. In the process of word formation a large number of words can be involved, as in

(5) "*Donaudampfschifffahrtsgesellschaftskapitänskajütentürschloss*" (60 letters, 9 words)[6].

[4]Latin has two additional cases: ablative – which marks motion away from something – and locative – which indicates a location.

[5]The Albanian language is a distinct Indo-European language which cannot be classified into any branch.

[6]The word is not really in use. However, it is grammatically correct.

The English translation for this word, given in [Voit, 2007], is "*The lock on the cabin door of the captain from the Danube steam-ship company*".

German allows lengthy nominal modifiers:

(6) "*Der während des Bürgerkrieges amtierende Premierminister*" (literally: *the during-the-civil-war office-holding prime minister*) - `www.wikipedia.org`.

To better understand the degree of inflection of all three languages, we present in Table 4.1 the declension of the noun "*man*", with definite and indefinite article.

With definite article			
Case	**Language**	**Singular**	**Plural**
nom.	ENG	the man	the men
	DEU	der Mann	die Männer
	RON	bărbatul	bărbaţii
acc.	ENG	the man	the men
	DEU	den Mann	die Männer
	RON	bărbatul	bărbaţii
dat.	ENG	(to) the man	(to) the men
	DEU	dem Mann	den Männern
	RON	bărbatului	bărbaţilor
gen.	ENG	(of) the man	(of) the men
	DEU	des Mannes	der Männer
	RON	bărbatului	bărbaţilor
voc.	RON	bărbatule!	bărbaţilor!
With indefinite article			
Case	**Language**	**Singular**	**Plural**
nom.	ENG	a man	some men
	DEU	ein Mann	-
	RON	un bărbat	nişte bărbaţi
acc.	ENG	a man	some men
	DEU	einen Mann	-
	RON	un bărbat	nişte bărbaţi
dat.	ENG	(to) a man	(to) some men
	DEU	einem Mann	-
	RON	unui bărbat	unor bărbaţi
gen.	ENG	(of) a man	(of) some men
	DEU	eines Mannes	-
	RON	unui bărbat	unor bărbaţi
voc.	RON	bărbate!	bărbaţi!

Table 4.1: Noun inflection.

As it can be seen from Table 4.1, taking into account both cases – with definite and indefinite article – there are four word forms for the noun in German, two in English and eight in Romanian. There are five forms for the definite article in German and one in

English. For Romanian the article is attached as an ending to the noun. For the indefinite article two forms are found for English and four for German or Romanian.

More exhaustive information about these languages is given in **Appendix B**, including an analysis of the language characteristics that might influence the MT process. More details about the Romanian grammar can be found in [Barbuţă et al., 2000] and [Cojocaru, 2003]. In [Motapanyane, 2000] comparative studies in Romanian syntax are presented. A concise description of the German grammar is presented in [Voit, 2007].

4.3 JRC-Acquis

One of the corpora used in this thesis is *the Joint Research Center Collection of the Acquis Communautaire* (JRC-Acquis), a freely available parallel corpus in 22 languages[7]. The corpus is built from the European Union (**EU**) documents mostly of legal nature. As mentioned in [Steinberger et al., 2006], it comprises the contents, principles and political objectives of the treaties, EU legislation, international agreements, acts and common objectives. The corpus and its documentation are freely available for research purposes on `http://wt.jrc.it/It/Acquis/` (last accessed on June 25th, 2011).

4.3.1 Motivation

Before describing the data, we discuss the motivation for the choice of this corpus. There are several aspects to be taken into account when deciding to use a parallel corpus:

- The language-pairs;
- Its use in other systems presented in the literature, to facilitate comparisons;
- Its domain and size;
- The available tools and (or) extra resources (e.g. sentence alignment).

Firstly, the choice of JRC-Acquis was motivated by the languages considered in this work: Romanian, German and English. As one of the goals is to analyze how the systems behave when changing the language-pair, it is necessary to have the same data for all language-combinations.

Romanian could be considered, until recently[8], an under-resourced language. Only few resources and tools have been developed for Romanian. [Tufiş et al., 2008a] and [Cristea and Tufiş, 2002] present an overview of the tools for Romanian. Bilingual resources which include Romanian are rare and, with few exceptions (such as [Vertan et al., 2005] or [Tufiş et al., 2008b]), relate only to English-Romanian. Although parallel corpora have been developed mostly after Romania joined the European Union in 2007, the choice of

[7]Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene and Swedish.

[8]More NLP tools and resources have been developed for Romanian after the adherence of the country to the European Union in 2007.

resources that contain **all** the three above-mentioned languages is still limited (see also **Appendix C**). While English-Romanian parallel corpora appeared in the 1990s (e.g. George Orwell's novel "**1984**", developed in the Multext-East project[9]), most German-Romanian resources have been created in the last few years.

Secondly, several MT results presented in the literature are based on JRC-Acquis and use (part of) these language-pairs: SMT experiments in [Ceauşu, 2008], [Ignat, 2009] or [Koehn et al., 2009] and EBMT in [Irimia, 2009]. The obtained results can be analyzed in report of what has been already published in the literature, including an investigation into how the MT approach and the use of additional linguistic resources influence the translation results. However, these experiments are not 100% comparable, as the systems do not use the same test data under the same (training) conditions.

As mentioned previously, in CBMT better results have been obtained with domain-specific corpora. Therefore, the third argument for using JRC-Acquis is that it is a domain-specific corpus. Nevertheless, it remains interesting as it contains a large vocabulary spectrum as EU laws and regulations cover different areas. Concerning this aspect, it refers to a wider domain than other corpora (with respect to vocabulary-size), but it is still domain specific (e.g. with respect to syntax). Evaluating the size and the number of language-pairs, at the time of starting this research, JRC-Acquis contained the largest amount of data. However, the size of bilingual subsets of JRC-Acquis differs significantly from language-pair to language-pair. Bilingual subsets in JRC-Acquis can have, at least for the languages we analyzed, even six times less aligned sentences when compared to the Europarl or the News Corpora used in recent investigations in the EuroMatrix project [Callison-Burch et al., 2009].

The fourth argument for choosing JRC-Acquis is that (sentence) alignments are available for all the language-combinations studied in this thesis: Romanian-German, German-Romanian, Romanian-English and English-Romanian.

4.3.2 Description

The Acquis Communautaire (**AC**) is the total body of European Union law applicable in the EU Member States. This collection of legislative texts contains texts written between the 1950s and today and it changes continuously. The texts are available in all official EU languages but Irish, i.e. in 22 languages. The Language Technology group of the European Commission's Joint Research Center and the Romanian Academy of Sciences[10] processed, aligned and encoded part of these texts and created the JRC-Acquis corpus, which is seen as "*an approximation of the Acquis Communautaire*".

The corpus consists of around 20 000 documents with an average of 47 million words per language[11]. It is XML encoded, following the Text Encoding Initiative Guidelines

[9] `http://nl.ijs/.si/ME/CD/docs/1984.html` - last accessed on January 21st, 2010.
[10] For Romanian and Bulgarian.
[11] Romanian is counted only with around 30 million words.

Language	No. texts	No. words (Text body)	No. words (Signatures)	No. words (Annexes)	Total no. words (Whole document)
German	23541	32059892	2542149	16327611	50929652
English	23545	34588383	3198766	17750761	55537910
Romanian (version 1)	6573	9186947	514296	11185842	20887085
Romanian (version 2)	19211	30832212	-	-	30832212

Table 4.2: JRC-Acquis statistics (Source: `http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/`) - (No = number).

Language pair	No. of documents	No. of links
German-Romanian	6558 docs	391972 links
German-English	23430 docs	1264043 links
English-Romanian	6557 docs	391334 links

Table 4.3: JRC-Acquis alignment statistics (docs=documents).

TEI P4[12] and contains two parts: the monolingual part (the legislative texts) and the bilingual part (231 language-pairs alignments). The alignment was created at paragraph level[13] using automatic tools, based on Vanilla[14] or HunAlign[15]. The paragraphs of the AC Corpus are usually short and contain one sentence or even only sub-sentential phrases, e.g. a noun-phrase. However, there are exceptions, when a paragraph means a complex or compound sentence or even several sentences separated by "." or ";".

From JRC-Acquis only the texts in Romanian, English and German are used in this dissertation. This sub-part of JRC-Acquis contains texts from 1958 until 2006 for German and English and until 2005 for Romanian, but there are years for which texts are missing completly. In the last version of the JRC-Acquis[16] a new Romanian corpus is integrated (see [Ceauşu, 2008]). As this version was not available at the time of conducting the experiments for this thesis and, not even today, sentence alignment information is available, we considered only the Romanian documents from the previous version (Version 2.2). Some statistics about the corpus and the alignments, for the languages studied, are given in Table 4.2 and Table 4.3.

The corpus is compiled from several documents. Each document is split into two parts: the TEI[17] header and the text itself. The header contains general information: a title, how many paragraphs the document has, the URL source, etc. The text is separated into three parts: the body, the signature and (sometimes) the annex: The body contains the

[12] `http://www.tei-c.org/Guidelines/P4/` - last accessed on June 27th, 2011.

[13] The HTML tag $< p >$ is used for the alignment.

[14] An implementation of the Gale & Church sentence alignment algorithm, 1993 [Gale and Church, 1993] - `http://nl.ijs/si/telri/Vanilla` - last accessed on June 27th, 2011.

[15] `http://mokk.bme,hu/resources/hunalign` - last accessed on June 27th, 2011, [Varga et al., 2005]

[16] Version 3.0, March 2009.

[17] `http://www.tei-c.org/index.xml` - last accessed on June 27th, 2011.

EU law; The signature includes the date, place and a list of person names and references to other documents; The annex is usually a plain text or a list of goods or addresses.

The document structure and its XML encoding is presented below (Source: `http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html` - last accessed on March 29th, 2010):

```
<TEI.2 id="jrcCELEX-LG" n="CELEX" lang="LG">
 <teiHeader lang="en" date.created="DATE">
  <fileDesc> ..... </fileDesc>
  <profileDesc> ..... </profileDesc>
 </teiHeader>
 <text>
  <body>
   <head n="1">Document Title</head>
   <div type="body">
    <p n="paragraph_number">... TEXT...</p>
     .....
   </div>
   <div type="signature">
    <p n="paragraph_number">... signature text...</p>
     .....
   </div>
   <div type="annex">
    <p n="paragraph_number">... annex text...</p>
     .....
   </div>
  </body>
 </text>
</TEI.2>
```

The initial alignment files contain the alignment for one language-pair. The data is also XML encoded (see below). The XML file contains a header including general information such as the file description and distribution rules. The alignment information is in the **text**-tag and is according to each document analyzed (the **div**-tag).

```
<TEI.2 id="jrc-en-ro" select="en ro">
 <teiHeader type="corpus" lang="en"
 date.created="2006-03-14" date.updated="13/07/2007">
  <fileDesc> .... </fileDesc>
  <encodingDesc> .... </encodingDesc>
  <profileDesc> .... </profileDesc>
  <revisionDesc> .... </revisionDesc>
 </teiHeader>
 <text select="en ro">
  <body>
   <div type="body" n="22002D0163" select="en ro">
    <p>19 parahraph links:</p>
    <linkGrp targType="head p" n="22002D0163" select="en ro" id="jrc22002D0163-en-ro"
```

```
     type="n-n" xtargets="jrc22002D0163-en;jrc22002D0163-ro">
     <link type="1:1" xtargets="2;2"/>
     ....
    </linkGrp>
   </div>
   ....
  </body>
 </text>
</TEI.2>
```

Before running the experiments for each MT system (see **Chapter 8.3**), we extracted the required format of the corpus and its alignments.

While running the experiments, part of the corpus was manually analyzed. Several sources of errors were found, such as wrong paragraph alignment (which leads to wrong translations in the corpus) and spelling errors. These types of errors were found both in the test and training data and might influence the output quality

More details on JRC-Acquis can be found in [Steinberger et al., 2006], [Ceauşu, 2008] and [Ignat, 2009].

4.4 RoGER

In order to analyze how the MT systems react to a smaller, but more accurate corpus, a second data source is used: RoGER[18]. We also included this corpus in experiments to test the behavior of the Moses-based SMT system trained during this research on out-of-domain data.

The RoGER corpus was compiled between 2005 and 2006 at the University of Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, in the Natural Language Systems Division [Gavrila and Elita, 2006]. It contains the specifications and the user's instructions for an electronic device. The motivation to create this corpus was that to the best of my knowledge, no multilingual resources, which satisfied our needs, were available at that time for the language-pairs considered.

4.4.1 Motivation

After employing the JRC-Acquis corpus for SMT and EBMT experiments, we used the RoGER corpus to analyze how the systems behave in the case of a smaller corpus. The small size could be compensated for by the correctness of the translations and the alignments provided in the corpus: the sentences have been manually aligned and checked.

Considering the motivation aspects mentioned in the **Subsection 4.3.1**, for the RoGER corpus it can be concluded that:

[18]RoGER = **Ro**manian - **G**erman - **E**nglish - **R**ussian.

- The language-pairs are the same as the ones selected from the JRC-Acquis corpus
- Sentence alignments have been provided;
- The corpus domain is even more restricted (e.g. vocabulary, syntactic structures) than in JRC-Acquis.

No previous experimental results with this corpus have been published in the literature. Also, to the best of my knowledge, no previous paper analyzes the behavior of SMT and EBMT systems using such a small corpus, for language-pairs employed in this thesis[19]. It is generally accepted that EBMT is better suited for small domains. Therefore, the usage of a small-size corpus is a setup which better fits the EBMT context.

4.4.2 Description

RoGER is a parallel corpus, aligned at sentence level. It is domain-restricted, as the texts are from a users' manual of an electronic device[20].

The languages included in the development of this corpus are Romanian, English, German and Russian. The corpus was manually compiled. It is not annotated and diacritics are ignored. The corpus was manually verified: the translations and the (sentence) alignments were manually corrected.

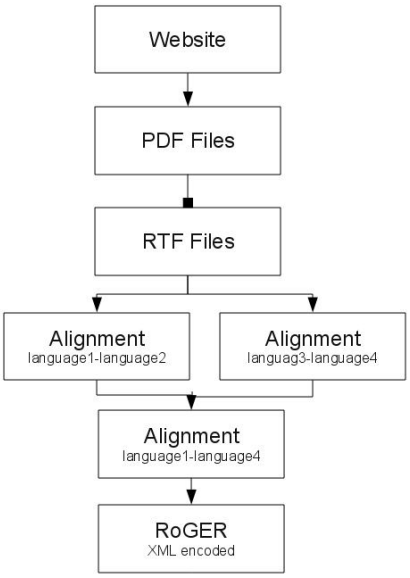

Figure 4.1: Building RoGER.

The initial PDF files of the manual were transformed into text (.RTF) files, where graphics and pictures were either left out (pictures around the text), or replaced with text

[19]Similar experiments, but for Serbian-English, are shown in [Popovic and Ney, 2006].

[20]As we could not obtain an official answer from the company that produces the electronic device, due to copyright conditions, the information about the electronic device (e.g. name, type, website) is left out.

Feature	English	Romanian	German	Russian
No. tokens	26096	25850	27142	22383
Vocabulary size	2012	3104	3031	3883
Vocabulary (*Word-frequency higher than two*)	1231	1575	1698	1904

Table 4.4: The RoGER corpus – Some statistics.

(pictures inside the text). The initial text was preprocessed by replacing numbers, websites and images with "*meta-notions*" as follows: numbers by NUM, pictures by PICT and websites by WWWSITE. In order to simplify the translation process, some abbreviations were expanded. The sentences were manually aligned, first for groups of two languages. Finally, the two alignment files obtained were merged, so that, after all, RoGER contained all four languages. The merged text files are XML encoded, as shown below:

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<sentences>
.................
  <sentence id=''1010''>
   <en>Press Options and some of the following options may be available .</en>
   <de>Druecken Sie Optionen . und einige der folgenden Optionen sind ggf.
       verfuegbar .</de>
   <ro>Apasati Optiuni dupa care unele din urmatoarele optiuni pot fi disponibile .</ro>
   <ru>...</ru>
  </sentence>
.......................
</sentences>
```

The corpus contains 2333 sentences for each language. More statistical data about the corpus is presented in Table 4.4. The average sentence length is eleven tokens for English, Romanian and German and nine for Russian. A tokens can be a lexical item, a punctuation sign or a number. More about the RoGER corpus can be found in [Gavrila and Elita, 2006]

4.5 Translation Challenges

In order to assess the validity of using such data for MT experiments, some parts of the two corpora has been analyzed from the point of view of the linguistic translation challenges they contain (see also **Chapter 2**). We did this analysis only for Romanian and English. Before showing the results, this section will present a small overview of the linguistic translation challenges.

Languages differ in the way they present the world. Therefore,

> "*translation must be sometimes a matter of approximating the meaning of a*

source language rather than finding an exact counterpart in the target language" [Kameyama et al., 1991].

The distinctions between SL and TL have been classified in [Barnett et al., 1991] and later in [Collins, 1998] into two categories: **translation divergences** and **translation mismatches**.

Translation divergence means that the same information appears in both SL and TL, but the structure of the sentence is different. Translation divergences are presented in the literature in [Dorr et al., 1999] and [Dorr, 1994]. In the case of a translation mismatch the information that can be extracted from the SL and TL sentence is not the same. Translation mismatches have received less attention in the literature (see [Kameyama et al., 1991]), but for CBMT approaches in general and for EBMT in particular, they are important, as they directly influence the translation process. Both challenges are described in Collins' work: "*Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*" [Collins, 1998].

The following subsections present the challenges for English and Romanian, found in both corpora. The divergences and mismatches presented are specific to the language-pair and to the corpus (domain), but the types of translation challenges are universal.

4.5.1 In JRC-Acquis

From JRC-Acquis a sample of 82 "paragraphs" has been extracted from the middle of the bilingual corpus for Romanian and English (texts from 1980) and manually analyzed. After splitting the paragraphs which contain more sentences separated by '.' or ';', the analyzed corpus consists of 110 sentences. The average sentence length is approximately 20 words. The sentences have been translated mostly in a one-to-one fashion, but translation divergences and mismatches (e.g. active voice translated with passive voice) were also encountered. The amount of divergence and mismatch was a factor to be investigated. Examples are given below. The following examples present the original texts in English and Romanian. These are followed by the direct translation into English of the original Romanian text. The translation is not necessarily correct.

(7) ***Examples of divergences from the corpus:***
CHANGES IN THE ARGUMENT STRUCTURE
article 3(1) shall be replaced by the following
la art. 3, se înlocuieşte alin. 1 cu următorul text
(* ENG *at Article 3, it is replaced the paragraph 1 by the following text*)

CATEGORY CHANGES
cameră ***cu uşă prevăzută cu broască***
lockable *room*
(* ENG *room* ***with door equipped with a lock***)

(8) ***Examples of translation mismatches:***
ANAPHORA
un spaţiu special amenajat pentru bălegar, dacă ***acesta*** *nu este evacuat imediat în mod igienic*
a specially prepared place for dung unless ***dung*** *is immediately and hygienically removed*
(In the English version no anaphora is encountered)
(* ENG *a specially prepared place for dung unless* ***this*** *is immediately and hygienically removed*)

IDIOMS, REPHRASING (REFORMULATIONS)
fără să aducă atingerea dispoziţiilor
notwithstanding
(* ENG *without modifying the dispositions*)
pentru
in the case of
(* ENG *for*)

The following specific phenomena were encountered while analyzing the 110 sentences:

- Divergences
  - Noun (**NN**) - adjective (**Adj**) inversion
  - Noun-Preposition-Noun (**NN-prep-NN**) translated as adjective-Noun (**Adj-NN**)
  - Subordinate clause translated as adjective
  - Different argument structure
  - Different type of articles
  - Voice change (for verbs)
- Mismatches
  - Extra information (the TL sentence is more explicit than the SL one)
  - Reformulations
- Wrong translation (due to incorrect alignment)

All these phenomena have a direct (negative) influence on the automatic evaluation scores, such as BLEU.

Although the corpus is domain restricted, the likelihood of at least one divergence or mismatch type occurring in a sentence is high. Only in approximately 10% of the sentences no phenomenon was encountered. Figure 4.2 shows the translation challenges found in JRC-Acquis in comparison with the ones extracted from RoGER.

Wrong or incomplete translations have a direct impact, first on the translation steps (in the specific case of EBMT on the alignment and recombination) and, in the end, on the output itself.

(9) **Example of incomplete human translation:**
RON: *c) mărfurile originare din spaţiul economic european (see) în sensul protocolului 4 la*

acordul see.
ENG: *c) goods originating in the european economic area (eea) within the meaning of protocol 4 to the agreement on the european economic area.*
(* ENG *the goods originating in the european economic area (eea) within the meaning of the protocol 4 to the eea agreement.*) – In the initial translation no abbreviation is used.

After running the word-alignment algorithm of the MT systems, "*on the economic area*" remains un-aligned and "*european*" is aligned to "*see*". This is not only due to the data size[21], but also to the human translation.

4.5.2 In RoGER

We analyzed 100 sentences from the middle of the RoGER corpus for English-Romanian. We noticed that the diversity of the challenges is reduced, while the number of challenges is sometimes higher compared to what had been encountered in JRC-Acquis, with up to five challenges in an example (a sentence and its translation).

Usually there is a one-to-one translation. Only in 12% of cases additional information appeared for one of the languages and in only 9% reformulations have been used. Two phenomena have been found most often: NN–prep–NN translated as NN–NN (or Adj–NN) and Adj–NN inversions.

Figure 4.2 shows the translation challenges encountered in both corpora.

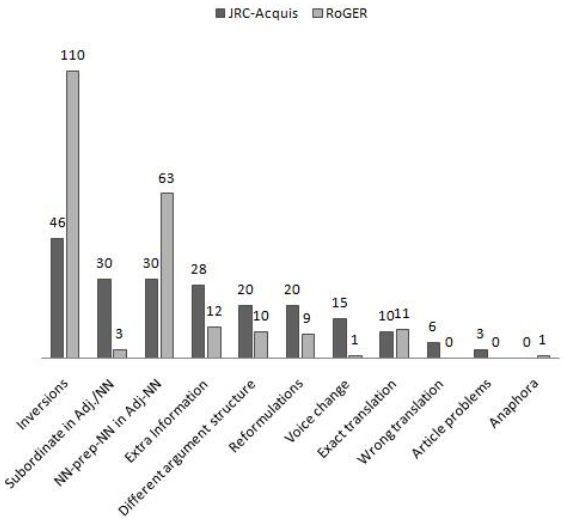

Figure 4.2: Translation challenges.

The average number of challenges in JRC-Acquis (1.89 challenges per sentences) is lower than the average number in RoGER (2.20 challenges per sentence) for the languages analyzed.[22] However, challenges with a more negative impact on the translation quality

[21] As GIZA++ is based on statistics, the corpus size has a direct influence on the word alignment.
[22] For the two corpora not always the same challenges are investigated.

(such as "Wrong translation", "Reformulations" etc.) appear more frequently in JRC-Acquis. The phenomenon encountered more often for the language-pair analyzed is noun-adjective inversions.

4.6 Chapter Summary

This chapter describes the motivation for choosing the corpora and the data used. Furthermore, the transfer challenges that have been found while manually analyzing part of the data are presented. A brief overview of the language characteristics for Romanian and German has been also provided at the beginning of this chapter. The information about the experimental settings – e.g. the amount of training and test data – will be described in **Chapter 8**.

Chapter 5

Overview of the Applications Used

In this chapter we will present the open-source applications (tools) that we employed for our MT experiments. The applications were used for creating the required word-alignment and language model of the EBMT system (e.g. SRILM, GIZA++), for adding linguistic information (e.g. RACAI Text Processing Web Services) or for running SMT experiments (e.g. Moses and Google Translate). The tools used for evaluation (BLEU/NIST and TER) will not be presented in this chapter, but will be described in **Chapter 8**. All the tools were used as black-box systems.

5.1 Moses

Moses (`http://www.statmt.org/moses/`[1]) is an SMT system that enables the user to automatically train translation models for a language pair, considering that the user has the required parallel aligned corpus. The development of Moses is mainly supported under the EuroMatrix[2], LetsMT[3] and EuroMatrixPlus[4] projects, funded by the European Commission under Framework Programme 6 and 7. It received additional support from the DARPA GALE and TC-Star projects and from several universities. The tool is licensed under the GNU Lesser General Public License (**LGPL**).

Among the features encountered in Moses, there are:

- phrase- and tree-based translation models,
- factored translation models, which allow the integration of linguistic and other information at the word level, and
- the decoding of confusion networks and word lattices, which enable easy integration with ambiguous upstream tools, such as automatic speech recognizers or morphological analyzers.

[1] Last time accessed on May 31st, 2010.
[2] `http://www.euromatrix.net/` - last accessed on June 27th, 2011.
[3] `http://www.letsmt.eu/` - last accessed on June 27th, 2011.
[4] `http://www.euromatrixplus.net/` - last accessed on June 27th, 2011.

More information about Moses can be found in [Koehn et al., 2007].

We employed Moses for developing an SMT system, using the corpora we have already presented in **Chapter 4**. Our Moses-based MT system follows the description and the parameter setting of the baseline architecture given for the EACL 2011 Sixth Workshop on SMT[5]. The exact parameters and training and testing steps can be found on the website of the workshop: `http://www.statmt.org/wmt11/baseline.html`[6].

We trained a phrase-based model that benefits from advanced features of the decoder, such as lexicalized reordering models. In the training we used SRILM for generating the language model and GIZA++ for the alignment (see **Subsection 5.3** and **5.4**, respectively).

In most of our SMT experiments, two changes were made to the system specification given at the Workshop on SMT:

- The tuning step was excluded.
- The language model (**LM**) order we considered is three as on the Moses website specification (`http://www.statmt.org/moses_steps.html`[7]), although in the specification of the workshop on SMT it was given as five. A reason for choosing the order three were the results shown in the presentation of the SMART[8] project [Rousu, 2008], in which it was stated that "*3-grams work generally the best*".

We used the original specification[9] of the Workshop on SMT only in one experimental setting for Romanian-German[10] and the JRC-Acquis corpus. The extent to which this setting influences the MT results will be shown in **Chapter 8**.

As the results of the original specification of the SMT system were not always better than the ones of our system setting (see **Chapter 8** for the results), we eliminated the tuning step. Another argument for this decision is that "*MERT[, the tuning approach in Moses,] can also be a relatively unstable training method, with different runs producing models of significantly different model quality*" [Cer, 2002, p. 102].

The initial XML encoded corpus files and the alignments were adapted to fit the description of the input files in Moses.

Before building the translation model (**TM**), the training data was preprocessed. After tokenizing the sentences, they were filtered out according to a sentence length criterion (the '*cleaning*' step)[11] and lowercased. The scripts for preprocessing the data are available on the website of the workshop. In the same way, the data for the LM was tokenized

[5]EACL 2011 Workshop on SMT: `http://www.statmt.org/wmt11/index.html` - last accessed on June 27th, 2011.

[6]Last accessed on May, 10th, 2011.

[7]Last accessed on June 27th, 2011.

[8]`www.smart-project.eu` - last accessed on June 27th, 2011.

[9]A tuning process based on Minimum Error Rate Training (**MERT**, [Och, 2003]) was included and the language model order was five. No clear description of the tuning data is provided on the Moses website.

[10]We considered only one direction of translation: Romanian-German.

[11]The maximum sentence length accepted was forty words, as suggested at the Sixth Workshop on SMT. The sentence length limit can be increased to 100 words. This is the maximum limit accepted by GIZA++.

and lowercased. The language model was built with SRILM, using the parameters recommended at the Workshop: "*interpolate* and *kndiscount*". The "*kndiscount*" uses Chen and Goodman [1996]'s modified Kneser-Ney discounting for n-grams of order n. The "*interpolate*" parameter causes the discounted n-gram probability estimates at the specified order n to be interpolated with lower-order estimates.

To train the TM, we ran the provided training script. We had as input the bilingual corpus in two text files: one for the SL, the other for the TL. Each line in the SL file has a corresponding line in the TL file. For the alignment we used the default heuristics given by the value "*grow-diag-final-and*" of the parameter "*-alignment*". It starts with the intersection of the two alignments and then adds additional alignment points. As previously mentioned, a reordering model for the decoder was used. By default, only a distance-based reordering model is included in the final configuration. Additional conditional reordering models may be built and they are conditioned on specified factors (in the source and target language). These learn different reordering probabilities for each phrase pair (or just the foreign phrase). The possible configurations can be found in the Moses manual [Koehn, 2010, p. 118]. We used a "*msd-bidirectional-fe*" model, which considers three different orientation types: **m**onotone, **s**wap and **d**iscontinuous. It is conditioned on both the SL and TL phrase ("*fe*"). The system considers the ordering of one phrase with respect to the previous one. Using the bidirectional model, also the ordering of the next phrase with respect to the current one is modeled.

For the experiments which used the original specification of the Moses system, the tuning data was also tokenized and lowercased. The tuning script was provided by Moses. At the end of the tuning step the new weights were inserted into the configuration file.

In order to run the system on the test sets, the test data was also tokenized[12] and lowercased. The data was decoded after filtering the model in order to fit into the memory[13].

Before evaluating the results, the output was transformed to fit the format of the reference translation[14] and the scoring tools[15].

5.2 Google Translate

For comparison reasons, we included another MT system in our experiments: Google Translate, an on-line MT system.

Google Translate (`http://translate.google.com`[16]) is a free statistically-based machine translation service, provided by Google Inc. to translate a section of text, document

[12] Only for JRC-Acquis.

[13] See the training and testing steps on `http://www.statmt.org/wmt11/baseline.html` - last accessed in May 2011.

[14] The initial output was recased and detokenized. Detokenization was done only for the JRC-Acquis corpus.

[15] Wrapping the transformed output in SGML (for evaluating with NIST/BLEU) or numbering the sentences (for the evaluation with TER).

[16] Last accessed on June 27th, 2011.

or webpage, from one source language into the target language. While Google Translate is classified as an SMT system on `Wikipedia.org`, on the Google support web page[17] it is stated only that it uses the "*state-of-the-art technology*" without reference to any specific MT approach.

The service was introduced, as it is known today, in 2007. Prior to 2007, a Systran[18]-based translator was used. Google Translate is based on the research conducted by Franz-Josef Och[19] [Och, 2005]. An exact description of its translation mechanism and data (corpora type and size) is, to the best of my knowledge, not publicly available.

Google Translate has been continuously developed. At the moment of writing[20], it was in the 20th stage of development[21] and included 57 languages. We translated our test data sets with the Google system at the beginning of our experiments and have not repeated the experiments.

5.3 The SRILM Toolkit

The SRI Language Modeling toolkit (**SRILM**) has been under development in the SRI Speech Technology and Research Laboratory since 1995 [Stolcke, 2002]. This thesis uses the version 1.5.7 of the SRILM toolkit for creating the LM of the Moses-based MT system and extracting the necessary information for the recombination step of the implemented EBMT systems. We also used it to extract statistics from the corpus (see **Chapter 8**).

SRILM is a collection of C++ libraries, executable programs and helper scripts, which supports the creation and evaluation of a variety of language model types based on n-gram statistics, as well as several related tasks, such as statistical tagging and manipulation of n-best lists and word lattices. It runs on the UNIX and Windows platforms. It is additionally applied in several fields, such as speech recognition, machine translation, tagging and segmentation and document processing.

The SRILM toolkit is freely available under an open source community license and can be downloaded from `http://www-speech.sri.com/projects/srilm/` (last accessed on June 27th, 2011). It is currently used in the research community for tasks requiring statistical language modeling It is also integrated or used in different NLP systems, such as Moses (see **Section 5.1**), Systran and MorphTagger[22].

[17] `http://translate.google.com/support/?hl=en` - last accessed on June 27th, 2011.
[18] `http://www.systran.co.uk/` - last accessed on June 29th, 2011.
[19] `http://research.google.com/pubs/och.html` - last accessed on August 19th, 2010.
[20] August 2010.
[21] The 20th stage was launched in June 2010.
[22] `http://www.cs.technion.ac.il/~barhaim/MorphTagger/` - last accessed on June 29th, 2011.

5.4 GIZA++

GIZA++ was developed by Franz Josef Och [Och and Ney, 2003] and is an extension of the program GIZA, which was part of the SMT toolkit EGYPT[23]. It can be used to train the IBM Models 1-5 [Brown et al., 1993] and an HMM word alignment model [Vogel et al., 1996]. The package also contains the source for the *mkcls* tool which generates the word classes necessary for training some of the alignment models.

GIZA++ can be freely used under the terms of GNU Public License (**GPL**) version 2 and is available on `http://code.google.com/p/giza-pp/`[24]. It is known to compile on Linux, Irix and SUNOS systems.

The version we used in this thesis is **1.0.2**. We needed GIZA++ to run the Moses-based SMT system and to obtain the word-alignments in the EBMT system(s) (see the system description in **Chapter 6**).

5.5 Text Processing Web Services

A collection of linguistic web services for Romanian and English is available on the website of the Research Institute for Artificial Intelligence of the Romanian Academy (RACAI) - `http://www.racai.ro/webservices/TextProcessing.aspx`[25]. It provides on-line web services for text processing [such as tokenization (ENG/RON), sentence splitting (ENG/RON), POS Tagging (ENG/RON) and lemmatization (ENG/RON)], factored translation and language identification. As described in [Tufiş et al., 2008a], the POS tagging is carried out with the TTL tool[26], a text preprocessing module developed in Perl. The tool, its components end evaluation[27] are presented in [Ion, 2007].

An output of the text processing tool is shown below:

(1) **Example of the output for the text processing web service:**
Input:
ENG: *do not end the call until told to do so .*
RON: *nu incheiati convorbirea pana nu vi se cere acest lucru in mod expres .*
Output:
ENG: *do|do|AUX2|Vaip2s not|not|NOT|Qz end|end|VINF|Vmn the|the|DM|Dd call|call|NN|Ncns until|until|CSUB|Cs told|tell|PAST|Vmis to|to|TO|Qn do|do|VINF|Vmn so|so|ADVE|Rmp .|.|PERIOD|PERIOD*
RON: *nu|nu|QZ|Qz incheiati|incheiati|V2|Vmip2s convorbirea|convorbire|NSRY|Ncfsry pana|pan ca|NSRY|Ncfsry nu|nu|QZ|Qz vi|tu|PPPD|Pp2-pd———w se|sine|PXA|Px3–a——*

[23] `http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/` - last accessed on June 27th, 2011.
[24] Last accessed on June 27th, 2011.
[25] Last accessed on June 27th, 2011.
[26] `http://www.clarin.eu/tools/ttl-tokenizing-tagging-and-lemmatizing-free-running-texts` - last accessed on June 27th, 2011.
[27] A precision between 96%-98% was reported [Ion, 2007, p. 19].

–w cere|cere|V3|Vmip3s acest|acest|DMSR|Dd3msr—e lucru|lucru|NSN|Ncms-n in|in|NSN|Ncms-n mod|mod|NSN|Ncms-n expres|Expres|NP|Np .|.|PERIOD|PERIOD

The text processing tool provides information comprised of the word, the lemma, the C-TAG ("*the Corpus Tag*") tag and the Morpho-Syntactic Descriptor (**MSD**)[28] tag. The C-TAG is a superset of the MSD tags.

For our experiments we used the web-service for text processing in order to extract POS information. We tested how POS information influences the translations results, for the RoGER corpus. The information used in the experiments is composed from the word and the C-TAG[29] tag.

[28] `http://nl.ijs.si/ME/V2/msd/` - last accessed on June 27th, 2011.

[29] The first tag after the lemma presented in Example (1).

Chapter 6

Lin-EBMT: a New EBMT System

This chapter will describe *Lin-EBMT*, the EBMT baseline system developed during this research. *Lin-EBMT* is a linear EBMT system, based on surface-forms, which uses as linguistic resources only the parallel aligned bilingual corpus. The approach developed is language independent, taking into account that necessary data (e.g. parallel corpus, alignment information) and tools are available for the language-pair under consideration. As it is implemented in Java 1.6[1], the system is platform-independent.

The need to develop an EBMT system was motivated by the fact that no EBMT systems were publicly available at the beginning of this research. That is why no EBMT system could be at that time used or further developed and no comparisons with other EBMT systems are presented in this dissertation[2]. Since Nagao's work [Nagao, 1984], several EBMT systems have been developed, but no (open source) resources were available until the end of 2009. After the Third Workshop on EBMT[3] in November 2009, open source EBMT (or hybrid) systems appeared, e.g. CMU EBMT System[4], OpenMaTrEx[5].

6.1 The System

Lin-EBMT is a linear EBMT system according to the system classification found in **Chapter 3**. It is based on surface-forms and uses no additional linguistic resources. The motivation to use no linguistic resources in addition to the parallel aligned bilingual corpus

[1] `http://www.oracle.com/technetwork/java/index.html` - last accessed on June 21st, 2011.

[2] A comparative study of empirical MT, using five empirical MT systems (the Moses-based SMT system and the two EBMT systems presented in this dissertation, one hybrid (EBMT-SMT) system (OpenMatrex) and Google Translate) is presented in an article submitted for publication in the first half of 2011: [**?**].

[3] The first two workshops on EBMT took place in 2001 and 2005.

[4] `http://sourceforge.net/projects/cmu-ebmt/` - last accessed on June 27th, 2011; the system was made available at the end of 2009.

[5] `http://www.openmatrex.org/` - last accessed on June 27th, 2011; the system was publicly available at the beginning of 2010. It is a open-source marker-driven EBMT system, which comprises two engines: one based on Marclator (`http://www.openmatrex.org/marclator/` - last accessed on June 27th, 2011), another on Moses.

is given by the fact that one of the languages involved, Romanian, is considered under-resourced[6]. The approach is language independent, provided that the necessary data (e.g. the parallel corpus) and tools (e.g. the tool for word alignment) are available.

6.1.1 Data Preparation

Lin-EBMT uses the same (preprocessed) training and test data as the Moses-based SMT system described in **Section 5.1**: the translation model data is used in the matching process and the language model data in extracting the information required in the recombination step. This way both corpus-based MT systems are based on the same (training and test) data. Therefore, we have a one-to-one comparison between the systems.

The training data was encoded in an XML file to fit the requirements of the EBMT system. The encoding is similar to the one found in RoGER. An example for an XML encoded sentence for English-Romanian is shown below[7].

```
<?xml version=''1.0'' encoding=''UTF-8''?>
<sentences>
.................
  <sentence id=''1010''>
   <en>Press Options and some of the following options
       may be available .</en>
   <ro>Apasati Optiuni dupa care unele din urmatoarele
       optiuni pot fi disponibile .</ro>
  </sentence>
.......................
</sentences>
```

While encoding the JRC-Acquis corpus in an XML file, modifications had to be made in the text in order to avoid formatting errors. For instance, the signs which represent the predefined entity references in XML[8], such as '&' or '<', had to be changed. More changes were involved in the transformation of the Romanian text. We encountered paragraph alignment errors[9] while making these modifications.

6.1.2 System Architecture

Before describing the main EBMT steps in more details, the system architecture will briefly be presented. Prior to the translation process, the training and test data are preprocessed as in the Moses-based SMT system and the files required for the translation, such as the word-index and the GIZA++ word alignments, are extracted.

[6]For more information on why is Romanian a lower resourced language, please see **Chapter 1**.

[7]The example does not contain diacritics, because the corpus does not include them.

[8]More on XML syntax can be found on `http://www.w3schools.com/xml/xml_syntax.asp` - last accessed on June 21st, 2011.

[9]Errors introduced by the Vanilla sentence aligner.

The Index

A word index for the SL data is used to reduce the search space in the matching step of the EBMT system. This approach is also found in other research papers, such as [Sumita and Iida, 1991] and [Smith and Clark, 2009][10].

In our approach, the word-index can be considered in fact a token[11] index, as it also contains punctuation signs and numbers. The information in the index is a pair of the form:

$$(Token, List\_of\_sentence\_ids).$$

The key in the index is represented by the token. The information attached is a list of ids of the SL sentences in the corpus which contain the token *Token*. The index is alphabetically sorted according to the key and is implemented as a "*Properties*" Java object, which has automatic procedures for searching and editing a value or for saving the information in an XML format. An excerpt from the index extracted from RoGER (SL English), for the tokens "*great*", "*exclusive*", "*non-modified*" and "*equipped*", is given below:

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
<comment>Index file for en</comment>
<entry key="great">2118</entry>
....
<entry key="exclusive">2189 2189</entry>
<entry key="non-modified">2187 2188</entry>
<entry key="equipped">2118</entry>
....
</properties>
```

Using the index, the search space is significantly reduced. For example, the search space for matching in the RoGER corpus is reduced down to 45.61%, 46.28% and 47.10% of the initial data size, when the SL is German, English and Romanian, respectively. For the JRC-Acquis corpus the index reduced the space down to approximately 20% - 35% of the initial data size depending on the source language. For reducing the matching search space even further, constraints can be used together with the index, such as considering only the sentence ids of the content words and ignoring, for example, prepositions and articles.

The Architecture

The main steps in *Lin-EBMT* are summarized below.

For each of the input sentences in the test data:

[10] [Smith and Clark, 2009] indexed every n-gram of length 1 to 5.

[11] A token can be a lexical item, a number, a punctuation sign, etc.

1. The tokens in the input are extracted: $\{token_1, token_2, ..., token_n\}$.
2. Using the token index, all sentence ids $\{sentenceId_1, ..., sentenceId_m\}$ that contain at least one token from the input are considered. The punctuation signs are ignored and the list of sentence ids contains no duplicates. The matching procedure is run only after the search space size is decreased. We obtain the 'reduced' corpus in this way.
3. Given the input sentence and the list of sentence ids $\{sentenceId_1, ..., sentenceId_m\}$, the matching procedure between the input and sentences in the 'reduced' SL-side of the corpus is run. If the input sentence is encountered in the corpus, the translation is found and the translation procedure is stopped. Else, the most similar sentences are extracted by using the similarity measure described in **Section 6.2.1** followed by the alignment and recombination steps.
4. Having the matched sentences which maximally cover the input, the corresponding alignments are extracted (see **Section 6.2.2**).
5. The output is generated using the "*bag of TL sequences*" obtained from the alignment (see **Section 6.2.3**).

These steps are graphically presented in Figure 6.1.

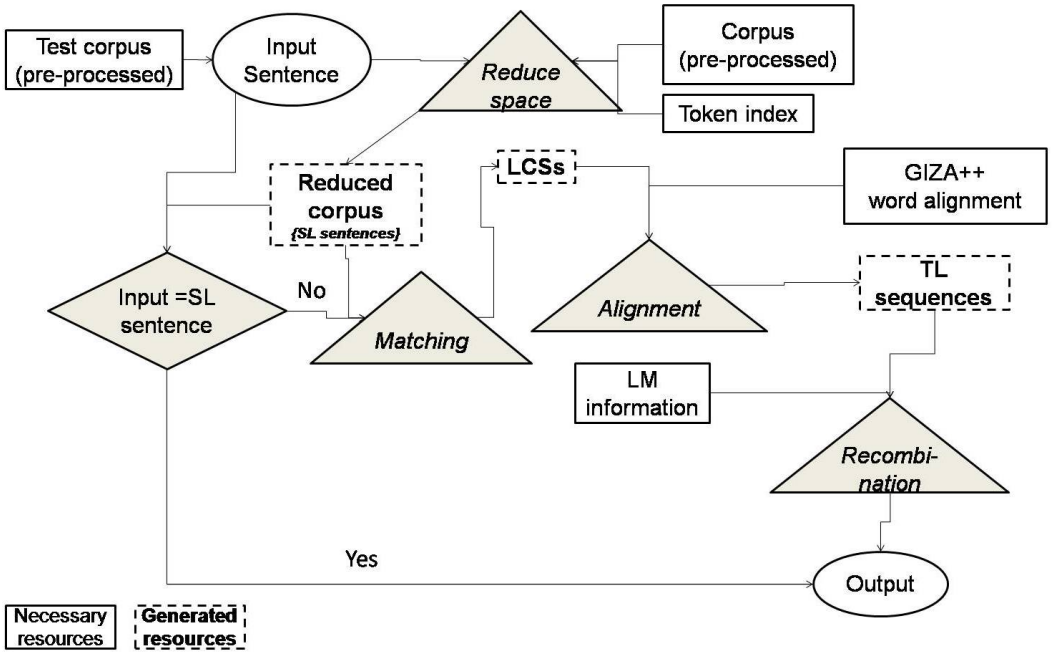

Figure 6.1: The *Lin-EBMT* system.

For the evaluation, the obtained translations had to be post-processed and formatted to fit the input requirements of the evaluation tools.

6.2 The EBMT Steps

The following subsections will present the main EBMT translation steps – matching, alignment and recombination –, in the case when the input sentence is not found in the corpus. The data, its quality and sparseness have a direct influence on the translation results in general and on each of the EBMT steps in particular.

6.2.1 Matching the Input

In the matching step, the input sentence is compared with the sentences extracted from the corpus after using the index. The algorithm tries to match the (whole) input with an entry in the corpus. In cases where this is not possible, it tries to match parts of the input with (parts of) the sentences in the corpus. In this first implementation, punctuation signs and out-of-vocabulary words are left aside before starting the matching algorithm[12].

The matching algorithm is recursive and follows the steps enumerated below:

1. Find the sentence in the corpus that matches the input best, using the similarity measure described in Formula 6.1. Keep this as part of the solution.
2. If the input is not fully covered, eliminate what has been already found and for the rest of the input return to step 1. Else stop the matching procedure: the result is found.

Algorithm 6.1 The matching algorithm.

Require: the input I, the ids of the sentences which have at least one token in common with the input (punctuation excluded): $\{sentenceId_1, ..., sentenceId_m\}$.

Ensure: the set of the M matched sentences together with the corresponding longest common subsequences: $Result = \{(sentenceId_i, LCS_i)\}$, $1 \leq i \leq M$, where $sentenceId_i$ is the id of the matched sentence S_i and LCS_i is the longest common subsequence between I and S_i.

$Result = \emptyset$
$copyI \leftarrow I$
$IDS \leftarrow \{sentenceId_1, ..., sentenceId_m\}$
while $copyI \neq \epsilon$ **do**
  {ϵ is the empty string.}
  $(id, LCS) \leftarrow findBestMatch(copyI, IDS)$
  $Result \leftarrow Result \cup \{(id, LCS)\}$
  $copyI \leftarrow remove(copyI, LCS)$
  $IDS \leftarrow IDS \setminus \{id\}$
end while

Algorithm 6.1 presents the matching procedure. The function $findBestMatch(copyI, IDS)$ (where $copyI$ is iniatlized with I - the input sentence) finds the sentence which best covers

[12]This approach might have a negative impact on the automatic evaluation results.

copyI and has as output the pair (id, LCS), where id is the id in the corpus of the matched sentence and LCS the longest common subsequence between *copyI* and the matched sentence. The function $remove(copyI, LCS)$ removes from *copyI* what has been already matched (the longest common subsequence). The variable *copyI* is initialized with I, the input sentence.

The matching procedure is a string-based approach, focusing on finding common substrings. By using the longest common subsequence (**LCS**), we hope to have a smaller number of elements to be recombined. The procedure is based on the Longest Common Subsequence Similarity (**LCSS**) measure we implemented during this research. The implementation uses a dynamic programming algorithm, similar to the one found in [Bergroth et al., 2000]. The initial LCS character-based algorithm is transformed into a token-based one, in which punctuation is ignored when comparing the strings. Although not fully relevant, since the training and test data is lowercased, the algorithm is implemented as case insensitive. Given two strings[13] - $s1$ and $s2$ - the LCSS measure is calculated as

$$LCSS(s1, s2) = LCSS_T(s1, s2) - P * noWords, \tag{6.1}$$

where

$$LCSS_T(s1, s2) = \frac{Length(LCS(s1, s2))}{Length(s1)}, \tag{6.2}$$

where

- $LCS(s1, s2)$ is the LCS between $s1$ and $s2$,
- $Length(s)$ is the number of tokens of a string s, and
- $noWords$ is the number of word-gaps found while comparing $LCS(s1, s2)$ to $s1$.

The formula introduces a penalty of $P = 0.01$ for each word-gap found. Word-gaps are tokens which included in $LCS(s1, s2)$ would create a continuous subsequence in $s1$. For example for the sentence $s1$ = "*Saving names and phone numbers (Add name)*" and the $LCS(s1, S2)$="*names and numbers*" the word-gap is "*phone*" and $noWords = 1$. To get the sequence $s2$ which best covers $s1$, it is first calculated a maximum value for $LCSS_T(s1, s2)$ and after a maximum for $LCSS(s1, s2)$.

This way we try to split the input sentence into a minimum number of sequences, so that the number of boundary friction problems is reduced. The choice of the sentence which provides the maximum value is not influenced by changing the value of the penalty P[14]: $LCSS(s1, s2)$ changes, but the maximum value is encountered for the same (matched) sentence.

For example, for the sentences

```
Input s1 = ‘‘Saving names and phone numbers ( Add name )’’
Sentence in the corpus s2 = ‘‘Erasing names and numbers’’
```

the longest common subsequence $LCS(s1, s2)$ is "*names and numbers*" and the value of similarity measure $LCSS(s1, s2)$ is $LCSS(s1, s2) = \frac{3}{7} - 0.01 * 1 = 0.4185$[15].

[13] A string can be seen as a sequence of tokens.

[14] As long as $P > 0$.

[15] The sentence $s1$ has 7 tokens, as punctuation (i.e. ‘(’, ‘)’) is ignored.

The matching procedure has as input the input sentence and the database of examples and gives as output the sentences that cover best the input. The output for this translation step is obtained recursively and choosing only one $LCSS(s1, s2)$ maximum value for each iteration of the algorithm. The process chooses the sentence which covers the input, with the least number of word gaps. This way it increases the chance to have a minimum number of sequences that should be recombined in order to form the output. Having less sequences as input for the recombination step, it should decrease the appearance of the boundary friction[16] problem.

$LCSS_T(s1, s2)$ has the value in the interval $[0, 1]$. 0 would indicate that the sentences are completely different. However, in this specific experimental setting, as the token index is used and the sentence ids are chosen so that at least one word appears in both $s1$ and $s2$, this situation is not possible. A value of 1 for $LCSS_T(s1, s2)$ shows that $s1$ and $s2$ are identical, a case in which also $LCSS(s1, s2)$ is 1.

Similarity measures in EBMT can be used for matching the input on an example database or for extracting similar sentence pairs as in template-based EBMT. As the goal in this thesis is to use this similarity measure to match the input on examples in a database, the similarity measure is not symmetric. On the basis of the use of similarity measures for extracting similar sentence pairs, a comparison a between symmetric version of LCSS and other similarity measures is presented in [Elita et al., 2007].

Similarity measures could introduce errors in the translation process. Concerning LCSS[17], errors might be introduced due to polysemous words, verbs with separable particle (especially for German), etc.

An example regarding the influence of the verbs with separable particle is the following: given the German input "*Ich sehe aus dem Fenster*" (ENG: "*I look out the window.*", RON: "*Privesc pe fereastră.*") and the sentences in the German-Romanian corpus:

1. "*Ich sehe gut aus.*" $\leftrightarrow$ "*Arăt bine.*"(ENG: "*I look good.*")

2. "*Ich sehe fern*" $\leftrightarrow$ "*Privesc la televizor.*" (ENG:*I watch TV.*)

the matching algorithm will choose as best match the first sentence, as the LCS is "*Ich sehe aus.*". The translation into Romanian for "*Ich sehe aus.*" would be "*Arăt*", which is semantically wrong considering the input sentence. No specific solutions have been implemented for these kind of problems. Additional linguistic information could improve the matching based on LCSS.

[16]For the definition of boundary friction please see **Chapter 2**.

[17]Not only LCSS introduces these kind of errors. This is a common problem for similarity metrics with no additional linguistic information.

6.2.2 Alignment

The required word alignment information is extracted at run-time[18] from the GIZA++[19] output obtained while running the Moses-based SMT system. From the two generated '*A3.final*' files, only the target-source language direction file is consulted for the implementation in this thesis. The '*A3.final*' file contains the final word-to-word alignment for each of the words in each line (in the same order as the input parallel aligned corpus). In our case a line represents a paragraph for JRC-Acquis and a sentence for RoGER.

The alignment procedure considers as input the matched sentences together with the corresponding LCSs (the output of the matching procedure). From the GIZA++ alignment information, the longest (possible) target language aligned subsequences are chosen for the recombination step: Let these sequences be $\{sequence_1, sequence_2, ..., sequence_N\}$.

Given two aligned SL and TL sentences we have for the SL matched words $w_{i_1}^{SL}, ...w_{i_k}^{SL}..., w_{i_n}^{SL}$ (the LCS) the aligned TL words $w_{j_1}^{TL}, ...w_{j_p}^{TL}..., w_{j_m}^{TL}$, where i_k $(1 \leq k \leq n)$ and j_p $(1 \leq p \leq m)$ represent the position of the respective words in the SL and TL sentence, respectively. The sequences do not necessary need to be continuous. A longest TL aligned subsequence represents a word sequence of the following form:

$$sequence = w_{j_r}^{TL} w_{j_r+1}^{TL} ... w_{j_s-1}^{TL} w_{j_s}^{TL} \tag{6.3}$$

where the words $w_{j_q}^{TL}$, for $r \leq q \leq s$ are all one after another in the TL sentence (continuous TL subsequences). The length of the word sequence *sequence* is $L = j_s - (j_r - 1)$, with $L \geq 1$. For the recombination step not all tokens $w_{j_p}^{TL}$ are considered, but all (possible) longest TL subsequences. For example, given the extracted LCS

(1) "*technical regulations standards*",

and the alignments:

(2) "*technical* ↔ *tehnice*" (position 8 in TL), "*regulations* ↔ *reglementările*" (position 7 in TL) and "*standards* ↔ *standarde*" (position 23 in TL),

the following word sequences are used in the recombination step:

(3) "*reglementările tehnice*" and "*standarde*".

As "*reglementările*" and "*tehnice*" follow one after another in the TL sentence, the sequence "*reglementările tehnice*" is further used in the recombination step.

Some examples from the GIZA++ '*A3.final*' files, for Romanian and English, in both directions, are given in the Examples (4) and (5).

[18]The GIZA++ information is already available when starting the translation process, but the required alignment information is computed at run time.

[19]For details about GIZA++, please see **Section 5.4**.

(4) English-Romanian:
Sentence pair (1) source length 2 target length 3 alignment score : 0.00255938
user 's guide
NULL ({ }) ghidul ({ 2 3 }) utilizatorului ({ 1 })
...
Sentence pair (43) source length 3 target length 4 alignment score : 0.00196744
changing the front cover
NULL ({ }) schimbarea ({ 1 }) capacului ({ 2 4 }) frontal ({ 3 })
...

(5) Romanian-English
Sentence pair (1) source length 3 target length 2 alignment score : 0.00222542
ghidul utilizatorului
NULL ({ }) user ({ 2 }) 's ({ }) guide ({ 1 })
...
Sentence pair (43) source length 4 target length 3 alignment score : 0.00223566
schimbarea capacului frontal
NULL ({ }) changing ({ 1 }) the ({ }) front ({ 3 }) cover ({ 2 })
...

In Examples (4) and (5) the indexes in curly brackets ('{', '}') represent the positions of the words in the sentences of the other language.

The choice of only the target-source language direction '*A3.final*' file is motivated by the need to avoid possible conflicts between the matching result and the alignment information, as alignment is not always consistent. For example, in the second sentence in Examples (4) and (5), the word "*capacului*" is aligned to "*the cover*" or to "*cover*". A more complex alignment algorithm, considering both files, could improve the alignment results. However, such an algorithm needs more information than only the one obtained from the sentences provided as output by the matching step. For example it could use word alignment information extracted from the whole training data. Such an approach is encountered in the word-alignment algorithm from the Moses-based SMT system and in [Smith and Clark, 2009].

Problems in the translation appear if, for example, for the matched sentences[20] there are cases when no GIZA++ word alignment is provided. Such cases can be avoided if constraints on the word alignment would be already verified in the matching step.

6.2.3 Recombination and Output Generation

Recombination, the last step of the EBMT system, has as input the "*the bag of word sequences*" $\{sequence_1, sequence_2, ..., sequence_N\}$ provided by the alignment step and as result the translation. A **word sequence** is represented by a ('longest') TL subsequence,

[20]The output of the matching step.

in which the TL tokens[21] $w_1 w_2 ... w_M$ appear one after another in the corresponding TL sentence. As they are extracted from a TL sentence in this specific order $w_1 w_2 ... w_M$, we consider that the order of these tokens is correct.

Possible errors can be introduced by the matching and alignment steps, such as a wrong matching of a polysemous word in the SL, rendering the wrong inflected form or no information available for the word alignment. Translations of some input words might be lost if no matching or alignment information is found. As these errors are introduced by previous steps they cannot be solved only by the recombination. The main challenge for the recombination step consists in finding the right word-sequence order by means of the information provided by matching and alignment.

The recombination algorithm is based on the monolingual distribution of bi-grams and on a "*recombination matrix*" $A_{N,N}$

$$A_{N,N} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{pmatrix},$$

which is defined as in the Definition 6.1.

Definition 6.1. *If the outcome of the alignment is N word-sequences {sequence$_1$, sequence$_2$, ..., sequence$_N$}, with sequence$_i$ = $w_{i_1} w_{i_2} ... w_{i_{last}}$ $(1 \leq i \leq N)$), and these word-sequences are not necessarily different, then A is a square matrix of order N that is defined as follows:*

$$A_{N,N} = (a_{i,j})_{1 \leq i,j \leq N} = \begin{cases} -3, & \text{if } i = j; \\ -2, & \text{if } i \neq j, \\ & count(w_{i_{last}} w_{j_1}) = 0; \\ \frac{2 * count(w_{i_{last}} w_{j_1})}{count(w_{i_{last}}) + count(w_{j_1})}, & \text{if } i \neq j, \\ & count(w_{i_{last}} w_{j_1}) \neq 0. \end{cases} \tag{6.4}$$

where count(s) represents the number of appearances of a token s in the corpus.

The bi-grams are formed from the last word of the sequence *sequence$_i$* – $w_{i_{last}}$ – and the first word of *sequence$_j$* – w_{j_1} $(1 \leq i, j \leq N)$. The value for the case "$i \neq j$, $count(w_{i_{last}} w_{j_1}) \neq 0$" (the sequence $w_{i_{last}} w_{j_1}$ is found into the corpus) is computed using the Dice coefficient [Dice, 1945], which returns a real value between 0 and 1. The results of the function *count*(*s*), where *s* represents one or two tokens, are obtained using SRILM. (For more details on SRILM, see **Chapter 5**, **Section 5.3**).

The value for the case $i = j$ is the lowest in recombination matrix to minimize the probability for repeating a sequence (i.e. *sequence$_i$* follows *sequence$_i$*). The value for the case when the sequence $w_{i_{last}} w_{j_1}$ is not found into the corpus ($count(w_{i_{last}} w_{j_1}) = 0$, $i \neq j$)

[21] In most of the cases a token is a word.

is lower than the minimum value of the Dice coefficient (i.e. it is a negative value), but higher than the one for the case $i = j$. Not finding the data in the corpus, it does not necessarily mean that the sequence is not valid in the target language. This could happen also due to data sparseness. The two values -2 and -3 could be changed in the matrix and other values which follow the rule below can be chosen:

$$\text{value for the case } i = j < \text{value for the case } count(w_{i_{last}} w_{j_1}) = 0 \ (i \neq j) < \text{value for the case } count(w_{i_{last}} w_{j_1}) \neq 0 \ (i \neq j)$$

The idea of representing the information in a matrix was initially motivated by the "*similarity matrix*" found in the sentence alignment algorithm presented in [Kit et al., 2002], which has already been presented in **Chapter 3**.

The recombination algorithm is based on finding the maximum value $a_{i,j}$ $(1 \leq i, j \leq N)$, '***combining***' $sequence_i$ and $sequence_j$, and deleting all the values from the matrix corresponding to $sequence_j$ – line and column j. When $sequence_i$ and $sequence_j$ **are combined**, they are concatenated and the values for the new element $sequence_i sequence_j$ are updated to the values in the matrix corresponding to $sequence_j$; $sequence_i sequence_j$ is replacing the position for the $sequence_i$. If the number of TL sequences provided by the alignment is greater than one, the recombination algorithm is repeated until the order of the matrix is one and the output has been obtained. With every repetition of the algorithm the order of the matrix is reduced by one. The maximum value for $a_{i,j}$ means that the probability that $sequence_j$ follows $sequence_i$ is the highest, given a certain corpus. This happens as the probability that w_{j_1} follows $w_{i_{last}}$ is the highest.

The information for building the recombination matrix is provided by 1- and 2-gram distributions extracted with the SRILM toolkit. The necessary values – the Dice coefficients – are calculated and, in order to reach this information fast, they are saved in the implementation as a '*Properties*' Java object and encoded in XML.

Without changing the recombination algorithm, the values in the recombination matrix can be changed. For example, instead of the information based on the Dice coefficient, another language model can be used. Results of such experiments will be described in **Chapter 9**, **Subsection 9.1.1**.

As the recombination is based on n-gram frequency information, data sparseness has a direct influence on the results.

6.3 Chapter Summary

This chapter described *Lin-EBMT*, the implemented EBMT baseline system. Detailed information on the experimental settings, the data used and the evaluation of the system will be presented in **Chapters 8** and **9**. A manual analysis of the results will be provided in **Chapter 10**.

While building the translation, the system uses the information found in the examples

extracted by the matching step. The auxiliary files used (e.g. GIZA++ files, the n-gram information) are built taking onto account the whole training data, before the translation process starts.

As mentioned in **Section 6.2.3**, the main challenge of the recombination step implemented as in *Lin-EBMT* is finding the right word-order. For the time being, only the information from the LM based on the Dice coefficient has been used. The information which can be extracted from the sentences produced as output by the matching step is ('*partly*')[22] left aside.

In the next chapter – **Chapter 7** – the recombination step will be extended by including also information from matched examples; $Lin - EBMT^{REC+}$, an extended version of *Lin-EBMT*, will be presented. The motivation for including information from matched examples will be presented in **Section 7.1**.

[22]We use the term '*partly*', as this information is directly involved in the alignment step, which has a direct influence on the recombination step.

Chapter 7

$Lin - EBMT^{REC+}$: *Lin-EBMT* Extended

$Lin - EBMT$, described in the previous chapter, is a linear EBMT system, which in the recombination step makes no use of the information directly extracted by the matching step. It employs only the output of the alignment, i.e. the '*bag of TL word sequences*'. From these word sequences, the output is formed by processing only the 2-gram information in the recombination matrix. This way, the information provided by the matching (the SL sentences and their translations) is lost, although helpful data for deciding the word order in the recombination step is still present. An example is presented in **Section 7.1**, where such data improves the output.

This chapter will describe $Lin-EBMT^{REC+}$, an extended version of the $Lin-EBMT$. In this extended system, implementation ideas from the template-based EBMT approach are incorporated into the recombination step. The previous two steps (matching and alignment) remain almost unchanged. The differences which appear between the two matching algorithms is that punctuation and out-of-vocabulary words (**OOV-words**) are also integrated in the output of $Lin - EBMT^{REC+}$. In this thesis the values in the recombination matrix of $Lin - EBMT$ are constrained by information extracted from templates.

Constraints in natural language processing play an important role, for example in constraint-based grammars[1]. Constraints restrict the possible values that a variable (or a feature) may take with respect to certain rules. In MT constraints have been used before: for example, in SMT influences of constraints are presented in [Canisius and van den Bosch, 2009] and [Cao and Sumita, 2010].

Before defining the templates and describing the approach, some remarks on terminology need to be made. From the terminology presented in **Chapter 3**, the terms
i) **template** for a generalized example and

[1]Information on constraint-based grammars can be found in [Pollard, 1996].

ii) **SL and TL side** for the SL and TL parts of a template (respectively) will be used further in this work.

7.1 Motivation

Lin-EBMT is a linear EBMT system, which in the recombination step makes use only of the word sequences provided by the alignment. The output is formed by employing only the 2-gram information extracted from the corpus and the recombination matrix. From the matching step, not only the SL sentences which best cover the input are obtained, but also the corresponding TL sentences. These TL sentences which contain the translations of the words or word-sequences in the input can provide important recombination information, such as word order constraints on the output. If only the recombination matrix is used, as in *Lin-EBMT*, or only an LM as in other linear systems, the information from the corresponding TL sentences is lost.

An example of how the translation output changes when constraints are used is shown below:

(1) Given the Romanian input
"*puteti memora imaginile si sunetele pentru a va personaliza telefonul .*"
(**The reference translation:** "*you can save the pictures and sounds for personalising your phone .*")
the following translations are provided by the two EBMT systems implemented in this dissertation:
The *Lin-EBMT* output:
"*phone and to you can save images polyphonic ringing personalize*"
The $Lin - EBMT^{REC+}$ output, when constraints are taken into account[2]:
"*you can save the phone . to personalize and tones images*"

Although the output of the $Lin - EBMT^{REC+}$ system is not entirely correct, the word order is better than the one in the output from *Lin-EBMT*.

In order to avoid the loss of word-order information, the recombination step can use ideas from template-based EBMT systems (see **Subsection 3.2.2**) by employing constraints extracted from templates. The SL-sides of the templates are built from the input sentence and the matched sentences. The corresponding TL sides are obtained by using word alignment information. Word-order information can be extracted from the TL sides. This information imposes specific constraints on the output formation. These constraints modify the values in the recombination matrix defined in the *Lin-EBMT* system.

As previously mentioned, $Lin - EBMT^{REC+}$ is the system which extends *Lin-EBMT* by making use of constraints and word-order information in the recombination step.

[2]For details on what constraints are, please see **Section 7.4**.

7.2 Template Definition

Before describing the template extraction algorithm and the use of the templates in the recombination step, this section will provide the definition of a template.

Our template follows partly the definition for "*translation patterns*" found in [McTait, 2001] and presented in **Section 3.2.2** (**Chapter 3**), but it considers only one general alignment information[3].

Before defining a template, the terms "*text fragment*" and "*variable*" need to be explained.

Definition 7.1. *A **text fragment** tf is a continuous series of one or more tokens, where a token can be a lexical item[4], a punctuation mark, a number etc.*

Definition 7.2. *A **variable** v is a placeholder for a text fragment.*

Definition 7.3. *Given TF the set of text fragments ($TF = \{e|e \text{ is a text fragment}\}$) and VAR the set of variables ($VAR = \{e|e \text{ is a variable}\}$), for which we have $TF \cap VAR = \emptyset$, and given $e_1, e_2 \in TF \cup VAR$, we define the operator $\oplus$ as a concatenation operator: $e_1 \oplus e_2 = e_1 e_2$.*

For a set X we use the notation $|X|$ for the cardinality of the set X (the number of elements of the set X). For SL and TL text fragment and variables, we define S and T as:

Definition 7.4. *$S = \oplus_{i=1}^{n} e_{SLi}$, where $n = |TF_{SL}| + |VAR_{SL}|$, $e_{SLi} \in TF_{SL} \cup VAR_{SL}$. TF_{SL} represents the set of SL text fragments and VAR_{SL} the set of SL variables.*

Definition 7.5. *$T = \oplus_{i=1}^{m} e_{TLi}$, where $m = |TF_{TL}| + |VAR_{TL}|$, $e_{TLi} \in TF_{TL} \cup VAR_{TL}$. TF_{TL} represents the set of TL text fragments and VAR_{TL} the set of TL variables.*

Definition 7.6. *Given A_v (the set of alignments for the variables[5]) and A_{tf} (the set of alignments for the text fragments[6]), where both A_v and A_{tf} include the cases when no alignment information[7] is available, we define the set of all alignments A_{all} as the union of A_v and A_{tf}: $A_{all} = A_v \cup A_{tf}$.*

Consequently, we can define a template as follows:

Definition 7.7. *A **template** is a triple $\{S, T, A_{all}\}$, in which S and T represent SL and TL text fragments separated by SL and TL variables, respectively (see Definition 7.4 and 7.5) and A_{all} is the alignment information as defined in Definition 7.6.*

[3] In [McTait, 2001] the alignment in a template was separated in alignment of the variables and alignment of the text fragments.

[4] A lexical item is a single word or chain of words which are the basic elements of a lexicon of one language.

[5] $A_v = \{e^{SL} \leftrightarrow e^{TL} | e^{SL} \in VAR_{SL} \cup \{NOALIGN\}, e^{TL} \in VAR_{TL} \cup \{NOALIGN\}\}$.

[6] $A_{tf} = \{e^{SL} \leftrightarrow e^{TL} | e^{SL} \in TF_{SL} \cup \{NOALIGN\}, e^{TL} \in TF_{TL} \cup \{NOALIGN\}\}$.

[7] The $NOALIGN$ value.

We call all text fragments and variables, which appear in a template, the **elements** of the template.

Corresponding aligned text fragments and variables are associated with a unique alignment number, which is the same on both sides of the template. When no alignments between SL and TL have been found, the variables, which find themselves in such cases, have the specific format **NOALIGN**. Due to the extraction algorithm (see **Section 7.3**), it is also possible to have SL text fragments with no correspondence in the TL. In this case, no text fragment on the TL side is marked with the same number as the one attached to the SL text fragment.

Another required definition is the one of the **operation of reduction**:

Definition 7.8. *Given a set of variables $v_i, v_{i+1}, ..., v_j$, with $j > i$: if these variables appear on both sides of a template as a sequence in the same order $v_i v_{i+1} ... v_j$, they are* ***reduced*** *in the template to one variable v_{i_j}. That means that the sequence of variables $v_i, v_{i+1}, ..., v_j$ is replaced by v_{i_j} on on both sides of the template.*

The **reduction** operation is applied for variables which appear in both SL and TL one after another, in the same order. The order is given by the alignment number. For example, if in both the SL and TL sides the sequence *VAR9 VAR10 VAR11* appears, this is reduced to *VAR9_11*.

No constraints (such as number of variables on a side, etc.) are imposed on the templates as sometimes encountered in previous works (see **Chapter 3**). Also, several variables can follow one after another, as not all the successive initially found variables can be '*reduced*'. The operation of reduction cannot always be done as the aligned variables might be split by text fragments or other variables. This happens due to m-to-n alignments or to inversions.

Following the syntax for regular expressions, a template can be expressed as follows:

$$((TF_{SL})^*(VAR_{SL})^*)^* TF_{SL}((TF_{SL})^*(VAR_{SL})^*)^* \leftrightarrow ((TF_{TL})^*(VAR_{TL})^*)^* \tag{7.1}$$

where

- TF_{SL} is an SL text fragment, VAR_{SL} an SL variable,
- TF_{TL} is a TL text fragment, VAR_{TL} a TL variable, and
- * is the Kleene operator[8].

It follows from Formula 7.1 that at least one text fragment should be present on the SL side. This constraint is not set for the TL side, since, due to the matching and alignment steps, it is possible that the only TF_{SL} in the SL side has no corresponding TF_{TL} on the TL side. This means that no word alignment information is available.

To better understand the definition a template, an example of a template extracted during the translation process is presented in Example (2). The character sequence "*&&*" is used in the template representation for delimiting the corresponding alignment number of a text fragment.

[8]Regular expressions rpresents the context in which $*$ was introduced by Stephen Kleene (1909-1994) to characterize certain automata and it means "zero or more".

(2) VAR1 more&&2&& VAR3 VAR4 NOALIGN5 VAR6_8 VAR9 VAR10 VAR11 VAR12 VAR13 VAR14 VAR15 ↔ VAR1 mai&&2&& multe&&2&& VAR4 VAR3 NOALIGN0 VAR6_8 VAR11 VAR9 VAR11 VAR10 VAR12 VAR14 VAR13 VAR15

In this example, the template elements are:

- Aligned variables - $VARnumber$: *VAR1* in SL and TL;
- Reduced variables - $VARnumber\_number$: *VAR6_8* in SL and TL;
- Not aligned variables - $NOALIGNnumber$: *NOALIGN5* in SL and *NOALIGN0* in TL;
- Aligned text fragments: *more&&2&&* in SL and *mai&&2&& multe&&2&&* in TL (*more* ↔ *mai multe*).

As has already been established, when elements of the template have the same type (e.g. text fragment, variable) and the same number attached on both the SL and TL sides, it means that these elements are aligned.

7.3 The Template Extraction Algorithm

During the translation process, the template extraction algorithm is applied to each test sentence in the test data set after the alignment step of *Lin-EBMT*. It is a run-time step in the translation process. It has as input the sentence to be translated, the matched sentences and their translations, the longest common subsequence (**LCS**) and the extracted GIZA++ alignments. A template is extracted for each matched sentence: function $getTemplate(LCS(I, S_i), S_i, T_i, A_{all_i}, I)$ in Algorithm 7.1. This template is afterwards reduced: function $reduce(TE_i^{'})$ in Algorithm 7.1. The output of the whole algorithm is the set of all reduced templates, which is further used to generate constraints in the recombination step.

Algorithm 7.1 The extraction algorithm for all templates.

Require: the input I, the number of matched sentences n, the matched sentences S_i and their translations T_i, the longest common subsequences $LCS(I, S_i)$ and the extracted alignments A_{all_i}, with $1 \leq i \leq n$.

Ensure: the set of reduced templates $TE = \{TE_i\}, 1 \leq i \leq n$

$TE \leftarrow \emptyset$
for $i \leftarrow 1$ **to** n **do**
  $TE_i^{'} \leftarrow getTemplate(LCS(I, S_i), S_i, T_i, A_{all_i}, I)$
  $TE_i \leftarrow reduce(TE_i^{'})$
  $TE \leftarrow TE \cup \{TE_i\}$
end for

The template extraction algorithm has two phases: a monolingual and a bilingual phase. The algorithm considers punctuation. The monolingual phase is realized only for the

source language, in contrast to other template extraction algorithms presented in the literature (see [McTait, 2001]). Before starting the monolingual phase, we extend the word alignment information for the matched sentences, so that each aligned sequence is marked either as a text fragment ($TEXT$) or as a variable (VAR). This extension is realized using the information provided by the longest common subsequence and the GIZA++ word-alignment.

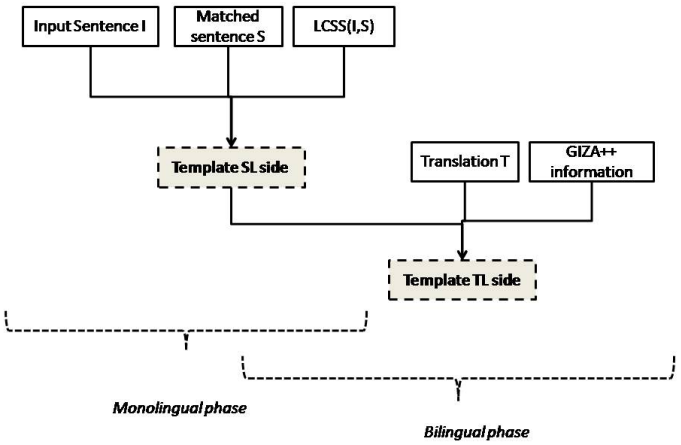

Figure 7.1: The template extraction algorithm.

Following the algorithm presented in Figure 7.1, a template TE can be extracted for the input sentence I and each matched sentence S. The information needed is the matched sentence S with its corresponding translation T, the longest common subsequence $LCS(I, S)$ and the extended alignment information A_{all}.

A template has two elements: the SL and the TL side. In other words a template can be seen as a pair $TE = (SLside, TLside)$.

The Monolingual Phase:
The monolingual phase of the algorithm has as output the SL side of the template ($SLside$). The common tokens between I and the SL matched sentence S (i.e. $LCS(I, S)$) are considered as text fragments on the SL side of the template. All the other tokens from S are represented by variables.

The Bilingual Phase:
Given the SL side of the template, the translation T and the extended alignment information A_{all}, the TL side of the template ($TLside$) can be obtained. The TL sequences aligned to the SL text fragments represent the TL text fragments. The remaining TL tokens are considered variables. The alignments between the SL and TL variables and text fragments are provided by the GIZA++ alignment. In case no alignment information has been encountered for some variables, these variables are of the type $NOALIGNnumber$. The variables from TL, which are not aligned, are also of the type $NOALIGNnumber$. If

no corresponding alignment number is found on the TL side, then SL text fragments are not aligned.

The extracted template is afterwards reduced (see Definition 7.8): if both template sides contain the same sequences of variables, these sequences are reduced to only variable on both template sides of the template. For example, given the input is "*press and hold clear to delete the characters more quickly .*" and the LCS "*to delete the characters .*", we suppose that the corresponding GIZA++ alignment for a matched sentence is:
pentru$_1$ **a**$_2$ **sterge**$_3$ simultan$_4$ toate$_5$ **caracterele**$_6$ cand$_7$ scrieti$_8$ un$_9$ mesaj$_{10}$,$_{11}$ apasati$_{12}$ optiuni$_{13}$ si$_{14}$ selectati$_{15}$ stergeti$_{16}$ textul$_{17}$.$_{18}$
NULL ({ }) **to** ({ 1 2 }) **delete** ({ 3 }) all ({ 5 }) **the** ({ }) **characters** ({ 6 }) at ({ 4 }) once ({ }) when ({ 7 }) writing ({ 8 }) a ({ 9 }) message ({ 10 }) , ({ 11 }) press ({ 12 }) options ({ 13 }) and ({ 14 }) select ({ 15 }) clear ({ 16 }) text ({ 17 }) **.** ({ 18 })
The template extracted is:
to&&1&& delete&&2&& VAR3 the&&4&& characters&&5&& VAR6 NOALIGN7 VAR8_18 .&&19&&
↔ pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8_18 .&&19&&

This illustrates that, as on both the SL and TL sides the variables from $VAR8$ to $VAR18$ occur one after another, they are reduced to $VAR8\_18$. The word sequences which are going to be employed in the recombination step and are extracted from this template, are the text fragments on the TL side "*pentru a sterge*" (ENG: "*to delete*"), "*caracterele*" (ENG: "*the characters*") and ".".

The template extraction algorithm is presented in Algorithm 7.2.

7.4 Extended Recombination Step

The recombination step in $Lin - EBMT^{REC+}$ builds the output almost in the same way as in $Lin - EBMT$. There are, however, differences in the values of the recombination matrix and in the way the maximum value is searched for.

From the extracted templates, word-order[9] rules are determined and a set C of constraints ($C = \{(w_i, w_j)\}$) is built. C contains no duplications. A constraint (w_i, w_j) imposes that the words w_i and w_j cannot appear one after another in the output as the sequence $w_i w_j$. Therefore, the value in the recombination matrix corresponding to the entry $w_i w_j$ is set to a negative value (i.e. **-2**), so that the possibility of choosing this combination as a maximum is reduced.

To the best of our knowledge, the inclusion of word-order constraints in a recombination step based on n-gram information has not been proposed in the literature.

We consider three types of constraints in this dissertation: First-Word-Constraints (C.1 constraints), TLSide-Template-Constraints (C.2 constraints) and Whole-Template-Constraints (C.3 constraints). The main motivation was to choose language-independent constraints. In this thesis we do not choose all possible constraints.

[9] In some cases a word is in this context a token (lexical item, punctuation sign, number).

Algorithm 7.2 The template extraction algorithm.

Require: $LCS(I, S)$, S,T, A_{all}, I.

Ensure: a template TE

```
Declare SLside and TLside as arrays of Strings
Initialize SLside and TLside with "NOALIGN"
counter ← 0
for i = 1 to A_all.size do
    Get the word alignment wa = A_all_i
    if wa.Type ==' TEXT' then
        SLside_i ← wa.getInfoSL() +' &&' + counter +' &&'
        Get the TL side information corresponding to SLside_i:
        TLside_i ← wa.getInfoTL() +' &&' + counter +' &&'
        {&& are characters for delimitation.}
        counter ← counter + 1
    else if wa.Type ==' VAR' then
        if there is no TL alignment then
            SLside_i ←' NOALIGN' + counter
        else
            SLside_i ←' VAR' + counter
            Get the TL side information corresponding to SLside_i:
            TLside_i ←' VAR' + counter
        end if
        counter = counter + 1
    else
        SLside_i ← ϵ
        {ϵ is the empty string.}
        Get the TL side information corresponding to SLside[i]:
        TLside_i ←' NOALIGN' + counter
        counter ← counter + 1
    end if
end for
return TE(SLside, TLside)
```

The First-Word-Constraint (C.1)

A First-Word-Constraint (**C.1**) refers to the first word of the output.

Constraint definition. *If a word $w_{SL_{first}}$ is the first word of the input and of the SL side of a template and in this specific template this word $w_{SL_{first}}$ is aligned to the first word on the TL side $w_{TL_{first}}$, then $w_{TL_{first}}$ is the first word of the output and no other words or word-sequences can precede it. Therefore, for all TL words w provided by the alignment, the constraint $(w, w_{TL_{first}})$ is added to the set of constraints C.*

As we consider for building the output not each TL word, but all longest TL sequences, not all the **C.1** constraints are used in further steps. Therefore, the algorithm for extracting C.1 constraints can be optimized in further work.

Below there is an example of a C.1 constraint:

(3) Given the input sentence: ***to*** *delete the characters more quickly press and hold clear.* and the template:
to&&1&& delete&&2&& VAR3 the&&4&& characters&&5&& VAR6 NOALIGN7 VAR8 18 .&&19&& ↔ pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8 18 .&&19&&
We have the word alignment for the first word: '**to**' ↔ '*pentru a*'
According to the **C.1** constraint, the first word of the output is '**pentru**' and no other words can precede it. This means that we build constraints of the form *(X,pentru)* (where X is a word which will be part of the output), such as **(a, pentru)**, **(sterge, pentru)**, **(caracterele, pentru)**.

TLSide-Template-Constraint (C.2)

TLSide-Template-Constraints (**C.2**) are deduced only from the TL side of each of the templates extracted by the algorithm presented in **Section 7.3**:

Constraint definition. *If on the TL side of a template the words w_1 and w_2 appear in the sequence $w_1[...]w_2$, then the sequence w_2w_1 is not allowed in the output formation. Therefore, the constraint (w_2, w_1) is added to the set of constraints C.*

(4) Given the TL side of a template
pentru&&1&& a&&1&& sterge&&2&& VAR6 VAR3 caracterele&&5&& VAR8 18 .&&19&&,
we can form **C.2** constraints, such as **(caracterele, sterge)** or **(sterge, a)**.

Drawbacks of using this type of constraints can appear if some words are repeated in the output. In this case the constraint could be too strong.

Whole-Template-Constraint (C.3)

The algorithm for extracting Whole-Template-Constraints (**C.3**) uses each of the templates, together with the input sentence and the alignment information. The align-

ment refers to the corresponding TL tokens (or token-sequences) of the tokens (or token-sequences) in the input.

Given the input sentence $I = \{t_{SL_1}...t_{SL_n}\}$ and the alignment information $t_{SL_i} \leftrightarrow t_{TL_i}$, where $1 \leq i \leq n$, we establish the following rules:

1. If t_{SL_i} is not aligned on the TL side, then t_{TL_i} has a generic value "*NOALIGN*": $t_{TL_i} \leftrightarrow NOALIGN$. This value is ignored in further steps;
2. If t_{SL_i} is an out-of-vocabulary word, then it is aligned to itself: That is $t_{TL_i} \leftrightarrow t_{SL_i}$.

t_{SL_i} and t_{TL_i} $(1 \leq i \leq n)$ represent tokens or sequences of tokens in the SL and TL, respectively.

We consider the template $TE = (SLside, TLside)$ and the text fragment t_{SL_k} in $SLside$ which is aligned to the TL text fragment t_{TL_k} (t_{TL_k} is on $TLside$): $t_{SL_k} \leftrightarrow t_{TL_k}$, $(1 \leq k \leq n)$.

Constraint definition. *If t_{SL_k} is preceded by the 'same' variables or text fragments as t_{TL_k}, then the TL aligned sequences $t_{TL_q}...t_{TL_r}$ corresponding to the SL sequences $t_{SL_s}...t_{SL_t}$ $(1 \leq s, t, q, r < k)$ appear in the output also before the t_{TL_k}. Therefore, constraints of the form (t_{TL_k}, t_{TL_j}), $1 \leq j \leq (k-1)$, are added to the set of constraints C.*

In the context of the **C.3** constraints, the '*same*' means that the variables and text fragments have the same alignment number. For clarity, it is also needed to be explained that $t_{SL_s}...t_{SL_t}$ are in the input before t_{SL_k}. The SL and the TL aligned sequences do not necessary have the indexes in the same order (see Example (5)). Therefore, we used different notations for indexes: s,t for the SL and q,r for the TL $(1 \leq s, t, q, r < k)$.

(5) Given the input "*press the button for five minutes*",
the alignments: '*press*'↔'*apasa*'; '*the button*'↔'*butonul*', '*for*'↔'*pentru*', '*five*'↔'*cinci*' and '*minutes*'↔'*minutes*',
and the template:
VAR 1 VAR 2 for&&3&& VAR4 .... ↔ VAR 2 VAR 1 pentru&&3&& VAR4 ....
We have in the input before the word '*for*' the words '*press*'↔'*apasa*' and '*the button*'↔'*butonul*'.
Following the previous definition, we can form **C.3** constraints, such as **(pentru, apasa)** or **(pentru, butonul)**.

For the recombination step, we define below the "**constrained recombination matrix**", which is an extended version of the recombination matrix from *Lin-EBMT* (see Definition 6.1):

Definition 7.9. *Given N word sequences $\{sequence_1, sequence_2, ..., sequence_N\}$ the outcome of the alignment, with $sequence_i = w_{i_1}w_{i_2}...w_{i_{last}}$ $(1 \leq i \leq N)$ which are not*

necessarily different, and a set of constraints $C = \{(w_i, w_j)\}$*, with* $1 \leq i, j \leq N$*, then* $A_{N,N}$ *is a square matrix of order* N *that is defined as follows:*

$$A_{N,N} = (a_{i,j})_{1 \leq i,j \leq N} = \begin{cases} -3, & \text{if } i = j; \\ -2, & \text{if } i \neq j,\ count(w_{i_{last}} w_{j_1}) = 0 \text{ or} \\ & (w_{i_{last}} w_{j_1}) \in C; \\ \frac{2 * count(w_{i_{last}} w_{j_1})}{count(w_{i_{last}}) + count(w_{j_1})}, & \text{if } i \neq j,\ count(w_{i_{last}} w_{j_1}) > 0, \\ & (w_{i_{last}} w_{j_1}) \notin C. \end{cases} \tag{7.2}$$

where $count(s)$ *represents the number of appearances of a token* s *in the corpus. We name the matrix* A *the "constrained recombination matrix".*

The bi-grams are formed from the last word of the sequence $sequence_i$ ($w_{i_{last}}$) and the first word of $sequence_j$ (w_{j_1}) with $1 \leq i, j \leq N$. The value for the case "$i \neq j$, and $count(w_{i_{last}} w_{j_1}) > 0$ and $(w_{i_{last}} w_{j_1}) \notin C$" is computed using the Dice coefficient and returns a real value between 0 and 1.

As in *Lin-EBMT*, the recombination algorithm of $Lin-EBMT^{REC+}$ is based on finding the maximum value $a_{i,j}$ ($1 \leq i, j \leq N$) in the constrained recombination matrix. If First-Word-Constraints cannot be applied, the algorithm follows the same steps as in *Lin-EBMT*: if the maximum value is $a_{i,j}$, $sequence_i$ and $sequence_j$ are combined and all values from the matrix corresponding to $sequence_j$ are deleted (i.e. line and column j). When $sequence_i$ and $sequence_j$ are combined, they are concatenated and the values for the new element $sequence_i sequence_j$[10] are updated to the matrix values corresponding to $sequence_j$. With every repetition of the algorithm the order of the matrix is reduced by one. In case the number of sequences given by the alignment is bigger than one, the recombination algorithm is repeated until the order of the matrix is one and the output is obtained.

The algorithm for finding the maximum value is different if First-Word-Constraints can be applied. In this case, given that the first word is w_{FIRST} in $sequence_p$ ($1 \leq p \leq N$), the first maximum value in the matrix is searched for in the row p as $a_{p,i}$ ($1 \leq i \leq N$). The algorithm continues by searching for the maximum value on the row corresponding to the previous word (sequence) found and incorporated in the output. This means that a maximum value is searched for in a specific row, and not, as in the case when no First-Word-Constraints are applied, in the whole matrix.

[10] Which replaces $sequence_i$.

For some experiments, the definition of the matrix has been changed as follows:

$$A_{N,N} = (a_{i,j})_{1 \leq i,j \leq N} = \begin{cases} -3, & \text{if } i = j; \\ -1, & \text{if } i \neq j,\ count(w_{i_{last}} w_{j_1}) = 0; \\ -2, & \text{if } i \neq j,\ (w_{i_{last}} w_{j_1}) \in C; \\ \frac{2 * count(w_{i_{last}} w_{j_1})}{count(w_{i_{last}}) + count(w_{j_1})}, & \text{if } i \neq j,\ count(w_{i_{last}} w_{j_1}) > 0, \\ & (w_{i_{last}} w_{j_1}) \notin C. \end{cases} \tag{7.3}$$

In this definition we make a distinction between the case when there is no entry in the language model ($count(w_{i_{last}} w_{j_1}) = 0,\ i \neq j$) and the case when constraints are set ($(w_{i_{last}} w_{j_1}) \in C,\ i \neq j$). For the case when no language model (**LM**) entry is found we set the value higher than the value for the case when a constraint is set. Finding no entry in the LM does not necessarily mean that the words are not allowed to appear in this order; It could just mean that the data is sparse. Setting a constraint on two words means that the words are not allowed to appear in that specific order. This is why the value in the matrix is lower than in the previous case.

In the experiment runs we will test the influence of each of the constraints and combinations of constraints on the output: see **Subsection 9.1.1, Chapter 9**.

7.5 System Architecture

Figure 7.2 presents the architecture of $Lin - EBMT^{REC+}$. Comparing it with the architecture of *Lin-EBMT* (Figure 6.1), the processes for creating the templates and the constraints are added. Two new generated resources are created: the templates and the constraints.

The recombination step differs in the two EBMT systems implemented. In $Lin - EBMT^{REC+}$ recombination follows the steps:

1. For each matched sentence, extract the template;
2. For each template, extract all possible constraints; and
3. Build the constrained recombination matrix and obtain the output.

We expect an improvement in the translation quality by including additional word-order information in the recombination. However, the changes in the recombination matrix could have a seldom impact on the results. Therefore, overall, it could be noticed (only) a small improvement in the evaluation scores. The rare influence on the results appears due to the corpus and the fact that only one best solution is considered in the recombination. Data might be sparse and in the matrix a lot of the values could be the same (for example -2). Therefore, this might not bring a change in the searching process for the maximum value. It could happen that the values which have been modified due to the constraints are not reached while searching for the maximum in the matrix.

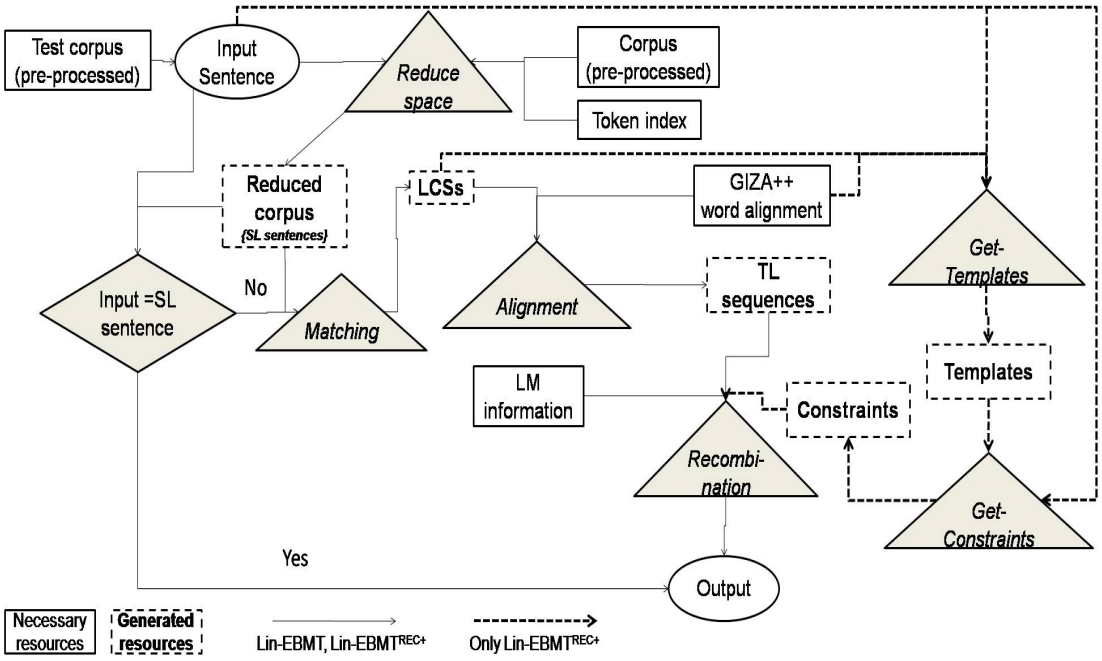

Figure 7.2: The $Lin - EBMT^{REC+}$ system.

7.6 Chapter Summary

This chapter presented $Lin - EBMT^{REC+}$, an extension of *Lin-EBMT*. This extended EBMT system combines ideas from linear EBMT systems and template-based ones. The main difference is in the recombination step. The other two EBMT steps – matching and alignment – remain almost unchanged: the implementation of $Lin - EBMT^{REC+}$ integrates punctuation and out-of-vocabulary words in the output.

The algorithm for template extraction and the method for generating the constraints are based only on surface forms. Additionally, the way constraints are included in the recombination step is language independent. For these reasons $Lin - EBMT^{REC+}$ can also be considered language independent.

The evaluation results for this system and how each type of constraints influences the translation results, when working with all language-pairs included in the thesis, will be shown in **Chapter 9**.

Chapter 8

Evaluation and Experimental Data

> "The success of a machine translation system can be measured according to two criteria: coverage and correctness." [Cicekli and Guvenir, 2001]
>
> "A good [correct] translation is one which conforms to the rules and idiom of the target language, while at the same time preserving - as much as possible - the meaning of the original". F. V. Eynde in "*Linguistic Issues in Machine Translations*" [Eynde, 1993]

MT evaluation has been accompanying MT research since the end of the 1950s[1]. It has been applied in several fields, such as system optimization, error analysis and system comparison. The complexity of MT evaluation is manifested in factors such as MT expressiveness and ambiguity, evaluation scope or availability of (other) tools.

Several evaluation methods have been proposed. Either manual or automatic, almost all use a predefined test data[2]. Each method introduces a significant bias in the evaluation process, as the process depends on how representative the test data is and on the number and quality of the reference translations.

After a short introduction on MT evaluation and a brief presentation of the metrics used, we will describe the experimental settings, the test and training data. The automatic evaluation results and their interpretation will be shown in **Chapter 9**. A manual analysis of a subset of the results will be presented in **Chapter 10**.

8.1 MT Evaluation

Over the last few years there has been a wide interest in automatic evaluation methods, as, in comparison to a manual evaluation, they provide quick results and do not need as much time, money and man-power.

Still, there are drawbacks when comparing automatic MT evaluation with a manual evaluation, such as a way of interpreting the automatic results. The importance of research

[1] One on the first papers on MT evaluation is "*Some psychological methods for evaluating the quality of translations*" by George A.Miller and J.G.Beebe-Center [A.Miller and J.G.Beebe-Center, 1956].

[2] The use of a predefined test data is an approach which is not encountered only in MT.

in MT evaluation has increased, as the "*standard*" BLEU score [Papineni et al., 2002] was analyzed more carefully and the research community started to criticize it. Current metrics are "*largely influenced by lexical choice and insensitive to reordering differences*" [Birch et al., 2010]. Part of the problems encountered (e.g. robustness, no details about the nature of an error) have been discussed in several papers, such as [Chan and Ng, 2008], [Giménez and Màrquez, 2007] and [Owczarzak et al., 2007]. An analysis of several evaluation scores is presented in [Callison-Burch et al., 2006].

The problems of automatic evaluation led to a development of an MT "*meta-evaluation*" research, which set several metric design considerations (see [Chan and Ng, 2008]), such as the intuitiveness of the interpretation of the result, permission of variations (synonyms, dependency) or correlation with human judgments.

There are several types of metrics for automatic MT evaluation: some are based on n-grams (e.g. BLEU, NIST), some are lexical similarities and use no external resources (e.g. WER, ROUGE[3]) and some employ semantic information (e.g. METEOR[4]). More information on evaluation and evaluation types[5] can be found in the report of the Euromatrix project: [EUROMATRIX, 2007]. An overview on (automatic) MT evaluation is also presented in the report of the "Framework for Machine Translation Evaluation in ISLE" project[6], in [Linares, 2008] and [Owczarzak, 2008]. Criteria for human evaluation are presented in [Chan and Ng, 2008]: adequacy, fluency, rank[7], constituent[8], etc. Further aspects on evaluation can be found in the ALPAC and ARPA reports (see [Linares, 2008]) and in [Vilar et al., 2006]. An analysis of automatic metrics, from the point of view of the correlation with human evaluation, has recently been carried out at the annual Workshop on SMT ([Callison-Burch et al., 2007] and [Callison-Burch et al., 2009]).

In order to overcome the disadvantages of previously used metrics, new methods have been developed. These methods make use of more linguistic information, such as syntactic similarity defined on shallow parsing results ([Popovic and Ney, 2007], [Linares, 2008]), on constituency structures ([Giménez and Màrquez, 2009],[Liu and Gildea, 2005]) and on dependency structures ([Amigó et al., 2006],[Mehay and Brew, 2007]).

In the next subsection we will show how translation results obtained with corpus-based MT (**CBMT**) approaches have been evaluated in the literature.

[3]Metric for MT and text summarization evaluation.

[4]METEOR is available only for English, German, Spanish, French, Czech (`www.cs.cmu.edu/~alavie/METEOR` - last accesed on June 22nd, 2011).

[5]Evaluation types: adequacy evaluation, diagnostic evaluation and performance evaluation.

[6]`http://www.issco.unige.ch/en/research/projects/isle/femti/` - last accessed on April 9th, 2010.

[7]This criterion refers to comparing and ranking different translations of an input sentence from best to worst.

[8]This criterion is based on human judges who have to rank the translations of some constituents from the parse tree of the input sentence.

8.1.1 Evaluation of Corpus-Based MT Systems

For the EBMT approach previous works do not really offer a clear framework for the evaluation. Sometimes evaluation results are presented, with no real description of the methodology ([Somers, 2003] and [Andriamanankasina et al., 2003]). There is no real specification for how the evaluation tests should be run (the method of evaluation) or how many and which type of test sentences should be used. In the literature, several approaches to evaluation have been identified. An overview of these approaches can be found in [Somers, 1999] and in Table 8.1.

System	Data size		Approach
	Training	Test	
[Sumita and Iida, 1991]	2450	100	Jackknife evaluation
[Frederking et al., 1994]			Comparison with other systems, no. of editing keystrokes
[Doğan, 2005]	970	100	BLEU
[McTait, 2003]	4858	1000 from 1858 / 2358	Random selection Measure based on Levenshtein Distance (**LD**) [Levenshtein, 1966]
[Sumita, 2001]	204108	500	Random selection
[Planas and Furuse, 2003]	7129 / 32526	50/75	Aid for human translation

Table 8.1: Evaluation approaches in EBMT.

In Table 8.1 it can be seen that the size of the test and training data are different and the evaluation approaches are (mostly) incomparable. The metric of evaluation also differs. For instance, Sumita [2001] uses coverage and accuracy for human evaluation; Sato [1992] only accuracy. In [Sumita, 2001] a scale with three values is applied for evaluating the 'coverage': '*Exactly*', '*Approximately*' and '*No output*'. For the 'accuracy' each translation is graded into one of four ranks ('*Perfect*', '*Fair*', '*Acceptable*' and '*Nonsense*') by a bilingual human translator who is a native speaker of the target language.

Both automatic evaluation and manual evaluation results are shown in [Watanabe and Sumita, 2003] and [Gough and Way, 2004]. The automatic evaluation metrics differ in the two papers mentioned before: WER, PER and BLEU in the former and BLEU, WER, SER, precision and recall in the latter. Watanabe and Sumita [2003] use as human evaluation subjective evaluation ranks ranging from A to D[9], judged by a native speaker for manually evaluating the translation results. Gough and Way [2004] performed a manual evaluation using '*intelligibility*'[10] and '*accuracy*'.

The number of (test and training) sentences also differs. In the analyzed literature the maximum number of test sentences was 5000 (see [Shirai et al., 1997]). In the majority of the experiments the tests were carried out on less than 500 sentences.

[9]The evaluation ranks in [Watanabe and Sumita, 2003]: **A** : perfect, **B** : fair, **C** : acceptable and **D** : nonsense.

[10]Intelligibility depends on the number of grammatical errors or mistranslations in the string.

For SMT it can be considered that an "*official*" evaluation framework is set by the EuroMatrix [11] and EuroMatrixPlus[12] projects and by the annual Workshops on SMT, where a shared task has been proposed and the necessary data has been provided. However, only a small number of language-pairs have been taken into account.

When both EBMT and SMT systems are involved, the same training and test data are usually used and BLEU is the automatic evaluation metric (see **Subsection 3.3**).

8.2 Automatic Evaluation Scores

We have evaluated the obtained translations using three (3) automatic evaluation metrics: BLEU, NIST and TER. All these metrics are based only on surface-forms.

The choice of these metrics for an automatic evaluation is motivated by the available resources (software, money, man-power, time) and, for comparison reasons[13], by the results reported in the literature. Some reasons for choosing several metrics are the fact that automatic metrics are not always correlating with the human judgements, and that each metric has advantages and disadvantages. Choosing several metrics gives us a better overview of the results.

A comparison between several metrics is presented in [Callison-Burch et al., 2009]. An analysis of the correlation of some automatic metrics with the human evaluation is shown in [Callison-Burch et al., 2007]: two cases have been analyzed: English as SL and language other than English as SL. In the first case, considering the overall correlation, from the eleven scores investigated, BLEU and TER were ranked 4th and 5th, respectively. The first three scores need additional linguistic information. In the second case, among 6 scores, BLEU and TER ranked first and second, respectively. The scores for an SL different than English are lower than the ones when English is SL. NIST has not been analyzed.

Due to lack of data, financial resources and further translation possibilities, in our experiments we compare the output with only one reference translation. No attempt was made to constrain or modify the test sentences on the basis of the length, inclusion in the training data or other characteristics. This way we ensured a realistic scenario, in which users just translate a text, without interfering themselves in the MT system or selecting a specific test data.

The evaluation methods as such are beyond the scope of this thesis, but they are considered as tools for the experiments. The metrics used for the evaluation are only briefly described in the subsections that follow.

[11] `http://www.euromatrix.net/` - last accessed on June 22nd, 2011. Project duration: 2006-2009.

[12] `http://www.euromatrixplus.net/` - last accessed on June 22nd, 2011. Project duration: 2009-2012.

[13] Only a partial comparison can be made, as a one-to-one comparison it is not possible, as the training and test data are not exactly the same.

8.2.1 BLEU

BLEU (**bil**ingual **e**valuation **u**nderstudy), one of the evaluation scores applied most frequently for MT evaluation, measures the number of n-grams of different lengths of the system output that appear in a set of references. More details about BLEU can be found in [Papineni et al., 2002].

Although criticized more recently, it is still important to calculate the BLEU score for comparison reasons[14], as for many previous developed systems it is the only evaluation measure available. The BLEU score is computed according to the following formula:

$$BLEU = BP * exp(\sum_{n=1}^{N} \frac{1}{N} log(p_n)) \tag{8.1}$$

where N is the maximum n-gram size and the brevity penalty BP is calculated as:

$$BP = min(1, e^{1-\frac{r}{c}}) \tag{8.2}$$

In Formula 8.2, c is the length of the corpus of hypothesis translations and r is the effective reference corpus length. The value for r is calculated as the sum of the single reference translation from the each set which is closest to the hypothesis translation.

Papineni et al. [2002] calculate the n-gram precision p_n as the sum over the matches for every hypothesis sentence S in the complete corpus C as:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} count(ngram)} \tag{8.3}$$

For the evaluation with BLEU, we used the twelfth version[15] of the NIST/BLEU implementation provided by `www.itl.nist.gov`.

8.2.2 NIST

The NIST score[16], described in [Doddington, 2002], is similar to the BLEU score in that it also uses n-gram co-occurrence precision. While BLEU uses a geometric mean of the n-gram precision, NIST calculates the arithmetic mean. Another difference is that n-gram precisions are weighted by the n-gram frequencies. The formula for NIST is

$$NIST = \sum_{n=1}^{n} BP * \frac{\sum_{All n-gram ThatCo-occur} info(ngram)}{\sum_{All n-gram InSystemOutput} 1} \tag{8.4}$$

where $info(ngram)$ is

$$info(ngram) = log_2 \frac{count((n-1)gram)}{count(ngram)} \tag{8.5}$$

[14] A one-to-one comparison is not possible, as the system has not been trained and tested using the same data. The comparison is only relative.

[15] "*mteval_v12*", as implemented on http://www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html - last accessed on April 18th, 2009. The BLEU / NIST implementation analyzes 1-grams to 9-grams.

[16] We used the same "*mteval_v12*" inplementation to calculate the NIST score.

In Formula 8.5 $count(ngram)$ represents the number of occurrences of the n-gram $ngram = w_1w_2...w_n$ in all the reference translations. $(n-1)gram$ represents the sequence $w_1w_2...w_{n-1}$.

BP is

$$BP = exp[\beta log^2 min(\frac{L_{sys}}{\overline{L}_{ref}}, 1)] \tag{8.6}$$

where L_{sys} is the length of the MT output, $\overline{L}_{ref}$ is the average number of words in a reference translation and β is chosen to make $BP = 0.5$ when $\frac{L_{sys}}{\overline{L}_{ref}} = \frac{2}{3}$.

Higher BLEU or NIST scores show better translation results.

8.2.3 TER

The TER (**t**ranslation **e**rror **r**ate)[17] score calculates the minimum number of edits needed to get from an obtained translation to the reference translations, normalized by the average length of the references. It considers insertions, deletions, substitutions of single words and an edit-operation which moves sequences of words. More information about TER is presented in [Snover et al., 2006].

$$TER = \frac{number\_of\_edits}{average\_of\_reference\_words} \tag{8.7}$$

The lower the TER scores are, the better the translation results are.

8.3 Experimental Settings and Data Description

In this section we present the experimental settings and the (training and test) data used.

Several parameters have been changed during the experiments: the MT system and approach, the language pair and the corpus (type and size) and several comparisons of the results have been carried out. The SMT and EBMT systems, trained and (or) developed during this research, are compared using the same training and test data. In some experiments part-of-speech (**POS**) information has been added.

When no other linguistic information is used, the obtained results have also been compared with the ones provided by Google Translate[18].

The experiments have been done using different corpora of different sizes: JRC-Acquis, RoGER and a sub-corpus of JRC-Acquis (JRC-Acquis$_{SMALL}$). We split the experiments in three experimental settings according to the training and test data involved.

A tabular overview of all experimental settings is shown in Table 8.2. The MT systems mentioned are:

- **Mb_SMT** - the Moses-based SMT system (no tuning involved; LM-order is 3);

[17] We used TER Version 7.25 as implemented on http://www.cs.umd.edu/ snover/tercom/ - last accessed on April 18th, 2009.

[18] For more details on Google Translate please refer to **Section 5.2**.

- **SMT_tuning** - the tuned version of **Mb_SMT**, having the LM-order 5;
- **SMT_POS** - the Moses-based SMT system **Mb_SMT**, when the data also contains POS information;
- **Google** - Google Translate, the on-line Google translation system;
- ***Lin-EBMT*** - the EBMT baseline system;
- ***Lin-EBMT*_POS** - *Lin-EBMT*, when POS information is added to the data;
- **EBMT_2:** $Lin - EBMT^{REC+}$, the extended EBMT system; and
- **EBMT_2_POS** - $Lin - EBMT^{REC+}$ with the data enriched with POS information.

RoGER **Experimental setting I**				
MT System	**ENG - RON**	**RON - ENG**	**DEU - RON**	**RON - DEU**
Experimental setting Ia				
Mb_SMT	x	x	x	x
Google	x	x	x	x
Lin-EBMT	x	x	x	x
EBMT_2	x	x	x	x
Experimental setting Ib				
SMT_POS	x	x	-	-
***Lin-EBMT*_POS**	x	x	-	-
EBMT_2_POS	x	x	-	-
JRC-Acquis **Experimental setting II**				
MT System	**ENG - RON**	**RON - ENG**	**DEU - RON**	**RON - DEU**
Experimental setting IIa				
Mb_SMT	x	x	x	x
Google	x	x	x	x
Lin-EBMT	x	x	x	x
EBMT_2	x	x	x	x
Experimental setting IIb				
SMT_tuning	-	-	-	x
JRC-Acquis$_{SMALL}$ **Experimental setting III**				
MT System	**ENG - RON**	**RON - ENG**	**DEU - RON**	**RON - DEU**
Mb_SMT	x	x	x	x
EBMT_2	x	x	x	x

Table 8.2: Experimental settings. (We mark with x the experiments run and with $-$ the settings for which no experiments have been made.)

The multitude of experiments show an ample view on the behavior of the CBMT approaches implemented in the dissertation. For all the experimental settings the evaluation is done automatically (**Chapter 9**) and in some cases the results are manually analyzed (**Chapter 10**).

Since general information about the corpora used has been presented in **Chapter 4**, in the following sections we will describe the data from the point of view of its separation into test and training data.

In order to better evaluate the influence of the data on the results, statistical information about the test and training data is needed. We analyzed three aspects:

1. The total number of tokens[19]. In this context a token represents a lexical item (word), a punctuation sign, a number etc. Inflected forms of a word are considered separate tokens.
2. The vocabulary size[20]: It represents all tokens, counted only once.
3. The average sentence length, which is calculated as $\frac{Total\_number\_of\_tokens}{Number\_of\_sentences\_in\_corpus}$.

8.3.1 Data for the Experimental Setting I (a+b)

This subsection will describe the data for the experiments based on the RoGER corpus. In order to run the experiments, the corpus has been formatted to fit the description of each of the MT systems. General information about RoGER has already been presented in **Chapter 4**.

We considered for this corpus two experimental settings:

1. Corpus with no additional linguistic information (Experimental setting Ia): all language pairs are included;
2. Corpus with additional POS information (Experimental setting Ib): only the language pair Romanian-English is used.

From the RoGER corpus (2333 sentences), 133 sentences (**Test RoGER**) were randomly extracted as the test data, the remaining (2200 sentences) being used as training data.

For the Experimental setting Ib, the data was annotated with POS information for Romanian and English. We annotated the corpus by means of the text processing web services described in **Section 5.5**. The POS information was concatenated to the word as **WORD+"*POS*"+POS**, where "*POS*" is a delimiter. A word together with its POS information (**WORD+"*POS*"+POS**) is considered during the translation as one token for both of the corpus-based MT approaches.

Statistical information on the data for Experimental setting Ia is shown in Table 8.3. The statistical information about the training and test data which contains POS information is presented in Table 8.4.

[19]The information is obtained using the "*Word Count*"-function from *OpenOffice*.

[20]The information is obtained using the SRILM tool, running the *ngram-count* function, with the parameter *-write1*).

Data SL	No. of tokens	Vocabulary size	Average sentence length
Experimental setting Ia (no additional linguistic information)			
English-Romanian			
Training	27889	2367	12.68
Test	1613	522	12.13
Romanian-English, Romanian-German			
Training	28946	3349	13.16
Test	1649	659	12.40
German-Romanian			
Training	28361	3230	12.89
Test	1657	604	12.46

Table 8.3: RoGER statistics.

Data SL	No. of tokens	Vocabulary size	Average sentence length
Experimental setting Ib (additional POS information)			
English-Romanian			
Training	27816	2815	12.64
Test	1610	564	12.11
Romanian-English			
Training	28954	4133	13.16
Test	1651	735	12.41

Table 8.4: RoGER statistics (additional POS information).

8.3.2 Data for the Experimental Setting II

In this experimental setting, JRC-Acquis is the corpus used for training and testing. In order to analyze the behavior of the **Mb_SMT** system when considering texts from a new domain, we also used part of the RoGER corpus for testing.

For running the experiments, the data was formatted according to the specifications of each of the MT systems. From the XML encoded JRC-Acquis monolingual documents and the alignment files, the SL and TL files for the Moses-based SMT systems were extracted. Also the XML input file for the EBMT systems was created.

We split this experimental setting into two categories, depending on the configuration of the Moses-based SMT system:

1. Experimental setting IIa: The order of the language model is three and no tuning step is used;

2. Experimental setting IIb: The order of the language model is five and tuning is included;

Experimental Setting IIa

In this experimental setting the training corpus is part of the JRC-Acquis (**Chapter 4**) and all four language combinations are taken into account. As already mentioned, two types of alignments are available on the JRC-Acquis website. They are automatically created using Vanilla or HunAlign. For our experiments we used the alignments realized with the Vanilla[21] aligner, decision also taken in [Ignat, 2009]. The alignment is realized at paragraph-level. The paragraphs are delimited by the $< p >$-tag from the initial HTML files. A *paragraph* in this case can be a sentence, a sub-sentential phrase (e.g. noun phrase), a complex or a compound sentence. To reduce possible errors, only one-to-one paragraph alignments have been used for our experiments[22]. Also Koehn et al. [2009] extract from JRC-Acquis a sub-corpus where sentences (paragraphs) are aligned in a one-to-one manner.

With respect to English and Romanian, only 336509 links[23] have been used from the total 391324 links ($< p >$-alignments) in 6557 documents, due to the one-to-one alignments. Because of the cleaning step of the SMT system[24], the number of one-to-one alignment links considered for the Language Model (**LM**) has been reduced to 240219 links for the Translation Model (**TM**). This represents 61.38% of the initial corpus.

For German and Romanian, from 391972 links in 6558 documents, only 324448 links have been considered for the LM. The TM size has been reduced to 238172 links - 60.76% of the initial corpus.

For this experimental setting we ran our experiments using test data from two different corpora: one is part of the JRC-Acquis corpus and the other is part of RoGER. Before training the system, 897 sentences (299 from the beginning, 299 from the middle and 299 from the end) have been removed from JRC-Acquis, in order to be used as test sets: the first test corpus. Sentences were removed from different parts of the corpus to ensure a relevant lexical, syntactic and semantic coverage. These sets of 299 sentences represent the data sets **Test 1**, **Test 2**, and **Test 3**, respectively. **Test 1+2+3** is formed from all 897 sentences. Not all test data sets have been used for the EBMT systems for all language pairs. More information about the use of the test data can be found in **Chapter 9**. The distribution of the out-of-vocabulary words (**OOV-words**) differs, as we extracted sentences from different parts of the corpus. This has a direct impact on the translation quality (see the analysis in **Chapter 10**).

In order to analyze the reaction of the **Mb_SMT** system to other types of text input, we used a second corpus, RoGER. From the 2333 sentences, we extracted 300 sentences from the middle of the corpus and used them as test data: **T_RoGER** data-set.

An exact description of the two corpora and the differences between them can be found

[21] See http://nl.ijs.si/telri/Vanilla/ - last accessed April 18th, 2009.

[22] Some M-to-N alignments in the corpus contain sometimes HTML tags.

[23] Paragraph alignments

[24] In the baseline system provided by the annually Workshops on SMT, sentences longer than 40 tokens are removed, see **Chapter 5**. This operation is called "*cleaning*".

in **Chapter 4**. For these experiments we avoided the use of other linguistic resources in order to be able to evaluate the robustness of a pure SMT-system against domain change. When changing the domain it is evident that out-of-vocabulary words (**OOV-words**), especially in domain specific vocabulary, play a major role (see the analysis in **Chapter 10**).

Statistical information on the training and test data is presented in Table 8.5.

Data	No. of tokens	Vocabulary size	Average sentence length
English - Romanian			
Training (SL)	3579856	39784	14.90
LM Romanian	9572058	81616	28.45
Test 1 (SL)	6424	1048	21.48
Test 2 (SL)	7523	735	25.16
Test 3 (SL)	5609	1111	18.76
Test 1+2+3 (SL)	19556	2345	21.80
T_RoGER (SL)	4474	635	14.91
Romanian-English			
Training (SL)	3386495	55871	14.10
LM English	9955983	55856	29.59
Test 1 (SL)	5672	1245	18.97
Test 2 (SL)	7194	923	24.06
Test 3 (SL)	5144	1355	17.20
Test 1+2+3 (SL)	18010	2717	20.08
T_RoGER (SL)	4561	876	15.20
German-Romanian			
Training (SL)	3256047	76600	13.67
LM Romanian	9122333	80484	28.12
Test 1 (SL)	5325	1140	17.81
Test 2 (SL)	10286	1439	34.40
Test 3 (SL)	5125	1292	17.23
Test 1+2+3 (SL)	20763	3000	23.15
T_RoGER (SL)	4550	782	15.17
Romanian-German			
Training (SL)	3453586	56219	14.50
LM German	8469146	121969	26.10
Test 1 (SL)	5432	1294	18.17
Test 2 (SL)	11488	1663	38.42
Test 3 (SL)	5317	1388	17.78
Test 1+2+3 (SL)	22237	3336	24.79
T_RoGER (SL)	4561	876	15.20

Table 8.5: JRC-Acquis statistics.

The degree of inflection and the vocabulary richness of the languages can be observed

also in the vocabulary size[25]: 76600 items for German, approximately 56000 for Romanian and 39784 for English. The total number of tokens[26] is smaller for German, compared to Romanian or English. A reason could be that German uses more compounds. More details on the languages have been shown in **Chapter 4**. Further information will be presented in **Appendix B**.

Experimental Setting IIb

This subsection describes the data used for the tuned SMT system (**SMT_tuning**), which has the LM-order five. The experiments are done using the JRC-Acquis data for Romanian-German.

With respect to the size of the data set for tuning, several possibilities have been found in the literature: for the Workshop on SMT in 2011 the tuning data size was set around 2500 sentences; it is reduced to 1000 sentences in the experiments presented in [Ignat, 2009]. In our experimental setting, 1000 sentences are randomly extracted for tuning from the initial 324448 sentences (the one-to-one alignments used for the language model in the experimental setting IIa). Therefore, 323448 sentences are used for the LM in this experimental setting. Only 237364 sentences are considered for the TM after applying the cleaning step in the SMT process. Further statistical information is presented in Table 8.6.

Corpus	No. of tokens	Vocabulary size	Average sentence length
Training - SL	3440687	56206	14.50
Tuning - SL	25060	2022	25.06
LM - TL	9097316	80471	28.13

Table 8.6: Corpora statistics for Experimental setting Ib.

No out-of-domain test data set was used in this experimental setting. The test data remains the same as in the previous experimental setting (Experimental setting IIa).

8.3.3 Data for Experimental Setting III

To analyze how the systems behave for another type of small corpus, 2333 sentences[27] have been extracted from the middle of the initial JRC-Acquis data[28] and form the JRC-Acquis$_{SMALL}$ corpus. From this data 133 sentences have been randomly selected as test data (**Test JRC-Acquis_Small**). The rest of 2200 remain as training data. JRC-Acquis$_{SMALL}$ was not manually verified or modified.

The training sentences, in contrast to the initial experimental setting with JRC-Acquis,

[25] Values in the second column of Table 8.5.
[26] Values in the first column of Table 8.5.
[27] The same size as RoGER.
[28] From the sentence 150001 to the sentence 152 333.

were not filtered considering the maximum sentence length of 40 words. The statistics on the training and test data are presented in Table 8.7.

Information	DEU - RON	ENG - RON	RON - DEU	RON - ENG
Training data				
No. tokens	69735	75405	75156	72170
Vocabulary size	5929	3578	6390	5581
Average sentence length	31.69	34.27	34.16	32.80
Test data				
No. tokens	3947	4434	4366	4325
Vocabulary size	1178	992	1320	1260
Average sentence length	29.67	33.33	32.82	32.51

Table 8.7: Statistics on the data for Experimental setting III.

8.4 Chapter Summary

In this chapter we briefly presented facts on MT evaluation and the automatic evaluation scores used for evaluating our translation results. We also described the training and test data for our three experimental settings.

In the following chapter, **Chapter 9**, we will show the automatic evaluation results and their interpretation. A manual analysis of a subset of the results will be presented in **Chapter 10**.

Chapter 9

Automatic Evaluation Results

Chapter 8 showed the experimental settings and the data used. As already mentioned, several parameters can be changed in our experiments: the language pair, the corpus type, the corpus size, the MT approach, the use of additional linguistic information or the use of additional steps in the training process of an SMT system. In this chapter, the evaluation results obtained with the metrics described in **Section 8.2** will be presented. Because the empirical approaches highly depend on the available training data, their strong point is not the coverage, but the correctness. In this perspective, the focus is comparing the correctness of the output obtained by the systems. The coverage of the **Mb_SMT** system was tested only with one test-data set when JRC-Acquis is used for training. This test data-set has been extracted from RoGER (**T_RoGER**).

Analyzing the experimental settings and the obtained results, several comparisons are made. The approaches (SMT vs. EBMT) are compared between themselves, using both a smaller (usually accepted as an EBMT framework) and a larger corpus (an SMT framework). The comparison is done using the same training and test data. In addition, part of the obtained results have been compared with the ones given by Google Translate. It is also analyzed how the systems react to different data-sets. In the case of the SMT approach and the JRC-Acquis corpus, an experiment is run to reveal how tuning influences the SMT results. Moreover, it is analyzed how additional linguistic information (i.e. POS) influences the translation in the RoGER corpus. To verify that the results for RoGER have not been only a casualness (due to the corpus type), we considered another small-size corpus: JRC-Acquis$_{SMALL}$

All these comparisons and the obtained results will be presented in the following sections.

9.1 Automatic MT Results

In this section we will present the automatic evaluation results for the three experimental settings.

9.1.1 Experimental Setting I (a+b)

First we will describe the experiments which helped us decide the configurations of the EBMT systems. We will show the comparative results for all MT systems and language pairs afterwards. Also the experiments with additional POS information will be presented in this subsection.

Influence of the Language Model

We have tested how different language models (**LMs**) influence the translation results. We ran experiments in which we changed in the recombination step of *Lin-EBMT* the LM based on the Dice coefficient into the values from the LM provided by **Mb_SMT**. In this case, the recombination matrix definition changed as follows:

Definition 9.1. *If the outcome of the alignment is N word-sequences $\{sequence_1, sequence_2, ..., sequence_N\}$, with $sequence_i = w_{i_1}w_{i_2}...w_{i_{last}}$ $(1 \leq i \leq N)$), and these word-sequences are not necessarily different, then A is a square matrix of order N that is defined as follows:*

$$A_{N,N} = (a_{i,j})_{1 \leq i,j \leq N} = \begin{cases} -3, & \text{if } i = j; \\ -2, & \text{if } i \neq j,\ count(w_{i_{last}}w_{j_1}) = 0; \\ VAL, & \text{if } i \neq j,\ count(w_{i_{last}}w_{j_1}) > 0. \end{cases} \tag{9.1}$$

*where VAL is the corresponding 2-gram value of the language model used in the **Mb_SMT** system.*

The LM in the Moses-based SMT system has been presented in **Chapter 5**. It uses Chen and Goodman [1996]'s modified Kneser-Ney discounting for n-grams of order n and the discounted n-gram probability estimates at the specified order n are interpolated with lower-order estimates.

The results for the this setting of *Lin-EBMT* are shown in Table 9.1.

Language pair	BLEU	NIST	TER
English-Romanian	0.2631	4.9495	0.6027
Romanian-English	0.2797	5.4224	0.5449
German-Romanian	0.2163	3.5582	0.6525
Romanian-German	0.2452	4.5336	0.6771

Table 9.1: Evaluation results for *Lin-EBMT* with the LM from **Mb_SMT**.

A comparative view on the influence of the LM in *Lin-EBMT* is presented in Table 9.2 (only BLEU scores) - boldface values are better..

Language pair	LM from Mb_SMT	LM based on the Dice coefficient
English-Romanian	0.2631	**0.2689**
Romanian-English	**0.2797**	0.2783
German-Romanian	0.2163	**0.2204**
Romanian-German	**0.2452**	**0.2452**

Table 9.2: Influence of LMs on *Lin-EBMT* (BLEU scores).

Overall (for all three evaluation metrics – see also Table 9.4) the system which uses the LM based on the Dice coefficient is better. However, the differences between the scores are quite small. For the Romanian-German direction of translation the results are the same, no matter the evaluation metric. The LM from **Mb_SMT** is better only in two cases: BLEU score for the Romanian-English direction of translation and TER score for the German-Romanian direction of translation. For further experiments we will consider the LM based on the Dice coefficient for both *Lin-EBMT* and $Lin - EBMT^{REC+}$.

Impact of Constraints

We have also tested how possible combinations of constraints and definitions of the constrained recombination matrix influence the evaluation results. We run these test for deciding an '*optimal*' setting for $Lin - EBMT^{REC+}$ in further experiments.

Before showing the results, the following notations need to be explained:

- **No C.**: No constraints are applied. *Lin-EBMT* is run, but out-of-vocabulary words and punctuation are included.
- **C. X**: Only constraint **X** is applied.
- **C. X+Y**: Both constraints **X** and **Y** are used.
- **C. 1+2+3**: All three constraints are integrated.
- **C. 1+2+3 1:2**: All three constraints are included, but it employs a different definition of the recombination matrix.

In these experiments we have usually employed the definition of the constrained recombination matrix shown in Formula 7.2. Only in the configuration **C. 1+2+3 1:2** we used Formula 7.3.

The evaluation for $Lin - EBMT^{REC+}$ is presented in Table 9.3 and a graphical representation of these results is shown in Figure 9.1.

We have obtained different best results for variant language-pairs and evaluation metrics. For further experiments we will consider for $Lin - EBMT^{REC+}$ the **C. 1+2+3 1:2** configuration which provided best translation results according to all the automatic evaluation metrics. This configuration rendered best results in 50% of the cases (six out of twelve cases) according to all three automatic metrics (boldface values in Table 9.3).

System	BLEU	NIST	TER
English – Romanian			
No C.	0.2997	5.4093	0.6046
C. 1	0.3067	5.5768	0.5930
C. 2	0.3042	5.4187	0.5991
C. 3	0.3083	**5.5836**	0.5906
C. 1+2	0.3062	5.5353	0.5930
C. 1+3	0.3083	**5.5836**	0.5906
C. 2+3	0.3073	5.5638	0.5882
C. 1+2+3	0.3073	5.5638	0.5882
C. 1+2+3 1:2	**0.3085**	5.5322	**0.5864**
Romanian – English			
No C.	0.3597	6.0586	0.5065
C. 1	0.3695	6.2694	0.5034
C. 2	0.3711	6.1625	**0.4984**
C. 3	0.3633	6.2415	0.5108
C. 1+2	**0.3712**	6.2879	0.5009
C. 1+3	0.3632	6.2355	0.5114
C. 2+3	0.3656	6.2620	0.5083
C. 1+2+3	0.3656	6.2620	0.5083
C. 1+2+3 1:2	0.3668	**6.2991**	0.5077
German – Romanian			
No C.	0.2643	4.5589	0.6428
C. 1	0.2658	**4.6935**	0.6422
C. 2	**0.2682**	4.6074	0.6409
C. 3	0.2627	4.6757	0.6422
C. 1+2	0.2654	4.6745	0.6428
C. 1+3	0.2627	4.6757	0.6422
C. 2+3	0.2633	4.6807	0.6422
C. 1+2+3	0.2633	4.6807	0.6422
C. 1+2+3 1:2	0.2646	4.6559	**0.6361**
Romanian – German			
No C.	0.2867	4.9792	0.6795
C. 1	0.2842	5.0664	0.6716
C. 2	0.2857	5.0253	0.6789
C. 3	0.2891	5.0622	0.6716
C. 1+2	0.2836	5.0591	**0.6698**
C. 1+3	0.2891	5.0622	0.6716
C. 2+3	0.2875	5.0593	0.6722
C. 1+2+3	0.2875	5.0593	0.6722
C. 1+2+3 1:2	**0.2894**	**5.0770**	0.6722

Table 9.3: Evaluation results for $Lin - EBMT^{REC+}$, when changing the constraints (C=constraint).

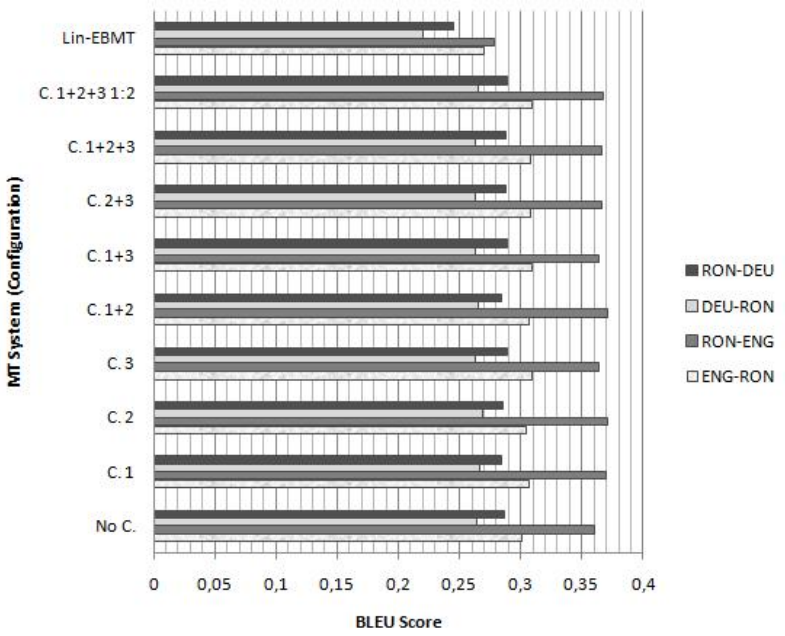

Figure 9.1: The Influence of constraints and constraint settings on $Lin - EBMT^{REC+}$. Comparison with *Lin-EBMT* – BLEU scores; corpus: RoGER.

Comparative Results

The results for the Experimental setting I (a+b) are presented in Table 9.4. We used all MT systems considered in this dissertation. For Romanian and English (both directions of translation), we also verified how POS information influences the translation results (Experimental setting Ib).

An analysis of these results will be shown in **Section 9.2**.

9.1.2 Experimental Setting II

Experimental Setting IIa

In this experimental setting we have tested the systems using different in-domain test data-sets and, for **Mb_SMT** only, also one out-of-domain data-set.

We run only two MT systems on all (five) test data-sets: **Mb_SMT** and Google Translate. For the two EBMT systems we did not run any out-of-domain tests. $Lin - EBMT^{REC+}$ was used only for **Test 1** for all language combinations. We have *Lin-EBMT* results for all in-domain data-sets for English-Romanian and German-Romanian directions of translation and only for **Test 1** for the other two language combinations.

The initial translations for the **T_RoGER** (out-of-domain) test data set contain diacritics due to the training data. However, the reference translation has no diacritics, as the whole RoGER corpus excludes them. The results shown in this subsection are the ones obtained after the diacritics had been eliminated from the initial translations. This

Experimental setting Ia				
Score	**Mb_SMT**	**Google**	***Lin-EBMT***	$Lin-EBMT^{REC+}$ **(C.1+2+3 1:2)**
English – Romanian				
BLEU	0.4386	0.4782	0.2689	0.3085
NIST	6.5599	6.9334	5.0787	5.5322
TER	0.3784	0.3565	0.5955	0.5864
Romanian – English				
BLEU	0.4765	0.5241	0.2783	0.3668
NIST	6.8022	7.4478	5.5313	6.2991
TER	0.3465	0.3087	0.5443	0.5077
German – Romanian				
BLEU	0.3240	0.2980	0.2204	0.2646
NIST	5.2643	5.2226	3.6371	4.6559
TER	0.5239	0.5530	0.6549	0.6361
Romanian – German				
BLEU	0.3405	0.3459	0.2452	0.2894
NIST	5.3140	5.5675	4.5336	5.0770
TER	0.5570	0.5769	0.6771	0.6722

Experimental setting Ib (additional POS information				
Score	**Mb_SMT_POS**	**Google**	***Lin-EBMT*_POS**	$Lin-EBMT^{REC+}$**_POS (C.1+2+3 1:2)**
English – Romanian				
BLEU	0.3879	-	0.2942	0.2916
NIST	5.8047	-	5.1641	5.0893
TER	0.4748	-	0.6402	0.6541
Romanian – English				
BLEU	0.4618	-	0.3624	0.3559
NIST	6.3533	-	6.1167	6.0039
TER	0.4000	-	0.5490	0.5751

Table 9.4: Evaluation results for RoGER.

way we ensured the compatibility with the reference translations. As English contains no diacritics, the initial results for English as TL have not been changed.

The TER, BLEU and NIST results for the Experimental setting IIa are presented in Tables 9.5, 9.6, 9.7, respectively. Boldface values show best results for a system, when using different test data-sets.

System	Test 1	Test 2	Test 3	Test 1+2+3	T_RoGER
English - Romanian					
Mb_SMT	0.5007	**0.4898**	0.5208	0.5023	0.7340
Google	0.4701	0.5330	**0.4563**	0.4908	0.4816
Lin-EBMT	0.8071	**0.6400**	0.7770	0.7326	-
$Lin-EBMT^{REC+}$	0.8389	-	-	-	-
Romanian - English					
Mb_SMT	0.5020	**0.3756**	0.4684	0.4457	0.7623
Google	0.4686	0.4531	**0.4379**	0.4541	0.3531
Lin-EBMT	0.7041	-	-	-	-
$Lin-EBMT^{REC+}$	0.7431	-	-	-	-
German - Romanian					
Mb_SMT	0.6200	**0.5905**	0.6438	0.6113	0.8311
Google	**0.6397**	0.6707	0.6642	0.6612	0.6299
Lin-EBMT	0.8339	**0.7865**	0.8224	0.8075	-
$Lin-EBMT^{REC+}$	0.8537	-	-	-	-
Romanian - German					
Mb_SMT	0.6437	**0.5588**	0.6791	0.6112	0.8637
Google	**0.5971**	0.6590	0.6576	0.6425	0.5689
Lin-EBMT	0.8432	-	-	-	-
$Lin-EBMT^{REC+}$	0.8562	-	-	-	-

Table 9.5: TER evaluation results for JRC-Acquis.

The analysis of the results will be shown in **Section 9.2**.

Experimental Setting IIb

In order to see how the tuning process influences the translation results, an experiment is run where the tuning step is included in the SMT system. This experiment follows the recommendations of the Workshop on SMT in 2011 for the "*baseline system*" completely (the **SMT_tuning** system). Therefore, also the LM order is set to 5. The data was evaluated with the same metrics. Table 9.8 shows the results compared with ones provided by the previous SMT experimental setting (no tuning, LM-order 3: **Mb_SMT system**). The boldface values are the better results in the comparison between these two system settings.

As it can be seen from Table 9.8, not in all cases the tuned MT system (**SMT_tuning**) is better than the un-tuned system. The NIST and the BLEU scores do not always correlate (for example the results for **Test 3**).

System	Test 1	Test 2	Test 3	Test 1+2+3	T_RoGER
English - Romanian					
Mb_SMT	0.3997	**0.4179**	0.3797	0.4015	0.0623
Google	0.4214	0.3947	**0.4740**	0.4263	0.3332
Lin-EBMT	0.1335	**0.3072**	0.1476	0.2125	-
$Lin-EBMT^{REC+}$	0.1572	-	-	-	-
Romanian - English					
Mb_SMT	0.2545	**0.5628**	0.4271	0.4255	0.0621
Google	0.2936	0.4359	**0.4422**	0.3909	0.4543
Lin-EBMT	0.0855	-	-	-	-
$Lin-EBMT^{REC+}$	0.1002	-	-	-	-
German - Romanian					
Mb_SMT	0.2955	**0.4244**	0.2884	0.3644	0.0357
Google	**0.2853**	0.2809	0.2740	0.2837	0.2165
Lin-EBMT	0.1374	**0.1818**	0.1347	0.1602	-
$Lin-EBMT^{REC+}$	0.1528	-	-	-	-
Romanian - German					
Mb_SMT	0.2953	**0.4411**	0.2939	0.3726	0.0271
Google	0.3277	0.3301	0.3208	**0.3332**	0.3031
Lin-EBMT	0.1271	-	-	-	-
$Lin-EBMT^{REC+}$	0.1537	-	-	-	-

Table 9.6: BLEU evaluation results for JRC-Acquis.

System	Test 1	Test 2	Test 3	Test 1+2+3	T_RoGER
English - Romanian					
Mb_SMT	6.6279	**6.8431**	6.3857	**7.4039**	2.7285
Google	6.5765	6.5040	7.2757	**7.5468**	5.8309
Lin-EBMT	4.2020	**5.7453**	4.3558	5.3809	-
$Lin-EBMT^{REC+}$	4.6487	-	-	-	-
Romanian - English					
Mb_SMT	3.8325	**7.6956**	6.8134	6.9261	2.7640
Google	4.3363	7.0324	**7.5547**	7.0521	7.2905
Lin-EBMT	2.2425	-	-	-	-
$Lin-EBMT^{REC+}$	2.8503	-	-	-	-
German - Romanian					
Mb_SMT	5.6135	6.1150	5.3053	**6.3704**	2.0498
Google	5.2557	5.3964	5.2857	**5.8486**	4.5360
Lin-EBMT	3.8631	**4.3497**	3.4341	4.3494	-
$Lin-EBMT^{REC+}$	4.0996	-	-	-	-
Romanian - German					
Mb_SMT	5.5629	6.1130	5.3215	**6.3531**	1.8351
Google	5.6889	5.3497	5.8416	**6.1419**	5.3058
Lin-EBMT	3.5431	-	-	-	-
$Lin-EBMT^{REC+}$	4.0764	-	-	-	-

Table 9.7: NIST evaluation results for JRC-Acquis.

Score	Test 1	Test 2	Test 3	Test 1+2+3
System 1: Mb_SMT Romanian – German, without tuning, LM order 3				
NIST	**5.5629**	6.1130	5.3215	6.3531
BLEU	**0.2953**	0.4411	**0.2939**	0.3726
TER	**0.6437**	0.5588	0.6791	0.6112
System 2: SMT_tuning Romanian – German, with tuning, LM order 5				
NIST	5.3808	**6.2644**	**5.3283**	**6.4213**
BLEU	0.2743	**0.4597**	0.2858	**0.3758**
TER	0.6608	**0.5391**	**0.6754**	**0.6052**
Difference (2nd-1st)				
NIST	-0.1821	**0.1514**	**0.0068**	**0.0682**
BLEU	-0.021	**0.0186**	- 0.0081	**0.0032**
TER	0.0171	**-0.0197**	**-0.0037**	**-0.0060**

Table 9.8: Evaluation results for **Mb_SMT** and **SMT_tuning**; corpus JRC-Acquis Romanian – German.

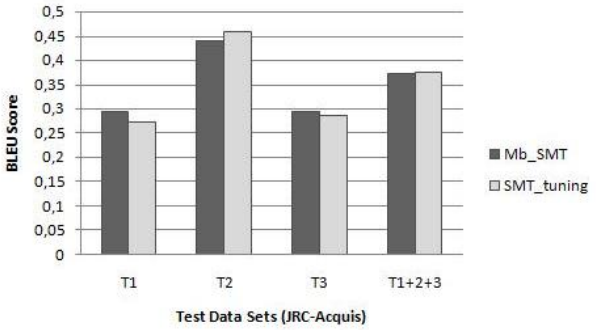

Figure 9.2: SMT with and without tuning – BLEU scores.

As the score differences are not always consistent, the automatic evaluation metrics have drawbacks (see **Chapter 8**) and in the overall test (**Test 1+2+3**) **SMT_tuning** is only slightly better than **Mb_SMT**, we decided to use in our experiments the **Mb_SMT** system.

9.1.3 Experimental Setting III

To ensure that the scores for the RoGER corpus (Experimental setting Ia) have not been obtained only due to the corpus type, we did the experiments for another type of corpus of the same size: JRC-Acquis$_{SMALL}$.

We translated the test data with **Mb_SMT** and $Lin - EBMT^{REC+}$ (**C. 1+2+3 1:2**) and evaluated the translation results with BLEU, NIST and TER.

To test the sensitivity of corpus-based MT to post-processing operations[1], we considered two frameworks:

[1]See the description of the Moses-based MT system in **Chapter 5**.

1. Including recasing and detokenization of the output in the evaluation step;
2. Excluding recasing and detokenization of the output in the evaluation step.

In both cases the reference translation was treated in the same way as the output: when the output was recased and detokenized, also the reference was post-processed in the same way.

The automatic evaluation results are shown in Table 9.9 and Table 9.10, respectively. Analyzing the results, it can be concluded that the post-processing steps (recasing and detokanization) affect negatively the automatic scores.

	DEU - RON		**ENG - RON**		**RON - DEU**		**RON - ENG**	
	1	**2**	**1**	**2**	**1**	**2**	**1**	**2**
BLEU	0.3051	0.2338	0.5359	0.3814	0.3279	0.2604	0.5573	0.4172
NIST	5.8001	5.0296	7.8833	6.8023	5.8781	5.3217	8.1515	7.6335
TER	0.5808	0.7029	0.3586	0.5852	0.5796	0.6977	0.3279	0.5293

Table 9.9: Evaluation results for JRC-Acquis$_{SMALL}$ (no recasing, no detokenization) - 1= **Mb_SMT**, 2=$Lin - EBMT^{REC+}$ (**C. 1+2+3 1:2**).

	DEU - RON		**ENG - RON**		**RON - DEU**		**RON - ENG**	
	1	**2**	**1**	**2**	**1**	**2**	**1**	**2**
BLEU	0.2811	0.2167	0.4801	0.3550	0.2926	0.2458	0.4904	0.3910
NIST	5.4036	4.6874	7.3328	6.4064	5.4104	4.9710	7.5224	7.1873
TER	0.6658	0.7652	0.5032	0.6803	0.6816	0.7736	0.4509	0.6036

Table 9.10: Evaluation results for JRC-Acquis$_{SMALL}$ (with recasing and detokenization) - 1= **Mb_SMT**, 2=$Lin - EBMT^{REC+}$ (**C. 1+2+3 1:2**).

The analysis of the results and a comparison with other experimental settings will be shown in **Section 9.2**.

9.2 First Considerations on the Results

In this section we will present some general considerations on the automatic results. Given the results presented in **Section 9.1**, several interesting aspects will be compared:

1. The behavior of each of the MT systems, when changing the test data-set for one training corpus. We test with in-domain data.
2. The behavior of each of the MT systems, when changing the (training) corpus: a larger vs. a smaller corpus. We test with in-domain data. The use of a larger corpus fits better into an SMT framework. A small(er) corpus is usually found in the EBMT approach.
3. The behavior of each of the MT systems, when POS information is added.
4. The MT approaches.

5. The behavior of each of the MT systems, when the language-pair changes.

We also analyzed how the **Mb_SMT** system behaves for out-of-domain test data.

As in the evaluation process the initial output is recased and for the JRC-Acquis it is also detokenized, errors might also be introduced by these steps (see also the results presented in **Section 9.1.3**).

9.2.1 Score Variation across Test Data-Sets of the Same Corpus, using In-domain Data

The first aspect to be compared is the variation of scores across sets of test data from the JRC-Acquis corpus, when using the same MT system. We consider in this section only the in-domain data-sets, i.e. **Test 1**, **Test 2**, **Test 3** and sometimes also **Test 1+2+3** .

The TER, BLEU, and NIST results have been presented in the Tables 9.5, 9.6 and 9.7, respectively. The score variations obtained for **Test 1**, **Test 2** and **Test 3** show how sensitive the empirical MT approaches are to the test data chosen.

A graphical representation of the BLEU scores for Lin-EBMT can be found in Figure 9.3.

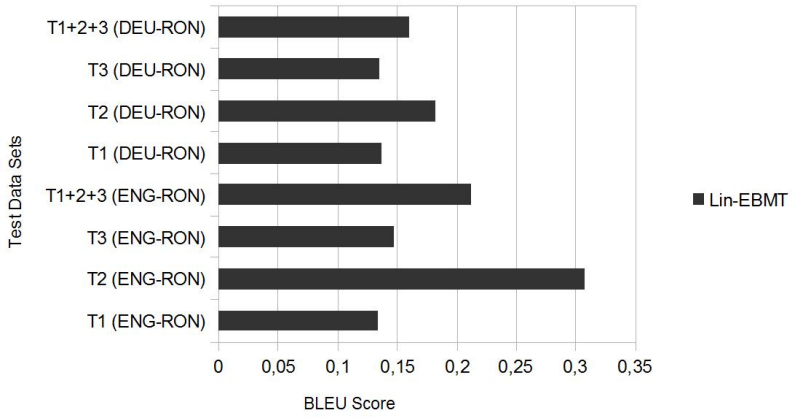

Figure 9.3: JRC-Acquis: BLEU scores (*Lin-EBMT*).

The graphical representation of the BLEU results for **Mb_SMT** and Google is shown in Figure 9.4.

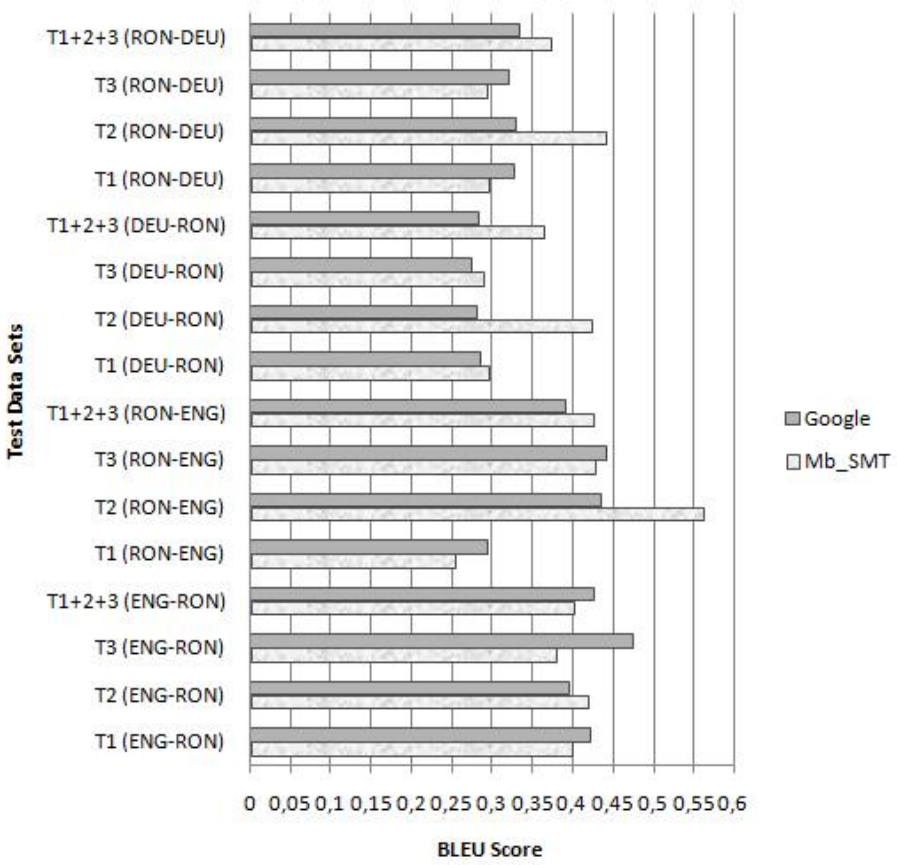

Figure 9.4: JRC-Acquis: BLEU scores (SMT and Google).

The best results are obtained for the test data-set **Test 2** in the case of the **Mb_SMT** system and of *Lin-EBMT*[2], across all automatic evaluation metrics.

For Google Translate, which has been trained on different data, the difference between the results is not as high and the best results are obtained for different test data-sets. The best Google results are obtained for English and Romanian (both directions of translation) for **Test 3** data-set. For Romanian and German the results change for each of the evaluation metric and direction of translation. As the best results differ according to the language pair, it indicates that Google uses different training data for various language-pairs.

Several parameters have influenced the results of the automatic evaluation for the MT systems we implemented or trained, such as the corpus, the way of creating the test data, the sentence length limitation (in the Moses-based translation model) or the variation of the paragraph length in the alignment.

We consider as the first reason for the results obtained with *Lin-EBMT* and **Mb_SMT** the corpus, which contains data from a lengthy time period: 1958-2006. Although the

[2]We chose for *Lin-EBMT* only the language combinations for which we had results for all (in-domain) test data-sets.

terminology might have changed, the languages involved (Romanian, German and English) have not suffered major transformations at the syntactic level.

Changes appeared in the orthography, at least for Romanian and German. An interesting observation for Romanian we extracted while analyzing the vocabulary file, is that for almost all the documents the orthography rules from 1993 have been applied, although orthography changed several times between 1958 and 2006: in 1964, 1993 and 2005. Only in a few cases the changes from 2005 are visible, i.e. the word "*niciunei*" ("none", in dative, having the form for feminine) appears two times, instead of the older form "*nici unei*". This can be explained by a possible later translation of the documents which contain the European regulations before 1993, as well as by scarce data for the documents after the change in 2005. An orthography reform for the German language took place in 1996[3]. The new rules had a strong impact as they were introduced in administration and schools. This change is noticeable also in the data we used for our experiments, e.g. in the corpus we found both forms for the word "*measure*": "*Massnahme*" and "*Maβnahme*". We have not any statistical analysis of the German data in this direction.

With regard to the variation of the paragraph length in the alignment, it was noticed that the one-to-one alignments vary considerably with respect to the length: sub-sentential phrases, simple sentences, complex sentences, etc.

The creation of the test data has also an influence on the results, as the data was extracted from different parts of the aligned corpus. As there is no equal distribution of sentences per year across the corpus, it might be that all sentences related to the EU-regulations from one year are only in the test data.

Out-of-vocabulary words (**OOV-words**) and differences in lexical semantics between years is also a source for the score variation. OOV-words have a direct influence on the translation quality. In order to better understand the results, the test data-sets have been analyzed with regard to the OOV-words[4]. An overview of the OOV-words in the different test data sets of the Experimental setting IIa is shown in Table 9.11[5].

As some of the OOV-words are produced due to spelling errors, the number of OOV-words would decrease if the data could be manually corrected. Therefore, the translation quality could increase. A closer inspection of the out-of-vocabulary words reveals that the words extracted from the RoGER corpus are usually correct (Table 9.12 - Experimental setting I). This happens due to the fact that RoGER has been created and corrected manually and that some words have been replaced with meta-words. In the case of the JRC-Acquis, due to segmentation errors and not-replacement of numbers, dates etc with meta-words, the extracted words are sometimes not correct or they are just numbers: "*2ev*", "*0155*", "***".

[3] `http://en.wikipedia.org/wiki/German_orthography_reform_of_1996` - last accessed on June 23rd, 2011.

[4] The calculation of the number of OOV-words is done comparing the 1-gram files obtained from training and test data, using SRILM (function *ngram-count*, with the parameter *-write1*). The comparison is not case sensitive, as the translation is also not case-sensitive.

[5] The analysis is for the MT systems we trained or implemented; Google Translate is excluded.

Corpus	No. of OOV-Words (% from vocabulary size)	Sentences in the corpus
Data for Experimental Setting IIa		
English-Romanian		
Test 1	33 (3.15%)	69 (23.07%)
Test 2	2 (0.27%)	134 (44.81%)
Test 3	96 (8.64%)	85 (28.42%)
Test 1+2+3	131 (5.59%)	288 (32.10%)
T_RoGER	93 (14.65%)	0 (0%)
Romanian-English		
Test 1	51 (4.10%)	69 (23.07%)
Test 2	7 (0.76%)	117 (39.13%)
Test 3	111 (8.19%)	81 (27.09%)
Test 1+2+3	169 (6.22%)	267 (29.76%)
T_RoGER	330 (37.67%)	0 (0%)
German-Romanian		
Test 1	69 (6.05%)	73 (24.41%)
Test 2	53 (3.68%)	121 (40.46%)
Test 3	187 (14.47%)	83 (27.75%)
Test 1+2+3	309 (10.30%)	277 (30.88%)
T_RoGER	295 (37.72%)	0 (0%)
Romanian-German		
Test 1	44 (3.40%)	76 (25.41%)
Test 2	97 (5.83%)	109 (36.45%)
Test 3	105 (7.56%)	79 (26.42%)
Test 1+2+3	246 (7.37%)	264 (29.43%)
T_RoGER	324 (36.99%)	0 (0%)

Table 9.11: Analysis of the test data sets (Experimental setting II).

Corpus	No. of OOV-Words (% from vocabulary size)	Sentences in the corpus
Data for Experimental Settings I(a+b) English-Romanian		
Test	60 (11.49%)	37 (27.81%)
Test (POS)	74 (13.12%)	37 (27.81%)
Romanian-English		
Test	84 (12.75%)	34 (25.56%)
Test POS	116 (15.78%)	34 (25.56%)
German-Romanian		
Test	101 (16.72%)	31 (23.30%)
Romanian-German		
Test	84 (12.75%)	34 (25.56%)
Data for Experimental Setting III English-Romanian		
Test	72 (7.25%)	38 (28.57%)
Romanian-English		
Test	129 (10.23%)	33 (24.81%)
German-Romanian		
Test	171 (14.51%)	41 (30.82%)
Romanian-German		
Test	160 (12.12%)	40 (30.07%)

Table 9.12: Analysis of the test data sets (Experimental settings I and III).

After a manual analysis of the extracted words for English-Romanian (Experimental setting IIa) and deleting all the words that have been incorrectly extracted, for the **Test 1+2+3** the number of OOV-words decreases to **101**, which means **4.31%** of the vocabulary size. For Romanian-German, a correction was needed due to spelling errors (RON: "*dreptulde*", correct: "*dreptul de*"; ENG: "*the right of*"). We also eliminated numbers from the OOV-list. In this case, the number of OOV-words was reduced to 120 (3.60%). For German-Romanian and Romanian-English, the number of OOV-words after the correction was 266 (8.87%) and 115 (4.23%), respectively. It was also noticed that some of the words are present in both data sets (training and test), although with different inflectional forms. Therefore, a lemmatizer could decrease the number of OOV-words and, indirectly, improve the translation results. Part of the OOV-words are names or numbers, thus a Named Entity Recognizer (**NER**) might have a positive effect on the output. When no NERs are available the inclusion of translation rules represents a solution.

The test scenario was kept as realistic as possible. Therefore, we have not excluded test sentences already in the training corpus: common users do not analyze the texts before translating them. We have also not preprocessed the data more[6] than what it has been indicated at the annual Workshops on SMT. The test data also has no restrictions with

[6]Such as special treatment for numbers.

regard to sentence length. The average sentence-length of the test data is usually higher than that of the training data, for both language pairs (see Table 8.5, **Chapter 8**.).

As the test data is not artificially created, but just extracted randomly from the corpus, it cannot be excluded that some of the test sentences are also part of the training data. The higher the number of such sentences, the better the translation results. The number of the sentences in both the test and training data is presented in Table 9.11 (for the Experimental setting IIa) and in Table 9.12 (Experimental settings I(a+b) and III). As expected[7], the lowest number of OOV-words and the highest number of test sentences included in the training data, for the systems developed during this research, are found in the **Test 2** data-set (Experimental setting IIa).

Although not relevant for this section, but interesting in general, the number of OOV-words and sentences included in both test and training data in the JRC-Acquis$_{SMALL}$ corpus is also shown in Table 9.12 (Experimental setting III).

The reference translation is sometimes wrong (error of the alignment in the corpus) for JRC-Acquis due to the automatic extraction of the test-sets. This has a negative impact on the automatic evaluations scores. Spelling errors, e.g. "*MisterNAME*" (no space) are another reason for a lower result.

The lower BLEU scores can also be explained by the fact that there is a serious flaw with BLEU's reliance on n-gram (surface forms) matching because it penalizes errors which are not strictly translation errors.

9.2.2 Score Variation, when Changing the Corpus

To study the behavior of each of the MT system we implemented (**Lin-EBMT**, $Lin - EBMT^{REC+}$ (**C.1+2+3 1:2**), **Mb_SMT**) on different corpora (JRC-Acquis, RoGER and JRC-Acquis$_{SMALL}$) we compare these systems on **Test 1**, **Test ROGER** and **Test JRC-Acquis_Small**.

Graphical views on the BLEU results are shown in Figures 9.6, 9.7 and 9.5. The BLEU scores for the JRC-Acquis corpus have been presented in Table 9.6, for RoGER in Table 9.4 and for JRC-Acquis$_{SMALL}$ in Table 9.10.

[7]Due to the highest evaluation scores.

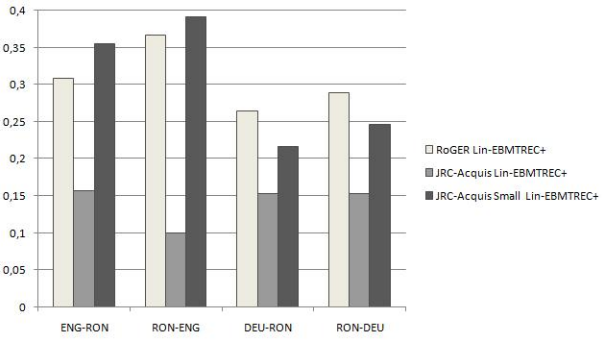

Figure 9.5: Variation of the BLEU scores, when changing the corpus ($Lin - EBMT^{REC+}$).

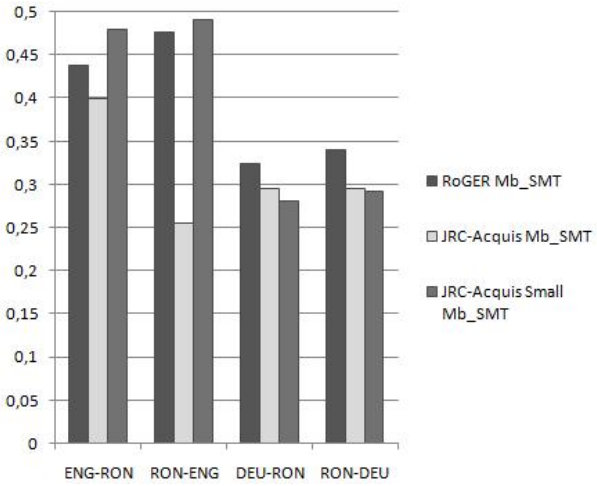

Figure 9.6: Variation of the BLEU scores, when changing the corpus (**Mb_SMT**).

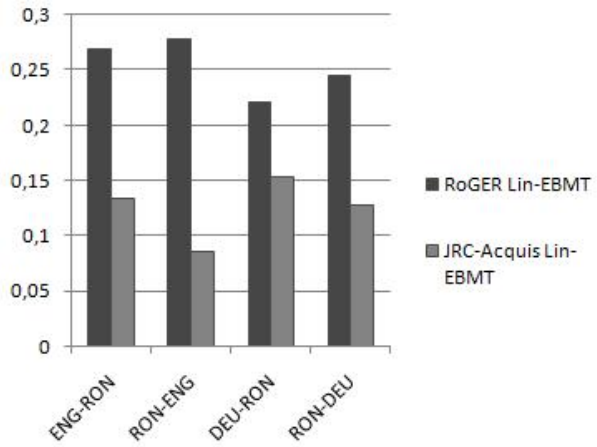

Figure 9.7: Variation of the BLEU scores, when changing the corpus (*Lin-EBMT*).

An improvement is found for all three MT systems and all language combinations for the RoGER corpus, although usually it is stated that a **large** corpus is needed for the

SMT approach. A reason for these positive results for all language combinations might be the corpus type (a manual of an electronic device) and the corpus compilation (it is manually created and corrected).

The results for JRC-Acquis$_{SMALL}$ of the **Mb_SMT** system are the best for English-Romanian (even over the ones for RoGER), but are the worst for German-Romanian.

We found the biggest gains for the smaller-size corpus for both *Lin-EBMT* and $Lin-EBMT^{REC+}$ systems, as opposed to **Mb_SMT**. The EBMT systems improved the results for all language pairs and both smaller corpora. $Lin-EBMT^{REC+}$ has a similar behavior as **Mb_SMT** for Romanian-English.

All these results show (again) how sensitive corpus-based MT approaches are to (test and training) data.

9.2.3 Influence of POS Information on Empirical MT Systems

We analyzed the influence of POS information on the different MT systems, using the RoGER corpus. The results obtained have been presented in Table 9.4. A graphical representation is shown in Figure 9.8.

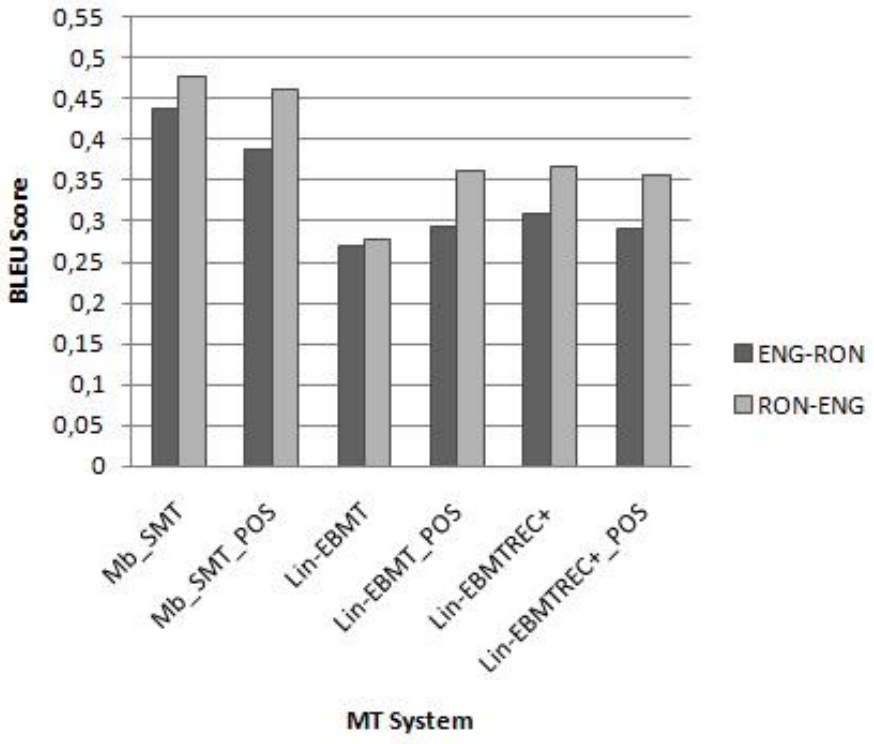

Figure 9.8: Influence of POS on the translation results (RoGER; BLEU scores).

It was noticed that the results for **Mb_SMT**, without POS information are better than the ones of **Mb_SMT**, with POS information. This could be due to an increase of data sparseness.

In the case of the *Lin-EBMT* system results are contradictory. An improvement occurred when POS information was added, when the systems have been evaluated with BLEU and NIST. The improvement could appear due to a more precise matching algorithm. On the contrary, the results are lower, when evaluating with TER.

If the translation is done with $Lin - EBMT^{REC+}\_POS$ (**C.1+2+3 1:2**), the results for the Experimental setting Ib (with POS information) are lower than the ones obtained when no additional linguistic information is used.

There are two reasons for these results: either POS information is affecting negatively the translations (for Mb_SMT and $Lin - EBMT^{REC+}$) or the automatic scores cannot capture the improvement. Therefore, we will manually analyze part of the results in **Chapter 10.**

In this dissertation we did not analyzed how errors in the POS tagging influence the translation results.

9.2.4 Comparing the MT Approaches

The translation results differ according to the evaluation metrics used for the following four MT systems: **Mb_SMT**, Google Translate, *Lin-EBMT* and $Lin - EBMT^{REC+}$ (**C.1+2+3 1:2**).

The evaluation scores have been presented in **Section 9.1**: for the JRC-Acquis corpus Tables 9.6 (BLEU), 9.7 (NIST), 9.5 (TER) and for the RoGER corpus Table 9.4. The comparison of the MT approaches for the JRC-Acquis corpus is shown in Figure 9.9 and for RoGER in Figure 9.10.

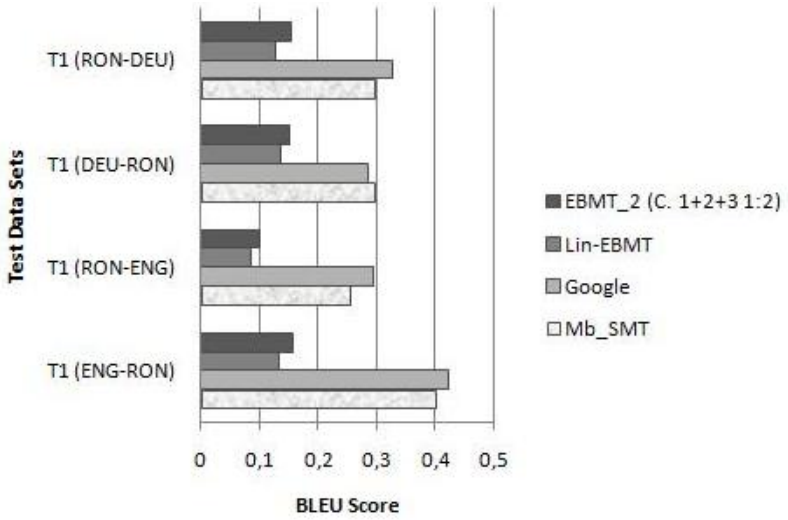

Figure 9.9: JRC-Acquis: BLEU scores.

In terms of overall scores, the SMT systems outperform the EBMT approaches.

We consider first the comparison between **Mb_SMT** and Google Translate, for all four

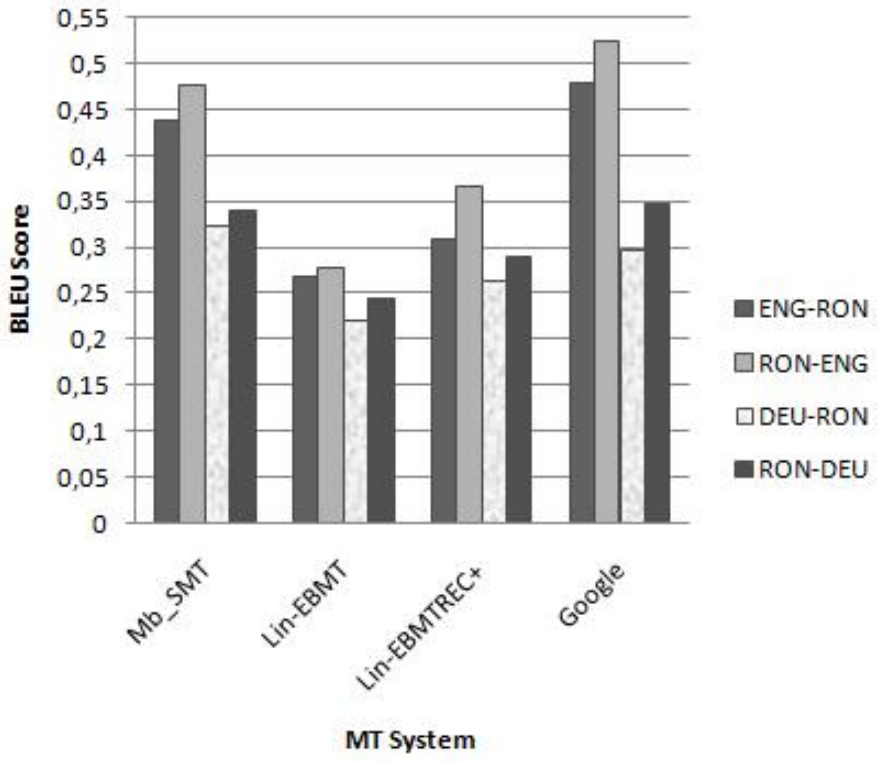

Figure 9.10: RoGER: BLEU scores.

test data-sets from JRC-Acquis, for all language-pairs and directions of translation. The SMT system based on Moses performs very similar to Google Translate. The similar performance with in-domain data somehow confirms the classification of Google Translate on `Wikipedia.org` as an SMT system. It is also highly probable that among the training data in Google also the JRC-Acquis corpus (or similar corpora) is included. As our training data-set is limited, compared to the one available for Google, we expect an increase of performance when using a larger training corpus. For BLEU and TER, in nine cases out of sixteen **Mb_SMT** is better than Google Translate. For NIST the results are mixed: eight cases are in favor of **Mb_SMT** and eight in favor of the Google system. In case of the RoGER test data-set, for English and Romanian, Google is better than the **Mb_SMT** system. For German-Romanian, the **Mb_SMT** system is better. In the case of Romanian-German, Google has better results for BLEU and NIST and lower for the TER score.

For NIST and BLEU metrics, *Lin-EBMT* has the lowest results when compared with the other three MT systems, for both corpora, all language-pairs and directions of translation. In terms of TER, *Lin-EBMT* is worse than **Mb_SMT** and Google Translate for both corpora, and than $Lin-EBMT^{REC+}$ (**C.1+2+3 1:2**) for the RoGER corpus. On JRC-Acquis, the TER scores for *Lin-EBMT* are better than the ones for $Lin-EBMT^{REC+}$ (**C.1+2+3 1:2**). For deciding which EBMT system is better for JRC-Acquis, as the automatic evaluation scores are not always correlating, we will manually analyze part of the results in Chapter 10.

For the **Test 2** data-set with English - Romanian as language-pair, the *Lin-EBMT* BLEU score is similar[8] to the one presented in [Irimia, 2009], where linguistic resources have been used.

$Lin-EBMT^{REC+}$ (**C.1+2+3 1:2**) performs worse compared to **Mb_SMT** and Google Translate.

A comparison between **Mb_SMT** and **SMT_tuning** has been already shown in **Subsection 9.1.2**. A clear conclusion for this case could not be drawn.

9.2.5 Influence of the Language Pair on Empirical MT Systems

Comparing all four MT systems (**Mb_SMT**, Google Translate, *Lin-EBMT* and $Lin-EBMT^{REC+}$ (**C.1+2+3 1:2**)) on the RoGER corpus, the best results have been obtained for Romanian-English, followed by English-Romanian. For German and Romanian the results are worse. They also differ, depending on the evaluation metric. While for BLEU and NIST the scores for Romanian-German are better than the ones for German-Romanian, for TER the relationship is reversed. The same behavior is encountered also for the JRC-Acquis$_{SMALL}$ corpus.

Evaluation on JRC-Acquis has only been done with **Test 1**, as only for this data set results are available for all language-pairs and MT systems. Again, the results are different for all four systems, for the three evaluation metrics.

An overview of the results is given in Table 9.13. We compared the results for JRC-Acquis **Test 1**, for each language pair and direction of translation. The three values in a cell of the table represent the ranking for BLEU, NIST and TER, in this specific order (BLEU / NIST / TER). The values from 1 to 4 represent the ranking positon of the system according to a score for the four language combinations: 1 the highest evaluation score, 4 the lowest evaluation score.

MT System	ENG - RON	RON - ENG	DEU - RON	RON - DEU
Mb_SMT	1 / 1 / 1	4 / 4 / 2	2 / 2 / 3	3 / 3 / 4
Google	1 / 1 / 2	3 / 4 / 1	4 / 3 / 4	2 / 2 / 3
Lin-EBMT	1 / 1 / 2	4 / 4 / 1	2 / 2 / 3	3 / 3 / 4
EBMT_2	1 / 1 / 2	4 / 4 / 1	2 / 2 / 3	3 / 3 / 4

Table 9.13: Influence of the language pair (JRC-Acquis).

We calculate an average value for each language pair. Best results are obtained for English-Romanian (average: 1.25), followed by German-Romanian (2.66) and Romanian-English (3). Romanian-German has the lowest average: 3.08. Overall for all systems, evaluation metrics and corpora best results were acquired as expected for the language-pair Romanian-English. The results are lower for the case when both SL and TL are inflected

[8] A one-to-one comparison is not possible, as the (test and training) data is not the same.

languages. Analyzing the results for RoGER and JRC-Acquis$_{SMALL}$ it can be concluded that matching for Romanian creates less problems as recombination on Romanian.

However, the results are depending on the data (see Table 9.14.) We calculated the values in Table 9.14, by considering the same approach as in Table 9.13 for all corpora. The values represent the ranking according to the average values.

MT System	ENG - RON	RON - ENG	DEU - RON	RON - DEU
JRC-Acquis	1	3	2	4
RoGER	2	1	4	3
JRC-Acquis$_{SMALL}$	2	1	4	3
Average	1.66	1.66	3.33	3.33

Table 9.14: Influence of the language pair. Comparison for all corpora.

Only the results for JRC-Acquis are somehow unexpected, as the results for German-Romanian are better than for Romanian-English. These results could be explained by problems in matching for Romanian for this specific test data-set or by bias of the automatic evaluation metrics chosen in this dissertation. We will further analyze part of the data manually in **Chapter 10** and **Appendix G**.

A graphical representation of the BLEU results is shown in Figure 9.11.

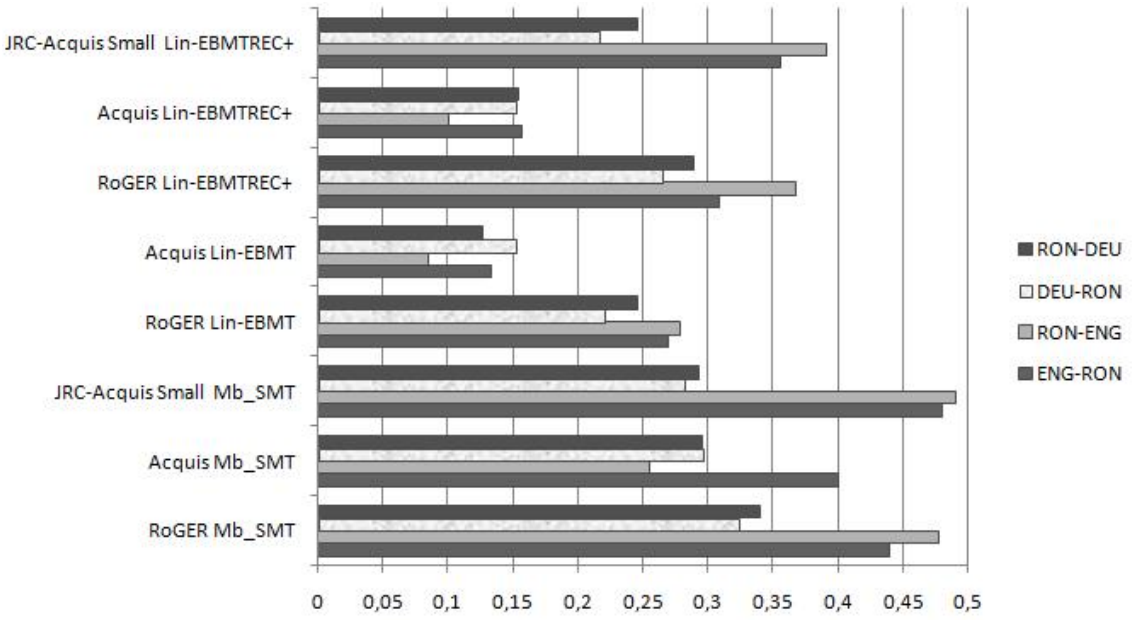

Figure 9.11: Changing the language-pair (BLEU scores, JRC-Acquis).

9.2.6 Testing with Out-of-domain Data

For the out-of-domain test data set, only the **Mb_SMT** system is used. As expected, the percentage of out-of-vocabulary words for **T_RoGER** is higher than the result for the

in-domain data set – see Table 9.11. Moreover, no test sentences are found in the training data.

The results for **Mb_SMT** are all considerably below the ones for in-domain data sets. Table 9.15 shows the BLEU scores for in-domain and out-of-domain data.

-	Mb_SMT		Google	
Language pair	**Test 1+2+3**	**T_RoGER**	**Test 1+2+3**	**T_RoGER**
English-Romanian	0.4015	0.0623	0.4263	0.3332
Romanian-English	0.4255	0.0621	0.3909	0.4543
German-Romanian	0.3644	0.0357	0.2837	0.2165
Romanian-German	0.3726	0.0271	0.3332	0.3031

Table 9.15: Comparison of the BLEU results: in-domain vs. out.of-domain test data.

For comparison reasons, the same data has been translated with Google. The Google results for **T_RoGER** are similar to the ones obtained for the JRC-Acquis test data. In three cases out of four the BLEU results of Google Translate for **T_RoGER** are lower than the ones obtained for the **Test 1+2+3** data set, but the differences are significantly smaller than the ones for the **Mb_SMT** system. Only in the case of Romanian-English the Google results are higher for **T_RoGER** than the ones for the JRC-Acquis data test.

On this data-set the performance of Google Translate is much better than the one of **Mb_SMT**. But Google Translate cannot be considered a reliable comparison as the system evolves dynamically by contributions of users and there is no detailed information about the architecture of the system. It is estimated that the training data is huge, comparable to the one used for the experiments reported in [Callison-Burch et al., 2009]. The Google BLEU score is very similar to the one in [Callison-Burch et al., 2009] when changing the domain.[9]

The Google Translate system is less sensitive to different types of data, i.e. the scores are close to each other for test data-sets from both corpora. This cannot be stated about **Mb_SMT**. In conclusion, the availability of a large training data (from different domains) could increase the coverage of a Moses-based SMT system.

As the BLEU scores for **Mb_SMT** are between 0.02 and 0.06 and the SMT systems usually outperformed the EBMT systems, we did not run the same experiments for the EBMT systems.

9.3 Chapter Summary

In this chapter we presented the experimental settings, the automatic evaluation scores obtained with BLEU, NIST and TER metrics and a first interpretation of the results. The

[9] In [Callison-Burch et al., 2009], experiments were run on English-French for a training corpus at least six times larger than the one available to us.

results show the high sensitivity of the corpus-based MT approaches to the (training and test) data.

To gain a better overview on the translation quality, a manual analysis of some of the translations will be presented in **Chapter 10**.

Chapter 10

Manual Analysis of the Results

We have presented in **Chapter 9** the first observations on the translation results, which can be extracted from the automatic evaluation scores. In this chapter, the methodology of the human analysis of the output will be described, followed in **Section 10.2** by the extracted conclusions. We will show the results of a manual analysis of the translations provided by the MT systems we implemented in this work.[1] This way, we get a better overview of the strengths and weaknesses of a specific corpus-based machine translation system.

10.1 Human Analysis: The Methodology

The aim of the manual analysis is to examine which phenomena have a negative influence on the scores of the automatic metrics, to evaluate how automatic results really correlate with translation quality (human judgment) and to determine the sources and types of errors for each of the MT system. Also a ranking of the systems is realized. Due to financial and time restrictions, we were not able to use several independent evaluators. Therefore the manual analysis was done only by the author.

The analysis of the translations requires an excellent knowledge of the (target) language, as specific morphological, syntactic and semantic errors need to be extracted. Therefore, we restricted the language-pairs to those which have Romanian as the target language, i.e. German-Romanian and English-Romanian. We used two corpora (JRC-Acquis and RoGER) and three MT systems – **Mb_SMT**, *Lin-EBMT*, and $Lin - EBMT^{REC+}$ (**C.1+2+3 1:2**) – in the analysis. For the English - Romanian RoGER data we also analyzed the influence of POS information on the translation results.

We investigated approximately 30% of the test data: 100 sentences from **Test 1** data-set and 50 sentences from **Test RoGER** data-set. We have considered two criteria in the process of choosing the test data sets for the manual analysis: its availability for all MT systems analyzed and an intermediate automatic evaluation score.

[1] Due to the language pair the **SMT_tuning** system was left aside in this analysis.

10. MANUAL ANALYSIS OF THE RESULTS

We adopted the criteria of a black-box user evaluation as described in [Dorr et al., 1999]. The manual analysis focuses on two aspects:

- System ranking;
- Sources and types of translation errors.

In order to rank the translations of the three MT systems, the translation output was analyzed from the point of view of **adequacy** and **fluency**[2]. Adequacy refers to the degree to which information in the original sentence is also communicated in the translation. Fluency refers to the degree to which the target sentence is well-formed according to the rules of the target language. The ranking of the MT systems is done following the instruction found in [Callison-Burch et al., 2009], i.e. "*Rank translations from Best to Worse relative to the other choices (ties are allowed).*" (1 = Best, 3 = Worst). Being ranked first, it does not mean that the system has a perfect translation; it only means that it provides the best translation, compared to the other systems.

In the analysis of the second aspect (sources and types of translation errors) we included morphological, syntactic and semantic errors, as well as other mistakes with minor impact on the understanding of the content, such as punctuation. The types of errors are split into three categories, according to their impact on the translation results:

Category I : high (negative) impact

(a) Translation incomplete *(e.g. text is totally missing; there are more than two OOV-words)*;

(b) Ungrammatical translation, translation not understandable *(e.g. severe word order problems, wrong prepositions, errors with impact on meaning)*;

(c) Wrong translation, in which the semantics of the source language is not preserved, as *the translation contains information not found in the input which changes the meaning of the initial sentence*;

Category II : moderate (negative) impact

(a) Translation incomplete *(e.g. maximum two OOV-words)*;

(b) Ungrammatical translation *(e.g. minor word order problems, wrong inflection, wrong prepositions, but with no real impact on the meaning)*;

Category III : (almost) no (negative) impact

(a) Wrong punctuation, wrong capitalization

(b) Additional information in the TL, with no real influence on the semantics, *e.g. a word appears twice.*

[2]Adequacy and fluency appeared the first time in the ARPA (Advanced Research Projects Agency) methodology, at the beginning of the 1990s. They are also used in more recent papers, such as [Callison-Burch et al., 2007].

Errors are counted only once in each of the categories, even if they appear more times in a sentence. Several types of errors can be found in a sentence.

There are several sources of errors in the translation, such as a process of the MT system or the data itself (e.g. errors in the corpus, the word alignment, the pre- and post-processing of the data).

The translation can be incomplete due to the OOV-words[3] or due to a wrong or absent word alignment (**Category I.(a)** or **II(a)**). For **Mb_SMT** and $Lin - EBMT^{REC+}$ the encountered OOV-words are left in the translation in the initial source language.

Errors of **Category I.(c)** can be induced by polysemantic words and wrong word alignments. Word order problems – **Category I.(b) or II.(b)** – can arise by adjective-noun inversions, a wrong position of the verb or in prepositional phrases.

Category II.(b) includes agreement errors, wrong tense or mood for verbs[4], wrong article, wrong POS, wrong case or case formation (e.g. genitive). The following agreement errors can be found: the agreement between subject and predicate (person and number), adjective or article and noun (number, case and gender) or adverb and verb (e.g. a temporal adverb of the past and a verb in the future tense).

10.2 The Results of the Human Analysis

Before analyzing the output, we compared the tokens[5] of the translations with those in the references. The results are shown in Table 10.1 in which "*Common tokens*" (**CT**) are tokens which the reference and the translation have in common and "*Ordered common tokens*" are common tokens between the translation and its reference, which have the same order in both sentences. For example, the following two sentences:

(1) I decided **to go** home **by** bus
We **go to** the theater **by** car.

have three "*common tokens*" (*to*, *go*, *by*) and two "*ordered common tokens*" (*go*, *by*).

The percentage values in Table 10.1 are calculated from the total number of tokens in the reference translation. The results for **Mb_SMT** are closer to the reference translation. When we use *Lin-EBMT* and $Lin - EBMT^{REC+}$, the results are again inconclusive: in six cases out of ten the results for $Lin - EBMT^{REC+}$ are better than those of *Lin-EBMT* – see boldface numbers in Table 10.1.

The following subsections discuss the results of the human analysis for the system ranking and the sources and types of translation errors.

[3] An overview of the percentage of the OOV-words, for all test data-sets is described in Table 9.11.
[4] Only if the semantics of the sentence does not change.
[5] In this context token means word, number or punctuation sign.

Description	Reference	*Lin-EBMT*	Mb_SMT	$Lin - EBMT^{REC+}$
DEU-RON, JRC-Acquis				
Total	1177	1097	1252	1226
Common tokens (CT)	-	602 (51.15%)	790 (67.12%)	621 (**52.76%**)
Ordered CT	-	452 (**38.40%**)	760 (64.57%)	428 (36.36%)
ENG-RON, JRC-Acquis				
Total	1252	1151	1370	1308
Common tokens	-	695 (**55.51%**)	960 (76.68%)	664 (53.04%)
Ordered CT	-	471 (**37.62%**)	918 (73.32%)	457 (36.50%)
DEU-RON, RoGER				
Total	495	361	464	414
Common tokens	-	225 (45.45%)	285 (57.58%)	252 (**50.91%**)
Ordered CT	-	184 (37.17%)	273 (55.15%)	209 (**42.22%**)
ENG-RON, RoGER				
Total	495	430	490	466
Common tokens	-	282 (56.97%)	352 (71.11%)	302 (**61.01%**)
Ordered CT	-	230 (46.46%)	343 (69.29%)	244 (**49.29%**)
ENG-RON POS, RoGER				
Total	490	461	472	480
Common tokens	-	258 (**52.65%**)	273 (55.71%)	257 (52.45%)
Ordered CT	-	205 (41.84%)	267 (54.49%)	211 (**43.06%**)

Table 10.1: Comparison between the translations and their references.

10.2.1 System Ranking

In ranking the systems adequacy and fluency have been indirectly involved. Although not fully relevant – as only one human evaluator was available- -, but still with possible impact on further research, the average results for adequacy and fluency are presented in Table 10.2. The evaluation scale for adequacy and fluency is the one described in [LDC, 2005]:

Adequacy: 1=None, 2=Little, 3=Much, 4=Most, 5=All.

Fluency: 1=Incomprehensible, 2= Disfluent, 3=Non-native, 4=Good, 5=Flawless

The results in Table 10.2 reconfirm the fact that **Mb_SMT** outperforms the EBMT systems. $Lin - EBMT^{REC+}$ performs better than *Lin-EBMT* for German-Romanian (both corpora). Similar results happen for English-Romanian when the data contains additional part-of-speech (POS) information.

The ranking results obtained for the three MT systems are shown in Table 10.3. From this information, we computed the percentage of cases corresponding only to the first place. The results achieved are presented in Table 10.4.

Analyzing only the data in the Table 10.4, it can be concluded that **Mb_SMT** is the first in most of the cases, result which is similar to the one obtained in the automatic evaluation.

Evaluation	**Mb_SMT**	*Lin-EBMT*	$Lin - EBMT^{REC+}$
ENG-RON, JRC-Acquis			
Adequacy	4.6	3.88	3.81
Fluency	4.26	3.37	3.38
DEU-RON, JRC-Acquis			
Adequacy	4.16	3.52	3.53
Fluency	4.07	3.23	3.31
ENG-RON, RoGER			
Adequacy	4.22	3.72	3.64
Fluency	4.08	3.5	3.44
DEU-RON, RoGER			
Adequacy	3.64	3.3	3.32
Fluency	3.54	3.2	3.2
ENG-RON POS, RoGER			
Adequacy	4.1	3.6	3.66
Fluency	3.74	3.14	3.3

Table 10.2: System analysis: adequacy and fluency (average values).

Place	**Mb_SMT**	*Lin-EBMT*	$Lin - EBMT^{REC+}$
ENG-RON, JRC-Acquis 1st	90	48	40
ENG-RON, JRC-Acquis 2nd	5	29	51
ENG-RON, JRC-Acquis 3rd	5	23	9
DEU-RON, JRC-Acquis 1st	93	47	47
DEU-RON, JRC-Acquis 2nd	7	28	38
DEU-RON, JRC-Acquis 3rd	-	25	15
ENG-RON, RoGER 1st	50	26	26
ENG-RON, RoGER 2nd	-	22	21
ENG-RON, RoGER 3rd	-	2	3
DEU-RON, RoGER 1st	41	25	27
DEU-RON, RoGER 2nd	7	23	19
DEU-RON, RoGER 3rd	2	2	4
ENG-RON POS, RoGER 1st	47	24	24
ENG-RON POS, RoGER 2nd	2	20	25
ENG-RON POS, RoGER 3rd	1	6	1

Table 10.3: System ranking (The values represent the number of times the system finds itself on the specified place.).

Data	Mb_SMT	*Lin-EBMT*	$Lin - EBMT^{REC+}$
DEU-RON RoGER	82%	50%	54%
ENG-RON RoGER	100%	52%	52%
ENG-RON RoGER POS	94%	48%	48%
DEU-RON JRC-Acquis	93%	47%	47%
ENG-RON JRC-Acquis	90%	48%	40%
All RoGER (no POS)	91%	51%	53%
All JRC-Acquis	91.5%	47.5%	43.5%
All DEU-RON	89.33%	48%	49.33%
All ENG-RON (no POS)	93.33%	49.33%	44%

Table 10.4: System ranking (First place) (The value is given in % of the total number of analyzed sentences).

$Lin - EBMT^{REC+}$ acquires better or similar results to those of *Lin-EBMT* on the RoGER data for the German - Romanian language pair. *Lin-EBMT* provides better results for English-Romanian on JRC-Acquis data. The overall results for RoGER and JRC-Acquis generally confirm the TER results presented in **Subsection 9.2.4**, **Chapter 9**.

Further ranking results are shown in **Appendix G**.

10.2.2 Sources and Types of Translation Errors

After manually analyzing the translations, we found several causes that negatively influenced the automatic evaluation scores, such as:

- A wrong translation as reference, probably due to paragraph alignment errors. We encountered this phenomenon only in the JRC-Acquis data for German - Romanian in 6% of the cases.
- An inexact translation in the reference, e.g. a translation using a noun in plural, although in the input it is singular.

Sometimes there is additional information in the reference translation, that does not specifically appear in the SL.

We also noticed that some of the translations were understandable from a human point of view, but these translations reformulated the reference translations (e.g. passive voice translated as active voice) or they contained small differences (e.g. an additional or a missing preposition, a different article, use of synonyms or different punctuation) – see Examples (2) and (3). As the automatic evaluation metrics are based on n-grams and surface forms, these aspects lead to a decrease of the automatic scores. Reformulations of the reference sentences have been encountered more frequently in the EBMT translations.

(2) **Input:** "*Verteilerlisten*" (ENG:*Distribution lists*)
Reference: "*Liste **de** distributie*"

Mb_SMT: "*Liste distributie*" The preposition "*de*" is missing. The translation is perfectly understandable, but the syntax is not fully correct.
Lin-EBMT*, $Lin - EBMT^{REC+}$:** "*Liste* ***de *distributie*"

(3) **Input:** *This menu is shown only if any info messages are received .*
Reference: *Acest meniu este afisat numai daca sunt receptionate mesaje informative .*
Mb_SMT: *Acesta meniu este afisat daca oricare masaje informative sunt primite .*
(ENG *The menu is shown only if any info messages are received .*)
Lin-EBMT: *Meniul este afisat numai daca primiti mesaje informative un acest*
(ENG *The menu is shown only if you receive info messages a this*)
$Lin - EBMT^{REC+}$: *Acest se afiseaza numai daca primiti mesaje informative .*
(ENG *This is shown only if you receive info messages .*)

In Example (3) it can be noticed that both SMT and EBMT systems use synonymous constructions, as "*receptionate*" (reference, ENG: "*received*") vs. "*primite*" (**Mb_SMT**, ENG: "*received*") or vs. the verb "*primiti*" (Lin-EBMT, $Lin - EBMT^{REC+}$, ENG: "*receive*") ;"*este afisat*" (reference, **Mb_SMT**, *Lin-EBMT*, ENG: "*is shown*") vs. "*se afiseaza*" ($Lin - EBMT^{REC+}$, ENG: "*shows itself*"). While the SMT output usually follows the SL syntax strictly, the EBMT systems reformulate the translation. All three systems introduce errors into the translation. Another example for synonyms used in the translation is the word "*regulations*', which is translated in the reference as "*regulamente*" and in the MT outputs as "*reglementările*".

An overview of the perfect translations (i.e. identical to the reference) and of the correct translations found in the analyzed data is shown in the Tables 10.5 and 10.6. By '*correct*' we mean correct from the point of view of the adequacy and fluency, but different from the reference translation. This type of translations are some kind of reformulations of the references.

System	DEU - RON	ENG - RON	ENG - RON with POS
Perfect translation			
Lin-EBMT	12	16	13
Mb_SMT	11	18	14
$Lin - EBMT^{REC+}$	12	17	14
Different, but correct translation			
Lin-EBMT	7	5	5
Mb_SMT	9	7	7
$Lin - EBMT^{REC+}$	7	5	5
Total			
Lin-EBMT	19 (38%)	21 (42%)	18 (36%)
Mb_SMT	20 (40%)	25 (50%)	21 (42%)
$Lin - EBMT^{REC+}$	19 (38%)	22 (44%)	19 (38%)

Table 10.5: RoGER: sentences translated correctly.

System	DEU - RON	ENG - RON
Perfect translation		
Lin-EBMT	30	35
Mb_SMT	32	37
$Lin - EBMT^{REC+}$	30	34
Different, but correct translation		
Lin-EBMT	9	7
Mb_SMT	13	24
$Lin - EBMT^{REC+}$	7	2
Total		
Lin-EBMT	39 (39%)	42 (42%)
Mb_SMT	45 (45%)	61 (61%)
$Lin - EBMT^{REC+}$	37 (37%)	36 (36%)

Table 10.6: JRC-Acquis: sentences translated correctly.

Between 36% and 50% of the RoGER sentences have a syntactically and semantically correct translation. The scores vary between 36% and 61% for the JRC-Acquis corpus. However, these JRC-Acquis scores are not fully relevant as, due to the used paragraph aligner, around 50% of the paragraphs are in fact NP- or VP-chunks with less than five words. Because of this, we cannot consider them as complete sentences.

As stated previously, most of the NP- or VP-chunks with just a few words are translated correctly by all systems. Exceptions are generated by OOV-words, which are not translated. The more complex the SL sentence is, the more problematic the translation is for all three MT systems.

In some cases the system provided no translation. For German - Romanian there were two cases for the RoGER data and seven[6] for the JRC-Acquis corpus, when the SMT output contains only SL words (untranslated text). For *Lin-EBMT* and $Lin - EBMT^{REC+}$ only one case in the RoGER data was found.

Excluding the sentences mentioned in Table 10.5, for the rest of the sentences, the SMT approach provides a better translation. In the EBMT systems, syntactic errors (e.g. word-order errors, which might appear due to recombination step) or vocabulary errors (which appear due to alignment and/or matching) have been most frequently encountered.

An overview of the the number of error cases found in the manual analysis is shown in Figure 10.1. The errors are included in **Categories I** and **II**. The values presented are normalized to the total number of sentences analyzed.

[6]It was seven times the same input.

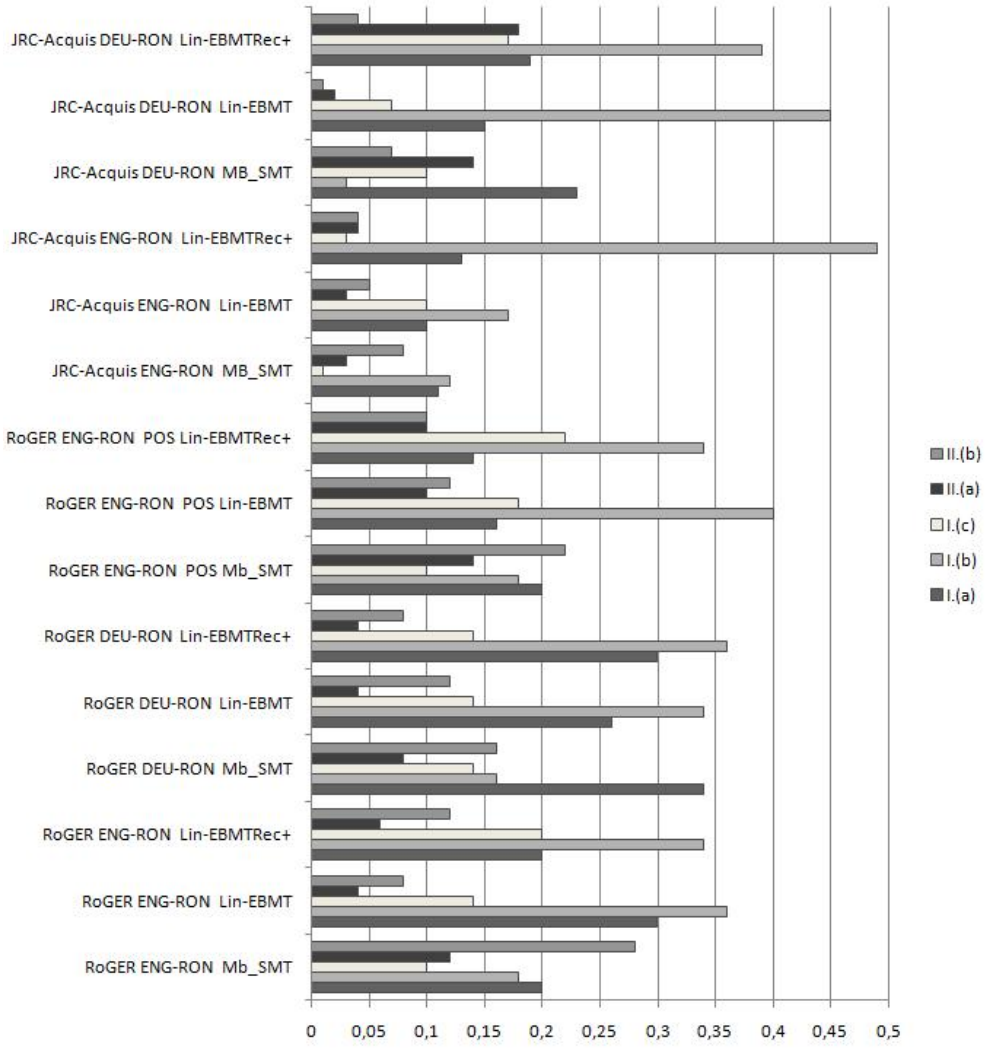

Figure 10.1: Errors in Categories I and II (*normalized values*).

Still, there are cases where the EBMT system provides a better translation, in which the output of the EBMT systems has a better syntax than the SMT translation (see Examples (4) and (5)).

(4) **SL:** *The EEA Joint Committee*
TL reference: *Comitetul mixt al SEE,*
SMT output: *SEE Comitetului mixt,*
(* ENG *EEA of the Joint Committee*)
***Lin-EBMT* output:** *Comitetului mixt SEE*
(* ENG: *of the EEA Joint Committee*)
$Lin-EBMT^{REC+}$ **output:** *Comitetului comun SEE,*
(* ENG: *of the EEA (Common)/Joint Committee*)

(5) **SL:** *Uebersicht ueber die Telefonfunktionen*
TL reference: *Prezentarea functiilor telefonului*

(ENG *Overview of the functions in the telephone*)
SMT output: *Prezentarea functiilor*
(ENG *Overview of the functions*)
***Lin-EBMT* output:** *Prezentarea functiilor telefonului*
$Lin - EBMT^{REC+}$ **output:** *Prezentarea functiilor telefonului*

Sometimes, the SMT system leaves a specific NP untranslated, considering it an OOV-word, although the EBMT systems find its translation. Such an example is the German word "*nummerntaste*" (ENG: "*the number key*") which is translated into Romanian by the EBMT systems as "*tasta numerica*", but is not translated by **Mb_SMT**.

For the cases under analysis, the errors found mostly fall into **Category II.(b)**: agreement and inflection problems, minor word order problems, wrong prepositions (e.g. "*in das Abkommen*" ("*into the Agreement*") translated as "*din acord*" ("*from the Agreement*")) or missing articles. There are also frequent **Category I.(a)** errors. In many cases, the **Mb_SMT** system does not translate some words or performs a word-for-word translation. This results in a translation where all of the words are translated, but one cannot always understand what the TL sentence actually means. Problems arose also in translating noun phrases of the form '**Adj.**–**NN**' or '**NN**–**NN**': e.g. "*Successful management*" as "*castigator gestionarea*" (ENG: "*winner management*"), "*Research fund*" as "*cercetare fondului*" (ENG: "*Research of the fund*"), instead of "*fond de cercetare*". Other problems which cannot be processed by the SMT system are multiple adjectives before a noun and subsentential chunks which have less similarity with the training data. Long dependencies are sometimes completely lost by the SMT system: occasionally the verb is no longer present.

Ungrammatical translations are produced more often by the EBMT systems, in which semantic or syntactic problems appear due to the word order and additional words which have nothing in common with the input. While the first error type is most likely introduced during the recombination step of the EBMT[7], the second appears due to the word-alignment or to a wrong matching decision in the EBMT system[8]. On the other hand, the EBMT systems produce translations which are usually rephrases of the references.

The errors of the EBMT systems mostly fall into **Category I**: word-order problems, wrong semantics as additional information is added to the translation, with no connection to the SL sentence, or information is fully missing.

While manually analyzing the results, we have noticed that a broad spectrum of problems can be identified for both language-pairs, when looking at complete sentences (and not only NP- or VP-chunks), such as wrong semantics and syntax or style problems. The types and frequency of the errors differ between the language pairs. There are more OOV-words for German-Romanian. This is due to the complexity of the German language: verb with

[7]Such errors sometimes also appear in the SMT translations and they are probably introduced by the LM.

[8]Such errors sometimes also appear in the SMT approach and they are probably introduced by the TM.

separated particle, more compound tenses, compounds, etc. This also has a direct influence on the word alignment with GIZA++. It was also noticed that for German-Romanian the verb in the translation is sometimes missing. This might be due to the German syntax (e.g. the distance between the auxiliary verb and main verb or the structure of the subordinate sentences) and word-alignments problems. For English-Romanian, it was observed that errors appear more often for prepositional phrases (PPs), noun-phrases (NPs) and for the nouns in genitive.

Some errors already enumerated (e.g. OOV-words) appear because of the limited training data. Due to the German compounds and syntax, an important source of errors is the word alignment. These errors can be solved by adding more data or a bilingual dictionary. The aspect of OOV-words is more interesting for German, as sometimes compounds or parts of compounds are not found. Among the OOV-words we found for example "*Forschungsfonds*" (ENG: "*Research funds*"); "*anpassungsprotokoll*" ("*the Protocol adjusting...*") translated as "***protocolul*** *anpassungs****protokoll***" instead of "*protocolul de adaptare*" (only half of the compound was found in the corpus, the other half being an OOV-word).

Errors from the **Category III (a)** do not appear as often in the translations of the **Mb_SMT** system, but they represent quite a common error for both EBMT systems. While *Lin-EBMT* almost[9] completely ignores punctuation, in $Lin - EBMT^{REC+}$ punctuation marks are usually in a wrong position. A different way of treating punctuation (e.g. rules, ignoring it in the main steps and adding it in the end or using priorities) could reduce the frequency of this error type. The errors of the SMT system are usually connected to the position of the parentheses ("*)*") and with the absence of the "*/*" character. Mistakes of the **Category III (b)** are encountered scarcely in the MT approaches under investigation.

For JRC-Acquis, the EBMT approaches had severe problems with translating numbers (i.e. law and paragraph numbers) and proper nouns (i.e. names of institutions). That is why it would be recommended either Named Entity Recognizers[10] (**NERs**) or some kind of translation rules should be integrated into the system.

The impact of the constraints in $Lin - EBMT^{REC+}$ could not be exactly determined by the manual analysis, the only relevant difference being noticed only when C.1 constraints have been applied. As stated, both systems have problems with word-order. However, a one-to-one comparison cannot be made, as the matching procedures are different.[11] Still, there are cases where the $Lin - EBMT^{REC+}$ results are better, such as the ones in Examples (6) and (7)

(6) **SL:** *To add a folder , press Options and select Add folder .*
TL reference: *Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .*

[9]The term "*almost*" is used, as sometimes punctuation marks appear in the translation, probably due to the word-alignment step.

[10]When available.

[11]One approach does not consider punctuation marks.

SMT output: *Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .*
***Lin-EBMT* output:** *Pentru a adauga un apasati Optiuni si selectati Adaugati dosarul dosar*
(* ENG *To add a press Options and select Add folder folder*)
$Lin-EBMT^{REC+}$ **output:** *Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .*

(7) **SL:** *(2) Die Entscheidung 2002/272/EG der Kommision vom 25.März 2002 zur Festlegung der Umweltkriterien für die Vergabe des Umweltzeichnis der Gemeinschaft für harte Bodenbeläge ist in das Abkommen aufzunehmen -*
(ENG *Commission Decision 2002/272/EC of 25 March 2002 establishing the ecological criteria for the award of the Community eco-label to hard floor-coverings(2) is to be incorporated into the Agreement,*)
TL reference: *Decizia Comisiei 2002/272/CE din 25 martie 2002 de stabilire a criteriilor ecologice pentru atribuirea etichetelor ecologice comunitare pentru pardoseli rigide2 trebuie încorporată în Acord,*
SMT output: *Decizia 2002 / 272 / CE a Comisiei din 25 martie 2002 de stabilire a criteriilor ecologice de acordare a etichetei ecologice comunitare pentru îmbrăcăminţile de pardoseală rigide (2) din acordul,*
(* ENG *Decision 2002/272/EC of the Commission of 25 March 2002 establishing the ecological criteria for awarding the Community eco-label to hard floor-coverings(2) from the Agreement,*)
***Lin-EBMT* output:** *2 declarare fi introduse în ce din Decizia Comisiei să permită introducerea procedurii 25 martie 2002 este de stabilire a criteriilor ecologice de atribuire a etichetei ecologice comunitare pentru acordurile 2002 272 îmbrăcăminţi de pardoseală rigide*
(* ENG *2 declaration be introduced in what from Commision Decision to allow the introduction of the procedure 25 March 2002 is to establishing the ecological criteria for the award of the Community eco-label for agreements 2002 272 hard floor-coverings*)
$Lin-EBMT^{REC+}$ **output:** *Decizia Comisiei în acordul de stabilire a criteriilor ecologice de atribuire a etichetei ecologice comunitare pentru îmbrăcăminţi de pardoseală rigide / 25 martie / CE din 2002. 2002*
(ENG *)*Commission Decision in the agreement for establishing the ecological criteria for the award of the Community eco-label to hard floor-coverings / 25 March / CE 2002. 2002*)

10.3 Chapter Summary

In this chapter we described a human analysis methodology and some of the observations made. As the analysis was carried out by a single human judge, the conclusions obtained can only be considered as a rough guide for further experiments and extensions of the systems.

Overall, the SMT approach outperformed the EBMT one. Still, there are cases when the output of the EBMT systems were better in comparison with the SMT translations. Putting together the good sides of both corpus-based approaches could represent a solution for improving the translation results, i.e. a hybrid machine translation approach.

Chapter 11

Conclusions

This final chapter contains our conclusions, an overview of the contributions and limitations of the study and perspectives for future work.

The first aim of this dissertation was to investigate the influence of word-order constraints on the translation results. We integrated the constraints in the recombination step of the linear baseline EBMT system (i.e. *Lin-EBMT*) we implemented. The system developed this way is $Lin - EBMT^{REC+}$. The constraints have been extracted using information from the template-based EBMT approach. For our experiments we used two language pairs, in both directions of translation: Romanian - English and Romanian - German. As we analyzed an under-resourced language, the systems we trained and developed have been kept as resource-free as possible, the implemented algorithms being based mainly on surface forms and corpus statistics.

The second main goal was to explore how example-based machine translation can be used when translating into or from an inflected under-resourced language, in this case Romanian. Since over the last few years the research community has concentrated its work more on the SMT approach, we compared our EBMT results to SMT ones. We used two parallel aligned corpora, of different sizes: a larger corpus (JRC-Acquis), closer to the SMT specifications and a smaller one (RoGER), which better fits the EBMT environment. To confirm the results obtained with the small-size corpus, a second small-size corpus (JRC-Acquis$_{SMALL}$) was added in our experiments. In some of the cases, results are compared to Google Translate, as this is the on-line MT system most widely used. Additionally we tested how POS information influences the translation results only for one corpus and one language pair (i.e. Romanian - English).

11.1 Contributions

In order to achieve the previously mentioned goals, several tasks have been done. These represent the contributions of this work:

Implementation of a linear EBMT system: *Lin-EBMT*

As at the moment of starting this research no open-source EBMT system was available, we implemented from scratch a baseline EBMT system: *Lin-EBMT*. The baseline system is platform- and language-pair independent, provided that a parallel aligned corpus for the language-pair exists and that the tools used for obtaining the needed intermediate information (e.g. word-alignment, LM information etc.) are available. The system was implemented using a minimum number of resources, as one of the languages in this thesis is under-resourced. Its implementation is based on the following steps

- Matching using a string-based similarity measure (LCSS - see **Section 6.2.1**);
- Alignment, base on the longest target language subsequence extracted from the GIZA++ result (**Section 6.2.2**);
- Recombination employing LM-information and the recombination matrix defined in **Section 6.2.3**.

The implementation fits into the framework of linear EBMT systems.

Implementation of a hybrid EBMT system, with influences from the linear and template-based approaches: $Lin-EBMT^{REC+}$

To avoid a possible loss of word-order information in the recombination step, we extended the EBMT baseline system *Lin-EBMT* by constraining the values in the recombination matrix with word-order information inspired by template-based EBMT approaches. Although constraints represent a well-known method which is used quite often in NLP, the use of constraints in an LM-based recombination step of an EBMT system is an innovative approach, which can open new paths in the domain of (example-based) machine translation. Three types of constraints were implemented: First-Word-Constraints (C.1), TLSide-Template-Constraints (C.2) and Whole-Template-Constraints (C.3) (see **Section 7.4**). $Lin-EBMT^{REC+}$ is platform- and language-pair independent under the same circumstances as *Lin-EBMT*. Both EBMT systems developed are easily adaptable for other language-pairs.

Creation of a parallel domain-restricted corpus RoGER

Parallel aligned corpora are useful for a multitude of cross-lingual applications. RoGER is a domain-restricted parallel aligned corpus, which includes four languages: Romanian, German, English, and Russian. RoGER was compiled together with my colleague **Natalia Eliţa** at the beginning of this research. RoGER represents a manual of an electronic device and it is manually aligned and corrected. The correction of the corpus has a direct impact on the results. This aspect allows a system developer to concentrate more on the application without worrying about the impact on the results of possible errors in the

data. RoGER can be used in applications which do not need large amounts of data or in a test phase for the other ones. The creation of the corpus was motivated by the lack of resources available for Romanian at that specific time. The whole description of the corpus has been presented in **Section 4.4**.

The experimental settings

The experimental settings in this thesis help analyzing the behavior of both corpus-based MT (**CBMT**) approaches in different settings. During the experiments presented in this research, the influence of several parameters have been investigated: the MT system and approach, the language pair, the corpus (type and size) and the test data type (in-domain and out-of-domain test data). We have compared the corpus-based MT approaches, while changing the above-mentioned parameters. The CBMT approaches (SMT and EBMT) have been directly compared, using the same training and test data. The results have also been examined in contrast to the ones provided by the Google Translate on-line MT system.

In the experiments two frameworks were considered: one with a larger corpus, closer to the SMT settings, another with a smaller corpus, which better fits the EBMT framework. The comparison is a one-to-one comparison, as the training and test data have been the same. Usually in the literature, EBMT and SMT are directly compared in a framework which better fits the SMT approach (where a large corpus is involved).

The experiments in this thesis were done in a realistic scenario, as no interference on the data of the corpus[1](such as choosing specific test sentences or correction of paragraph alignments in JRC-Acquis) was made. In some experiments also the influence of additional linguistic information, i.e. POS, has been studied. We used two language pairs, in both directions of translation: Romanian - English and Romanian - German.

11.2 Limitations of the Study

In this section we refer to limitations in the implemented EBMT systems, such as:

- Choosing only one best result in the matching procedure sometimes does not lead to the best solution from a global point of view.
- Errors introduced in the translation by the alignment step, as the TL sequences that form the output are extracted from the GIZA++ files using only the alignment information from the matched sentences. Also no verification for the word-alignment is done during the matching procedure. The Moses-based SMT system is less affected by bad word alignments than the EBMT system, as Moses is likely to choose the best alignment in the whole corpus data, while the EBMT is making use only of the matched examples.

[1]Abstraction from some characteristics of the RoGER corpus.

- The consideration of only one best solution in the recombination algorithm (a limitation in *Lin-EBMT*, which is taken over also in $Lin - EBMT^{REC+}$). This way valuable information could be lost, as sometimes a local maximum value does not necessarily mean a global maximum value.

- Types of constraints: only three types of constraints have been used in $Lin - EBMT^{REC+}$. However, several ways for the extension of the constraint-types are possible. More information in this direction will be presented in **Section 11.3.1**.

11.3 Further Work

In this section we will discuss several directions for further work having as a starting point the research presented in this thesis.

11.3.1 Extending the EBMT System

In this dissertation we concentrated our efforts in implementing an EBMT baseline system and testing how constraints in the recombination step influence the translation results. However, we are aware that our system(s) have limitations (see **Section 11.2**) and that several extensions can be made. In this subsection we present some possible extensions of the systems, which could improve the translation results:

- The use of restrictions in the matching procedure, similar to the ones found in [McTait, 2003] (e.g. length constraints, word frequency constraints) and the integration of word-alignment information. Using word-alignment information in the matching procedure decreases the risk of not having matched sequences aligned in the further translation steps.

- The use of NERs, where available, to translate better proper names, numbers, etc. Where no NERs are available, rules could be integrated.

- Adding various criteria for reordering the words in recombination by extending the types of constraints used in this dissertation, such as the integration of new constraints (e.g. Last-Word-Constraints), or the use of weights and priorities for the constraints. In our approach we set priority only to the First-Word-Constraints. Another approach would be to search for constraints which are motivated linguistically. However, this way the system might not remain language independent, as differnt motivations could appear for various languages.

- For both matching and recombination a ranking approach of several possible solutions could be considered. These results should be examined in order to decide if considering several options brings better results than, for example, extending the types of constraints.

- Dictionaries might be attached in order to correct possible word alignment mistakes and to improve the initial GIZA++ results. If no dictionaries are available, adding more data could ameliorate the initial GIZA++ alignment.

11.3.2 Extending the Manual Analysis

Manual evaluation plays an important role in MT, as the "automatic measures are an imperfect substitute for human assessment of translation quality" [Callison-Burch et al., 2010]. The approach is used either for evaluating an MT system, for extracting the translation error types and sources of errors or for validating an automatic MT evaluation metric.

The manual analysis in this research was done by a single human judge and it included only a few sentences. This happened due to man-power, time and money limitations. In order to have a better overview of the MT approaches and their advantages and disadvantages, the manual evaluation should be extended to more data and more human judges should be involved. Having more information, more relevant conclusions on the results could be drawn.

11.3.3 Other Directions

Among other possible directions there are the use of different data, the integration in a hybrid translation environment and comparisons with other EBMT systems.

In this thesis three languages have been analyzed: Romanian, German and English. As SMT experiments were run for quite a large number of languages (e.g. [Ignat, 2009]), it would be interesting to test how the EBMT systems implemented during this research behave for different language pairs. The questions which appear are "would the affirmation *the results are better when un-inflected languages are used* be confirmed or it is the choice of the test and training data that has the biggest impact on the translation results?" Further experiments with different corpora (type or size) could also be of an interest.

Each of the MT approaches has its own strengths and weaknesses. Over the last few years hybrid approaches have been implemented in order to obtain better translation results: rule-based approaches together with statistical models ([Eisele et al., 2008]), example-based with SMT ([Smith and Clark, 2009]), etc. More details have already been presented in **Section 2.3**.

An interesting further research would address issues such as how the $Lin-EBMT^{REC+}$ could be integrated into a hybrid framework and which influence would it have on the results. As one of the languages in this thesis is under-resourced, a hybrid framework including SMT and EBMT could be interesting. For a language-pair where both languages are not under-resourced, also an integration in an RBMT framework could be possible.

As discusses in **Chapter 6**, at the end of 2009 open-source EBMT systems appeared, but no such system was used in this research. A comparison between $Lin-EBMT^{REC+}$

and such systems could contribute to a better understanding on the strengths and weaknesses of each of the systems and eventually lead to a way for combining more approaches in order to improve the translation results. Comparisons between $Lin - EBMT^{REC+}$ and OpenMatrex (`www.openmatrex.org`) are presented in [Gavrila and Elita, 2011a] and [Gavrila and Elita, 2011b].

———* * *———

In this thesis we showed how corpus-based MT approaches behave when having a lower-resourced inflected language (i.e. Romanian) as a source language or as target language. For Romanian - German, we tested the behavior of the same systems when both SL and TL languages are inflected and one of them is lower-resourced. We investigated the influence on the translation results of word-order constraints extracted from the template-based EBMT approach and integrated in the recombination step of the baseline linear EBMT system. For the language pair English - Romanian we also tested the impact of part-of-speech information on the translation quality.

Although the SMT system outperforms the EBMT system in all experiments, the behavior of the systems when changing the parameters in the experimental settings confirm the big impact the training and test data have on both of the CBMT approaches. It was also noticed that the difference between results of the approaches decreases when a smaller corpus is used. We also showed that constraints improve the translation, although a clear decision which constraint-combination works best could not be taken. Both CBMT approaches worked better for shorter sentences.

Appendix A

A Tabular Overview of Existing EBMT Systems

This appendix presents an overview of existing EBMT systems, showing the language-pairs and the size and type of the corpora used in the translation process. Several EBMT systems found in Table A.1 have been also discussed in [Somers, 1999]. All these EBMT systems have been presented in research papers, but are not available as (open source) software. This is why no comparison between these systems and the systems developed in this thesis have been possible.

The size of the parallel aligned corpus in an EBMT system differs. In [Somers, 1999] it varies between **7** and **726 406** sentences, but the size increased up to the WWW over the last few years (see Table A.1).

Several languages are used in EBMT translation. However, from the 30 systems presented in [Somers, 1999], English (ENG) is used as SL and TL in almost half of the cases: in 14 and 15 cases, respectively. Among other languages used as SL there are: Japanese (Jap) – ten times, Spanish (Spa) and French (Fre) – each two times, German (DEU) and Irish - each one time. As target language, Japanese appears five times, Spanish, French, German and Turkish (Tur) two times (each), and Urdu and Serbo-Croatian one time (each).

Table A.1 is an extension of the information found in [Somers, 1999].

System	Languages	Train/Test size	Corpus type
[Grefenstette, 1999]	DEU → Spa DEU → ENG Spa → ENG	WWW (AltaVista)	
wEBMT [Way and Gough, 2003]	Fre and ENG	WWW	
[Smith and Clark, 2009]	Fre and ENG	Europarl	Law domain
[Hutchinson et al., 2003]	ENG → Fre	960 000	
MSR-MT [Brockett et al., 2002]	ENG → Jap	596 000 (238 sentences)	Microsoft documentation and student dictionary
[Doi et al., 2005a], [Doi et al., 2005b]	Jap → ENG	404 022	
[Liu et al., 2006]	ENG → Chi	262 060	
[Gough and Way, 2004]	ENG → Fre	207 468	
[Sumita, 2001]	Jap and ENG	204 108	
[Way and Gough, 2005]	ENG ↔ Fre	203 529	
[Gough and Way, 2004]	ENG → Fre	203,529 / 3,939	Sun TM
[Lepage and Denoual, 2005]	Jap ↔ ENG	160 000 (510 test)	C-STAR (travel)
[Watanabe and Sumita, 2003]	Jap, Chi, Korean, ENG	152 169 and 10148 / 4846 - 510 considered 6	
[Brown et al., 2003]	Fre → ENG	100 000 (10*100 sentence fragments)	
CTM [Sato, 1992]	ENG → Jap	67 619	
[Mandreoli et al., 2002]	ENG → Ita	34 550 / 421	Technical manual
[Aramaki and Kurohashi, 2004]	Jap → ENG	20 000 (500 sentences 16 refs)	
[Brown, 2001]	Fre → ENG Spa → ENG	19 730 1M words	
HARMONY	ENG → Jap	12 000	

(Cont.)

Table A.1 – Continued

System	**Languages**	**Train/Test size**	**Corpus type**
[Franz et al., 2000]			
[Feiliang et al., 2007]	Chi → Jap	10 083	
Chunky [Engel, 2000]	DEU → ENG	10 000	
[Gough and Way, 2003]	ENG → Fre	3885 / 200	user-guide controlled language, part of the Sun TM
[McTait, 2001]	ENG → Fre ENG → Spa	3 000 (1000 test) and 600 (500)	WHO AFI titles ScanWorkX manual
[McTait and Trujillo, 1999]	ENG → Spa	3 000	
ATR [Sumita and Iida, 1991]	Jap → ENG	2 550	
[Saha and Bandyopadhyay, 2005]	ENG → Bengali	2 000	
Gaijin [Veale and Way, 1997]	ENG → DEU	1 836	
[Doğan, 2007]	ENG ↔ Tur	970 (100 sentences)	
(S/D) TTL [Cicekli and Guvenir, 1996], [Cicekli and Guvenir, 1998], [Cicekli and Guvenir, 2001], [Cicekli and Guvenir, 2003]	ENG ↔ Tur	747	
TTL [Oz and Cicekli, 1998]	ENG ↔ Tur	488	
EDGAR [Carl, 1999]	DEU → ENG	303	
[McLean, 1992]	ENG → Fre	32	
CAPMT [Aramaki et al., 2004]	Jap → ENG	not described exactly 240 sentences (4 refs)	(NHK news corpus[1])

(Cont.)

Table A.1 – Continued

System	Languages	Train/Test size	Corpus type
ReVerb [Collins et al., 1996], [Collins and Cunningham, 1996], [Collins, 1998]	ENG → DEU		Corel Draw Manual
[Al-Adhaileh and Kong, 1999]	ENG → Malay		
METIS-II [Dirix et al., 2005]	Dutch → ENG		
[Maruyama and Watanabe, 1992]	Jap → ENG		
[Kaji et al., 1992]	Jap → ENG		
[McTait, 2001]	Fre, Spa and ENG		
LFG-DOT [Way, 2001]	Fre and ENG		
SimTran [Watanabe, 1992]	Jap → ENG		
PalmTree [Watanabe and Takeda, 1998]	ENG → Jap		
[Irimia, 2009]	ENG ↔ RON		JRC-Acquis

Table A.1: Overview of EBMT systems.

[1] 40000 articles with 5.2 sentences per article for Japanese and 7.4 for English

Another EBMT system is presented in [Markantonatou et al., 2006], which uses several source languages (Greeck, Spanish, Dutch and German) and English as TL.

In the systems presented in this **Appendix** new languages have been encoutered, such as Chinese (**Chi**), Bengali, Malay, Dutch or Romanian. Leaving aside the languages used in [Watanabe and Sumita, 2003] and including the ones in [Markantonatou et al., 2006], the distribution of the SL and TL languages is presented in Table A.2: 53.5% from the SL and 42.8% from the TL are still represented by English.

Language	**SL**	**TL**
English (ENG)	30	24
Japanese (Jap)	9	7
French (Fre)	5	11
German (DEU)	4	2
Dutch	2	-
Spanish (Spa)	3	4
Chinese	1	1
Romanian (RON)	1	1
Greek	1	-
Turkish (Tur)	-	3
Malay	-	1
Italian (Ita)	-	1
Bengali	-	1

Table A.2: The number of times a language is used as a SL and a TL in Table A.1.

A. A TABULAR OVERVIEW OF EXISTING EBMT SYSTEMS

Appendix B

A Selective Analysis of the Languages Used

In this appendix we present a selective analysis of the characteristics of the languages used in this dissertation to motivate the hypothesis that "*the languages are morphologically and syntactically different enough, in order to make the MT process challenging*". Some translation challenges found in the corpora have already been presented in **Chapter 4**.

The choice of the aspects under investigation is motivated mainly by the phenomena found in the corpora used[1] and by the fact that these aspects might represent a challenge for an MT system. This appendix is not aiming at providing a complete overview of the descriptions of the three languages, the differences and similarities between them. Also, not all translation challenges have been described. As English is one of the languages most widely used in NLP applications, the focus is on Romanian and German.

Figure B.1 presents an overview of the position of the three languages inside the Indo-European language family. Although English and German are both West Germanic languages, they belong to different branches: the branch of Anglo-Frisian languages[2] and the High German branch[3] respectively. Romanian is an Eastern Romance language.

Romanian is a Romance language; its grammar and basic vocabulary are closely related to those of its relatives Italian, Spanish, Catalan etc. It has influences from Slavic languages, Hungarian and Turkish. Alboiu and Motapanyane [2000] describe Romanian as "*a hybrid between Romance and Balkan languages, and many of its peculiarities can be understood only with reference to equivalent paradigms in Romance and Balkan*". Romanian preserved, in contrast to most other Romance languages, the 3-gender system from Latin and it is a highly inflected language. Another Latin element that has survived in Romanian while having disappeared from other Romance languages is the morphological

[1] For example: aspects related to tense and mood are not really relevant, as in the corpora used, due to their domain restriction, this aspects are limited.

[2] More specifically, English belongs to the English branch, also found under the name of Insular Anglo-Frisian or Anglic.

[3] More exactly German belongs to the Central German branch.

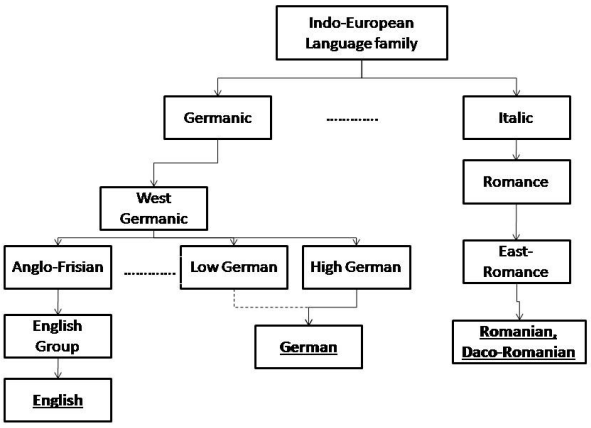

Figure B.1: The Indo-European languages.

case differentiation in nouns, albeit reduced to only three forms (nominative/accusative, genitive/dative and vocative) from the original seven. An often mentioned characteristics of Romanian is that it is the only Romance language where definite articles are attached to the end of the noun as enclitics (as in Bulgarian, Macedonian and Albanian). As in all Romance languages, Romanian verbs are highly inflected according to e.g. person, number, tense and mood. Inside one sentence there is no predefined position for the verb, adjectives can be situated before or after the noun, but the semantics might be different in each case. Pronoun-elliptic sentences are possible, as in other Romance languages. These are only some aspects that could make difficult the adaptation of language technology systems for other languages.

German is a Germanic language, which is also inflected. It also has a 3-gender system and well defined inflection classes. A special feature is represented by the verb particles: the separation of the particle from the verb inside the sentence and the challenge that the particle can also be in different contexts a preposition or an adverb. The verb changes its meaning depending on the particle: different particles lead to different meanings for the verb. Word order is generally less rigid than in Modern English. The position of a noun in a German sentence has no bearing on its being a subject, an object or another argument. On the contrary, in a declarative sentence in English if the subject does not occur before the predicate the sentence could well be misunderstood. From the syntactic point of view, rules establish the position of the verb in the main and subordinate clause in German. Another characteristic of the language is that it can contain embedded relative clauses. Like most Germanic languages, German forms noun compounds where the first noun modifies the second. Unlike English or Romanian, where newer compounds or combinations of longer nouns are often written in open form with separating spaces, German nearly always uses the closed form without spaces. German allows arbitrarily long compounds, but such long

words are rarely used in every-day language. Another interesting aspect for German is that it permits lengthy nominal modifiers.

From the multitude of aspects that can be analyzed (e.g. verb conjugation, the correspondences between tenses and moods, multiple negation, pronouns or degrees of comparison), we will present only five of them in this appendix:

1. Noun inflection
2. Compounds
3. Verbs with a separable particle
4. Word order
5. Genitive formation

All five aspects influence the results of an MT system for the languages used in this thesis.

B.1 Noun Inflection

In this section we describe the noun inflection, with respect to the lack of article or existence of the definite or indefinite article. We also look at the agreement between noun and adjective and the inflected forms of the adjectives.

Unlike English, which has lost almost all forms of inflection for nouns and adjectives, German and Romanian still inflect nouns, adjectives and pronouns for four grammatical cases: nominative, accusative, dative and genitive. Romanian also presents the vocative case.

Romanian and German nouns are categorized into three genders: masculine, feminine and neuter. The Romanian neuter gender consists of masculine forms (for singular) and feminine ones (for plural). Contrary to strongly inflected languages, German marks case on the article rather than on the noun, though especially the difference between plural and singular is expressed by suffixes. The adjective inflection depends not only on the number, gender and case of the noun it modifies, but also on whether the indefinite article, definite article or no article is used with it. In contrast to Romance languages, adjectives are only declined in the attributive position. Predicative adjectives are not declined and are indistinguishable from adverbs. An overview of the noun and adjective inflection in German is shown in Table B.1.

Romanian presents a more complicated system for the inflection of nouns and adjectives. The definite article occurs as enclitic to the noun or adjective and display different forms for gender and number. The morpheme for the case may attach to the definite article. Some examples can be found in Table B.2.

Various classes of adjectives may either precede or follow the noun. Adjectives preceding the noun carry the enclitic article and case morphology - see Table B.3 and Table B.4. There are also adjectives that can be positioned either after or before the noun, such as "*acesta*" (ENG: "*this-A*") or "*biet*" (ENG: "*poor*"): "*băiatul acesta*" (boy-the this-A)

Case	Singular			Plural
	Masculine	**Feminine**	**Neuter**	
nom.	der trockene Wein	das kühle Bier	die warme Milch	die guten Getränke
	ein trockener Wein	ein kühles Bier	eine warme Milch	gute Getränke
	kein	kein	keine	keine
	trockener Wein	kühles Bier	warme Milch	guten Getränke
	trockener Wein	kühles Bier	warme Milch	gute Getränke
acc.	den trockenen Wein	das kühle Bier	die warme Milch	die guten Getränke
	einen trockenen Wein	ein kühles Bier	eine warme Milch	gute Getränke
	keinen	kein	keine	keine
	trockenen Wein	kühles Bier	warme Milch	guten Getränke
	trockenen Wein	kühles Bier	warme Milch	gute Getränke
dat.	dem trockenen Wein	dem kühlen Bier	der warmen Milch	den guten Getränken
	einem trockenen Wein	einem kühlen Bier	einer warme Milch	guten Getränken
	keinem	keinem	keiner	keinen
	trockenen Wein	kühlen Bier	warme Milch	guten Getränken
	trockenem Wein	kühlem Bier	warmer Milch	guten Getränken
gen.	des trockenen Weines	des kühlen Bieres	der warmen Milch	der guten Getränke
	eines trockenen Weines	eines kühlen Bieres	einer warmen Milch	guter Getränke
	keines	keines	keiner	keiner
	trockenen Weines	kühlen Bieres	warmen Milch	guten Getränke
	trockenen Weines	kühlen Bieres	warmer Milch	guter Getränke

Table B.1: Noun and adjective inflection in German.

Masculine			
Case	**Article**	**Singular**	**Plural**
nom.-acc.	indefinite	un prieten	nişte prieteni
	definite	prietenul	prietenii
dat.-gen.	indefinite	unui prieten	unor prieteni
	definite	prietenului	prietenilor
Feminine			
Case	**Article**	**Singular**	**Plural**
nom.-acc.	indefinite	o poveste	nişte povesşti
	definite	povestea	poveştile
dat.-gen.	indefinite	unei povesşti	unor poveşti
	definite	poveştii	poveştilor
Neuter			
Case	**Article**	**Singular**	**Plural**
nom.-acc.	indefinite	un tablou	nişte tablouri
	definite	tabloul	tablourile
dat.-gen.	indefinite	unui tablou	unor tablouri
	definite	tabloului	tablourilor

Table B.2: Noun inflection in Romanian.

but not * "*acesta băiat*" (this-A boy); bietul băiat (poor-the boy), but not "*băiatul biet*" (boy-the poor).

Masculine		
Case	**Singular**	**Plural**
nom.-acc.	un bun prieten bunul prieten	nişte buni prieteni bunii prieteni
dat.-gen.	unui bun prieten bunului prieten	unor buni prieteni bunilor prieteni
Feminine		
Case	**Singular**	**Plural**
nom.-acc.	o minunată poveste minunata poveste	nişte minunate poveşti minunatele poveşti
dat.-gen.	unei minunate poveşti minunatei poveşti	unor minunate poveşti minunatelor poveşti
Neuter		
Case	**Singular**	**Plural**
nom.-acc.	un frumos tablou frumosul tablou	nişte frumoase tablouri frumoasele tablouri
dat.-gen.	unui frumos tablou frumosului tabloul	unor frumoase tablouri frumoaselor tablouri

Table B.3: Adjective before the noun, with definite and indefinite article in Romanian.

Masculine		
Case	**Singular**	**Plural**
nom.-acc.	un prieten bun prietenul bun	nişte prieteni buni prietenii buni
dat.-gen.	unui prieten bun prietenului bun	unor prieteni buni prietenilor buni
Feminine		
Case	**Singular**	**Plural**
nom.-acc.	o poveste minunată povestea minunată	nişte poveşti minunate poveştile minunate
dat.-gen.	unei povesşti minunate poveştii minunate	unor poveşti minunate poveştilor minunate
Neuter		
Case	**Singular**	**Plural**
nom.-acc.	un tablou frumos tabloul frumos	nişte tablouri frumoase tablourile frumoase
dat.-gen.	unui tablou frumos tabloului frumos	unor tablouri frumoase tablourilor frumoase

Table B.4: Adjective after the noun, with definite and indefinite article in Romanian.

Some of the nouns also appear in the vocative in Romanian, e.g. "*prietene*" (ENG: "*friend*").

In English, the adjective is not inflected and a noun has only two forms - singular and plural (see Table B.5). Unlike German and Romanian, English does not have a grammatical gender, although some nouns denote feminine or masculine animate objects, e.g. "*lion*" and "*lioness*".

Case	Singular	Plural
nom.-acc.	the/a tall girl	the/some tall girls
dat.-gen.	to/of the/a tall girl	to/of the/some tall girls

Table B.5: Adjective and nouns in English.

These aspects pose difficulties in the MT system such as finding the correct inflected form in German or Romanian (for example when having English as SL) or the adjective-noun inversion: in English and German the adjective precedes the noun, in Romanian it can appear as antecedent or precedent of the noun.

B.2 Compounds

An important aspect in the comparison of these three languages is represented by the compounds. In German, compounds are encountered quite often. Compounds in German are normally written as single words, without spaces or other word boundaries. They can be made up of two or more parts. Sometimes there are coordinated compound constructions. In a few cases, the compounds are hyphenated. Some examples are shown below:

1. *Regierungskonferenz* - ENG: *intergovernmental conference*, RON: *conferinţa interguvernamentală*
2. *Rollstuhl* - ENG: *wheelchair*, RON: *scaun cu rotile*
3. *Fremndsprachenkenntnisse* - ENG: *knowledge of foreign languages*, RON: *cunoştinţe de limbi straine*
4. *See- und Binnenhäfen* - ENG: *sea and inland ports*, RON: *porturi la mare şi interne*
5. *Kosovo-Konflikt* - ENG: *Kosovo conflict*, RON: *conflictul din Kosovo*
6. *Völkermord* - ENG: *genocide*, RON: *genocid*

Noun composition occurs most often with two (or more) nouns, but can also take place with other parts of speech, such as:

1. Noun+Noun: *Liebeskummer* - ENG: *love sickness*, RON: *probleme în dragoste*
2. Adjective+Noun: *Rotwein* - ENG: *red wine*, RON: *vin roşu*
3. Noun+Adjective: *bildschön* - ENG: *picture-perfect*, RON: *drăguţ ca într-o pozã*
4. Verb+Noun: *Boxhandschuhe* - ENG: *boxing-gloves*, RON: *mănuşi de box*
5. Preposition+Noun: *Vorvertrag* - ENG: *preliminary contract*, RON: *contract preliminar*

Compounds can also be formed by: Adjective+Adjective: *hellblond* - ENG: *light-blond*, RON: *blond deschis* or Adverb+Verb: *wiedersehen* - ENG: *meet again*, RON: *revedea*

The main meaning of a compound in German is determined by the last added word. The longest German word verified to be actually in (albeit very limited) use is *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*, which, literally translated, is "*beef labelling supervision duty assignment law*" [from *Rind* (*cattle*), *Fleisch* (*meat*), *Etikettierung(s)* (*labelling*), *Überwachung(s)* (*supervision*), *Aufgaben* (*duties*), *Übertragung(s)* (*assignment*), *Gesetz* (*law*)].

The ways of forming compounds in Romanian are similar to the ones for German:

1. Noun (Nominative)+Noun (Nominative): *câine-lup* (ENG: *wolf dog*),
2. Noun (Nominative)+Noun (Genitive): *floarea-soarelui* (ENG: *sunflower*)
3. Noun (Nominative)+Preposition+Noun: *cocoş-de-munte* (ENG: *capercaillie*)
4. Adjective+Noun (Nominative) or Noun (Nominative) +Adjective: *bunăstare* (ENG: *welfare*), *argint-viu* (ENG: *quicksilver*)
5. Numeral+Noun (Nominative) or Noun (Nominative) +Numeral: *trei-fraţi* (* ENG: *Three brothers* , the name of a plant)
6. Noun+Verb or Verb+Noun: *gură-cască* (ENG: *gaper*), *pierde-vară* (ENG: *lazybones*)
7. Noun formed from an imperative sentence: *nu-mă-uita* (* ENG: *Do not forget me*, name of a plant.)

The parts that form the compound can be written together or separated by a hyphen. However, Romanian does not have compounds formed with so many words as in German.

There are three forms of compounds in English: the closed form, in which the words are melded together (such as *secondhand*, *childlike*, *makeup*, *notebook*), the hyphenated form (such as *daughter-in-law*, *six-pack*, *six-year-old*), and the 'open form' (such as *post office*, *real estate*, *middle class*).

The biggest difficulty for the MT system in this case is when the SL is German: splitting the compound and finding the right translation. Also, it is difficult to obtain the required compound in German, when German is the TL. Usually in this case the translation is done using constructions, and not compounds, unless, for example, the word-alignment algorithm finds the right corresponding compound. Compounds also represent a challenge for the word-alignment.

B.3 Verbs with a Separable Particle

German contains many verbs that have a separable particle (prefix) that can be attached to its root. The particle stays together with the main verb only in some specific cases, such as in the infinitive form. The particle changes the meaning of the main verb, as in Example (1).

(1) The verb "*machen*"(ENG: to do, to make, to create) can attach, among others, the particles "*ab-*", "*auf-*" and "*zu-*" and form:
abmachen (ENG: to agree, to arrange)
aufmachen (ENG: to open)
zumachen (ENG: to close)

The particle may represent a preposition (e.g. "***über****laufen*" ENG: "*to flow over*") or an adverb (e.g. "***hin****legen*" ENG: "*to put something down*"). They represent a challenge in MT as the verb and its particle are usually not together and the distance between them can be quite significant. Also the particle is in some cases ambiguous.

A similar phenomenon can be considered the case of the phrasal verbs in English. However, the phenomenon is simpler than in German, as the distance between the verb and the particle is usually smaller. Phrasal verbs are idiomatic expressions, combining verbs and prepositions (or adverbs) to make new verbs whose meaning is often not obvious from the dictionary definitions of the individual words. An example is the verb "*get*" - see Example (2)

(2) The verb "*get*". On `www.usingenglish.com` (last accessed on June 2nd, 2011) there are 66 phrasal verbs with "*get*".
to get
to get over
"*I hope you will get over your operation quickly.*"
"*Work hard, and get your examination over with.*"
to get along
"*Why don't you two get along? You're always arguing.*"
to get up
"*They got up a list of two hundred people who were opposed to the local council's plans.*"

This phenomenon is not relevant for Romanian.

B.4 Word Order

Word order plays an important role in translation.

In English the meaning is usually derived from the word order: the first noun is the subject and the second the object:

(3) "*The girl eats the fish*".

If the word order is changed as in

(4) "*The fish eats the girl*",

the sentence meaning is totally changed. In German, due to the changes in form for each case, both word orders are accepted and have the same meaning:

(5) "*Das Mädchen isst den Fish*" oder "*Den Fish isst das Mädchen*".

For this specific example, for Romanian the word oder is exactly as in English:

(6) "*Fata mănâcă peştele*".

Romanian has a relatively free word-order. The inflected forms and the use of prepositions or other ways of marking the syntactic role of the words allow changes in the word order without changing the meaning. The subject does not always have to be present in the sentence, as in "*Mergem sa mâncăm*" (ENG *(We) go to eat.*): The pronoun "*we*" (ROM: "*noi*") does not appear in the sentence. The ending of the verb gives us the necessary information about the subject.

The word order in German is, generally speaking, not as fixed as in English. However, there are specific word order rules for the verb, such as:

- The main verb must be the second element in the main clause: "*Ich fliege oft nach Rumänien*" ("*I fly often to Romania*"). The subject is placed as close as possible to the verb.
- The past participle, the infinitive or other verb parts must always be the last element in the main clause: "*Heute können wir einen Tee trinken gehen*" (*Today we can go to drink a tea*). The main and the last verb form a kind of verbal bracket around the rest of the sentence.
- The main verb must be the last element in a subordinate clause: "*Ich fliege oft nach Rumänien, weil meine Eltern dort wohnen*" (* "*I fly often to Romania, because my parents there live*").
- In questions or imperatives, the sentence starts with the verb.

In English and German the adjective usually occurs before the noun it modifies. As already seen in **Section B.1**, in Romanian the adjective appears either before the noun or after it. Depending on its position, the adjective may include the definite article. The English "*the pretty girl*" can be expressed as "*fata frumoasă*" (ENG: "*girl-the pretty*") or as "*frumoasa fată*" (ENG: "*pretty-the girl*"), depending on the position of the adjective.

B.5 Genitive Formation

In this section we refer to the genitive formation of the nouns. The genitive case is used to show possession. In German it has its own special form, e.g "*die Mutter des Kindes*" (ENG: "*the child's mother*"). Posession can also be expressed by dative, especially in more casual speech: "*die Mutter von dem Kind*" (ENG: "*the mother of the child*"). An '*-s*' is simply added to the end of the name if the identity of the possessor is specified, as in "*Claudias Buch*" (ENG: "*Claudia's book*").

In English, the genitive is formed with "*'s*" (e.g. "*the child's mother*") or using the preposition "*of*" (e.g. "*the toy of the child*").

In Romanian, the genitive can be formed in two distinctive ways: using the genitive form of the article or using the "possessive article" (see Table B.6).

Masculine		
Case	**Singular**	**Plural**
nom.-acc.	al	ai
dat.-gen.	-	alor
Feminine		
Case	**Singular**	**Plural**
nom.-acc.	a	ale
dat.-gen.	-	alor

Table B.6: Possessive article in Romanian.

Some examples of the two ways of forming genitive are shown below:

- 1. With the definite article: *floarea femeii / femeilor* (ENG: *the woman's / the women's flower*) - *flower-the woman-the-GEN. / women-the-GEN.*
  2. With the indefinite article: *floarea unei femei / unor femei* (ENG: *a woman's / some women's flower*) - *flower-the a-GEN woman-GEN. / a-GEN women-GEN*
- With the possessive article[4]:

  1. *o floare a copilului / a unui copil* (ENG: *a flower of the child's / of a child's*
  2. *acest creion al copilului / al unui copil* (ENG: *this pencil of the child's / of a child's*
  3. *florile roşii ale copilului / ale unui copil* (ENG: *the red flowers of the child's / of a child's*
  4. *primii paşi ai copilului / ai unui copil* (ENG: *the first steps of the child's / of a child's*)

The definite or indefinite article is also used in the case when the possessive article appears in the formation of the genitive. The possessive article agrees in gender and number with the first noun, which denominates the object that is possessed.

One of the challenges for an MT system is represented by the correct choice of the possessive article, when Romanian is the target language.

——* * *——

It can be concluded from this overview that "*the languages are morphologically and syntactically different enough, in order to make the MT process challenging*"

[4]In some works the possessive article is known under the name of pre-genitive particle [Motapanyane, 2000].

Appendix C

Minor Parallel Corpora

Parallel aligned corpora play an important role in corpus-based machine translation (**CBMT**). Although the most widely used corpus in statistical MT (**SMT**) is the Europarl, it cannot be used for the experiments presented in this dissertation, as it does not contain all analyzed languages: Romanian, English and German[1].

This appendix briefly describes other corpora that might be of interest for further experiments. Before we present these corpora, the term "*minor*" in the title needs to be explained to avoid mis-understandings: this term refers to the importance and impact of these corpora on the experiments presented ONLY in this thesis. It has no qualitative meaning and it does not judge the possibility of using them in other experiments.

Parallel aligned corpora that include all three languages or at least two of them are presented in the sections below. The list is not exhaustive.

C.1 OPUS

OPUS (`http://opus.lingfil.uu.se/`[2]) is a (growing[3]) collection of multilingual (sub-) corpora, which contains translated open source documents available on the Internet. The corpus files have been encoded in Unicode UTF8 and are sentence aligned for all possible language pairs. As the alignments have been automatically performed, possible errors might be encountered. The alignments have been done using a length-based approach found in [Gale and Church, 1993] and they are stored in the XCES format[4].

Among the (sub-)corpora included in OPUS there are: EMEA, EUconst, Europarl, OpenOffice, KDE, KDE4, KDEdoc, PHP, OpenSubs, SPC and SETIMES.[5] Several tools

[1] Status at the moment of running the experiments.

[2] Last accessed on June 27th, 2011.

[3] The OPUS collection is continuously growing: the latest (sub-)corpus included in the collection is SETIMES (April 2010).

[4] XCES is the XML version of the Corpus Encoding Standard. More details on `http://www.xces.org/` - last accessed on June 27th, 2011.

[5] Status: Summer 2010.

and linguistic resources have been created for some of the (sub-)corpora in OPUS, such as dictionaries extracted with GIZA++. More details about OPUS are found in [Tiedemann and Nygaard, 2004] and [Tiedemann, 2009].

Not all of the (sub-)corpora contain all three languages we used in this research. The ones that include all three languages are described in Table C.1. From these, EMEA (part of OPUS version 3) is the one with the largest number of sentences [Tiedemann, 2009][6]. For the language pairs where Romanian is included, its size is larger even than the one of JRC-Acquis Version 2.2. Still, the latest version of JRC-Acquis for Romanian (Version 3.0) contains more data than EMEA, but no alignment information is publicly available[7]. We have not used EMA for our experiments as manually analyzing a small part of the data, we noticed that not all sentences have been translated in Romanian, i.e. part of the sentences have been left in the initial language

Corpus	**Number of sentences**		
	Romanian-English	**German-Romanian**	**English-Romanian**
EMEA	1038722	1064107	1163348
KDE	95717 / 73392	31885	78028 / 51515
KDE4	66473 / 59382	61352	169286/ 106169
KDEdoc	94	210	3030
PHP	36199	33335	42250
OpenSubs	305259	18382	76007

Table C.1: OPUS Overview: sub-corpora which contain all three languages: Romanian, German and English.

EMEA

EMEA is a parallel aligned corpus in 22 languages[8] and can be found on `http://opus.lingfil.uu.se/EMEA.php`[9]. It contains documents from the European Medicines Agency (`http://www.emea.europa.eu`[10]). The data is automatically aligned. As a medical corpus its vocabulary is more restricted than the one of JRC-Acquis. Although it is stated in the literature that EMEA is sentence aligned, we noticed by manually analyzing a small part of the corpus that in some cases a "*sentence*" means only a noun-phrase or a number. Table C.2 shows several statistics on sub-corpora of EMEA.

(1) Examples of "sentences" in the corpus:

- *European Medicines Agency*
- *EMEA/ H/ C/ 471*
- *EUROPEAN PUBLIC ASSESSMENT REPORT (EPAR)*

[6]Status: Summer 2010.
[7]Status: July 2010.
[8]Status: February 2009.
[9]Last accessed on June 27th, 2011.
[10]Last accessed on June 27th, 2011.

No. Sentences	No. tokens		Vocabulary size		Average sentence length	
	1	2	1	2	1	2
1=English, 2=German						
10k	138492	128526	4812	6656	13	12
25k	321309	298862	7297	10768	12	11
50k	666791	621263	11166	17952	13	12
75k	982788	917264	14603	24943	13	12
100k	1301916	1214944	17528	30648	13	12
1=English, 2=Romanian						
10k	148441	159101	4955	6366	14	15
1=German, 2=Romanian						
10k	128165	149640	6655	6268	12	14

Table C.2: Statistics on sub-corpora of EMEA.

- *This document is a summary of the European Public Assessment Report (EPAR).*
- *It explains how the Committee for Medicinal products for Human Use (CHMP) assessed the studies performed , to reach their recommendations on how to use the medicine .*

C.2 SEE-ERA.net

The SEE-ERA.net corpus, described in [Tufiş et al., 2008b], contains the SEnAC Corpus (SEE-ERA.net Administrative Corpus) and the SEnLC Corpus (SEE-ERA.net Literary Corpus). The initial corpus included Bulgarian, Greek, Romanian, Serbian, Slovenian and English. Three languages have been added: French, German and Czech.

The SEE-ERA.NET Resources Webpage[11] contains an English-Romanian corpus. The corpus is tokenized, POS-tagged and lemmatized. It contains 60389 translation units (TUs)[12]. The text is part of the JRC-Acquis corpus. The data seems to be annotated with tools found on the RACAI Text Processing Webservices webpage (`http://www.racai.ro/webservices/textProcessing.aspx`[13]). No data for German has been found[14].

C.3 Other Corpora

Other corpora which include at least two of the languages are:

- Rada Mihalcea's parallel corpus (Romanian - English): This is a news corpus.

[11] `http://www.racai.ro/ReaserachActivity/WebServicesandResources/SEEERANETResources/tabid/131/Default.aspx` - last accessed on June 27th, 2011.

[12] A TU is a paragraph in the JRC-Acquis sense.

[13] Last accessed on June 27th, 2011.

[14] Status: August 2009.

The translations are sometimes incomplete. `www.cs.unt.edu/~rada/downloads.html`[15].

- EU Official Journal (`http://eur-lex.europa.eu/JOIndex.do`[16]), a multilingual legal text in 22 European languages (`http://apertium.eu/data`[17]). German and English were taken into account in 1998; Romanian was introduced in 2007.
- The Romanian-English-Russian corpus from `www.azi.md`[18]: This is a news corpus. Sometimes the translations are incomplete.
- The De-News corpus (German - English): This is a news corpus. Sometimes the translations are incomplete. `www.iccs.informatics.ed.ac.uk/~pkoehn/publications/de-news`[19].
- Europarl is the corpus most widely used in the MT research community, but it contained no Romanian texts when running the experiments. Among its 11 languages German and English are included. `www.statmt.org/europarl`[20]. Romanian was included in the seventh release of the Europarl corpus[21], but no Romanian-German corpus is available.
- "Specialized" corpora [Steinberger et al., 2006]: "1984" by George Orwell and the Bible

[15]Last accessed on June 27th, 2011.
[16]Last accessed on June 27th, 2011.
[17]Last accessed on June 27th, 2011.
[18]Last accessed on June 27th, 2011.
[19]Last accessed on June 27th, 2011.
[20]Last accessed: June 2011.
[21]Status: May 2012.

Appendix D

Excerpts from the Corpora Used

D.1 JRC-Acquis

In this section an excerpt from the JRC-Acquis corpus for Romanian - English is presented. The format is the one of the input in the EBMT systems implemented in this dissertation.

<?xml version="1.0" encoding="UTF-8"?>
<sentences>
.....
<sentence id="2514">
<ro>- beneficiarii pensiilor de invaliditate , pensiilor pentru limita de vârsta sau pensiilor de urmas platite de banca europeana de investitii .</ro>
<en>- persons receiving disability , retirement or survivors ' pensions paid by the european investment bank .</en>
</sentence>
<sentence id="2515">
<ro>articolul 5</ro>
<en>article 5</en>
</sentence>
<sentence id="2516">
<ro>regulamentul nr . 8 / 63 / euratom , 127 / 63 / cee3 se abroga .</ro>
<en>regulation no 8 / 63 euratom , 127 / 63 / eec (1) is hereby repealed .</en>
</sentence>
<sentence id="2517">
<ro>prezentul regulament este obligatoriu în toate elementele sale si se aplica direct în toate statele membre .</ro>
<en>this regulation shall be binding in its entirety and directly applicable in all member states .</en>
</sentence>
<sentence id="2518">

<ro>public în domeniul transportului feroviar , rutier si naval</ro>
<en>the council of the european communities ,</en>
</sentence>
<sentence id="2519">
<ro>având în vedere decizia consiliului din 13 mai 1965 privind armonizarea unor prevederi cu efecte asupra concurentei în domeniul transportului feroviar , rutier si naval1 ,</ro>
<en>having regard to the council decision of 13 may 1965 (1) on the harmonisation of certain provisions affecting competition in transport by rail , road and inland waterway ;</en>
</sentence>
<sentence id="2520">
<ro>având în vedere propunerea comisiei ,</ro>
<en>having regard to the proposal from the commission ;</en>
</sentence>
<sentence id="2521">
<ro>având în vedere avizul parlamentului european2 ,</ro>
<en>having regard to the opinion of the european parliament (2) ;</en>
</sentence>
<sentence id="2522">
<ro>având în vedere avizul comitetului economic si social3 ,</ro>
<en>having regard to the opinion of the economic and social committee (3) ;</en>
</sentence>
.....
<sentence id="2524">
<ro>întrucât este oportun sa se prevada ca întreprinderile de transport nu pot sa prezinte cereri de eliminare a obligatiilor de serviciu public decât daca aceste obligatii atrag pentru ele dezavantaje economice stabilite conform metodelor comune definite în prezentul regulament ;</ro>
<en>whereas it is appropriate to provide that transport undertakings may apply for the termination of public service obligations only where such obligations involve them in economic disadvantages determined in accordance with common procedures defined in this regulation ;</en>
</sentence>
<sentence id="2525">
<ro>întrucât este necesar sa se aplice prevederile prezentului regulament la orice caz nou de obligatii de serviciu public definite în prezentul regulament , care pot fi impuse unei întreprinderi de transport ;</ro>
<en>whereas the provisions of this regulation should be applied to any new public obligation as defined in this regulation imposed on a transport undertaking ;</en>
</sentence>
<sentence id="2526">
<ro>întrucât comisia trebuie sa poata obtine din partea statelor membre toate informatiile utile cu privire la aplicarea prezentului regulament ;</ro>
<en>whereas the commission must be able to obtain from member states all relevant information concerning the operation of this regulation ;</en>

</sentence>
<sentence id="2527">
<ro>adopta prezentul regulament :</ro>
<en>has adopted this regulation :</en>
</sentence>
.....
<sentence id="2529">
<ro>1 . statele membre elimina obligatiile inerente notiunii de serviciu public , definite în prezentul regulament , impuse în domeniul transporturile feroviare , rutiere si navale .</ro>
<en>1 . member states shall terminate all obligations inherent in the concept of a public service as defined in this regulation imposed on transport by rail , road and inland waterway .</en>
</sentence>
.....
</sentences>

D.2 RoGER

In this section an excerpt from the RoGER corpus for German - Romanian is presented. The format is the same as in **Section D.1**.

<?xml version="1.0" encoding="UTF-8"?>
<sentences>
<sentence id="1">
<de>bedienungsanleitung</de>
<ro>ghidul utilizatorului</ro>
</sentence>
<sentence id="2">
<de>konformitaetserklaerung</de>
<ro>declaratie de conformitate</ro>
</sentence>
<sentence id="3">
<de>wir , die nameprod corporation , erklaeren voll verantwortlich , dass das produkt npl - num den bestimmungen der folgenden direktive des rats der europaeischen union entspricht : num .</de>
<ro>noi , firma nameprod corporation declaram pe proprie raspundere ca produsul npl - num este in conformitate cu prevederile urmatoarei directive a consiliului : num .</ro>
</sentence>
<sentence id="4">
<de>alle rechte vorbehalten .</de>
<ro>toate drepturile rezervate .</ro>
</sentence>
<sentence id="5">

<de>der inhalt dieses dokuments darf ohne vorherige schriftliche genehmigung durch nameprod in keiner form , weder ganz noch teilweise , vervielfaeltigt , weitergegeben , oder gespeichert werden .</de>
<ro>este interzisa reproducerea , transferul , distribuirea si stocarea unor parti sau a intregului continut al acestui material fara permisiunea prealabila a firmei nameprod .</ro>
</sentence>
<sentence id="6">
<de>nameprod , nameprod connecting people , xpress - on und pop - port sind marken oder eingetragene marken der nameprod corporation .</de>
<ro>nameprod , nameprod connecting people , xpress - on si pop - port sunt marci comerciale sau marci inregistrate ale nameprod corporation .</ro>
</sentence>
<sentence id="7">
<de>andere in diesem handbuch erwaehnte produkt - und firmennamen koennen marken oder handelsnamen ihrer jeweiligen eigentuemer sein .</de>
<ro>alte nume de produse si de firme mentionate aici pot fi nume comerciale sau marci comerciale apartinind proprietarilor respectivi .</ro>
</sentence>
<sentence id="8">
<de>nameprod tune ist eine tonmarke der nameprod corporation .</de>
<ro>nameprod tune este o marca de sunet a corporatei nameprod .</ro>
</sentence>
<sentence id="9">
<de>nameprod behaelt sich deshalb das recht vor , ohne vorherige ankuendigung an jedem der in dieser dokumentation beschriebenen produkte aenderungen und verbesserungen vorzunehmen .</de>
<ro>ca atare , nameprod is rezerva dreptul de a face modificari si imbunatatiri oricarui produs descris in acest document fara notificare prealabila .</ro>
</sentence>
<sentence id="10">
<de>nameprod ist unter keinen umstaenden verantwortlich fuer den verlust von daten und einkuenften oder fuer jedwede besonderen , beilaeufigen , mittelbaren oder unmittelbaren schaeden , wie immer diese auch zustande gekommen sind .</de>
<ro>in nici un caz nameprod nu va fi raspunzatoare pentru nici un fel de pierderi de informatii sau de venituri sau pentru nici un fel de daune speciale , incidente , subsecvente sau indirecte , oricum s - ar fi produs .</ro>
</sentence>
<sentence id="11">
<de>der inhalt dieses dokuments wird so praesentiert , wie er aktuell vorliegt .</de>
<ro>continutul acestui document trebuie luat " ca atare " .</ro>
</sentence>
<sentence id="12">

<de>nameprod uebernimmt weder ausdruecklich noch stillschweigend irgendeine gewaehrleistung fuer die richtigkeit oder vollstaendigkeit des inhalts dieses dokuments , einschliesslich , aber nicht beschraenkt auf die stillschweigende garantie der markttauglichkeit und der eignung fuer einen bestimmten zweck , es sei denn , anwendbare gesetze oder rechtsprechung schreiben zwingend eine haftung vor .</de>
<ro>cu exceptia cazurilor prevazute de legea aplicabila , nici un fel de garantii , explicite sau implicite , incluzind , dar fara a se limita la garantiile implicite de vandabilitate si adecvare la un scop anume , nu se ofera in legatura cu acuratetea , corectitudinea sau continutul acestui document .</ro>
</sentence>
.....
<sentence id="14">
<de>die verfuegbarkeit bestimmter produkte variiert je nach region .</de>
<ro>este posibil ca nu toate produsele sa fie disponibile in zona dumneavoastra .</ro>
</sentence>
<sentence id="15">
<de>wenden sie sich an einen nameprod vertragspartner in ihrer naehe .</de>
<ro>va rugam sa consultati pentru aceasta cel mai apropiat dealer nameprod .</ro>
</sentence>
<sentence id="16">
<de>entsorgen sie die akkus entsprechend den behoerdlichen vorgaben .</de>
<ro>va rugam sa aruncati acumulatoarele la deseuri , conform dispozitiilor in vigoare .</ro>
</sentence>
<sentence id="17">
<de>inhalt</de>
<ro>cuprins</ro>
</sentence>
<sentence id="18">
<de>sicherheitshinweise</de>
<ro>pentru siguranta dvs .</ro>
</sentence>
<sentence id="19">
<de>allgemeine informationen</de>
<ro>informatii generale</ro>
</sentence>
.....
<sentence id="22">
<de>multimedia - mitteilungsdienst (mms)</de>
<ro>serviciul de mesaje multimedia (mms)</ro>
</sentence>
.....
</sentences>

Appendix E

Translation Examples

In this appendix we present some examples of the translations obtained with **Mb_SMT**, *Lin-EBMT* and $Lin - EBMT^{REC+}$.

E.1 JRC-Acquis

German - Romanian

Input: Beschluss des Gemeinsamen EWR-Ausschusses
Reference: DECIZIA COMITETULUI MIXT AL SEE
Mb_SMT: Decizia Comitetului mixt ewr-ausschusses
Lin-EBMT: Ale Decizia (AAA1)
$Lin - EBMT^{REC+}$: Decizia ale ewr-ausschusses

Input: Nr. 163/2002
Reference: nr. 163/2002
Mb_SMT: Nr. 163 / 2002
Lin-EBMT: Nr 163 2002
$Lin - EBMT^{REC+}$: Nr. / 2002 163

Input: vom 6. Dezember 2002
Reference: din 6 decembrie 2002
Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Din 6 decembrie 2002

Input: zur Änderung des Anhangs II (Technische Vorschriften, Normen, Prüfung und Zertifizierung) des EWR-Abkommens
Reference: de modificare a anexei II (Regulamente tehnice, norme, testare şi certificare) la Acordul SEE
Mb_SMT: De modificare a anexei II (reglementări tehnice, standardelor, verificarea şi certificarea în acordul SEE)
Lin-EBMT: Reguli standarde anexa II şi de modificare a examinării tehnice rezerva certificării membre ale Acordului privind SEE $Lin - EBMT^{REC+}$: De modificare a anexei II), şi (rezerva

certificării. Tehnice standardelor membre ale Acordului privind SEE

Input: DER GEMEINSAME EWR-AUSSCHUSS -
Reference: COMITETUL MIXT AL SEE,
Mb_SMT: 4. Ewr-ausschuss,
Lin-EBMT: Organizarea comună din
$Lin - EBMT^{REC+}$: Ewr-ausschuss comune,

Input: gestützt auf das Abkommen über den Europäischen Wirtschaftsraum, geändert durch das Anpassungsprotokoll zum Abkommen über den Europäischen Wirtschaftsraum, nachstehend %quot%Abkommen%quot% genannt, insbesondere auf Artikel 98,
Reference: având în vedere Acordul privind Spaţiul Economic European, modificat de protocolul de adaptare a Acordului privind Spaţiul Economic European, denumit în continuare "acordul", în special art. 98,
Mb_SMT: Având în vedere Acordul privind Spaţiul Economic European, modificat de protocolul anpassungsprotokoll la Acordul privind Spaţiul Economic European, denumit în continuare "acordul", în special art. 98,
Lin-EBMT: Având în vedere Acordul privind Spaţiul Economic European 98 denumit în continuare acordul privind acordurile europene în modificată în special art
$Lin - EBMT^{REC+}$: Având în vedere acordul european, în special art la europene, la ", a acordului în continuare denumit privind Spaţiul Economic" 985 privind Spaţiul Economic European mărfurile anpassungsprotokoll acordul

Input: (1) Anhang II des Abkommens wurde durch den Beschluss des Gemeinsamen EWR-Ausschusses Nr. 13/2001 vom 23. Februar 2001(1) geändert.
Reference: (1) Anexa II la acord a fost modificată de Decizia Comitetului mixt al SEE nr. 13/2001 din 23 februarie 20011.
Mb_SMT: (1) Anexa II la acord a fost constituit prin Decizia Comitetului mixt ewr-ausschusses nr 13 / 2001 din 23 februarie 2001 (1).
Lin-EBMT: Anexa II a 1 din acord nr 13 23 2001 prin Decizia 2001 1 a fost
$Lin - EBMT^{REC+}$: (1) nr. În a fost prin Decizia 2001 / 13 din acord urmează anexa II din 23 februarie 2001 modifică ewr-ausschusses

Input: (2) Die Sechsundzwanzigste Richtlinie 2002/34/EG der Kommission vom 15. April 2002 zur Anpassung der Anhänge II, III und VII der Richtlinie 76/768/EWG des Rates zur Angleichung der Rechtsvorschriften der Mitgliedstaaten über kosmetische Mittel an den technischen Fortschritt(2) ist in das Abkommen aufzunehmen -
Reference: (2) A douăzeci şi şasea Directivă a Comisiei 2002/34/CE din 15 aprilie 2002 de adaptare la progresul tehnic a anexelor II, III şi VII la Directiva Consiliului 76/768/CEE privind apropierea legislaţiilor statelor membre cu privire la produsele cosmetice2 trebuie integrată în acord,
Mb_SMT: (2) A DOUĂZECIŞIŞASEA DIRECTIVĂ A COMISIEI 2002 / 34 / CE din 15 aprilie 2002 privind adaptarea la progresul tehnic a anexelor II, III şi VII la Directiva 76 / 768 / CEE a Consiliului de apropiere a legislaţiilor statelor membre referitoare la produsele cosmetice (2) din acordul,
Lin-EBMT: III 2 din acord de adaptare la progresul tehnic a anexelor II şi VII la Directiva Con-

siliului nr.76 768 CEE privind armonizarea legislaţiei statelor membre în domeniul produselor cosmetice 34 CE din 15 aprilie 2002 2 A DOUĂZECIŞIŞASEA DIRECTIVĂ A COMISIEI 2002
$Lin-EBMT^{REC+}$: (CE) şi VII la Directiva Consiliului nr.76 / 768 / CEE privind armonizarea legislaţiei statelor membre în domeniul produselor cosmetice din 15 aprilie 2002 / 2002 / 34 de adaptare la progresul tehnic a anexelor II, III acordurile 2. Directiva Comisiei 2 a douăzecişişasea

English - Romanian

Input: Decision of the EEA Joint Committee
Reference: DECIZIA COMITETULUI MIXT AL SEE
Mb_SMT, and *Lin-EBMT*: Decizia Comitetului mixt SEE,
$Lin-EBMT^{REC+}$: Decizie Comitetului SEE

Input: No 163/2002
Reference: nr. 163/2002
Mb_SMT: Nr. 163 / 2002
Lin-EBMT: Iunie 1983 2002
$Lin-EBMT^{REC+}$: No / 2002 163

Input: of 6 December 2002
Reference: din 6 decembrie 2002
Mb_SMT, *Lin-EBMT*, and $Lin-EBMT^{REC+}$: Din 6 decembrie 2002

Input: amending Annex II (Technical regulations, standards, testing and certification) to the EEA Agreement
Reference: de modificare a anexei II (Regulamente tehnice, norme, testare şi certificare) la Acordul SEE
Mb_SMT: De modificare a anexei II (reglementările tehnice, standardele de testare şi de certificare la Acordul SEE)
Lin-EBMT: Anexei II şi de modificare reglementările tehnice efectuate de certificare standarde Acordul SEE testare
$Lin-EBMT^{REC+}$: (standardele Spaţiul Economic European regulamente tehnice, şi certificarea anexa II) Acordul testarea

Input: THE EEA JOINT COMMITTEE,
Reference: COMITETUL MIXT AL SEE,
Mb_SMT: SEE Comitetului mixt,
Lin-EBMT: Comitetului mixt SEE
$Lin-EBMT^{REC+}$: Comitetului Comun SEE,

Input: Having regard to the Agreement on the European Economic Area, as amended by the Protocol adjusting the Agreement on the European Economic Area, hereinafter referred to as %quot%the Agreement%quot%, and in particular Article 98 thereof,
Reference: având în vedere Acordul privind Spaţiul Economic European, modificat de protocolul de adaptare a Acordului privind Spaţiul Economic European, denumit în continuare "acordul", în special art. 98,
Mb_SMT: Având în vedere Acordul privind Spaţiul Economic European, modificat de protocolul

de modificare a Acordului privind Spaţiul Economic European, denumit în continuare "acordul", în special art. 98,
Lin-EBMT: Denumit în continuare Acordul Europene, în special în acordul modificate având în vedere Acordul privind Spaţiul Economic European de 98 modificării Protocolului
$Lin-EBMT^{REC+}$: Având în vedere acordul european, în special 98 Europene, denumit în continuare "a la Spaţiul Economic," întrucât Acordul privind Spaţiul Economic,, modificat de protocolul modificării acordul

Input: (1) Annex II to the Agreement was amended by Decision of the EEA Joint Committee No 13/2001 of 23 February 2001(1).
Reference: (1) Anexa II la acord a fost modificată de Decizia Comitetului mixt al SEE nr. 13/2001 din 23 februarie 20011.
Mb_SMT: (1) Anexa II la acord a fost modificată de Decizia Comitetului mixt al SEE nr. 13 / 2001 din 23 februarie 2001 (1).
Lin-EBMT: Comitetului pentru a fost modificată de Decizia 2001 13 februarie 2001 23 din anexa II la 1 Acord SEE
$Lin-EBMT^{REC+}$: (1) nr. / 2001 2001 23 februarie. Anexa II modifică prin Decizia 1 proiectul comun acord a fost Comitetul 13 SEE

Input: (2) Twenty-sixth Commission Directive 2002/34/EC of 15 April 2002 adapting to technical progress Annexes II, III and VII to Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products(2) is to incorporated into the Agreement,
Reference: (2) A douăzeci şi şasea Directivă a Comisiei 2002/34/CE din 15 aprilie 2002 de adaptare la progresul tehnic a anexelor II, III şi VII la Directiva Consiliului 76/768/CEE privind apropierea legislaţiilor statelor membre cu privire la produsele cosmetice2 trebuie integrată în acord,
Mb_SMT: (2) A DOUĂZECIŞIŞASEA DIRECTIVĂ A COMISIEI 2002 / 34 / CE din 15 aprilie 2002 privind adaptarea la progresul tehnic a anexelor II, III şi VII la Directiva Consiliului 76 / 768 / CEE privind apropierea legislaţiilor statelor membre referitoare la produsele cosmetice (2) este inclus în acord,
Lin-EBMT: Anexelor II şi VII la Directiva Consiliului 76 III este de adaptare la progresul tehnic a Comisiei 2002 768 34 15 aprilie 2002 2 în a douăzecişişasea 2 ce a Directivei CEE privind apropierea legislaţiilor statelor membre cu privire la produsele cosmetice incorporated
$Lin-EBMT^{REC+}$: (2), şi VII la Directiva Consiliului 76 / 768 / CEE privind apropierea legislaţiilor statelor membre cu privire la produsele cosmetice a incorporated în de adaptare la progresul tehnic a douăzecişişasea anexelor II, III directivei este Comisiei 2002 / 34 15 aprilie 2002 / CE din a

Romanian - German

Input: COMITETUL MIXT AL SEE,
Reference: DER GEMEINSAME EWR-AUSSCHUSS -
Mb_SMT: Der den Gemischten Ausschuß,
Lin-EBMT: EWR
$Lin-EBMT^{REC+}$: Der Ausschuss - dem EWR eingesetzten Gemischten Ausschuß

Input: având în vedere Acordul privind Spaţiul Economic European, modificat de protocolul de adaptare a Acordului privind Spaţiul Economic European, denumit în continuare "acordul", în special art. 98,
Reference: gestützt auf das Abkommen über den Europäischen Wirtschaftsraum, geändert durch das Anpassungsprotokoll zum Abkommen über den Europäischen Wirtschaftsraum, nachstehend %quot%Abkommen%quot% genannt, insbesondere auf Artikel 98,
Mb_SMT: Gestützt auf das Abkommen über den Europäischen Wirtschaftsraum, geändert durch das Protokoll zur Anpassung des Abkommens über den Europäischen Wirtschaftsraum, nachstehend% quot% Abkommen% quot%), insbesondere auf Artikel 98,
Lin-EBMT: Gestützt auf das Abkommen über den Europäischen Wirtschaftsraum Europäischen Vereinbarung das Protokoll über nachfolgend geändert insbesondere auf Artikel des Abkommens mit von den Europäischen Wirtschaftsraum genannt 98.
$Lin-EBMT^{REC+}$: Gestützt auf das Protokoll über den Europäischen Wirtschaftsraum nachfolgend% quot% des% quot% genannt, insbesondere auf Artikel 98 Europäischen Vereinbarung mit, geändert des Abkommens, Anpassungsmaßnahmen über den Europäischen Wirtschaftsraum,

Input: întrucât:
Reference: in Erwägung nachstehender Gründe:
Mb_SMT, *Lin-EBMT*, and $Lin-EBMT^{REC+}$: In Erwägung nachstehender Gründe:

Input: (1) Anexa II la acord a fost modificată de Decizia Comitetului mixt al SEE nr. 13/2001 din 23 februarie 20011.
Reference: (1) Anhang II des Abkommens wurde durch den Beschluss des Gemeinsamen EWR-Ausschusses Nr. 13/2001 vom 23. Februar 2001(1) geändert.
Mb_SMT: (1) Anhang II der Übereinkommen wurde geändert durch die Entscheidung des Gemischten Ausschusses des EWR Nr. 13 / 2001 des Rates vom 23. Februar 20011.
Lin-EBMT: 625H 1. Februar 2001 eingereicht hat bei Vorlage 23 10 13 der für den Wirkstoff Clefoxydim Überkommens eingesetzten Gemeinsamen vertreten 2001
$Lin-EBMT^{REC+}$: (1). Hat bei der dem 23. Februar Vorlage 625H) eingereicht 2001 2001 eingesetzten Gemischten Ausschuß für den Wirkstoff Clefoxydim (10 13

Romanian - English

Input: DECIZIA COMITETULUI MIXT AL SEE
Reference: Decision of the EEA Joint Committee
Mb_SMT: Decision of the EEA Joint Committee
Lin-EBMT: 3 the Committee the Decision EEA Joint
$Lin-EBMT^{REC+}$: 3 the joint EEA decision the Committee

Input: nr. 163/2002
Reference: No 163/2002
Mb_SMT: No 163 / 2002
Lin-EBMT: No 163 2002
$Lin-EBMT^{REC+}$: Nr. / 2002 163

Input: din 6 decembrie 2002
Reference: of 6 December 2002

Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Of 6 December 2002

Input: de modificare a anexei II (Regulamente tehnice, norme, testare şi certificare) la Acordul SEE
Reference: amending Annex II (Technical regulations, standards, testing and certification) to the EEA Agreement
Mb_SMT: Amending Annex II (technical regulations, standards, testing and certification. In the EEA Agreement
Lin-EBMT: Technical standards performance testing certification Regulations and to the EEA Agreement amending Annex II
$Lin - EBMT^{REC+}$: Regulations to amend Annex II, and certification testing standards, in the EEA agreement on technical

Input: COMITETUL MIXT AL SEE,
Reference: THE EEA JOINT COMMITTEE,
Mb_SMT: The Joint Committee of the relevance,
Lin-EBMT: The Joint Committee EEA
$Lin - EBMT^{REC+}$: The Committee, joint of EEA

E.2 RoGER

German - Romanian

Input: Den vollstaendigen Text der Konformitaetserklaerung finden Sie unter : WWWSITE .
Reference: O copie a declaratiei de conformitate poate fi gasita pe pagina de Internet WWWSITE .
Mb_SMT: Vollstaendigen text , de Declaratie de conformitate , consultati : WWWSITE .
Lin-EBMT: Consultati WWWSITE text Declaratie de conformitate
$Lin - EBMT^{REC+}$: Text , consultati Declaratie de conformitate WWWSITE . Vollstaendigen

Input: NAMEPROD entwickelt entsprechend seiner Politik die Produkte staendig weiter .
Reference: NAMEPROD duce o politica de dezvoltare continua .
Mb_SMT: NAMEPROD entwickelt uzati pozitia politik de produkte staendig si .
Lin-EBMT: Ca nu NAMEPROD sa stiintifice
$Lin - EBMT^{REC+}$: NAMEPROD . Ca nu staendig sa politik stiintifice

Input: Uebersicht ueber die Telefonfunktionen
Reference: Prezentarea functiilor telefonului
Mb_SMT: Prezentarea functiilor
Lin-EBMT and $Lin - EBMT^{REC+}$: Prezentarea functiilor telefonului

Input: Verteilerlisten
Reference: Liste de distributie
Mb_SMT: Liste distributie

Lin-EBMT and $Lin - EBMT^{REC+}$: Liste de distributie

Input: Loeschen von Namen und Nummern
Reference: Stergerea numelor si a numerelor
Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Stergerea numelor si a numerelor

Input: Dienstnummern
Reference: Numere de servicii
Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Numere servicii

Input: Sicherheitshinweise
Reference: Pentru siguranta dvs .
Mb_SMT: De siguranta
Lin-EBMT and $Lin - EBMT^{REC+}$: Pentru siguranta Dvs .

Input: Wenden Sie sich an Ihren Netzbetreiber oder Diensteanbieter , um Informationen zur Verfuegbarkeit und zu den Nutzungsvoraussetzungen von GPRS zu erhalten .
Reference: Contactati operatorul Dvs de retea sau furnizorul Dvs de servicii in legatura cu disponibilitatea si abonamentul la serviciul GPRS .
Mb_SMT: Contactati operatorul Dvs de retea sau furnizorul Dvs de servicii pentru a pentru informatii referitoare la disponibilitatea si la modalitatea de GPRS .
Lin-EBMT: Pentru informatii referitoare la disponibilitatea de si la modalitatea abonare la acesta contactati operatorul pentru mai si in , contactati operatorul Dvs retea sau furnizorul Dvs servicii GPRS pentru a
$Lin - EBMT^{REC+}$: Contactati furnizorul Dvs de servicii pentru a . Pentru informatii referitoare la disponibilitate si modalitatea abonare la GPRS pentru retea sau

English - Romanian

Input: A copy of the Declaration of Conformity can be found from WWWSITE .
Reference: O copie a declaratiei de conformitate poate fi gasita pe pagina de Internet WWWSITE .
Mb_SMT: O copie a Declaratie de conformitate poate fi found din WWWSITE .
Lin-EBMT: O copie cartii de pot fi Declaratie de conformitate WWWSITE
$Lin - EBMT^{REC+}$: O copie fi . Cartii de Declaratie de conformitate WWWSITE found

Input: NAMEPROD operates a policy of continuous development .
Reference: NAMEPROD duce o politica de dezvoltare continua .
Mb_SMT: NAMEPROD operates un policy de continuous development .
Lin-EBMT: NAMEPROD o a celular functioneaza folosind semnale
$Lin - EBMT^{REC+}$: NAMEPROD . O a continuous celular functioneaza folosind semnale policy development

Input: Erasing names and numbers
Reference: Stergerea numelor si a numerelor

Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Stergerea numelor si a numerelor

Input: Info numbers
Reference: Numere de informatii
Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Numere informatii

Input: Service numbers
Reference: Numere de servicii
SMT: Numere servicii
Lin-EBMT, and $Lin - EBMT^{REC+}$: Numere de servicii

Input: To - do list
Reference: De rezolvat
Mb_SMT: Pentru a - nu lista
Lin-EBMT: Lista
$Lin - EBMT^{REC+}$: De

Input: The loudspeaker has been activated , see Loudspeaker on page NUM .
Reference: Difuzorul a fost activat , a se vedea Difuzorul la pagina NUM .
MbSMT: Difuzorul a fost Activata , consultati Difuzorul la pagina NUM .
Lin-EBMT: Consultati la pagina NUM a fost Difuzorul Difuzor activate
$Lin - EBMT^{REC+}$: Consultati Difuzorul la pagina NUM . A fost Difuzor activate

Input: To add a folder , press Options and select Add folder . **Reference:** Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .
Mb_SMT: Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .
Lin-EBMT: Pentru a adauga un apasati Optiuni si selectati Adaugati dosarul dosar
$Lin - EBMT^{REC+}$: Pentru a adauga un dosar , apasati Optiuni si selectati Adaugati dosarul .

Romanian - German

Input: Pentru siguranta dvs .
Reference: Sicherheitshinweise
Mb_SMT, *Lin-EBMT*, and $Lin - EBMT^{REC+}$: Sicherheitshinweise

Input: Etichetele contin informatii importante pentru service si pentru alte actiuni de asistenta a clientilor .
Reference: Die Aufkleber enthalten wichtige Informationen fuer den Service und die Kundenunterstuetzung .
Mb_SMT: Im contin Informationen importante fuer Service und fuer andere actiuni von asistenta zu clientilor .
Lin-EBMT: Fuer und andere Wichtige Sicherheitshinweise enthalten Service im
$Lin - EBMT^{REC+}$: Um und , um Wichtige Sicherheitshinweise andere enthalten . Service im actiuni asistenta clientilor

Input: Pentru a schimba codul si a seta telefonul sa solicite acest cod , consultati Setari sigu-

ranta la pagina NUM .
Reference: Wie Sie den Code aendern und das Telefon so einstellen , dass es Sie zur Eingabe des Codes auffordert , erfahren Sie unter Sicherheitseinstellungen auf Seite NUM .
Mb_SMT: Um den Brieftaschen-Code zu aendern und legen Sie fest , dass dieses falschen finden Sie unter Sicherheitseinstellungen auf Seite NUM .
Lin-EBMT: Informationen , Sie , damit das Telefon finden unter auf Seite NUM unter Sie Sicherheitseinstellungen der dieses aendern
$Lin-EBMT^{REC+}$: Informationen , damit das Telefon , finden Sie unter der Sie unter Sicherheitseinstellungen auf Seite NUM . dieses aendern ,

Input: Puteti memora imaginile si sunetele pentru a va personaliza telefonul .
Reference: Sie koennen die Bilder und Toene speichern , um das Telefon individuell anzupassen .
Mb_SMT: Sie koennen Bilder und Ruftoene in zu speichern , um die personaliza das Telefon .
Lin-EBMT: Fuer und das Telefon Sie Speichern Anpassen des Profils
$Lin-EBMT^{REC+}$: Um die Bilder und das Telefon Sie speichern . Ruftoene in anpassen des Profils

Input: Contactati operatorul Dvs de retea sau furnizorul Dvs de servicii in legatura cu disponibilitatea si abonamentul la serviciul GPRS .
Reference: Wenden Sie sich an Ihren Netzbetreiber oder Diensteanbieter , um Informationen zur Verfuegbarkeit und zu den Nutzungsvoraussetzungen von GPRS zu erhalten .
Mb_SMT: Wenden Sie sich an Ihren Netzbetreiber oder Diensteanbieter sich bezueglich der Verfuegbarkeit und abonamentul auf den GPRS .
Lin-EBMT: An Ihren Netzbetreiber oder Diensteanbieter auf wenden sich bezueglich Verfuegbarkeit und GPRS
$Lin-EBMT^{REC+}$: An Ihren Netzbetreiber oder Diensteanbieter . wenden sich bezueglich Verfuegbarkeit GPRS und auf abonamentul

Romanian - English

Input: Daca selectati Nu , mesajele expediate nu sunt memorate .
Reference: If you select No , the sent messages are not saved .
Mb_SMT: If you select not , the sent messages are not saved .
Lin-EBMT: The sent messages are saved if you select not
$Lin-EBMT^{REC+}$: If no , select messages are saved the sent .

Input: Tastati adresa , apasati PICT pentru un punct , apoi apasati OK .
Reference: Key in the address , press PICT for a dot , and press OK .
Mb_SMT: Key in the address , press PICT for a dot , and press OK .
Lin-EBMT: Key in address PICT for a dot and press OK press
$Lin-EBMT^{REC+}$: Key in address , press PICT for a dot , and press ok .

Input: Tastati parola si apasati OK .
Reference: Key in the password and press OK .
Mb_SMT, *Lin-EBMT*, and $Lin-EBMT^{REC+}$: Key in the password and press OK .

Input: Pentru tastarea numerelor , consultati Efectuarea unui apel la pagina NUM .

E. TRANSLATION EXAMPLES

Reference: To key in the numbers , see Making a call on page NUM .
Mb_SMT: For tastarea numbers , see Making a call on page NUM .
Lin-EBMT: To see call on page NUM Making a numbers Keying letters
$Lin - EBMT^{REC+}$: To making a numbers , see call on page num . keying letters

Appendix F

Technical Information

To have a better overview of the environment in which the experiments have been made, this appendix contains a brief overview on the technical conditions of the experiments.

All experiments have been run on a computer with Ubuntu as operating system. The technical characteristics of the computer were: two Inter(R) Pentium(R) 4 CPUs 3.00 Ghz, frequency 2800.000 Mhz, L2 cache 2048 KB, memory 3015 MiB, 1 GB main memory, 75 GB hard-disk.

While running the experiments, it was noticed that the Java-based systems (the EBMT systems) were slower than the $C++$ based ones (the Moses-based systems). The heap size for running the Java program was between 256M and 1664M (Java parameters -Xmx1664m -Xms256m). For a test data-set from JRC-Acquis the translation time for **Mb_SMT** was some hours[1] and for **Lin-EBMT** several days[2]. Between 18 and 21 minutes were necessary for *Lin-EBMT* to translate the 133 test sentences from RoGER, for each direction of translation. Less time was needed by the **Mb_SMT** system.

Concerning the translation time, on the whole, $Lin-EBMT^{REC+}$ requires less time than *Lin-EBMT*, although additional time is needed for the extraction of the constraints. This happens due to the changes in the recombination matrix: the search algorithm in the recombination matrix changes when First-Word-Constraints are set.

The programming language has a major influence and Java is known to be slower than $C++$, when the $C++$ code is compiled under certain circumstances (`http://verify.stanford.edu/uli/java_cpp.html` - last accessed on June 27th, 2011). A comparison of the two programming languages can be found on `http://en.wikipedia.org/wiki/Comparison_of_Java_and_C%2B%2B`[3]. An evaluation from the time-performance point of view is shown on `http://verify.stanford.edu/uli/java_cpp.html`[4]. Optimization methods for Java can be applied to improve speed (time, resources), but in this thesis no

[1] From two to four hours.

[2] The exact duration cannot be established, as problems with the server were encountered during the translation process. Sometimes more than one week was needed.

[3] Last accessed on June 27th, 2011

[4] Last accessed on June 27th, 2011.

methods have been implemented in this direction. Moreover, the algorithm of the EBMT systems (e.g. the comparisons in LCSS, finding the solution in the recombination matrix) slows down the translation process.

Appendix G

Additional Ranking Results

In this appendix we extend the information from **Chapter 10** by presenting additional ranking results for **Mb_SMT**, *Lin-EBMT* and $Lin - EBMT^{REC+}$. We ranked the systems manually by analyzing part of the translations which have Romanian as TL. More information on the ranking methodology has been presented in **Chapter 10**.

Table G.1 shows how often a system was classified on a specific place. The value is calculated as percentage (%) from the total number of analyzed sentences.

G. ADDITIONAL RANKING RESULTS

Data	**Place**	**Mb_SMT**	*Lin-EBMT*	$Lin-EBMT^{REC+}$
DEU-RON RoGER	1st	82%	50%	54%
DEU-RON RoGER	2nd	14%	46%	38%
DEU-RON RoGER	3rd	4%	4%	8%
ENG-RON RoGER	1st	100%	52%	52%
ENG-RON RoGER	2nd	-	44%	42%
ENG-RON RoGER	3rd	-	4%	6%
ENG-RON RoGER, with POS	1st	94%	48%	48%
ENG-RON RoGER, with POS	2nd	4%	40%	50%
ENG-RON RoGER, with POS	3rd	2%	12%	2%
DEU-RON JRC-Acquis	1st	93%	47%	47%
DEU-RON JRC-Acquis	2nd	7%	28%	38%
DEU-RON JRC-Acquis	3rd	-	25%	15%
ENG-RON JRC-Acquis	1st	90%	48%	40%
ENG-RON JRC-Acquis	2nd	5%	29%	51%
ENG-RON JRC-Acquis	3rd	5%	23%	9%
All language-pairs RoGER (no POS)	1st	91%	51%	53%
All language-pairs RoGER (no POS)	2nd	7%	45%	40%
All language-pairs RoGER (no POS)	3rd	2%	4%	7%
All language-pairs JRC-Acquis	1st	91.5%	47.5%	43.5%
All language-pairs JRC-Acquis	2nd	6%	28.5%	44.5%
All language-pairs JRC-Acquis	3rd	2.5%	24%	12%
Both corpora DEU-RON	1st	89.33%	48%	49.33%
Both corpora DEU-RON	2nd	9.33%	34%	38%
Both corpora DEU-RON	3rd	1.33%	18%	12.67%
Both corpora ENG-RON (no POS)	1st	93.33%	49.33%	44%
Both corpora ENG-RON (no POS)	2nd	3.33%	34%	48%
Both corpora ENG-RON (no POS)	3rd	3.33%	16.67%	8%

Table G.1: System ranking.

References

Mosleh H. Al-Adhaileh and Tang Enya Kong. Example-based machine translation based on the synchronous sstc annotation schema. In *Proceedings of the MT-Summit VII*, pages 244–249, Singapore, 1999. 31, 142

Gabriela Alboiu and Virginia Motapanyane. *Comparative Studies in Romanian Syntax*, chapter The Generative Approach to Romanian Grammar: an Overview, pages 1–48. Elsevier, 2000. Editor: Virginia Motapanyane. 145

Juan Carlos Amengual, Jose Miguel Benedi, Francisco Casacuberta, Asuscion Castano, Antonio Castellanos, Victor Jimenez, David Llorens, Andreas Marza, Moises Pastor, Federico Prat, Enrique Vidal, and Juan Miguel Vilar. The eutrans-i speech translation system. *Machine Translation*, 15(1/2):75–103, 2000. 9

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. Mt evaluation: Human-like vs. human acceptable. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 17–24, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P06/P06-2003`. 84

George A.Miller and J.G.Beebe-Center. Some psychological methods for evaluating the quality of translations. *Mechanical Translation*, 3(3):73–80, December 1956. 83

Tantely Andriamanankasina, Kenji Araki, and Koji Tochinai. *EBMT of POS-ragged sentences via inductive learning*, pages 225–254. Kluwer Academic Publishers, 2003. ISBN 1-4020-1400-7 (hardback), 1-4020-1401-5 (paperback). Editors: Michael Carl and Andy Way. 85

Eiji Aramaki and Sadao Kurohashi. Example-based machine translation using structural translation examples. In *Proceedings of the International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2004]*, pages 91–94, Kyoto, Japan, 2004. 32, 140

Eiji Aramaki, Sadao Kurohashi, Hideki Kashioka, and Hideki Tanaka. Example-based machine translation without saying inferable predicate. In *Proceedings of the IJCNLP 2004 (the 1st International Joint Conference on Natural Language Processing)*, pages 38–45, 2004. 141

Doug J. Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, and Louisa Sadler. *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, 1994. ISBN: 1855542-17x. 11

Ion Barbuţă, Armenia Cicala, Elena Constantinovici, Teodor Cotelnic, and Alexandru Dirul. *Gramatica uzuala a limbii romane*. Litera Educational, 2000. The book is in Romanian. 40

REFERENCES

Jim Barnett, Inderjeet Mani, Paul Martin, and Elaine Richard. Reversible machine translation: What to do when the languages don't line up? In *Proceedings of the Workshop on Reversible Grammars in Natural Language Processing, ACL-91*, pages 61–70, Berkeley, California, 1991. University of California. 47

Lasse Bergroth, Harri Hakonen, and Timo Raita. A survey of longest common subsequence algorithms. In *Proceedings of the Seventh International Symposium on String Processing and Information Retrieval - SPIRE 2000*, pages 39–48, A Curuna, Spain, September 2000. ISBN: 0-7695-0746-8. 62

Alexandra Birch, Miles Osborne, and Phil Blunsom. Metrics for mt evaluation: Evaluating reordering. *Machine Translation*, 24:15–26, March 2010. ISSN 0922-6567. doi: http://dx.doi.org/10.1007/s10590-009-9066-5. URL `http://dx.doi.org/10.1007/s10590-009-9066-5`. 84

Chris Brockett, Takako Aikawa, Anthony Aue, Arul Menezes, Chris Quirk, and Hisami Suzuki. English-japanese example-based machine translation using abstract semantic representations. In *Proceedings of the COLING 2002 Workshop*, Taiwan, October 2002. 140

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, 1990. ISSN 0891-2017. 12

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics - Special issue on using large corpora: II*, 19:263–311, June 1993. ISSN 0891-2017. URL `http://portal.acm.org/citation.cfm?id=972470.972474`. 11, 12, 23, 55

Ralf D. Brown. Example-based machine translation in the pangloss system. In *Proceedings of the 16th conference on Computational linguistics*, pages 169–174, Morristown, NJ, USA, 1996. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992628.992660. 23, 25

Ralf D. Brown. Transfer-rule induction for example-based translation. In *Recent Advances in Example-Based Machine Translation*, pages 1–11. Kluwer Academic, 2001. 25, 140

Ralf D. Brown, Paul N. Bennett, Jaime G. Carbonell, Rebecca Hutchinson, and Peter Jansen. Reducing boundary friction using translation-fragment overlap. In *Proceedings of MT Summit IX*, pages 24–31, 2003. 24, 140

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the EACL*, pages 249–256, 2006. 84

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *In Proceedings of ACL-2007 Workshop on Statistical Machine Translation*, StatMT '07, pages 136–158, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1626355.1626373`. 84, 86, 122

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March, 30-31 2009. 41, 84, 86, 119, 122

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 17–53, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 978-1-932432-71-8. URL `http://portal.acm.org/citation.cfm?id=1868850.1868853`. 2, 137

Sander Canisius and Antal van den Bosch. A constraint satisfaction approach to machine translation. In *Proceedings of the 13th Annual Conference of the EAMT*, pages 182–189, Barcelona, May 2009. 69

Hailong Cao and Eiichiro Sumita. Filtering syntactic constraints for statistical machine translation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 17–21, Uppsala, Sweden, July, 11-16 2010. 69

Jaime G. Carbonell, Teruko Mitamura, and Eric Nyberg. The kant perspective: A critique of pure transfer (and pure interlingua, pure statistics, . . .). In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Empiricist vs. rationalist methods in MT*, pages 225–235., Montreal, June 1992. CCRIT-CWARC. 10

Michael Carl. Inducing translation templates for example-based machine translation. In *Proceedings of MTSummit VII*, pages 250–258, 1999. URL `citeseer.ist.psu.edu/carl99inducing.html`. 22, 26, 29, 35, 141

Michael Carl. A model of competence for corpus-based machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 997–1001, Morristown, NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992730.992792. 22

Michael Carl. A system-theoretical view of ebmt. *Machine Translation*, 19(3-4):229–249, 2005a. 22

Michael Carl. Re: [mt-list] phrasal smt vs ebmt. http://www.mail-archive.com/mt-list@eamt.org/msg00777.html, February 2005b. Last accessed 11 November 2008. 22

Michael Carl, Leonid L. Iomdin, and Oliver Streiter. Towards dynamic linkage of example-based and rule-based machine translation. In *ESSLLI '98 Machine Translation Workshop*, Saarbrucken, 1998. 18, 35

Alexandru Ceauşu. Colectarea şi procesarea documentelor româneşti ale corpusului jrc-acquis. In *In Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, pages 125–130, Iaşi, Romania, November 2008. Publisher: Ed. Univ. Alexandru Ioan Cuza, ISSN: 1843-911X. 41, 42, 44

Daniel Cer. *Parameterizing Phrase Based Statistical Machine Translation Models: An Analytic Study*. PhD thesis, University of Colorado, 2002. 52

Yee Seng Chan and Hwee Tou Ng. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P08/P08-1007`. 84

REFERENCES

John Chandioux. Meteo, an operational system for the translation of public weather forecasts. *American Journal of Computational Linguistics*, microfiche 46:pp.27–36, 1976. 10

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318, Santa Cruz, California, 1996. Overview on Smoothing Techniques. 53, 98

Ilyas Cicekli and Altay Guvenir. Learning translation templates from bilingual translation examples. *Applied Intelligence*, 15(1):57–76, 2001. 25, 26, 27, 83, 141

Ilyas Cicekli and Altay Guvenir. *Recent Advances in Example-based Machine Translation*, chapter Learning Translation Templates From Bilingual Translation Examples, pages 225–286. Kluwer Acad. Publ., 2003. 26, 141

Ilyas Cicekli and H. Altay Guvenir. Learning translation rules from a bilingual corpus. In *Proceedings of the 2nd International Conference on New Methods in Language Processing (NeMLaP-2)*, pages 90–97, Ankara, Turkey, September 1996. 141

Ilyas Cicekli and Halil Altay Guvenir. Learning translation templates from examples. *Information Systems*, 23(6):353–363, 1998. 22, 27, 141

William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI, Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78, August 9-10 2003. 15

Dana Cojocaru. *Romanian Grammar*. SEELRC, 2003. URL `http://www.seelrc.org:8080/grammar/pdf/stand_alone_romanian.pdf`. Published Online. 40

Brona Collins. *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*. PhD thesis, Trinity College, Dublin, 1998. 22, 36, 47, 142

Brona Collins and Padraig Cunningham. Adaptation guided retrieval in ebmt: A case-based approach to machine translation. In I. Faltings B Smith, editor, *Lecture Notes in Artificial Intelligence*, volume 1168, pages 91–104. Springer Verlag, 1996. URL `citeseer.ist.psu.edu/collins96adaptationguided.html`. 142

Brona Collins, Padraig Cunningham, and Tony Veale. An example-based approach to machine translation. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pages 1–13, Montreal, Quebec, 1996. 142

Lambros Cranias, Harris Papageorgiou, and Stelios Piperidis. A matching technique in example-based machine translation. In *Proceedings of the 15th conference on Computational linguistics*, pages 100–104, Morristown, NJ, USA, 1994. Association for Computational Linguistics. 14, 15

Dan Cristea. Romanian language technology and resources go to europe. Presentated at the FP7 Language Technology Informative Days, January, 20-11 2009. To be found at: ftp://ftp.cordis.europe.eu/pub/fp7/ict/docs/language-technologies/cristea_en.pdf - last accessed on 10.04.2009. 33, 34

Dan Cristea and Dan Tufiş. Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române. In Ofelia Ichim and Florin Teodor Olariu, editors, *Identitatea Limbii şi Literaturii*

Române în Perspectiva Globalizării, pages 211–234. Academia Română, Institutul de Filologie Română "A. Philippide", Editura Trinitas, Iaşi, May 2002. ISBN 973-8179-12-2. 40

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, July 1945. 66

Peter Dirix, Ineke Schuurman, and Vincent Vandeghinste. Metis-ii: Example-based machine translation using monolingual corpora - system description. In *Proceedings of the Workshop on Example-Based Machine Translation Hosted by MT SUMMIT X*, pages 43–50, Thailand, Phuket, September 16 2005. 142

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. 87

Takao Doi, Hirofumi Yamamoto, and Eiichiro Sumita. Graph-based retrieval for example-based machine translation using edit-distance. In *Proceedings of the Workshop Example-Base Machine Translation at MT Summit X*, pages 51–58, September 2005a. 24, 36, 140

Takao Doi, Hirofumi Yamamoto, and Eiichiro Sumita. Example-based machine translation using efficient sentence retrieval based on edit-distance. *Proceedings of the ACM Transactions on Asian Language Information Processing (TALIP)*, 4(4):377–399, 2005b. ISSN 1530-0226. doi: http://doi.acm.org/10.1145/1113308.1113310. 140

Bonnie J. Dorr. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633, December 1994. 47

Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A survey of current paradigms in machine translation. *Advances in Computers*, 49:2–68, 1999. 8, 47, 122

Hande Doğan. Learnıng part-of-speech tagged translation templates from bilingual translation examples. In *ML Workshop,*, Bilkent University , Ankara, Turkey, 2005. URL `http://www.cs.bilkent.edu.tr/~guvenir/courses/cs550/Workshop/Hande_Dogan.pdf`. 85

Hande Doğan. Example based machine translation with type associated translation examples. Master's thesis, Bilkent University, January 2007. 141

Hiroshi Echizen-ya, Kenji Araki, Yoshio Momouchi, and Koji Tochinai. Effectiveness of layering translation rules based on transition networks using inductive learning with genetic algorithms. In *MT 2000: Machine Translation and Multilingual Applications in the New Millennium*, Exeter, England, November, 20-22 2000. 29

Andreas Eisele. Hybrid architectures for machine translation. Presentation at the Second Machine Translation Marathon, May 2008. URL `http://www.mt-archive.info/MTMarathon-2008-Eisele-ppt.pdf`. Wandlitz, Berlin, Germany; 29 slides. 17

Andreas Eisele, Christian Federmann, Hands Uszkoreit, Herve Saint-Armand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen. Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of the 12th EAMT Conference*, pages 27–34, Hamburg, Germany, September 2008. 18, 137

REFERENCES

Natalia Elita, Monica Gavrila, and Cristina Vertan. Experiments with string similarity measures in the ebmt framework. In *Proceedings of the RANLP 2007 Conference*, Bulgaria, September 2007. Poster. 63

Ralf Engel. Chunky: An example based machine translation system. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-2000)*, volume 4, pages 426–429, Beijing, 2000. 141

EUROMATRIX. 1.3: Survey of machine translation evaluation. Deliverable 1.3, December 2007. URL `http://www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf/view`. Euromatrix EU Project. 84

Frank Van Eynde, editor. *Linguistic Issues in Machine Translation*. Pinter Publishers, London and New York, 1993. 8, 83

Ren Feiliang, Zhang Li, Hu Minghan, and Yao Tianshun. Ebmt based on finite automata state transfer generation. In *TMI-2007: Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 65–74, Sweden, September, 7-9 2007. 141

Alexander Franz, Keiko Horiguchi, Lei Duan, Doris Ecker, Eugene Koontz, and Kazami Uchida. An integrated architecture for example-based machine translation. In *Proceedings of the 18th conference on Computational linguistics*, pages 1031–1035, Morristown, NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992730.992799. 141

Robert Frederking, Serghei Nirenburg, David Farwell, Stephen Helmreich, Eduard Hovy, Kevin Knight, Stephen Beale, Constantine Domashnev, Donalee Attardo, Dean Grannes, and Ralf Brown. Integrating translations from multiple sources within the pangloss mark iii machine translation system. In *Technology Partnerships for Crossing Linguistic Barrier: proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 73–80, Columbia, Maryland, 1994. 35, 85

William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, 1993. 42, 155

Monica Gavrila and Natalia Elita. Roger - un corpus paralel aliniat. In *Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române Workshop Proceedings*, pages 63–67, December 2006. Workshop held in November 2006, Publisher: Ed. Univ. Alexandru Ioan Cuza, ISBN: 978-973-703-208-9. 3, 44, 46

Monica Gavrila and Natalia Elita. Comparing cbmt approaches using restricted resources. In Gorka Labaka and Maite Melero, editors, *Proceedings of the International Workshop on Using Linguistic Information for Hybrid Machine Translation LIHMT and of the Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation*, pages 43–49, Barcelona, Spain, November 2011a. ISBN 978-84-615-2995-7. 138

Monica Gavrila and Natalia Elita. Comparing cbmt approaches for german-romanian. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for ComputerScience and Linguistics (LTC 2011)*, pages 435–439, Poznan, Poland, November 25-27 2011b. ISBN 978-83-932640-1-8. 138

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogeneous mt systems. In *Proceedings of WMT 2007 (ACL'07)*, pages 256–264, Prague, June 2007. 84

Jesús Giménez and Lluís Màrquez. On the robustness of syntactic and semantic features for automatic MT evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 250–258, Athens, Greece, March 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W09/W09-0x40`. 84

Nano Gough and Andy Way. Controlled generation in example-based machine translation. In *Proceedings of the MT Summit IX*, pages 133–140, New Orleans, LA., 2003. 25, 141

Nano Gough and Andy Way. Robust large-scale ebmt with marker-based segmentation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 95–104, Baltimore, MD, 2004. 23, 25, 85, 140

T. R. G. Green. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, 18(4):481 – 496, 1979. 28

Gregory Grefenstette. The world wide web as a resource for example-based machine translation. In *Proceedings of Aslib Conference on Translating and the Computer*, London, November 1999. 23, 140

Declan Groves and Andy Way. Hybrid data-driven models of machine translation. *Machine Translation*, 19(3-4):301–323, 2005. ISSN 0922-6567. doi: http://dx.doi.org/10.1007/s10590-006-9015-5. 18

Nikolaus P. Himmelmann. *Language Typology and Language Universals: An International Handbook*, chapter 62. Articles, pages 831–841. Walter de Gruyter, 2001. Published online in 2008. 38

Eduard Hovy. Re: [mt-list] phrasal smt vs ebmt. http://www.mail-archive.com/mt-list@eamt.org/msg00777.html, February 2005. Last accessed 11 November 2008. 22

John Hutchins. Two precursors of machine translation: Artsrouni and trojanskij. *International Journal of Translation*, 16(1):11–31, Jan-June 2004. 8

John Hutchins. Towards a definition of example-based machine translation. In *Proceedings of the Workshop on Example-Based Machine Translation, hosted by MT-Summit X*, September, 12-16 2005a. 14, 16, 21

John Hutchins. Example-based machine translation: a review and commentary. *Machine Translation*, 19(3-4):197–211, 2005b. ISSN 0922-6567. doi: http://dx.doi.org/10.1007/s10590-006-9003-9. 16, 21

W. John Hutchins. Machine translation: a concise history. In Chan Sin Wai, editor, *Computer aided translation: Theory and practice*, 2007. 8

W. John Hutchins and Harold L. Somers. *An Introduction to Machine Translation*. Academic Press, London, 1992. ISBN: 0-12-362830-X. 8

Rebecca Hutchinson, Paul N. Bennett, Jaime Carbonell, Peter Jansen, and Ralf Brown. Maximal lattice overlap in example-based machine translation. Technical report, School of Computer Science, Carnegie Mellon University, 2003. CMU-CS-03-138. 24, 140

REFERENCES

Camelia Ignat. *Improving Statistical Alignment and Translation Using Highly Multilingual Corpora.* PhD thesis, INSA - LGeco- LICIA, Strasbourg, France, June, 16th 2009. It can be found on: http://sites.google.com/site/cameliaignat/home/phd-thesis - last accessed on 3.08.09. 33, 34, 41, 44, 92, 94, 137

Kenji Imamura. Hierarchical phrase alignment harmonized with parsing. In *Natural Language Processing Pacific Rim Symposium*, pages 377–384, 2001. 18

Radu Ion. *Word Sense Disambiguation Methods Applied to English and Romanian.* PhD thesis, Research Institute for Artificial Intelligence of the Romanian Academy, 2007. 55

Elena Irimia. Ebmt experiments for the english-romanian language pair. In *Proceedings of the Recent Advances in Intelligent Information Systems*, pages 91–102, 2009. ISBN 978-83-60434-59-8. 33, 34, 41, 117, 142

Daniel Jones. Non-hybrid example-based machine translation architectures. In *Proceedings of TMI-92*, pages 163–172, Montreal, 1992. 22

Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. Learning translation templates from bilingual text. In *Proceedings of the 14th conference on Computational linguistics*, pages 672–678, Morristown, NJ, USA, 1992. Association for Computational Linguistics. 26, 27, 142

Megumi Kameyama, Ryo Ochitani, and Stanley Peterson. Resolving translation mismatches with information flow. In *Proceedings of the 29th Annual Meeting of ACL*, pages 193–200, Berkeley, 1991. 47

Martin Kay. *An Introduction to Machine Translation*, chapter Foreword, pages xi–xiii. Academic Press, 1992. Editors: W. John Hutchins and Harold L. Somers. 7

Satoshi Kinoshita, Akira Kumano, and Hideki Hirakawa. Improvement in customizability using translation templates. In *Proceedings of the 15th International Conference on Computation Linguistics: COLING 1994*, pages 25–31, Kyoto, Japan, August, 5-9 1994. 26, 27

Chunyu Kit, Haihua Pan, and Jonathan J. Webster. *Translation and Information Technology*, chapter Example-Based Machine Translation: A New Paradigm, pages 57–78. Chinese U of HK Press, 2002. http://personal.cityu.edu.hk/ ctckit/papers/EBMT-review-CUHK.pdf. 14, 15, 22, 24, 67

Philipp Koehn. *Moses. Statistical Machine Translation System. User Manual and Code Guide.* University of Edinburgh, August 2010. 53

Philipp Koehn and Chris Callison-Burch. Statistical machine translation. Course at the 20th European Summer School in Logic, Language and Information, August 2005. 12

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073445.1073462. 12

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic, June 2007. 52

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for europe. In *Proceedings of the MT Summit XII*, pages 65–72, Ottawa, Canada, August 2009. 33, 41, 92

Gorka Labaka, Nicolas Stroppa, Andy Way, and Kepa Sarasola. Comparing rule-based and data-driven approaches to spanish-to-basque machine translation. In *Proceedings of the Machine Translation Summit XI*, pages 297–304, Copenhagen, Denmark., September 2007. 17

LDC. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. revision 1.5. Technical report, Linguistic Data Consortium, 2005. URL `http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf`. 124

Lilian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 65–72, 2001. 15

Yves Lepage. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics*, pages 728–734, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/980845.980967. 30

Yves Lepage. Languages of analogical strings. In *Proceedings of the 18th conference on Computational linguistics*, pages 488–494, Morristown, NJ, USA, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. doi: http://dx.doi.org/10.3115/990820.990891. 30

Yves Lepage and Etienne Denoual. The 'purest' ebmt system ever built: no variables, no templates, no training, examples, just examples, only examples. In *Proceedings of Second Workshop on Example-Based Machine Translation, MT Summit X*, pages 81–90, September 16 2005. 29, 30, 31, 140

Yves Lepage and Guilhem Peralta. Using paradigm tables to generate new utterances similar to those existing in linguistic resources. In *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, pages 243–246, 2004. 30

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. 85

M. Paul Lewis, editor. *Ethnologue: Languages of the World*. Tex.: SIL International., Dallas, sixteenth edition edition, 2009. Online version: http://www.ethnologue.com/. xvii, xviii, 37

Jesús Ángel Giménez Linares. *Empirical Machine Translation and its Evaluation*. PhD thesis, Departament de Llenguatges i Sistemes Informatics Universitat Politecnica de Catalunya, 2008. 84

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Workshop On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/or Summarization*, 2005. 84

REFERENCES

Zhanyi Liu, Haifeng Wang, and Hua Wu. Example-based machine translation based on tree-string correspondence and statistical generation. *Machine Translation*, 20(1):25–41, March 2006. 23, 140

Adam Lopez. Statistical machine translation. *ACM Comput. Surv.*, 40(3):1–49, 2008. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1380584.1380586. 12

Christos Malavazos and Stelios Piperidis. Application of analogical modelling to example based machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics: COLING 2000*, pages 516–522, Saarbruecken, Germany, 1999. 28

Christos Malavazos, Stelios Piperidis, and George Carayannis. Towards memory and template based translation synthesis. In *Proceedings of the MT 2000: Machine Translation and Multilingual Applications in the New Millenium*, pages 1.1–1.8., Exeter, England, November, 20-22 2000. 28

Federica Mandreoli, Riccardo Martoglia, and Paolo Tiberio. Searching similar (sub)sentences for example-based machine translation. In *SEBD*, pages 208–221, 2002. 14, 24, 140

Stella Markantonatou, Sokratis Sofianopoulos, Vassiliki Spilioti, George Tambouratzis, Marina Vassiliou, and Olga Yannoutsou. Using patterns for machine translation (mt). In *Proceedings of the European Association for Machine Translation*, pages 239–246, Oslo, Norway, June 19-20 2006. 143

Hiroshi Maruyama and Hideo Watanabe. Tree cover search algorithm for example-based translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 173–184, 1992. 142

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. Structural matching of parallel texts. In *31st Annual Meeting of the Association for Computational linguistics*, pages 23–30, Columbus, Ohio, 1993. 14, 32

Ian J. McLean. Example-based machine translation using connectionist matching. In *Proc. of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, pages 35–43, Montreal, Canada, 1992. 36, 141

Kevin McTait. *Translation Pattern Extraction and Recombination for Example-Based Machine Translation*. PhD thesis, Centre for Computational Linguistics, Department of Language Engineering, PhD Thesis, UMIST, 2001. 14, 15, 22, 26, 27, 36, 71, 74, 141, 142

Kevin McTait. *Translation Patterns, Linguistic Knowledge and Complexity in an Approach to EBMT*, volume Recent Advances in Example-based Machine Translation, pages 307–338. Kluwer Acad. Publ., 2003. 11, 26, 85, 136

Kevin McTait and Arturo Trujillo. A language-neutral sparse-data algorithm for extracting translation patterns. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, pages 98–108, 1999. 141

Dennis Mehay and Chris Brew. Bleuatre: Flattening syntactic dependencies for mt evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Sweden, 2007. 84

Virginia Motapanyane, editor. *Comparative Studies in Romanian Syntax*, volume 58 of *North-Holland Linguistic Series*. Elsevier, 2000. 40, 154

Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *Proceedings of the international NATO symposium on Artificial and human intelligence*, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc. ISBN 0-444-86545-4. 13, 15, 21, 23, 27, 57

Sergei Nirenburg, Constantine Domashnev, and Dean J. Grannes. Two approaches to matching in example-based machine translation. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-93).*, pages 47–57, 1993. 15

Eric H. Nyberg and Teruko Mitamura. The kant system: fast, accurate, high-quality translation in practical domains. In *Proceedings of the 14th conference on Computational linguistics - Volume 3*, COLING '92, pages 1069–1073, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/993079.993131. URL `http://dx.doi.org/10.3115/993079.993131`. 11

Franz Josef Och. *Statistical Machine Translation: from Single-World Models to Alignment Templates*. PhD thesis, Facultaet fuer Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfaelischen Technischen Hochschule Aachen, 2002. 10

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, 2003. 52

Franz Josef Och. Statistical machine translation: foundations and recent advances. Tutorial at MT Summit X, September 2005. URL `http://www.mt-archive.info/MTS-2005-Och.pdf`. Phuket, Thailand. 54

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. 12, 55

Bjarne Orsnes, Bradley Music, and Bente Maegaard. Patrans: a patent translation system. In *Proceedings of the 16th conference on Computational linguistics*, pages 1115–1118, Morristown, NJ, USA, 1996. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/993268.993379. 11

Miles Osborne. *Encyclopedia of Machine Learning*, chapter Statistical Machine Translation. Springer, 2010. URL `http://www.statmt.org/ued/?n=Public.Publications`. Editor: Claude Sammut and Geoffrey I. Webb. 12

Karolina Owczarzak. *A Novel Dependency-Based Evaluation Metric for Machine Translation*. PhD thesis, Dublin City University School of Computing, 2008. 84

Karolina Owczarzak, Josef van Genabith, and Andy Way. Labelled dependencies in machine translation evaluation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 104–111, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0214`. 84

Zeynep Oz and Ilyas Cicekli. Ordering translation templates by assigning confidence factors. In *AMTA*, pages 51–61, 1998. URL `citeseer.ist.psu.edu/oz98ordering.html`. 27, 141

REFERENCES

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation*, pages 311 – 318, Philadelphia, Pennsylvania, 2002. Publisher: Association for Computational Linguistics Morristown, NJ, USA. 32, 84, 87

Aaron B. Phillips and Ralf D. Brown. Cunei machine translation platform: System description. In Mikel L. Focada and Andy Way, editors, *Proceedings of the 3rd International Workshop on Example-Based Machine Translation*, pages 29–36, Dublin, Ireland, November 2009. 18

Emmanuel Planas and Osamu Furuse. *Recent Advances in Example-Based Machine Translation*, chapter Formalizing Translation Memories, pages 157–188. Kluwer Academic Publishers, 2003. ISBN 1-4020-1400-7 / 1-4020/1401-5. 85

Carl Pollard. The nature of constraint-based grammar. Talk at the Pacific Asia Conference on Language, Information, and Computation, December 1996. URL `http://195.113.2.21/knihovna/pollard_talk1.pdf`. 69

Maja Popovic and Hermann Ney. Statistical machine translation with a small amount of bilingual training data. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: Strategies for developing machine translation for minority languages*, pages 25–29, Genoa, Italy, May 2006. 45

Maja Popovic and Hermann Ney. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W07/W07-0707`. 84

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in speech recognition*, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1-55860-124-4. URL `http://portal.acm.org/citation.cfm?id=108235.108253`. 18

Juho Rousu. Workpackage 3 advanced language models. Online, January 2008. URL `http://www.smart-project.eu/files/SMART-Y1-review-WP3.v1.pdf`. SMART Project. 52

Diganta Saha and Sivaji Bandyopadhyay. A semantics-based english-bengali ebmt system for translating news headlines. In *MT Summit X second workshop on example-based machine translation*, pages 125–133, Phuket, Thailand, 2005. 141

Felipe Sanchez-Martinez, Mikel L. Focada, and Andy Way. Hybrid rule-based-example-based mt: Feeding apertium with sub-sentential translation units. In Mikel L. Focada and Andy Way, editors, *Proceedings of the 3rd International Workshop on Example-Based Machine Translation*, pages 11–18, Dublin, Ireland, November 2009. 18

Satoshi Sato. Ctm: an example-based translation aid system. In *Proceedings of the 14th conference on Computational linguistics*, pages 1259–1263, Morristown, NJ, USA, 1992. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/992424.992477. 24, 85, 140

Satoshi Sato. Example-based translation of technical terms. In *Proceedings of TMI-93*, pages 58–68, 1993. 24

Satoshi Sato. Mbt2: a method for combining fragments of examples in example-based translation. *Artificial Intelligence*, 75(1):31–49, 1995. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/0004-3702(94)00063-7. 31, 32

Satoshi Sato and Makoto Nagao. Towards memory-based translation. In *Proceedings of the 13th conference on Computational linguistics*, pages 247–252, Morristown, NJ, USA, 1990. Association for Computational Linguistics. ISBN 952-90-2028-7. doi: http://dx.doi.org/10.3115/991146.991190. 22

Reinhard Schaeler, Michael Carl, and Andy Way. *Example-based Translation in a Hybrid Integrated Environment*, chapter Chapter 3, pages 83–114. Kluwer Academic Publishers, 2003. Editors: Michael Carl and Andy Way. 17

Anja Schwarzl. *The (Im)Possibilities of Machine Translation*. European University Studies: Series XIV, Angla-Saxon Language and Literature. PETER LANG, 2001. 8, 10

Satoshi Shirai, Francis Bond, and Yamato Takahashi. A hybrid rule and example based method for machine translation. In *Natural Language Processing Pacific Rim Symposium '97: NLPRS-97*, pages 49–54, Phuket, 1997. URL `citeseer.ist.psu.edu/shirai97hybrid.html`. 85

Michael Simard. Re: [mt-list] phrasal smt vs ebmt. http://www.mail-archive.com/mt-list@eamt.org/msg00777.html, February 2005. Last accessed 11 November 2008. 22

James Smith and Stephan Clark. Ebmt for smt: A new ebmt-smt hybrid. In Mikel L. Forcada and Andy Way, editors, *Proceedings of the 3rd Intyernational Workshop on Example-Based Machine Translation*, pages 3–10, Dublin, Ireland, November, 12-13 2009. 18, 33, 59, 65, 137, 140

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, August, 2006. 88

Harol Somers, Sandipan Dandapt, and Sudip Kumar Naskar. A review of ebmt using proportional analogies. In Mikel L. Focada and Andy Way, editors, *Proceedings of the 3rd International Workshop on Example-Based Machine Translation*, pages 53 – 60, November 2009. 31

Harold Somers. New paradigms in mt: the state of play now that the dust has settled. In *Proceedings of the 10th European Summer School on Logic, Linguistics and Information (ESSLLI), Workshop on machine translation*, pages 22–33, Saarbruecken, August, 24-28 1998. 15

Harold Somers. *Review Article: Example-based Machine Translation*, volume Machine Translation 14, pages 113–157. Kluwer Acad. Publ., 1999. http://stp.ling.uu.se/ ebbag/somers.pdf. 14, 15, 16, 85, 139

Harold Somers. *Example-Based Machine Translation*, volume Handbook of Natural Language Processing, pages 611–628. Marcel Dekker, Inc., 2000a. 13, 16

Harold Somers. *Handbook of Natural Language Processing*, chapter Machine Translation, pages 329–346. Marcel Dekker Inc, 2000b. Editors: Robert Dale and Hermann Moisl and Harold Somers. 8

Harold Somers. *An Overview of EBMT*, volume Recent advances in Example-based Machine Translation, pages 3–57. Kluwer Acad. Publ., 2003. 15, 21, 22, 85

REFERENCES

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Daniel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, Genoa, Italy, May, 24-16 2006. 40, 44, 158

Andreas Stolcke. Srilm - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, pages 901–904, Denver, Colorado, September 2002. 54

Eiichiro Sumita. Example-based machine translation using dp-matching between word sequences. In *Proceedings of the workshop on Data-driven methods in machine translation*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118037.1118038. 15, 29, 85, 140

Eiichiro Sumita and Hitoshi Iida. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 185–192, Morristown, NJ, USA, 1991. Association for Computational Linguistics. 23, 59, 85, 141

Eiichiro Sumita, Yasuhiro Akiba, Takao Doi, Andrew Finch, Kenji Imamura, Michael Paul, Mitsuo Shimohata, and Taro Watanabe. Ebmt, smt, hybrid and more: Atr spoken language translation system. In *Proceedings of IWSLT-2004: ICSLP-2004 satellite workshop, International Workshop on Spoken Language Translation – Evaluation Campaign on Spoken Language Translation*, pages 13–20, 2004. 18

Gregor Thurmair. Comparing rule-based and statistical mt output. In *LREC-2004. Workshop: The amazing utility of parallel and comparable corpora*, pages 5–9, 2004. 17

Jörg Tiedemann. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing (vol V)*, pages 237–248, John Benjamins, Amsterdam/Philadelphia, 2009. 156

Jörg Tiedemann and Lars Nygaard. The opus corpus - parallel and free. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1183–1186, Lisbon, Portugal, May, 26-28 2004. 156

Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Comput. Linguist.*, 29:97–133, March 2003. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120103321337458. URL `http://dx.doi.org/10.1162/089120103321337458`. 12

Dan Tufiş, Radu Ion, Alexandru Ceauşu, and Dan Ştefăneşcu. Racai's linguistic web services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008*, Marrakech, Morocco, May 2008a. ELRA - European Language Resources Association. ISBN 2-9517408-4-0. 40, 55

Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev. Building language resources and translation models for machine translation focused on south slavic and balkan languages. In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, Dubrovnik, Croatia, September 25-28 2008b. ISBN 978-953-55375-0-2. In Marko Tadiz, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.). 40, 157

Davide Turcato and Fred Popowich. What is example-based machine translation? In Michael Carl and Andy Way, editors, *Proceedings of the Workshop on Example-Based Machine Translation, hosted by MT-Summit VIII*, page 43–48, September, 18-22 2001. 21

Joseph P. Turian, Luke Shen, and Dan I. Melamed. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX*, pages 386–393, New Orleans, LA, 2003. 32

Hans Uszkoreit. Research avenues:a new hype? a paradigm shift? Presented at the FP7 Language Technology Informative Days, January, 20-11 2009. To be found at: ftp://ftp.cordis.europe.eu/ - last accessed on 10.04.2009. 1

Vincent Vandeghinste and Scott Martens. Top-don transfer in example-based mt. In Mikel L. Focada and Andy Way, editors, *Proceedings of the 3rd International Workshop on Example-Based Machine Translation*, pages 69–76, November, 12-13 2009. 32

Daniel Varga, Peter Halacsy, Andras Kornai, Viktor Nagy, Laszlo Nemeth, and Viktor Tron. Parallel corpora for medium density languages. In *In Recent Advances in Natural Language Processing (RANLP 2005) (2005)*, pages 590–596, 2005. 42

Tony Veale and Andy Way. Gaijin: A bootstrapping approach to example-based machine translation. In *Proceedings of International Conf.- Recent Advances in Natural Language Processing*, pages 239–244, Tzigov Chark, Bulgaria, 1997. 16, 28, 141

Cristina Vertan and Vanessa Espin Martin. Experiments with matching algorithms in example based machine translation. In *In Proceedings of the International workshop "Modern approaches in Translation Technologies", in conjunction with RANLP*, pages 42–45, September 2005. 15

Cristina Vertan, Walther von Hahn, and Monica Gavrila. Designing a parole/simple german-english-romanian lexicon. In *Language and Speech Infrastructure for Information Access in the Balkan Countries Workshop Proceedings - RANLP 2005*, pages 82–87, Borovets, Bulgaria, September 2005. 40

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error analysis of statistical machine translation output. In *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, pages 697–702, Genova, Italy, May 2006. 84

Stephan Vogel, Hermann Ney, and Christoph Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841. Association for Computational Linguistics, 1996. doi: http://dx.doi.org/10.3115/993268.993313. URL `http://dx.doi.org/10.3115/993268.993313`. 55

Heike Voit. *PONS Grammatik kurz & buendig Deutsch*. Ernst Klett Sprachen GmbH, 2007. 39, 40

Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag, 2000. 9

Hideo Watanabe. A similarity-driven transfer system. In *Proceedings of the 14th conference on Computational linguistics*, pages 770–776, Morristown, NJ, USA, 1992. Association for Computational Linguistics. 14, 31, 32, 142

Hideo Watanabe. A model of a bi-directional transfer mechanism using rule combinations. In *Machine Translation 10*, pages 269–291, 1995. 31, 32

REFERENCES

Hideo Watanabe and Koichi Takeda. A pattern-based machine translation system extended by example-based processing. In *Proceedings of the 17th international conference on Computational linguistics*, pages 1369–1373, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/980691.980792. 142

Tano Watanabe and Eiichiro Sumita. Example-based decoding for statistical machine translation. In *MT-Summit*, pages 410–417, 2003. 18, 85, 140, 143

Andy Way. Translating with examples. In *Proceedings of MT Summit VIII Workshop on Example-Based Machine Translation*, pages 66–80, Santiago de Compostela, Spain, 2001. 142

Andy Way and Nano Gough. webmt: developing and validating an example-based machine translation system using the world wide web. In *Computational Linguistics*, volume 29, number 3, pages 421–457, Cambridge, MA, USA, 2003. MIT Press. doi: http://dx.doi.org/10.1162/089120103322711596. 25, 36, 140

Andy Way and Nano Gough. Comparing example-based and statistical machine translation. *Natural Language Engineering*, -(11):295–309, 2005. doi: 10.1017/S1351324905003888. Cambridge University Press. 32, 36, 140

Warren Weaver. *Machine translation of languages: fourteen essays*, chapter Translation, pages 15–23. Technology Press of the Massachusetts Institute of Technology, Cambridge, Mass. and John Wiley & Sons, Inc., New York, 1955. Written 15 July 1949. Editors: William N. Locke and A. Donald Booth. 11

Dekai Wu. Mt model space: statistical versus compositional versus example-based machine translation. *Machine Translation*, 19:213–227, 2005. 22

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbst verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Hamburg, Juli 2011

(Monica Roxana Gavrilă)