Universität Hamburg
MIN-Fakultät
Department Informatik

A Computational Model for the Influence of Cross-Modal Context upon Syntactic Parsing

Doctoral Thesis

submitted by
Patrick McCrae
of Dublin

Hamburg, March 2010

Genehmigt von der MIN-Fakultät, Fachbereich Informatik
der Universität Hamburg auf Antrag von

Erstgutachter	Prof. Dr.-Ing. Wolfgang Menzel (Betreuer) Fachbereich Informatik Universität Hamburg
Zweitgutachter	Prof. Dr. Christopher Habel Fachbereich Informatik Universität Hamburg
Externer Gutachter	Prof. Dr. Maosong Sun Department of Computer Science Tsinghua University, Beijing

Hamburg, den 07. Juli 2010 (Tag der Disputation)

Prof. Dr. Horst Oberquelle
Leiter des Fachbereichs Informatik

To O.M.

Abstract

Ambiguity is an inherent property of natural language. Its most prominent manifestations comprise syntactic ambiguity, lexical ambiguity, scope ambiguity and referential ambiguity. Considering the high frequency with which ambiguity occurs in unrestricted natural language, it is surprising how seldom ambiguity causes misunderstandings. Most linguistic ambiguities in inter-human communication even pass unnoticed, mainly because human cognition automatically and unconsciously attempts to resolve ambiguity. A central contribution to this automatic and unconscious disambiguation is made by the integration of non-linguistic information from cognitively readily available sources such as world knowledge, discourse context or visual scene context. While a large body of behavioural investigations into the interactions between vision and language has been accumulated, comparatively few computational models of those interactions have been reported.

The focus of this thesis is to motivate, specify and validate a computational model for the cross-modal influence of visual scene context upon natural language understanding and the process of syntactic parsing, in particular. We argue for a computational model that establishes cross-modal referential links between words in the linguistic input and entities in a visual scene context. Cross-modal referential links are assigned on the basis of conceptual compatibility between the concepts activated in the linguistic modality and the concepts instantiated in visual context. The proposed model utilises the thematic relations in the visual scene context to modulate attachments in the linguistic analysis.

In contrast to the majority of extant computational models for the interaction between vision and language, our model is motivated by an integrated theory of cognition. We base our model architecture on the cognitive framework of Conceptual Semantics, an overarching theory of cognition and language processing by Ray Jackendoff. In our model, we adopt the central tennet of Conceptual Semantics that all cross-modal interactions of non-linguistic modalities with language are mediated by Conceptual Structure, a single, uniform representation of linguistic and non-linguistic semantics. Conceptual Structure propagates the influence of the non-linguistic modalities into syntactic representation via a syntax-semantics interface. The purpose of this interface is to map between the syntactic and the semantic representation by means of representational correspondence rules.

Our model implements central aspects of the cognitive architecture in Conceptual Semantics. We encode the semantic information for all entities, be they linguistic or non-linguistic in nature, on a single level of semantic representation. In particular, the semantic part of linguistic analysis and visual scene information are included in this representation. The semantic preferences arising from visual context constrain the semantic part of linguistic analysis. The semantic part of linguistic analysis, in turn, constrains syntactic analysis via the syntax-semantics interface. In this way, our model achieves a semantically mediated propagation of non-linguistic visual scene information into syntactic representation.

We validate our model's context integration behaviour under a range of experimental conditions. The integration of visual scene context as a hard constraint on linguistic analysis enforces an absolute dominance of visual context information over linguistic analysis. As a result, hard integration can lead to a contextualised linguistic analysis that violates linguistic well-formedness preferences in order to be semantically compatible with the modelled visual context. Integrating visual context information as a soft constraint on linguistic analysis affords cognitively more plausible results. Soft integration permits to achieve a balance between conflicting linguistic and contextual preferences based on the strength of the individual preferences. Under soft integration, our model also diagnoses which aspects of linguistic analysis are in conflict with visual context information. Diagnosis constitutes an important cognitive capability in the situated cognition of natural systems. The ability to diagnose cognitive input permits the effective identification of which parts of that input are incorrect, inconsistent or incompatible with pre-existing top-down expectations and thus enables a more specific and adequate response to that input.

We further demonstrate our model's robustness to conceptual underspecification in the contextual representation. Our experiments show that the integration of conceptually underspecified context representations still provides valuable information to support the process of syntactic disambiguation. The capability of processing conceptually underspecified semantic information is a relevant feature with regards to the handling of perceptual uncertainty and perceptual ambiguity.

The implementation of our model centres around WCDG2, a weighted-constraint dependency parser for German. We encode situation-invariant semantic knowledge including semantic lexical knowledge and world knowledge in terms of concepts in an OWL ontology (A-Box). Situation-specific visual scene information is encoded in context models that assert instantiations of concepts from the ontology joined by thematic relations. The contextual constraints upon the semantic part of linguistic analysis are communicated to the parser in the form of score predictions for semantic dependency assignments in the linguistic analysis. These score predictions are computed by a predictor component prior to parse time and are accessed by the parser at parse time. The predictor computes its prediction scores based on the input sentence and the visual scene information in the context model. The primary objective of the predictor component is to veto all semantic dependencies in the input sentence that are incompatible with the asserted visual context information.

The implementation of our model for the cross-modal influence of visual scene context upon linguistic processing is also subject to a number of significant limitations. The most severe of these with regards to the objective of modelling vision-language interaction are the unidirectionality of the implemented vision-language interaction, our non-incremental approach to linguistic processing and the limited scope of the semantic part of linguistic analysis. We discuss these limitations in detail and point out directions for further research to address them.

In summary, the model presented in this thesis is the result of an interdisciplinary research effort whose main objective was to bring together a suitable theory of cross-modal cognition and methods of natural language engineering. While this work cannot claim to have bridged the gap between the disciplines in its entirety, the presented results constitute an encouraging first step towards achieving the ambitious overall goal. The outcome of this research is a cognitively motivated model implementation that achieves selective modulations of syntactic attachments based on representations of visual scene context by mediation of a single shared representation of linguistic and non-linguistic semantics.

Zusammenfassung

Ambiguität ist eine inhärente Eigenschaft natürlicher Sprache, deren häufigste Ausprägungen syntaktische oder strukturelle Ambiguität, lexikalische Ambiguität, Scopus-Ambiguität und referenzielle Ambiguität umfassen. In Anbetracht der großen Häufigkeit, mit der Ambiguität in natürlicher Sprache vorkommt, ist es verwunderlich, wie selten Ambiguität tatsächlich Missverständnisse verursacht. Die meisten Ambiguitäten in menschlicher Kommunikation werden nicht einmal bemerkt, vorwiegend weil die menschliche Kognition automatisch und unbewusst versucht, Ambiguitäten aufzulösen. Einen zentralen Beitrag zu dieser automatischen und unbewussten Disambiguierung leistet die Integration von nicht-sprachlichen Informationen aus kognitiv zugänglichen Quellen wie Weltwissen, Diskurskontext oder visuellem Szenenkontext. Während eine Vielzahl von verhaltenspsychologischen Untersuchungen zu Interaktionen zwischen Sehen und Sprache vorliegen, wurde bisher nur eine vergleichsweise geringe Zahl von computationellen Modellen beschrieben.

Der Kern dieser Arbeit beinhaltet die Motivation, Spezifizierung und Validierung eines computationellen Modells für den cross-modalen Einfluss von visuellem Szenenkontext auf das Verstehen natürlicher Sprache im Allgemeinen — und den Prozess des syntaktischen Parsings im Besonderen. Wir stellen ein computationelles Modell vor, das cross-modale Referenzbeziehungen zwischen Worten im sprachlichen Input und Entitäten im visuellen Kontext herstellt. Die cross-modalen Referenzbeziehungen werden dabei zugewiesen basierend auf der Konzeptkompatibilität zwischen den sprachlich aktivierten Konzepten und den Konzepten, die im visuellen Kontext instanziiert wurden. Das vorgestellte Modell nutzt thematische Relationen im visuellen Szenenkontext, um Anbindungen der sprachlichen Analyse zu beeinflussen.

Im Gegensatz zu der Mehrzahl der bestehenden computationellen Modelle ist unser Modell durch eine umfassende Theorie der menschlichen Kognition motiviert. Die Architektur unseres Modells basiert auf dem kognitiven Framework der Konzeptuellen Semantik (*Conceptual Semantics*), einer weitreichenden Theorie zu Kognition und Sprachverarbeitung von Ray Jackendoff. In unserem Modell folgen wir der zentralen Annahme der Konzeptuellen Semantik, dass alle cross-modalen Interaktionen von nicht-sprachlichen Modalitäten mit Sprache durch die Konzeptuelle Struktur (*Conceptual Structure*) vermittelt werden. Bei der Konzeptuellen Struktur handelt es sich um die zentrale Repräsentation sprachlicher und nicht-sprachlicher Semantik. Die Konzeptuelle Struktur reicht den Einfluss der nicht-sprachlichen Modalitäten in die syntaktische Repräsentation über eine Schnittstelle zwischen Syntax und Semantik weiter. Die Aufgabe dieser Schnittstelle ist es, syntaktische und semantische Repräsentationen über Korrespondenzbeziehungen aufeinander abzubilden.

Unser Modell implementiert zentrale Aspekte der kognitiven Architektur aus der Konzeptuellen Semantik. Wir repräsentieren die semantische Information aller Entitäten, seien sie sprachlicher oder nicht-sprachlicher Natur, auf ein und derselben Repräsentationsebene. Insbesondere werden der semantische Teil der sprachlichen Analyse sowie visuelle Szeneninformationen in dieser Repräsentation abgebildet. Die semantischen Präferenzen, die sich aus dem visuellen Kontext ergeben, beschränken den semantischen Teil der sprachlichen Analyse. Der semantische Teil der sprachlichen Analyse wiederum beschränkt die syntaktische Analyse über die Syntax-Semantik-Schnittstelle. Auf diese Weise erzielt unser Modell die semantisch vermittelte Propagation nicht-sprachlicher visueller Szeneninformation in die syntaktische Repräsentation.
Wir validieren das Verhalten des vorgestellten Modells hinsichtlich der Integration von kontextueller Information unter verschiedenen experimentellen Bedingungen. Die Integration von visuellem Szenenkontext als harte Beschränkung der sprachlichen Analyse erzwingt eine absolute Dominanz der visuellen Kontextinformation über die sprachliche Analyse. Wir beobachten, dass die harte Integration zu einer kontextualisierten Analyse des sprachlichen Inputs führen kann, die Regeln sprachlicher Wohlgeformtheit verletzt, um semantische Kompatibilität mit dem modellierten visuellen Kontext zu erzielen. Die Integration von visueller Kontextinformation als weiche Beschränkung der sprachlichen Analyse hingegen ergibt kognitiv plausiblere Resultate. Weiche Integration gestattet konfligierende sprachliche und kontextuelle Präferenzen basierend auf ihrer Gewichtung gegeneinander abzuwägen. Weiche Integration eröffnet in unserem Modell auch die Möglichkeit der Diagnose, um festzustellen, welche Aspekte der sprachlichen Analyse mit der visuellen Kontextinformation im Konflikt stehen. Die Fähigkeit zur Diagnose ist eine wichtige kognitive Fähigkeit natürlicher Systeme im Rahmen von kontextuell eingebundener Wahrnehmung und Interaktion. Diagnose ermöglicht zu erkennen, welche Teile eines kognitiven Inputs inkorrekt, inkonsistent oder inkompatibel mit bestehenden Top-Down-Erwartungen ist, und ermöglicht so, angemessen und effektiv auf diesen Input zu reagieren. Wir demonstrieren weiterhin die Robustheit unseres Modells gegenüber konzeptueller Unterspezifikation in der Repräsentation von visuellem Kontext. Unsere Experimente zeigen, dass die Integration von konzeptuell unterspezifizierten Kontextrepräsentationen dennoch wertvolle Informationen liefern kann, um den Prozess der syntaktischen Disambiguierung zu unterstützen. Die Fähigkeit, konzeptuell unterspezifizierte semantische Information verarbeiten zu können, ist eine wichtige Systemeigenschaft für die Modellierung von perzeptueller Unsicherheit und perzeptueller Mehrdeutigkeit.

Im Mittelpunkt der Implementierung des Modells steht WCDG2, ein Dependenzparser des Deutschen auf Basis eines gewichteten Constraint-Formalismus. Situationsunabhängiges semantisches Wissen wie semantisches lexikalisches Wissen und Weltwissen sind durch Konzepte abgebildet, die die Konzepthierarchie einer OWL-Ontologie definieren. Situationsspezifische Szeneninformation bilden wir in Kontextmodellen ab, die Instanziierungen der Konzepte aus der Ontologie und thematische Relationen zwischen diesen Konzeptinstanzen beinhalten. Die kontextuellen Präferenzen, die sich aus dem modellierten visuellen Szenenkontext ergeben, werden

dem Parser in Form von Bewertungsvorhersagen für die Zuweisung von semantischen Dependenzen in der sprachlichen Analyse übergeben. Diese Bewertungsvorhersagen werden von einer Prädiktor-Komponente vor der Parsezeit berechnet; der Parser greift dann zur Parsezeit auf diese Bewertungsvorhersagen zu. Die Berechnung der Vorhersagen durch den Prädiktor erfolgt basierend auf dem eingegebenen Satz und der visuellen Szeneninformation im Kontextmodell. Die Hauptaufgabe des Prädiktors ist es dabei, all jene semantischen Dependenzen durch Vergabe schlechter Bewertungen zu verbieten, die inkompatibel mit der visuellen Kontextinformation sind.

Die Implementierung unseres Modells für den cross-modalen Einfluss von visuellem Szenenkontext auf die sprachliche Verarbeitung unterliegt auch einer Vielzahl von nicht unerheblichen Einschränkungen. Aus unserer Sicht sind drei dieser Einschränkungen hinsichtlich des Modellierungszieles besonders schwerwiegend: 1) die Unidirektionalität der implementierten Sehen-Sprache-Interaktion, 2) das Fehlen von Inkrementalität in der sprachlichen Verarbeitung und 3) die begrenzte sprachliche Abdeckung im semantischen Teil der sprachlichen Analyse. Wir diskutieren diese Einschränkungen im Detail und zeigen Ansätze auf, diesen Einschränkungen im Rahmen weiterführender Forschungsansätze zu begegnen.

Zusammenfassend kann gesagt werden, dass diese Arbeit das Resultat eines interdisziplinären Forschungsansatzes darstellt, dessen Hauptziel es war, eine geeignete Theorie der cross-modalen Kognition mit entsprechenden Methoden der Sprachtechnologie zusammen zu führen. Auch wenn diese Arbeit nicht den Anspruch erhebt, dieses Ziel in vollem Umfang erreicht zu haben, so sind die vorgestellten Ergebnisse doch vielversprechende erste Schritte in Richtung der Erreichung dieses ehrgeizigen Gesamtzieles. Das Ergebnis dieser Arbeit ist die Implementierung eines kognitiv motivierten Modells, das anhand von visuellem Szenenkontext in der Lage ist, selektiv syntaktische Anbindungen zu beeinflussen. Die Beeinflussung der syntaktischen Verarbeitung erfolgt dabei durch Vermittlung einer zentralen Repräsentation von sprachlicher und nicht-sprachlicher Semantik.

Acknowledgements

First and foremost, I would like to express my sincere thanks to my supervisor Wolfgang Menzel for his outstanding support throughout the three and a half years of my PhD candidature. Working with Wolfgang has been a truly rewarding experience, both intellectually and personally. He always had an open ear for me when I needed it and proved a continual source of inspiration and encouragement. Of the countless fruitful discussions we had, many have fundamentally influenced the direction that this research project has taken. His insightful guidance was a tremendous help for me in conducting this research.

I am also grateful to Christopher Habel for his continual focus on the cognitive dimension of my work. Over the years, Christopher has been an inexhaustible source of profound, thought-provoking questions and food for thought. I particularly thank him for his repeated critical and inspiring reviews of the intermediate stages of my work. His comments over the years were invaluable directions on the way towards maturing and completing this research project.

Many thanks are also due to the other members of the CINACS Graduate Research Group for the many inspiring discussions on issues of cross-modality. I especially thank my Chinese co-supervisor Maosong Sun from Tsinghua University, Beijing, for his constructive and helpful comments regarding the importance of perceptual uncertainty in cross-modal cognition. I am further grateful for the continual funding of my research by the DFG (German Research Foundation) under grant number GRK-1247/1.
I also would like to dedicate warm thanks to the first generation CINACS Informatics students Tian Gan, Sascha Jockel, Martin Weser, Cengiz Acartürk and Christian Graf, with whom we frequently debated a whole range of hot scientific issues, not infrequently over an exotic dinner or a nice mug of steaming coffee.

There are many other people whose help and support I was glad to accept in some way or the other on the way towards completing this thesis. Thanks are due to all of them, and I hope to list a good part of them here.

My sincere thanks go to Kilian Foth, undoubtedly the most knowledgeable person alive regarding the standard implementation of WCDG. The patience and ease with which he provided his always extremely enlightening answers to questions concerning the technical intricacies of the WCDG syntax parser were as astounding as they were helpful.

I am also indebted to my former student assistants Christopher Baumgärtner, Yvonne Küstermann and Rörd Hinrichsen, all of whom have done an amazing job at supporting me during the implementation phase of this research project. I am happy to note that Christopher has abandoned his promising career as an image processor in favour of his own research project on the interaction between visual context and syntactic parsing.

Thanks also to Pine Eisfeld for sending over all the way from England that ever so important and equally unexpected giant bar of *Cadbury's Diary Milk Chocolate* as extra brain fuel on the finishing straight of compiling this thesis.

I also extend a big, colourful '*Thank you!*' to Bianca Ehlebracht for her much appreciated advice on the layout and typography of this thesis.

Many, many thanks are owed to the good dozen of lovely people who were brave enough to take on the daunting challenge of proof-reading some chapters from the pre-final draft of this thesis. Kris, Isabelle, Christopher, Lu, Patrick, Kilian, Christian, Hadya, Lidia, Martin, Jan-Christian, Niels, Jason, Thinh, and Katharina ... your feedback was a great help and a much needed support in getting all the t's crossed and the i's dotted as submission deadline was approaching just that tad too rapidly.

I also would like to thank all the other wonderful people who have supported me, in one way or another, during my time as a PhD student but whom I have failed to mention here. Rest assured that your support and help were much appreciated. I presumably should have taken more of your advice than I did. All the remaining errors in this thesis, unquestionably, are mine, and mine alone.

To conclude, I would like to express my gratitude the consortium of the CINACS Graduate Research Group headed by Jianwei Zhang for making the — not unchallenged — decision of giving a mature student such as myself the opportunity to participate in CINACS. It was one of the intellectually most stimulating and enriching periods of my life.

Last, and by no means least, I want to thank my entire family — and especially my wife Kirsa, my children Bennet and Linnea and my mother as well as my in-laws Dagmar and Eberhard — for being there for me always ... and for putting up with me in the course of the past few years on those not infrequent occasions when my mind was revolving around thematic role inferences, semantic grammar modelling and lines of Java code more than anything else. Without their love and support, none of this work would have been possible.

Contents

List of Figures

List of Tables

Part I

Model Motivation

Chapter 1

Introduction

A prominent feature of natural language is the occurrence of ambiguity. Ambiguity denotes the fact that a single linguistic entity gives rise to more than one interpretation. The sources of ambiguity are manifold and comprise *lexical ambiguity*, *syntactic* or *structural ambiguity*, *referential ambiguity* and *scope ambiguity* as foremost representatives. Examples for these types of ambiguity are:

Lexical Ambiguity
They ***read*** *a book.*
'read' can be either present or past tense.

Structural Ambiguity
Flying planes *can be dangerous.*
'planes' can either be the direct object of 'flying' or the subject of 'can'.

Referential Ambiguity
He *is a friend of mine.*
Without disambiguating context it is unknown which entity in the real world 'He' is referring to.

Scope Ambiguity
There was a name tag beside ***every*** *plate.*
The quantifier 'every' can take wide or narrow scope such that there may have been a single name tag beside all plates or a separate name tag beside each plate.

Linguistic enquiry leads to the realisation that ambiguity is an inherent property of natural language rather than a defect; as such, it contributes to the linguistic norm rather than constituting an exception to that norm. Despite the omnipresence of ambiguity, language-mediated communication between humans is surprisingly successful in general, even when ambiguities remain without explicit or conscious resolution. We consider an ambiguity resolved if the number of its possible interpretations has been reduced down to precisely one. Relative to the frequency of their occurrence, misunderstandings resulting from the above types of ambiguity are quite rare. This begs the question as to the nature of the cognitive processes that account for the comparative robustness and effectiveness of human natural language understanding in the presence of ambiguity.

In principle, three approaches for processing linguistic ambiguity are conceivable:

1. Attempt ambiguity resolution and succeed.
   In this case, disambiguation can be achieved either by adopting suitable defaults in linguistic decision making or by the automatic and unconscious incorporation of additional sources of information. Including the additional information permits to constrain utterance interpretation, which results in the dismissal of invalid interpretations. Plausible candidates for such additional sources of information are discourse context, world knowledge and immediate visual scene context.

2. Attempt disambiguation and fail.
   If disambiguation according to 1 failed and the resolution of the ambiguity is indispensable for achieving a communicatively adequate level of utterance understanding, linguistic processing must attract attention to signal for help in disambiguation. In this case, the inability to arrive at a single uniform interpretation blocks the process of understanding and may trigger appropriate communicative strategies to resolve the ambiguity interactively. In contrast to the other two options, the ambiguity has surfaced into consciousness in this case.

3. Do not attempt disambiguation.
   An ambiguity that still permits to attain a level of understanding which is appropriate in the given communicative situation may remain unresolved. The corresponding linguistic entity then continues to be processed in its semantically underspecified form and may be resolved at a later stage when sufficient information is available for its disambiguation.

Findings from psycholinguistic suggest that human language understanding in fact involves a mixture of the three strategies: Ambiguities whose resolution is not essential for the overall comprehension of the utterance or the speech act may be left unresolved and seem to be processed in their semantically underspecified, "*good-enough*" form (Ferreira et al., 2002; Christianson et al., 2006; Ferreira and Patson, 2007). For ambiguities whose resolution is essential to the given communicative situation, disambiguation is attempted by access to information from readily available sources such as discourse context, world knowledge or immediate visual scene context. If successful, the resolution of these ambiguities proceeds automatically, i.e., without any conscious effort. Finally, ambiguities essential for understanding which cannot be resolved need to be addressed consciously. Typically, this involves clarification strategies that are compatible with the pragmatic constraints of the current communicative situation. The majority of linguistic ambiguities is handled by strategies 1 and 3 such that the presence of ambiguity in human communication is rarely even consciously noticed.

Given that the production and understanding of linguistic utterances by humans is always embedded in some form of context (e.g., Crain and Steedman, 1985; Gee, 2001), the automatic integration of extra-sentential context information plays a significant role in *situated language comprehension*. Yet, in the implementation of computational language analysis systems, contextual influences upon linguistic analysis and language understanding still constitute one of the most widely disregarded factors. As a result, the majority of parsers today still proceed sentence by sentence and compute their linguistic analyses in complete contextual isolation.

The focus of this thesis therefore is on the modelling of linguistic ambiguity resolution as part of natural language understanding based on information from immediate visual scene context as an extrasentential and non-linguistic source of information. As an example for our modelling focus, consider Sentence 1.1, taken from Tanenhaus et al. (1995). This syntactically ambiguous instruction can be parsed to afford either of the structural representations 1.1.1.Syn or 1.1.2.Syn. Each of these structural representations corresponds to a semantically distinct interpretation which we represent by the conjunction of predicates in 1.1.1.Sem and 1.1.2.Sem, respectively. In the absence of a biasing context, both interpretations are equally acceptable; each interpretation has a valid structural representation such that a decisive disambiguation on syntactic grounds alone cannot be achieved. A purely syntactic parser needs to incorporate additional information in order to arrive at a qualified structural decision.

(1.1) *Put the apple on the towel in the box.*

(1.1.1.Sem) $put\_on(Apple, Towel) \wedge in(Towel, Box)$

(1.1.1.Syn) [Put [the apple]$_{NP}$ on [the towel [in the box]$_{PP}$]$_{NP}$]$_{S}$.

(1.1.2.Sem) $put\_in(Apple, Box) \wedge on(Apple, Towel)$

(1.1.2.Syn) [Put [[the apple]$_{NP}$ [on the towel]$_{PP}$]$_{NP}$ in [the box]$_{NP}$]$_{S}$.

The integration of suitable context information can help constrain the linguistic analysis of Sentence 1.1 to support the formation of interpretational preferences. Context provides support to linguistic analysis if referential links between contextual and linguistic entities are established; otherwise, the context is perceived as unrelated to the utterance. Visual scene context can contribute to disambiguation if words in the sentence are found to refer to entities in the visual scene. Tanenhaus et al. (1995) observed that humans, when presented with an ambiguous sentence in a visual scene context, automatically attempt to establish referential links between linguistic and visual entities, i.e., humans assume that the sentence *makes reference to* the co-present visual scene and hence attempt to match linguistic entities and entities in visual context across modalities.

Once we know which words refer to which entities in the visual scene, the relations between referents in the visual scene can enrich our knowledge of relations between linguistic entities. Contextual support of disambiguation is achieved if the knowledge from the visual scene imposes additional constraints on the set of acceptable linguistic interpretations. Contextual constraints do not effect the complete dismissal of an interpretation; rather, they influence the degree of an interpretation's acceptability in the given context. Acceptability hence is a graded and context-dependent phenomenon (Crain and Steedman, 1985).

Interpreting Sentence 1.1 in the presence of a visual scene context containing a single apple and a towel which is lying in a box will provide a strong bias in favour of Interpretation 1.1.1.Sem. Conversely, a visual scene context containing an apple resting on a towel beside an empty box will afford a preference for Interpretation 1.1.2.Sem. The preferred sentence interpretation is the one which most closely aligns with the visually perceived state of affairs. A modification to the visual scene context can therefore modulate the linguistic interpretation and hence the corresponding syntactic analysis. This is an evident example of the influence of non-linguistic visual scene context upon linguistic decision making – and syntactic analysis, in particular.

Considering the importance of visual context in situated language understanding it is surprising to see how few successful computational modelling approaches have been reported for this phenomenon. In the extant models, the problem of integrating cross-modal context into language processing is primarily perceived as an engineering challenge rather than as an issue of cognitive process modelling. Consequently, the implementation focus of those models is on *observational adequacy* rather than on the adequate modelling of cognitively plausible processes of human cognition and natural language understanding. Nor do the existing models attempt to integrate into the context of a more comprehensive theoretical framework of human cognition. With the work presented in this thesis, it is our intention to make a first step towards bridging the gap between cognitive theory and methods of natural language engineering. We aspire to do so by deriving requirements for our computational model from two sources: behavioural observations of cross-modal interactions in human language processing and an integrated theory of human cognition. In approaching the modelling challenge from a cognitive as well as from a language-engineering perspective, we aim to design and implement a model that — apart from exhibiting observationally adequate behaviour — also meets important cognitive requirements of natural systems and, as such, can be argued for within the framework of a general theory of human cognition.

1.1 Line of Argument and Central Claims

From the large span of interaction phenomena between vision and language, we select the influence of visual scene understanding upon linguistic processing as the topic of this thesis. We use the term *visual understanding* in a broad sense to comprise the entire process of visual perception from the initial stages of sensory processing to the higher stages of visual processing and interpretation. We use the term *linguistic processing* to denote the processes of semantic and syntactic analysis in the context of natural language understanding. One of the central questions to be addressed in this thesis is how inherently non-linguistic information from a visual scene context can affect linguistic processing — and the resolution of syntactic ambiguity, in particular. The primary objective of this work is to motivate, implement and evaluate a model for the influence of visual understanding upon linguistic processing based on an existing syntax parser implementation. Our modelling approach is structured into three main steps: 1) the identification of key findings from the literature and the derivation of suitable modelling requirements from those findings, 2) the integration of the collected requirements into a coherent and implementable computational model, and 3) the critical evaluation of that computational model's implementation.

The line of argument and the central claims in this thesis can be summarised as follows: There is significant empirical evidence to suggest that visual and linguistic processing proceed in parallel and strongly interact with each other in the course of their progress (Cooper, 1974; Tanenhaus et al., 1995; Spivey et al., 2002). Experimentally observed eye-movement patterns support the interpretation that humans continually seek to establish reference between linguistic and visually perceived entities (Tanenhaus et al., 1995; Spivey et al., 2002). A critical factor in establishing cross-modal reference is the degree of conceptual compatibility between the concepts activated linguistically and concepts activated visually (Cooper, 1974; Huettig et al., 2006). A cognitively motivated model of the cross-modal matching between linguistic and visual entities must therefore link the representations of linguistic and visually perceived entities to the corresponding concepts. Furthermore, the model must permit to evaluate the conceptual compatibility between different concepts.

An integrated theoretical account of the interaction between non-linguistic information and linguistic processing is provided by Jackendoff's theory of Conceptual Semantics which provides a representationalist account of cognition (Jackendoff, 1983). Each modality creates its own, domain-specifically encoded representation such that modalities are informationally encapsulated and cannot directly interact with each other (Jackendoff, 1996). For this reason, the representations resulting from visual understanding and syntactic processing cannot interact with each other directly. According to Conceptual Semantics, there are two indirect ways in which modalities can interact with each other: either via an interface which maps between the modalities' representational codes based on correspondence rules or via a mediating shared level of representation which is constrained by the interacting representations. Conceptual Semantics centres around the hypothesis that cross-modal interactions with language are all mediated by a single, uniform level of semantic representation which encodes concepts, concept instances and semantic relations between concept instances (*Conceptual Structure Hypothesis*). This uniform representation of semantics is constrained by syntax and visual understanding. The representations of syntax and visual understanding interact with the mediating semantic representation via representational interfaces.

Our model of linguistic processing seeks to implement this mediation between linguistic and non-linguistic information via a shared semantic representation. In line with Conceptual Semantics, our model treats visual context as a source of additional, non-linguistic information that gives rise to constraints on the set of acceptable semantic interpretations of linguistic input. The constraints of visual context propagate into syntax via the interface between the syntactic and semantic levels of representation.

Our model implementation centres around a constraint-based parser that permits the integration of additional constraints – such as visual context compliance – into its linguistic processing capabilities. We augment the parser's syntactic processing capabilities with a semantic level of representation that interfaces with the syntactic level via correspondence rules. The semantic level of representation is constrained to comply with both syntax and the semantic representation of visual context. The representation of visual context consists of ontological concept instances between which semantic relations have been defined. Contextual constraints enforce the compliance of the shared semantic representation with visual context. We hence achieve a semantically mediated propagation of visual context information into syntax: visual context constrains the semantic representation of linguistic semantics which, in turn, interacts with syntactic representation. To show the effectiveness of our model, we evaluate its disambiguating capabilities under a number of different contextual conditions.

1.2 Thesis Structure

The overall structure of this thesis reflects the structure of our approach and hence breaks down into three main parts: the outline of the model motivation in Part I, the detailed description of the proposed model and its computational implementation in Part II and the discussion of the experimental results from model validation as well as the summary of the overall conclusions in Part III.

The model motivation in Part I begins with the introduction provided in this chapter to delineate the thesis topic and to define the topical focus of the thesis. Chapter 2 reviews the state of the art, both in behavioural research and in computational modelling. We present central publications from the current body of literature on the interaction between vision and language and provide an overview over extant modelling efforts. A small number of more recent modelling implementations are discussed in detail.
An important constraint to our model is the requirement of its integrability into a more general theory of cognition. To this end, Chapter 3 introduces Ray Jackendoff's Conceptual Semantics as a theoretical framework which offers an integrated account of the cross-modal interaction between vision and language.
Chapter 4 motivates the use of WCDG, a weighted-constraint dependency-parser, as the component for linguistic processing in our model. The chapter also outlines the benefits and limitations of approaching natural language parsing as a constraint-satisfaction problem. Chapter 4 concludes our model motivation and the collection of modelling requirements.

Part II provides an in-depth description of our modelling decisions and the implementation-specific aspects of the proposed model. We begin with a detailed description of the functional enhancements to the WCDG parser in Chapter 5. These functional extensions were needed to enable the integration of visual context information into linguistic processing.

Another important aspect of our model is the representation of situation-invariant semantic knowledge and situation-specific visual scene knowledge. We describe our modelling decisions regarding the representation of these types of knowledge in Chapter 6. The chapter also outlines the role of the reasoner in our model and describes the types of inferences it draws.

The PPC is the central component in our model which enables the cross-modal influence of visual context upon linguistic processing. We described it in detail in Chapter 7. We outline how fundamental cognitive processes in the cross-modal interaction between vision and language such as *grounding* and *cross-modal matching* are implemented in our model and how visual context information can exert an effect upon linguistic processing.

In Part III, finally, we report the behaviour of our model under various experimental conditions. The capability to perform semantic parsing constitutes a key prerequisite for our model implementation. Chapter 8 describes a pre-experiment in which the coverage of the semantic extension to WCDG's standard grammar in our model is evaluated on a corpus of unrestricted natural language.

Chapter 9 discusses the first application of our model implementation. The aim of this experiment is to demonstrate that an influence of visual scene information upon syntactic parsing can be enforced in our model. This chapter offers a discussion of the results obtained from enforcing an absolute dominance of visual context over linguistic analysis by integrating contextual information via hard integration constraints.

In the subsequent chapters we report successive refinements to the initial context integration approach. The first improvement is provided by turning the context integration constraints into soft constraints on linguistic analysis. Constraint relaxation permits to balance contextual against linguistic preferences such that the absolute dominance of visual context over linguistic analysis is resolved. As a consequence of constraint relaxation, our model can process and diagnose conflicts between linguistic and contextual preferences. The effects of constraint relaxation upon linguistic analysis and syntactic disambiguation are reported in Chapter 10.

Chapter 11 discusses the importance of grounding for the cross-modal influence of visual context upon linguistic processing. In these experiments we release the assumption that linguistic and visual modality provide information of the same degree of conceptual specificity. In that chapter we investigate the effect upon syntactic parsing that results from integrating conceptually underspecified representations of visual scene context.

Part III of the thesis concludes with Chapter 12 which contains a summary of the central findings and conclusions of this thesis as well as an outlook to future directions of research.

The appendix to this thesis provides additional material to complement the examples given in the argumentative parts of this thesis. Concretely, it contains the list of all requirements collected, the concept hierarchy used in context modelling, mathematical derivations of some of the more complex formulae quoted, the sentences studied in the experimental runs as well as all the parse trees for the reported experiments and the empirical data based on which the graphs were plotted.

Chapter 2

Cross-Modal Interactions between Vision and Language

The scientific investigation of cross-modal interactions between vision and language has been intensifying continually since the report of the first linguistically relevant studies in the 1970s (e.g., Cooper, 1974, 1976; McGurk and MacDonald, 1976, 1978). A comprehensive view of the spectrum of these interactions needs to integrate insights from psycholinguistics, cognitive neuroscience, cognitive psychology, linguistics and cognitive science. It is the purpose of this chapter to provide a phenomenological overview over some of the central aspects of the cross-modal interactions between vision and language. We cite influential empirical reports that form a major source of motivation for the modelling attempt described in this thesis. In the course of our discussion of the literature we identify relevant requirements for the implementation of a computational model. The empirical observations presented in this chapter are intended to serve as a fact basis that an integrated theory of cognition needs to account for. One such theory will be discussed in Chapter 3.

This chapter begins with establishing the distinction between the cross-modal interactions in sensory and representational modalities in Section 2.1. From there we proceed with a focus on the interaction between vision and language, and outline cross-modal interaction phenomena at word and sub-word level in Section 2.2. Following the course of historical development in the field, we discuss the findings of some very influential studies on the interaction between vision and language comprehension at the level of linguistically more complex units such as phrases and entire sentences in Section 2.3. Section 2.4 reviews investigations aiming to illucidate the nature of the mental representations underlying the cross-modal interaction with language. Section 2.5 provides an overview of existing computational modelling efforts for the cross-modal interaction between vision and language.

2.1 Sensory versus Representational Modalities

For simple auditory-visual stimuli such as combinations of light flashes and beeps, multisensory integration has been reported to commence as early as visual cortical processing, about 46 ms after stimulus onset (Molholm et al., 2002). In comparison, the cross-modal interactions with the cognitively higher levels of linguistic processing such as language understanding occur at a much later period in time. EEG studies reveal that specific brain responses to lexical, syntactic and semantic features of linguistic input are observed in the order of magnitude of one to several hundred milliseconds after stimulus onset. These latencies can be accounted for by considering that the linguistic information must first be extracted and decoded from the sensory input via which it has been received in the auditory, visual or haptic modality. Interactions with language understanding hence build on the results of sensory processing and consequently must be temporally posterior to the onset of sensory processing in the sensory input modality.[1] Multisensory integration, on the other hand, occurs during early and cognitively lower-level sensory processing. The empirically observed and significant temporal differences in cross-modal integration responses provide a first indication of the qualitative difference between the cross-modal interactions of purely sensory and linguistic stimuli.

The categorisation of sensory stimulation is performed based on of the physical parametrisation of its sensorially detectable properties such as brightness, loudness, pressure, temperature, duration etc. If the information encoded in the stimulus is non-symbolic in nature, stimulus categorisation results in the formation of a direct link between the internal representation of the stimulus and the conceptual category it activates. If, on the other hand, the stimulus encodes symbolic information, its categorisation results in the identification of the encoded symbol. The retrieval of the symbol's meaning is a separate process. In contrast to the linguistic symbols which do carry a meaning, non-symbolic percepts have no intrinsic meaning. It is in this respect, that cognitive processing of a purely sensory stimulus differs from that of a sensory stimulus which encodes symbols with an intrinsic meaning, such as language. We refer to a modality that encodes and processes the latter type of stimuli as a *representational modality*.Other, non-linguistic examples of representational modalities are spatial, musical or visual scene understanding. In all of these, low-level sensory perception provides input which, upon categorisation of the encoded symbols, is processed further in higher cognitive processes. We henceforth refer to a stimulus evoking purely sensory simulation that encodes exclusively non-symbolic information as a *sensory stimulus*. A stimulus evoking sensory stimulation which encodes symbolic information is referred to as a *representational stimulus*. A special subset of representational stimuli are *linguistic stimuli* in which the encoded information consists of linguistic symbols.

[1] This is not to say, however, that sensory and linguistic processing occur in strict temporal succession; nor do they proceed in complete isolation of each other.

Processing a linguistic stimulus results in the categorisation of its sensory input as consisting of discrete[1] linguistic building blocks or *atoms* in a temporal sequence. For spoken language, these atoms are the identified phonemes; in reading and touch-reading, they are the individual letters perceived. Combinations of these atoms form arbitrary linguistic symbols, be they morphemes or words, that combine "*rulefully*" (Harnad, 1990) to make up an utterance. Each of these arbitrary linguistic symbols carries its own meaning that it contributes to the process of evaluating the utterance's overall meaning. The categorisation of a linguistic stimulus hence gives rise to a discrete symbolic representation.

The diverse nature of the information encoded in different modalities — be they sensory or representational in nature — begs the question of whether — and if so, how — different modalities can interact with each other at all. An integrated account of cross-modal interaction with language must be expected to provide an answer to this question. The general theory of cognition discussed in Chapter 3 does indeed offer an account of these phenomena.

In the further course of this thesis we refer to an early cross-modal interaction at the stage of sensory processing as *multisensory integration*. We continue to use the more general term *cross-modal interaction* for any type of interaction in which two modalities mutually affect each other. For a strictly unidirectional effect of one modality upon another we adopt the term *cross-modal influence*.

Both multisensory integration and cross-modal interactions between representational modalities serve the purpose of minimising the amount of incompatible information in cognition. How this goal is achieved, differs depending on the type of modalities that interact.

In the sensory modalities, multisensory integration produces a single, informationally fused percept from multimodal sensory input whenever possible.[2] When the information obtained from the different modalities is compatible with each other, multisensory integration gives rise to superadditive neural response patterns and produces a robust integrated percept of the different sensory inputs. This is observed, for example, in cases where and auditory and a visual stimulus temporally and spatially co-occur within well-defined temporal windows (e.g., Wallace et al., 1998).

In cases in which the information in the modalities is cross-modally incompatible, sensory processing still attempts to form a single, uniform percept from the sensory input. The physical parameters of that percept are chosen such that the overall perceptual conflict between the modalities is minimised. Interestingly, the percepts

[1] This holds true even if the sensory input via which language is received is encountered as a – more or less – continuous stream of input. Typical examples are the continuity of human-generated speech or the continuous flow of movements in the production of sign-language.

[2] A discussion of the boundary conditions under which multisensory integration occurs is beyond the scope of this thesis. Suffice it to say here that certain spatio-temporal constraints apply in order for multisensory integration to occur. Meredith et al. (1987), e.g., investigate the temporal constraints on stimulus co-occurrence in order for multisensory integration to occur.

thus generated do not truthfully represent the sensory input anymore; they are indeed *sensory illusions* created by our brain to satisfy the overall cognitive goal of reducing the perceptual conflict that arises from the incompatibility of the sensory inputs. Classic examples for this type of cross-modal conflict resolution by multisensory integration are visual capture phenomena such as the *ventriloquist effect* or the *Shams illusion*. In the ventriloquist effect, the presence of a dominant visual stimulus influences the spatial localisation of a co-occurring auditory stimulus (e.g., Bertelson and Aschersleben, 1998). In the Shams illusion, the perceived number of visual stimuli is modulated by a co-occurring auditory stimulus (Shams et al., 2002).

In representational modalities, cross-modal integration effects do not occur as part of sensory processing but during the subsequent stages of interpreting already classified symbolic input. To achieve cross-modal integration, an interpretation is generated in which the information from the different modalities is unified into a coherent overall interpretation. As an example, consider a situation in which a deictic pronoun is used in the linguistic modality and a potential referent can be inferred from a pointing gesture in the process of visual understanding. If the properties of the identified referential candidate are compatible with the referent properties expected based on the pronoun, then the integrated interpretation will treat the deictic pronoun and the pointing gesture as co-referential. If visual understanding provides several referential candidates that give rise to equally acceptable interpretations, further referential disambiguation may be required.

If the interpretation of the entities from visual and linguistic processing are incompatible, e.g., because of an apparent number or gender disagreement of the deictic pronoun with the referential candidate pointed at, an alternative interpretation of the multimodal information needs to be found which removes – or at least minimises – these conflicts. Cognitive strategies for conflict resolution can be to initiate a visual search for an alternative referent or to re-analyse the linguistic input in search of an alternative, compatible interpretation (e.g., Spivey et al., 2001).

If no acceptable interpretation can be found, alternative communicative or perceptual strategies may be triggered, depending on which modality's input appears more reliable. These alternative strategies can be an attempt to either disambiguate the linguistic input, e.g., by means of clarification questions, or to improve the quality of cross-modal perception, e.g., by modification of the visual perspective.

2.2 Cross-Modal Interaction at Word and Sub-Word Levels

One of the earliest reported – and presumably most widely known – examples for a cross-modal interaction between vision and language is the *Stroop effect* which refers to the interference between a word's meaning and the time it takes to respond to the colour in which the word is printed. In his very influential and frequently cited study, Stroop (1935) investigated subjects' performance on two tasks: the reading aloud of colour words printed in coloured ink (Experiment 1) and the naming of the ink colour in which colour words were printed (Experiment 2). Experiment 1 did not produce any significant interference between reading speed and the colour in which the colour words were printed. Experiment 2, on the other hand, revealed a substantial increase in response time on ink colour naming for words that denoted a colour different from the ink colour they were printed in. Notably, this interference persisted even with training on the task.

Modern cognitive psychology emphasises the role of attention in the Stroop effect (MacLeod, 1991, p. 187). In the literature, the most common — though not undisputed (MacLeod, 1991, p. 188) — explanation for the effect and its inherent asymmetry is the *relative-speed-of-processing account*. According to this account, words are read and comprehended faster than colours are named. In the Stroop experiments, the two processes compete with each other to trigger a response (*response-competition*). The focus of attention determines which response is desired. Hence, the observed interference between the two processes is larger when the focus of attention is on the completion of the slower process: By the time colour naming is performed, the result of the faster word reading process is already available. The response to its outcome needs to be suppressed in order to permit the response of the attended-to slower process to come through. Clearly, this suppression is not required when attention is directed to the output of the faster process. In that case, the attended process returns a result before the slower process has completed, so no inhibition is required.

From the perspective of a cross-modal interaction between vision and language, the relative-speed-of-processing account is somewhat unsatisfactory as it grounds on the assumption that the two processes, word reading and colour naming, occur independently of each other and only differ in the time they require to trigger a response. This account effectively adopts a modular view on processing in the Fodorian sense.[1] The relative-speed-of-processing account also cannot explain two important additional observations related to the Stroop effect:

1. The gradience effect of semantic distance upon the strength of the observed Stroop interference reported by Dalrymple-Alford (1972): words that do not denote a colour themselves but are associated with a colour, such as the word *sky*, produce a stronger interference on the colour-naming task than words that are completely colour-neutral. Their effect is not as strong, however, as that of incongruent colour words proper.

2. Stroop facilitation as reported by Dunbar and MacLeod (1984) and others: when colour word and ink colour coincide, response times for ink-colour naming are slightly faster than in the control conditions. The observed effect is smaller than the response delays in the incongruent cases, but still has been shown to be statistically significant.

[1] The modularity of the human language faculty goes back to Fodor (1983). Modules in the Fodorian sense are informationally encapsulated cognitive units that process information individually and in parallel. The interaction between modules is restricted to an interaction via their input and output, i.e., modules cannot interact with each other in the course of their processing. Modules process their input bottom-up in a strict feed-forward manner such that the higher-level cognitive functions, which Fodor labels *central processes*, do not influence lower-level processing. Modules process their input automatically, fast and domain-specifically. According to Fodor, each module is associated with a fixed neural architecture and hence exhibits characteristic breakdown patterns.

Successful attempts to model a large number of observations associated with the Stroop effect computationally have been reported (e.g., Roelofs, 2003). However, to date there is no unanimously accepted account of the effect that can explain *all* related findings listed in MacLeod (1991)[1] and Roelofs (2003).
For the purpose of this thesis, suffice it to say that the Stroop effect is the result of an only partly understood complex and asymmetric interaction of reading, visual perception, attention and action at word-level. The interference, facilitation and semantic gradience effect observed in the colour-naming task support the interpretation that at some stage of visual and linguistic processing semantic representations arising from different modalities are involved in the cross-modal interaction.

With the advent of eye tracking technology in the early 1970s, the interactions between vision and language have become considerably more accessible to scientific enquiry. The first use of eye tracking technology to study interactions between vision and language was reported by Cooper (1974). Cooper used a camera to monitor eye movements of subjects who were simultaneously exposed to visual stimuli in the form of object depictions and auditory linguistic stimuli. This experimental procedure subsequently became known as the *visual-world paradigm.*[2] Cooper showed that spoken word semantics influenced subjects' fixation patterns on co-present visual stimuli. More specifically, Cooper found that from a selection of nine co-present visual stimuli subjects preferably fixated those that were either direct depictions of referents denoted by the words presented auditorily or depictions of items semantically related to the words' referents. Cooper concluded that the eye movement patterns are a reflection of the on-line activation of word semantics from speech.
Huettig et al. (2006) point out that Cooper did not control for the type of semantic interaction that gave rise to the observed cross-modal effect. The fixation preference on the semantically related visual stimuli could have arisen from either associative relatedness or genuine semantic similarity. While associative relatedness (e.g., *piano* and *practice*)[3] does not necessarily link concepts from the same semantic category, semantic similarity holds between members of the same semantic category only (e.g., *trumpet* and *piano*)[4]. The challenge in conducting word-based association task experiments is to disentangle associative relatedness from semantic relatedness. This differentiation becomes important in the light of Huettig and Altmann (2005)'s findings that the degree of conceptual relatedness between concepts activated in vision and language has an effect upon the strength of the influence of word semantics upon fixation patterns. Huettig and Altmann (2005) found that

[1] This milestone paper provides an extremely detailed and comprehensive review of the first five decades of research on the Stroop effect.

[2] Cooper had the methodological foresight to realise that this novel technique constituted an experimental paradigm whose "*linguistic sensitivity (...) together with its associated small latencies suggests its use as a practical new research tool for the real-time investigation of perceptual and cognitive processes*".

[3] This example is taken from Nelson et al. (1998), a list of association norms for more than 5,000 English word primes and their associated targets. The lists are based on the responses of over 6,000 participants.

[4] This is a carefully constructed example from Huettig et al. (2006) of a semantically related word pair that is not associatively related according to Nelson et al. (1998).

fixation patterns are influenced by semantically related stimuli – but not by associatively related stimuli. We therefore expect conceptual similarity to play an important role in cross-modal matching as discussed in further detail in Section 3.7. The findings of Cooper (1974) and Huettig and Altmann (2005) support the view that the interaction between vision and language is mediated by a representation of linguistic meaning. We formulate this as modelling requirement R1.

Requirement R1

In a model for the interaction between visual context and linguistic understanding, the cross-modal interaction must be mediated by a representation of linguistic meaning.

Another famous and frequently cited interaction between vision and language is the *McGurk effect*. We briefly discuss it here to make clear why we disregard it in the collection of requirements for our model. In the McGurk effect, the visual perception of lip movements and lip shapes interacts systematically with the auditory perception of concurrently presented phones (McGurk and MacDonald, 1976, 1978). In their classical experiment, McGurk and MacDonald auditorily presented subjects with the phone `/ba/` dubbed onto a video of a mouth producing the phone `/ga/`. Subjects reported to hear the phone `/da/`. The McGurk effect hence occurs at the level of individual phones, i.e., at sub-word level. Exploiting the systematicity of the cross-modal effect, Massaro and Stork (1998) report the illusion also to occur for larger phonetic units such as an entire sentence: exposing subjects to `/My bab pop me poo brive/` auditorily and `/My gag kok me koo grive/` visually induced the auditory percept of *My dad taught me to drive.*[1] The perceived auditory percept can simply be predicted on the basis of concatenating the individual cross-modal interactions at phoneme level. The McGurk effect has been studied extensively and is observed robustly over a wide range of languages and different conditions such as speaker-gender incongruence between visual and auditory modalities and others.

The important difference between the McGurk effect and the interactions between vision and language observed in Stroop's experiments is that the McGurk effect is based on an an *early* interaction between vision and the *auditory perception of speech*. The McGurk interaction affects the perception of phones rather than any later – and cognitively higher – stages of processing that involve language comprehension. This view is consistent with the observation that the McGurk effect also occurs with single syllables, pseudo-words and non-words, none of which have a meaning or semantic representation that could form the basis of this interaction.

The McGurk effect has widely been interpreted as a bottom-up integration of incongruent cross-modal stimuli. More recent studies suggest, however, that top-down effects in the form of sentence context and word semantics can also modulate the strength of the effect (Windmann, 2004; Ali, 2007). The question whether the influence of vision upon audition in the McGurk effect is due to a bottom-up integration

[1] Massaro and Stork also showed that the corresponding unimodal stimuli on their own were unintelligible: The majority of the subjects gave an accurate phonemic description of the non-sensical audio input and were unable to extract a meaning by lipreading the video when either stimulus was presented in isolation.

in stimulus identification or due to an expectation-modulated interaction in stimulus discrimination — or a combination thereof — has not been answered conclusively.[1] In analogy to the combination of bottom-up and top-down processes believed to operate during visual object recognition, we hypothesise that the McGurk effect also results from a convergence of bottom-up and top-down processes acting in parallel. In the context of this thesis we classify the effect as a primarily sensory phenomenon which can experience top-down modulation under special conditions. The robustness of the effect in the absence of expectation- or knowledge-driven top-down effects further supports the interpretation in terms of a bottom-up integration. As such, we choose to exclude it from further consideration in our model of the influence of visual context understanding upon linguistic processing.

Summarising the cross-modal interactions between vision and language at word and sub-word levels, we can say that the Stroop effect and Cooper's visual world experiments provide convincing evidence for the involvement of a semantic representation in the cross-modal interaction between visual and linguistic processing. Cooper's experiments suggest that the interaction between modalities is such that visual processing aims to identify entities in visual context which are conceptually related to the concepts activated linguistically. Huettig and Altmann refined this view to show that only semantic relatedness gives rise to the effect. The observations of the Stroop effect suggest that the degree of conceptual overlap between the concepts processed in each modality has an effect on the ease with which certain tasks can be performed. For tasks exhibiting a Stroop effect, conceptual congruence results in task facilitation and conceptual incompatibility results in an interference.

The following section investigates the effect of non-linguistic information obtained from visual understanding upon the processing of more complex linguistic structures such as phrases and entire sentences.

2.3 Cross-Modal Interaction at Phrase and Sentence Level

In another milestone investigation into the interactions between vision and language, Tanenhaus et al. (1995) recorded subjects' eye movements during the concurrent auditory-visual presentation of syntactically ambiguous sentences in the presence of visual scene depictions. Tanenhaus et al. found that subjects' eye movements were tightly time-locked with the unfolding of the linguistic stimulus, i.e., eye movements progressed to linguistically relevant referents in strict temporal alignment with the mentioning of the corresponding entities in the linguistic modality. The observed latencies between word end and the fixation of the correct object were in the range of 145 ms when the visual scene contained no other object with a phonetically similar name and around 230 ms when there was an object with a similar name. Previously, Matin et al. (1993) had reported the average time needed to compute a saccadic eye movement at around 200 ms. Tanenhaus et al. hence concluded that, when no

[1] See Section 3.6 for a discussion of the processes of *discrimination* and *identification* in the bottom-up grounding of sensory perception.

other object with a similar name was present, subjects must have integrated the information from the auditory stimulus and the visual scene to accomplish object identification *prior to* hearing the end of the word, i.e., prior to completing the perception and processing of the respective auditory stimulus. Visual distractor objects which exhibited no referential connection to the linguistic stimulus — such as a pencil for the sentence *Put the apple on the towel in the box.* — had no observable effect upon fixation patterns.

In their central experiment, Tanenhaus et al. presented subjects with locally ambiguous instructions of structure $\texttt{V NP PP}_1\{\texttt{PP}_2\}$. The prepositional phrase ($\texttt{PP}_1$) could be interpreted either as a modifier to the sentence initial verb (V) or to the noun phrase object (NP). The local ambiguity was resolved either by the unfolding of $\texttt{PP}_2$, in which case $\texttt{PP}_1$ was interpreted as a modifier to NP, or by the end of the sentence, in which case $\texttt{PP}_1$ was interpreted as a modifier to V.

If initial syntactic processing were modular in the Fodorian sense (see Footnote 1 on page 15) — and as such informationally encapsulated against visual scene information —, no effect of visual scene context on early syntactic processing and hence on eye fixations should be observable. Tanenhaus et al. found, however, that subjects' fixation patterns differed significantly depending on whether the visual scene contained a single or two possible referents for the NP. In the case of a single potential referent for NP in the visual scene, $\texttt{PP}_1$ was initially interpreted as a modifier to V. This initial interpretation had to be revised when a $\texttt{PP}_2$ subsequently followed. In the case of two potential referential candidates for NP in the visual scene, $\texttt{PP}_1$ was always initially interpreted as a modifier for NP.

The authors interpret the observed eye movements as direct reflections of the progress of syntactic processing. The different fixation behaviours induced by the difference in visual contexts show that the same transient ambiguity of $\texttt{PP}_1$ can give rise to different syntactic starting hypotheses. The authors interpret this as evidence for an access to visual context information during the earliest moments of linguistic processing. Their observations provide substantial support for the hypothesis of a *close* and *continual* interaction between visual and linguistic processing. A continual interaction between visual and linguistic processing is postulated by the proponents of strongly interactive models of sentence processing that contrast with the modular processing architecture suggested by Fodor.

From this most influential investigation we extract a number of modelling requirements related to the interplay between visual and linguistic processing. The observation of successive eye movements to linguistically relevant referents in synchrony with the unfolding of the linguistic stimulus shows that linguistic processing progresses over time and is *incremental*.

Requirement R2

In a model for the interaction between visual context and linguistic understanding, linguistic processing must be incremental.

As revealed by the strong time correlation between eye fixations and linguistic processing, the interactions between the two modalities occur in close temporal alignment.

Requirement R3

A model for the interaction between visual scene context and linguistic processing must be based on temporally synchronised interactions between the visual modality and linguistic processing.

The immediate interactions between visual and linguistic processing observed by Tanenhaus et al. support a strongly interactive model of sentence processing based on continual cross-modal interactions at parse time. These interactions enable what Tanenhaus et al. refer to as the "*rapid and nearly seemless integration of visual and linguistic information*".

Requirement R4

A model for the interaction between visual scene context and linguistic processing must be based on continual interactions between non-linguistic information and linguistic processing.

Tanenhaus et al.'s experimental findings further support the view that the interaction between visual and linguistic processing is bidirectional. We capture this as two separate requirements, one for each direction of the interaction. Given the same syntactic material in different visual scene contexts, fundamentally different fixation patterns were observed. This is clear evidence for the influence of visual scene context upon the early stages of linguistic processing.

Requirement R5

A model for the interaction between visual scene context and linguistic processing must include the influence of visual understanding upon linguistic processing.

The experiments also demonstrate the influence of linguistic upon visual processing: the mention of linguistic entities in the auditory input immediately directed eye fixations to the corresponding referent in the visual scene.

Requirement R6

A model for the interaction between visual scene context and linguistic processing must include the influence of linguistic processing upon visual understanding.

From the fact that referentially unrelated visual distractor objects had no observable effect upon linguistic processing we conclude that referentially unrelated visual information remains neutral with respect to linguistic processing.

> **Requirement R7**
> *In a model for the interaction between visual scene context and linguistic processing, referentially unrelated visual context information must leave linguistic processing unaffected.*

2.4 Information in the Mind & Information in the World

Open to this point is what the mental substrate for the interaction between vision and language might be. Interestingly, both Cooper and Tanenhaus et al. observed *anticipatory eye movements*, i.e., eye movements to visually represented entities prior to the complete unfolding of the corresponding linguistic stimulus. However, neither author offers a discussion of which form of information may specify the targets of these eye movements or in what form that driving infomation might be encoded mentally. Based on the extremely fast eye movements to the depictions of the visual scene, visual search can be excluded as a possible explanation for the anticipatory eye movements. The speed and precision with which the eye movements were executed in the setting of Tanenhaus et al. suggest that information about the target position must already be available when planning the eye movement. It is therefore likely that the anticipatory eye movements are driven by a mental – and thus inherently internal – representation of visual scene information.[1] Clearly, the creation of this representation must have preceded the onset of the linguistic stimulus and the corresponding eye movements in time.

Altmann (2004) reports experimental evidence in favour of such a mental representation as the basis for the interaction between vision and language. Altmann employed the *blank-screen paradigm*, a variation of the visual-world paradigm in which after an exposure of about 5 seconds the visual stimulus is removed shortly before the onset of the linguistic stimulus. Given sufficiently small inter-stimulus intervals of about 1 second, Altmann observed eye movements similar to those obtained in the visual-world condition, even in the absence of the corresponding visual stimulus. Altmann concludes that the observed eye movements result from an interaction between language and a stored mental representation of the visual scene. He argues that eye movements are not based on the actual location of the item but on the position of the item in the representation of the scene. This interpretation also permits to explain the anticipatory eye movements observed by Cooper and Tanenhaus et al. since the information of where a given entity is located in the visual field can be encoded in the mental representation of the visual scene.[2]

[1] We expand further on the mental representations resulting from visual perception in Section 3.1.

[2] In our view, additional experimental investigation is required for a full understanding of the nature of the spatial references to the visual scene. The setup in Altmann (2004) does not clarify rigorously whether the spatial references represented mentally encode the location of absolute points in space or just their position relative to the perceiving subject. It remains open whether the spatial references to the

Altmann's findings support a representationalist view of visual scene perception and provide evidence for the hypothesis that the cross-modal interaction with language occurs based on the mental representation of the visual scene. We capture this as requirement R8:

> **Requirement R8**
>
> *In a model for the interaction between visual scene context and linguistic processing, linguistic processing interacts with a representation of the visual scene context.*

Altmann (2004) argues for the existence of a mental representation of visual scene context and proposes that subjects have access to the information about the location of objects in that representation. The question arises which other information the mental representation of the visual scene holds and how informationally rich that representation is. Is every object perceived in the visual field also represented with the totality of information known about it — or are there cognitive mechanisms at work that strive to reduce the amount of internalised information in favour of ease and efficiency of encoding and processing?

To gain further insight into the level of detail with which objects in the visual field are represented mentally, some authors have consulted findings from change blindness research (e.g., O'Regan, 1992; Spivey et al., 2004). *Change blindness* refers to subjects' inability to detect sudden changes to the visual field if these occur during blinks (O'Regan et al., 2000) or saccades (McConkie and Currie, 1996) or are accompanied by minor visual distractors such as short image flickers (Rensink et al., 1997). However, changes to the visual field that occur in the course of a fixation uninterrupted by blinking or saccadic eye movements are noticed immediately (Yantis and Jonides, 1990).

Spivey et al. (2004) attribute change blindness to the informational sparseness of the mental representation of a visual scene. According to Spivey et al., the mental representation of a visual scene only encodes a fraction of the information that can be extracted from the visual scene itself, leaving a significant part of the information in the outside world, waiting to be accessed if and when needed. Access to that information proceeds via *spatial indices* or *pointers* in the mental representation that themselves are informationally poor but point to the spatial location of the object in the visual scene from which the information can be retrieved via oculomotor ac-

presented objects are encoded in terms of egocentric or allocentric coordinates. In all studies discussed so far, subjects remained in a constant spatial relation to the site of visual stimulus presentation throughout the experiment. A discrimination between egocentric or allocentric spatial frames of reference could not be achieved in that way. To achieve a clarification, it would be necessary to examine how eye movements are affected in the blank-screen paradigm by systematic changes to the spatial relation of the site of stimulus presentation to the subject (or vice versa) in the interstimulus interval. Based on evidence from spontaneous eye movements in visual imagery as reported by Brandt and Stark (1997), Mast and Kosslyn (2002) or Spivey and Geng (2001), we would expect spatial references to be encoded in terms of coordinates in an egocentric frame of reference. If this is indeed the case, subjects' fixation patterns should remain unaffected by changes to the spatial relation between the site of stimulus presentation and the subject in the interstimulus interval.

tivity.[1] The concept of spatial indices goes back to Pylyshyn in the late 1980s (for more recent accounts of spatial indices, see Pylyshyn, 2000, 2001). Spivey et al.'s argument of *sparse encoding* means that the visual scene becomes an externalised part of the individual's memory which can be retrieved by moving the eye to the corresponding pointer target.[2]
A spatial index minimally needs to contain information about the spatial location of its reference object as well as basic labelling information on when and how to access the reference object. Spivey et al. (2004, pp. 169) review a number of experiments that are consistent with the interpretation that object attributes such as colour are *not* encoded in the pointer information. This permits the conclusion that the pointers to objects in the visual scene indeed are informationally sparse and do not encode the full object information which readily is available from the visual scene.

In contrast, Simons and Ambinder (2005) argue that the observation of change blindness as such does not warrant the conclusion of representational sparseness. According to Simons and Ambinder, change detection to the visual environment requires a comparison of the present state of the visual scene with a mental representation of the preceding state and the identification of differences between the two. Simons and Rensink (2005) describe this comparison as comprising the following steps: encoding the preceding state, retaining that representation, perceiving the current state, accessing the representation of the preceding state and comparing the current state with the representation of the preceding state. An overall failure to detect change may therefore arise from a failure at any one of these steps. Simons and Rensink point out that, before valid conclusions can be drawn from the change blindness experiments, a number of experimental conditions still need to be controlled for to ascertain change blindness indeed results from the effects that Spivey et al. attribute it to. We hence cannot draw final conclusions on the granularity of visual scene representation and hence choose not to formulate any requirements on this aspect of modelling.

2.5 Extant Computational Models

With the availability of eye tracking, the interaction between vision and language understanding has been studied empirically *in extenso* for some decades now. Given the large body of behavioural findings, it is quite surprising that only a comparatively small number of computational models of the interaction between vision and language have been reported. Maybe less surprisingly, none of the extant models cover the *full* scope of interactions between vision and language that is known from

[1] Spivey et al.'s definition of spatial indices only addresses pointers to concrete physical objects in the visual scene. This definition leaves unanswered the question of the granularity of reference. Furthermore, Spivey et al. do not discuss whether – and if so, how – references to higher aggregates of individual objects or more abstract correlations between objects are represented, accessed and processed internally.

[2] The philosophical implications of this argument in the context of the mind-brain dualism are that the mind does indeed have access to more information than the brain physically holds.

the behavioural investigations.

According to Roy and Mukherjee (2005), models of vision-language interactions can be categorised qualitatively by the type of information provided by the visual modality: *intention-related* and *situation-related* information. Intention-related visual information conveys sender intention in the act of producing the linguistic signal, be it lip movements in speech production, or gestures in sign language. This type of visual information is exploited as visual input to systems for audio-visual speech or gesture recognition. Situation-related visual information, on the other hand, is information *about* the immediate visual scene in which the linguistic stimulus is produced and typically contains references to entities or situations in the visual scene. Roy and Mukherjee point out that visual context comprises both intention-related and situation-related visual information. Since we focus on situation-related visual information in the context of this thesis, we limit our discussion to extant model implementations which incorporate situation-related visual information.[1]

2.5.1 Historical Overview

Historically one of the first systems ever to combine natural language understanding with different levels of non-linguistic representation was Winograd's SHRDLU reported in Winograd (1971).[2] SHRDLU was a dialogue system for English capable of answering questions and executing commands in a blocks world based on knowledge representations of semantic information and context. A heuristic understanding component combined syntactic analysis with context information and world-knowledge to determine actual sentence meaning. While the system did not incorporate computer vision as such, it was capable of manipulating internal knowledge representations of the spatial arrangement of different objects.

André et al. (1988) describe the implementation of SOCCER, a system for the generation of natural language descriptions for dynamically evolving visual football scenes. The linguistic descriptions arise from the recognition of situation instances in the visual scene. In contrast to earlier work, SOCCER generates its descriptions *in parallel* to the incremental processing of the visual scene rather than in retrospect. Retrospective generation essentially is a sequential process of the linguistic re-encoding of previously extracted visual information. SOCCER performs linguistic planning while the process of visual extraction is still ongoing. As a result, changes in the output of visual recognition can still have a limited effect upon language generation. The extent to which changes in visual information dynamically influence language generation cannot be assessed based on the information given in André et al. (1988).

[1] Another classifications of computational models for the integration of linguistic and visual information is provided in the review article Srihari (1995). Here, systems are classified into those accepting only unimodal and those that accepting bimodal input. For a historical review of extant modelling efforts, we consider this classification less helpful.

[2] His PhD thesis was subsequently re-published with minor changes as Winograd (1972).

The system XTRACK reported in Koller et al. (1991) adopts a similar approach in the automated characterisation of motion trajectories in traffic scenes captured by a stationary camera. The central achievement of this implementation lies in its extraction of characteristic scene features and the subsequent mapping of detected motion trajectories onto one of approximately ninety different motion verbs.
Brown et al. (1992)'s speech activated manipulator SAM is a robotic system with sensory capabilities that is controlled via natural spoken language. SAM obtains world information from two sensors and from conversation and fuses that input to perform actions in a blocks world. The robot understands about 1041 semantically meaningful English natural language sentences with a vocabulary of about 200 words. Speech recognition is constrained by a finite state grammar and augmented with a domain-specific semantic analysis to arrive at a single interpretation for the linguistic utterance. Integration of linguistic and sensory information such as object shape, height, size, location and colour is performed when both processing streams are complete. Information fusion is additive across modalities. Conflicting information is resolved interactively with the human controller of the system.
Srihari and Burhans (1994) describes PICTON, a system for extracting linguistic information from image captions to guide a computer vision system in image understanding. PICTON employs a natural-language-processing module to generate constraints for subsequent in image understanding. A language-image interface then fuses the information from the encapsulated processes of natural language and image processing by applying the constraints from linguistic processing upon the hypotheses generated by the image-understanding module. The system is applied to the domain of face recognition in newspaper articles.

For the model implementations that have been reported in the last decade we now provide somewhat more detailed discussions.

Model 1. A Bayesian network implementation for the integration of speech and image understanding: Socher et al. (1996); Socher (1997); Wachsmuth et al. (1999); Socher et al. (2000)

Model 2. A model implementation of visual context priming achieved via an online influence of visual context upon the language model underlying speech recognition: Roy and Mukherjee (2005)

Model 3. A connectionist model for the anticipation and assignment of thematic roles in a visual world context: Mayberry et al. (2005a,b, 2006)

Model 4. A robotic system for incremental language processing with tight perceptual and motor integration: Brick and Scheutz (2007)

We briefly introduce these models now and discuss their strengths and weaknesses. Aspects for discussion include the implementation's suitability to our task of modelling the influence of immediate visual scene context upon linguistic processing, the model's scalability, the generality of the context representations employed, the mechanistic transparency of vision-language integration and the model's integrability into a more general theory of cross-modal cognition.

2.5.2 Model 1: A Bayesian Network Implementation for the Integration of Speech and Image Understanding

Socher et al. (2000) report a model for image understanding based on three components: speech understanding, image understanding and a Bayesian network as integrating inference machine. The system fuses visual information from a 3D-camera with linguistic information from the automated recognition of spoken instructions to identify objects in the visual scene and to carry out simple instructions. The domain is limited to manipulations performed on objects from a wooden toy construction kit. Typical instructions are '*Give me the X.*' or '*Take the X.*' where *X* is the specification of a domain object. Visually, objects are identified based on their type, colour and spatial relations to other objects. Visual object recognition results from a hybrid approach in which a neural network generates object hypotheses which are then either confirmed or rejected based on information from a semantic network.[1]
To model uncertainty and errors in the sensor input, Socher et al. enrich the qualitative object representations with probabilistic information that expresses the reliability of the hypothesised object properties. The system achieves a translation of numerical, sensory input into qualitative, symbolically encoded object information that is accessible to reasoning under uncertainty in the Bayesian network. Combining the input from vision and speech, the Bayesian network computes the most plausible overall interpretation of the situated natural language instruction and performs the according action. Vision-language integration in this model is late in the sense that both signals are first processed individually and then fused into an integrated cross-modal percept.
The accuracy of the system for real data is reported at 92.5% when using idealised, i.e. recognition-error-free, input data. In more realistic scenarios in which both visual and linguistic modalities are afflicted with sensory error, the rates for object identification vary between 70% and 86.3%.[2][3]

Socher et al.'s Model 1 achieves a convincing late integration of modularly processed visual and linguistic information for image understanding. Visual object recognition works well in the modelled domain but requires a more generalised knowledge base representation of object attributes to ensure scalability and applicability to other domains. The two most salient limitations of the approach with regards to our modelling objective are the limitation of visual information to object recognition and the late integration of vision and language. While establishing object co-reference between modalities is an important part of modelling the cross-modal interaction between vision and language, the recognition of situations and thematic roles of the participants in a scene is another significant output of visual understanding. Since the domain selected for this model only comprises static spatial relationships

[1] Details about this approach are given in Socher (1997).

[2] Result precisions are quoted as reported in the original article.

[3] Further extensions to this system have been reported by Bauckhage et al. (2002) and others. However, the mechanisms of cross-modal integration in these extensions do not differ substantially from those in the implementation described here.

of objects this aspect of vision-language interaction has not been considered. Most relevantly from the point of view of linguistic processing, the interaction between vision and language does not occur interactively at parse time. The processes of linguistic and visual processing exhibit no interaction prior to their integration in the Bayesian network.

As for the mechanistic transparency of the cross-modal integration, the Bayesian network behaves like a black box whose associations are formed during training. The model is not argued for in the context of a general theory of cross-modal cognition.

2.5.3 Model 2: A Model for the Effect of Visual Attention upon Speech Recognition

With FUSE, Roy and Mukherjee (2005) report the successful modelling of the effect of visual attention upon speech recognition. The reported model consists of four main components: a speech decoder, a visual scene analyser, a language-driven visual attention module and a language model driven by visual context. The model is applied to a constrained scene description task in which Lego® blocks of specific colour and size have to be identified given visual and linguistic input.

As with most automated speech recognition systems today, the output of the speech decoder depends on a statistical language model. Typically, these language models are invariant to cross-modal context and hence result in modular, contextually encapsulated processing of the speech input. In this model, however, the likelihoods expressed for word recognition in the language model vary with the input provided by the visual scene analyser and the visual attention component. Furthermore, visual attention is directed to those elements in the visual scene which have been extracted during the early stages of speech recognition. The result is a system in which speech recognition drives visual attention, visual attention dynamically influences the language model and the language model enhances the expectation for the recognition of certain words or combinations of words based on visual context. This cycle of propagated influences can be interpreted as an implementation of a bi-directional interaction between vision and language via 1) a top-down influence of visual attention upon speech recognition and 2) a top-down influence of speech recognition upon visual attention.

The average speech recognition error is defined as the percentage of words that are classified incorrectly in the auditory modality. In the absence of a visual context, the system achieves a speech recognition error of 24.3%. The introduction of a visual context effects an improvement of 31% and reduces the average speech recognition error down to 16.7%. The average error rates for object recognition improve by 41% from 24.4% in the absence of a visual context down to 14.3% in the presence of a visual context. The bottom line for random identification was an error rate of 90%.

In view of the reported reductions in speech and object recognition, the system performs rather well. More impressive, in our view, than the numerical results of this model's reported speech and object recognition accuracies is its cognitively plausible architecture which allows to integrate a top-down influence of speech upon vision and of vision upon speech into incremental linguistic processing. The immediate

influence of both modalities upon each other at the time of processing is highly significant in that it constitutes the first computational model of an early, non-modular integration of vision and language.
Based on the information provided in Roy and Mukherjee (2005) it is difficult to judge whether the system can scale up. To be able to maintain the central benefit of this model – namely the mutual influence of the two modalities upon each other at the time of processing – it needs to be ensured that the effect of information extracted from the visual domain can be propagated into the language model at the time of linguistic processing. Otherwise, the cyclic effect of speech recognition upon visual attention upon the language model upon speech recognition breaks down. It remains questionable whether this can be achieved when removing the strong domain restrictions of this model and extending its linguistic scope to unrestricted natural language input. Especially the trained statistical component for enhancing the probability of certain word combinations in the language model may not scale arbitrarily.

This model's major limitation with regards to our modelling objective, which it shares with all the other models discussed here, is its limitation of visual understanding to the level of object recognition and inter-object spatial relations. In contrast with some of the earlier work described in the historical overview in Section 2.5.1, no situation recognition is performed. Furthermore, and in marked contrast with Model 1, the system provides no reasoning capabilities for handling possible conflicts between the results of visual and linguistic understanding. While capable of replicating significant behavioural properties of natural systems in vision-language integration, the model itself is not argued for in the context of a specific theory of cross-modal cognition.

2.5.4 Model 3: A Connectionist Model of Anticipation in Visual Worlds

Mayberry et al. (2005a) present a simple recurrent artificial neural network that is capable of making highly accurate thematic role assignments in the course of incremental sentence processing given an input sentence as well as visual scene information. A simple recurrent network is chosen because it exhibits three attractive properties: automatic development of expectations *prior to* the completion of processing, seamless integration of input from multiple sources and instances of non-monotonic hypothesis revision that are reminiscent of human re-analysis behaviour during incremental linguistic processing.
Two different implementations of the network are evaluated on input material which previously was studied in eye-tracking experiments with human subjects. Input to the network are the sentence as well as the visual context information. The representation of visual context encodes AGENT and THEME relations for all participants in the visual scene but no grammatical information such as case or gender. The network's output indicates which of the two nouns in the input sentence of limited structural variance is predicted to be the AGENT and which one the THEME. The network was trained on 1,000 sentences over 15,000 epochs, which is reported to have taken about two weeks on a regular PC. Input sentences are generated from a lexicon of 326 words on which a number of morphological and lexical simplifications have been imposed to facilitate training and testing.

The first version of the model displays imperfect anticipation rates of 96% and 95% for two sets of unambiguous sentences. The error is attributed to incorrect token identification. On two other sets of ambiguous sentences, the model reaches disambiguation accuracies of 100% and 98%, respectively. Most relevant with regards to our modelling objective, however, is the model's performance in the fifth experiment, in which visual scene information was set to dominate in case of conflicting linguistic and visual inputs. The best results for this condition exceed 99% accuracy in thematic role anticipation during incremental sentence processing and reaches 100% at sentence end.

Model 3 makes highly accurate predictions on the assignment of thematic roles given visual scene information and a relatively short German input sentences built from a lexicon of toy size. In addition to using incremental sentence processing, the system provides very accurate thematic role anticipations during sentence processing as hypothesised by various models of incremental sentence processing.

In the studied examples, the assignment of thematic roles was a binary syntactic structural decision. While the thematic role anticipation and final assignments are performed with high accuracy, the complexity of the linguistic task is substantially lower than that of building up full syntactic and semantic representations for an input of initially unknown structure. Given the long training time of two weeks for the comparatively small number of sentences of moderate lexical and syntactic complexity, it seems exceedingly unlikely that this model scales up to be able to process arbitrary representations of visual scene contexts in combination with unrestricted natural language input.

The context representations used were non-declarative and encoded thematic relations between entities in terms of weights in the network's hidden layers. The use of the connectionist approach results in a loss of mechanistic transparency for the process of cross-modal integration. Since the internal representation of the model is purely activation-based, the system cannot perform any symbolic reasoning operations. It hence remains unclear how the performance of this model – which, for the reported domain, unquestionably is impressive – generalises and integrates into a more comprehensive cognitive account of the cross-modal integration for vision and language in natural systems.

2.5.5 Model 4: A Model of Incremental Sentence Processing with Tight Perceptual and Motor Integration

Brick and Scheutz (2007) report RISE, a robotic system capable of integrating sensory information from binocular camera vision into the processing of spoken natural language instructions at parse time. The system performs incremental syntactic and semantic parsing in parallel and additionally integrates visual scene information from a block world to constrain the set of sentence interpretations. Furthermore, pragmatic constraints are imposed to guide syntactic decisions of phrase closure.

A notable feature of RISE is that it can anticipate the selection of referents and their communicative function *before* the completion of linguistic processing, e.g., it can decide whether a referent is the *operand* or the *destination of movement* while still processing a given *move*-instruction. Brick and Scheutz use a set-based approach to determine the number of possible referents for an identified word in visual context. Unambiguous reference is established when the referent set size is unity. When unambiguous cross-modal reference has been established, syntactic phrase closure is effected based on the pragmatic consideration that referential overspecification is similarly undesirable as in human-to-human communication. Any additional modifiers detected in speech are interpreted as referring to a new referent.

While the performance of the system has not been evaluated formally, the examples discussed make a convincing case for the online integration of referential cross-modal information into incremental natural language understanding. This model performs an early identification of referents based on visual information and elementary reasoning. The coupling of linguistic and visual understanding with action allows the system to effect actions and so-called *back-channel responses* such as immediate changes of gaze to unambiguously identified referents, prior to completing the analysis of the linguistic input.

With its link between cross-modal processing and action this model goes one step further than Model 2 in that the results of incremental linguistic and visual processing do not only influence each other but can even trigger new, externally perceivable system actions. This approximates human behaviour during verbal communication which comprises continual responses on various levels to the incrementally processed linguistic input. We consequently consider Model 4 a prime candidate for an attempt to model the closely time-locked influence of vision upon language during incremental processing as reported by Tanenhaus et al. (see Section 2.3).

Brick and Scheutz (2007) leave open to what extent the model supports the reverse direction of the interaction between vision and language, i.e., the attention-mediated influence of language upon vision. None of the quoted examples show a revision of visual processing based on linguistic input which nurtures the suspicion that visual processing in this model is still modular.

From the quoted system outputs we conclude that the influence of linguistic processing upon vision is implemented in terms of triggered actions such as visual system queries to verify or disambiguate linguistic information. The system's remarkable benefit of identifying referents prior to their complete linguistic specification can only be maintained if these queries are performed in a time frame that still permits the subsequent back integration into incremental linguistic processing. While, at first glance, this appears to be more of an engineering than a conceptual challenge, it may have significant influence on the system's capability to scale up. In the absence of definitive time scale information for the system outputs, however, no factually substantiated prediction can be made here.

In summary we can say that the historical development of the model implementations for the interaction between vision and language as discussed in this section reflects the level of understanding of the cross-modal interaction: while early models were based on a predominantly modular view of visual and linguistic processing with late integration of modalities encapsulated in the Fodorian sense, more recent models shift towards increasingly interactive realisations that incorporate an integration of vision and language during the earlier stages of processing. The driving force behind

this trend is the insight that human language processing is inherently incremental and highly interactive already during the early stages of linguistic processing as demonstrated by Tanenhaus et al.
The implementations discussed here only work for a restricted domain and have an uncertain potential for upward scalability. More importantly, the extant models are limited in their use of the information they extract from the visual modality: So far, visual information is only used to establish cross-modal co-reference at object-level and for the extraction of the spatial relations that these objects engage in. While some earlier computational models address the recognition and categorisation of situations from visual scenes, none of the more recent models discussed here incorporate these aspects.

2.6 Chapter Summary

In this chapter we have presented a number of significant experimental findings that permit important conclusions as to the interaction between the processing of visual stimuli and linguistic processing in the context of natural language understanding. The Stroop effect demonstrated that the interaction between vision and language was automatic and mediated by lexical semantics. Cooper's visual world experiments provided further evidence for a semantic mediation in the interaction between vision and linguistic processing. By careful control of experimental stimuli, Huettig et al. refined our understanding as to which kind of semantic interaction leads to cross-modal interaction with language: cross-modal interaction with language is mediated by semantic category similarity rather than by associative relatedness. Tanenhaus et al. argue convincingly for a bidirectional, incremental and closely time-locked interaction between vision and language. Finally, Altmann provides experimental support for a representational view of the cross-modal interaction with language. All of these findings have been formulated as requirements for our model of cross-modal interaction of vision with language.

This chapter has also provided an overview over extant historical and more recent computational implementations of the interaction between vision and language. Socher et al. use a Bayesian network to achieve reasoning capabilities in late cross-modal integration. Roy and Mukherjee report a successful implementation of visual priming in speech recognition during incremental parsing. Mayberry et al. describe a successful – but presumably not scalable – connectionist system for anticipating binary thematic role assignment decisions during incremental sentence processing. Finally, Brick and Scheutz provide a remarkable account of coupling robot action with incremental and contextually aware sentence processing. All of the discussed models, however, restrict themselves to the extraction of object information and spatial relations. They fail to utilise visual context to extract information about the thematic roles that the recognised entities take in the context of an observed situation. As for visual understanding, this constitutes another, higher level of complexity which is still to be integrated into modelling. Furthermore, none of the reported models are motivated by or integrated into a general, implementation-independent theory of cross-modal cognition.

In the next chapter, we set out to discuss a general theory of cognition that attempts to account for the mechanisms that enable a cross-modal interaction between non-linguistic and linguistic modalities. We intend to base the specification of our model upon requirements from these three sources: 1) the body of experimental findings presented in this chapter, 2) the discussion of extant models in this chapter and 3) the cognitive theory to be presented in the following chapter.

Chapter 3

Conceptual Semantics — An Integrated Theory of Cognition

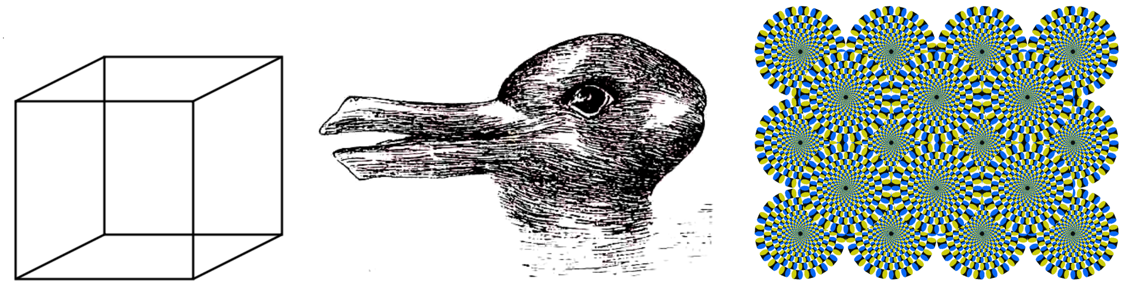
The interaction of non-linguistic modalities with language is a mental process that occurs quite effortlessly in our brains. We therefore know the effects of cross-modal interaction from our own experience. Yet, we are mostly unaware of the mechanisms that underly this interaction. The preceding chapter has provided important experimental findings about the interaction between vision and language. What we are missing to this point is a unified cognitive theory capable of providing an integrated account of the observed phenomena.

In this chapter we outline Ray Jackendoff's theory of Conceptual Semantics in as far it pertains to the cross-modal interaction of non-linguistic modalities with language. Conceptual Semantics takes a representationalist view of cognition and offers a perspective on the interaction between non-linguistic modalities and language.

We begin this chapter with an argument in favour of representationalism as a prerequisite to the discussion of Conceptual Semantics. In Section 3.2 we introduce important constituents of the cognitive architecture Jackendoff develops in the context of his theory of Conceptual Semantics. Section 3.3 discusses to what extent encoding is representation-specific. Sections 3.4 and 3.5 describe the elements for semantic representations by addressing Conceptual Structure and thematic relations. Sections 3.6 and 3.7 outline the fundamental issues of grounding and cross-modal matching in the interaction between linguistic and non-linguistic modalities. Throughout the course of our discussion, we continue to identify further modelling requirements, now from the perspective of an overarching theory of cognition.

3.1 Representationalism

The questions whether our senses show us reality or just a filtered projection thereof has intrigued philosophers since antiquity. The systematic study of perceptual illusions has been used extensively to gain insight into how perception is represented and processed in our minds. In cognitive psychology, a large number of visual illusions are known that induce multistable or even apparently dynamic visual percepts as the result of visual ambiguity. Famous representatives from this class of visual



(a) The Necker cube. (b) Jastrow's duck-rabbit. (c) Apparently rotating circles.

Figure 3.1: Examples of visual illusions in which a constant visual stimulus results in a multistable or even dynamic visual percept.

illusions are the Necker cube in Figure 3.1 (a), Jastrow's duck-rabbit illusion[1] in Figure 3.1 (b), or the apparently rotating circles in Figure 3.1 (c). All of these illusions have in common that a temporally invariant, static visual stimulus produces a non-static visual percept.

The occurrence of perceptual illusions as such — and the observed perceptual dynamics resulting from a static stimulus in particular — are important arguments to support the view that we do not actually perceive the world *as it is* but only the way that *our senses tell us about this world*. This characterisation of cognition is advocated by the school of *representationalism*. Its central tennet is that human cognition and consciousness are based on internal mental representations of the world in the mind of the perceiver rather than the real world itself. While causally connected, the real world and its mental representation clearly are distinct from each other.

Mental representations are construed based on input from the sensory modalities in combination with the results of subsequent processing by the higher cognitive faculties. Cognitively experienceable in the view of representationalism is only what has been mentally represented before.

Other well-known, though not necessarily fully understood, cognitive phenomena providing support for a representationalist view of cognition are visual mental imagery, dreams and hallucinations. For all of these, subjects report mental states that are very similar – if not identical – to the states that result from the regular sensory perception of the corresponding external stimuli. The mental image of one's office chair, for example, is largely congruent with the actual visual percept attained when looking at that chair in the real world. Representationalism holds that the information about this chair is encoded as a mental representation. It is the very same representation that is activated irrespective of whether we are visually perceiving or just imagining that very chair.

[1] Wittgenstein (1953, p. 22) also discusses this illusion, albeit on the basis of a graphically somewhat simplified version. For this reason, some sources in the literature, e.g. Jackendoff (1983, p. 25), refer to the illusion as *Wittgenstein's* duck-rabbit.

In the subsequent development of our model for the interaction between non-linguistic modalities and linguistic processing we follow Jackendoff (and many others) in adopting a representationalist view of cognition. We consequently treat the representations of linguistic and visual understanding as representational modalities.

3.2 Levels of Representation

Any theory for the interaction of non-linguistic modalities with linguistic processing will need to account for how the different levels of linguistic description interact with non-linguistic levels of representation. Jackendoff (1983)'s theory of Conceptual Semantics is a representationalist approach to such an integrated account of cognition and language processing. We now outline the central elements of Jackendoff's theory in as far as they pertain to our modelling objectives. We also identify further modelling requirements for the interaction between non-linguistic modalities and language that arise from Conceptual Semantics.
Jackendoff (1983) proposes two distinct – though connected – levels of mental representation for syntax and semantics. This proposal is in fulfilment of the general requirement for a semantic theory identified by Katz and Fodor (1963) which stipulates that the description of a language shall be split into two distinct levels of description, namely syntax and semantics. We adopt this as a requirement for our model of the cross-modal interaction with language:

Requirement R9

A model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics must contain distinct levels of representation for syntax and semantics.

Katz and Fodor make a point about excluding phonology from consideration in their treatment of the requirements for a comprehensive description of a language.[1] Jackendoff, however, in his description of the linguistic system, includes a phonological level of representation which can act as an input channel to syntactic representation. Since the overall objective of this thesis is to argue for a model of the cross-modal interaction with language — and the syntactic level of representation in particular — we henceforth limit our description of Conceptual Semantics to the interaction between syntactic representation and non-linguistic modalities. We consequently disregard the interaction between the phonological and syntactic levels of representation in the discussion of Jackendoff's theory and follow Jackendoff in assuming that the contents of syntactic representation are independent of the modality from which the represented input has been obtained. A syntactically relevant piece of information will therefore be represented in the same way syntactically, irrespective of whether it has been received from the processing of speech, text, sign language or a non-linguistic modality.

[1] Katz and Fodor (1963, p. 172) mark the difference between a *full description of a language*, which comprises syntax and semantics, and a *full theory of speech*. Katz and Fodor argue that the latter must also cover phonological aspects in order to be able to account for speech production and recognition.

Jackendoff argues that every mental representation must comply with a finite set of representation-specific well-formedness rules (WFRs), which he assumes to be universal and innate (Jackendoff, 1983, p. 17). We capture the representational constraint on the well-formedness rules as requirement R10:

> **Requirement R10**
>
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the set of permissible representations on a given level of representation must be defined by a finite set of well-formedness rules.*

3.3 Representational Modularity

Jackendoff (1992, p. 4) argues that every level of mental representation is encoded in its own *"language of the mind"*. Since representations are dedicated to a specific cognitive domain such as *syntax* or *semantics*, they are domain-specific. As a result, most of their representational primitives cannot be shared with other levels of representation. For encoding and processing, the mind therefore dedicates a separate module to each representation. Since each level of representation is encoded uniquely, it is informationally encapsulated in the Fodorian sense (Fodor, 1983). For two distinct cognitive modules M_A and M_B, cognitive module M_A can only decode and process information from representation R_A — but not from M_B's representation R_B.

The key difference between Jackendoff's and Fodor's notion of modules is, that Jackendoff's representational modules are characterised by the distinctness of the representation they process whereas Fodor's modules are characterised by the cognitive function the modules provide. The latter implies that the cognitive functions themselves are modular, which makes them inaccessible to online interaction during processing. The modularity of cognitive functions cannot account for the observed immediate and incremental cross-modal interaction phenomena described in Section 2.3.

Representational modularity does permit interactions between representations in the course of processing. Levels of representation can interact with each other via *representational interfaces* that translate between different representational encodings. This translation process is governed by *correspondence rules* that stipulate which information from a given source representation is encoded in its target representation and how. Fodorian modules do *not* permit such an online-interaction since they are strict input/output modules: a new input is only accepted once processing of the previous input has completed. We summarise these aspects of representational modularity in the following modelling requirements.

According to Jackendoff, cognitively different information is encoded in different languages of the mind that give rise to different, domain-specific representations.

> **Requirement R11**
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the encoding of each representational level is domain-specific.*

Each individual representation is domain-specific and hence processed by a separate module. As input, this module only accepts the one representation that it specialises on.

> **Requirement R12**
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the processing on each level of representation is representationally encapsulated.*

Since every module is encoded in its own language of the mind it can only interact with another module by mapping the different representations onto each other according to correspondence rules.

> **Requirement R13**
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the mapping between representations is achieved by correspondence rules.*

In the view of Conceptual Semantics, two different levels of representation can only interact via the interface between them. It is in this interface that the correspondence rules between the representations are applied.

> **Requirement R14**
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the interaction of levels of representation via representational interfaces occurs online, i.e., at the time of linguistic processing.*

As for the cross-modal interaction with language, two important questions arise at this point:

Q1. Do non-linguistic representations interact directly with the linguistic levels of representation or is their interaction mediated by intervening levels of representation?

Q2. Do representational modalities interact with the linguistic levels of representation via the same mechanisms as sensory modalities do?

Jackendoff's theory provides an answer to these questions by introducing an additional level of mental representation, *Conceptual Structure*. Conceptual Structure is a complex and highly expressive mental level of representation through which all non-linguistic modalities, both sensory and representational, can interact with language. We describe Jackendoff's notion of Conceptual Structure in more detail in the following Section.

3.4 Conceptual Structure

Humans are very well capable of speaking about their sensory perception and of converting verbal instructions into motor action. Jackendoff (1983, pp. 16) attributes this capability to the existence of a single level of mental representation at which the information conveyed by language interacts with both information from other peripheral sensory systems as well as with information subsequently conveyed to the motor system. Jackendoff (1983, p. 17) states this in his *Conceptual Structure Hypothesis*.

> **Jackendoff's Conceptual Structure Hypothesis**
>
> "There is a *single* level of mental representation, *conceptual structure*, at which linguistic, sensory and motor information are compatible".[1]

Jackendoff does not provide a conclusive argument for why the interaction between non-linguistic and linguistic modalities be mediated by a *single* and *unified* level of representation. While this claim is both plausible and attractive from the point of view of representational efficiency, other representational architectures involving more than one level of semantic representation are indeed conceiveable. Jackendoff himself concedes that "*there is no logical necessity for the existence of such a unified level—as there is for the existence of individual interfaces between modalities. At worst, however, the Conceptual Structure Hypothesis is a plausible idealisation; at best, it is a strong unifying hypothesis about the structure of the mind*" (Jackendoff, 1983, p. 17). Jackendoff continues to maintain his view of Conceptual Structure as the single unified level of semantic representation in his later work, but admits that this view is not uncontended (Jackendoff, 1996).

If we accept Conceptual Structure as the single, uniform level of semantic representation, the question arises what information precisely is encoded in it such that an interaction between non-linguistic and linguistic representations can occur. Based on extensive linguistic evidence, mainly from the unsatisfactory mapping of first order logic expressions of simple sentences like *'The book is lying on the table.'* to syntax and a detailed linguistic discussion of categorisation judgements, Jackendoff (1983, Chapters 4–6) concludes that Conceptual Structure must encode information about types, i.e. concepts, and tokens, i.e. individuals, about taxonomic concept relations, relational predicates between concepts and between concepts and indivi-

[1] The textual emphasis is Jackendoff's.

duals. Moreover, Conceptual Structure must provide the capability to evaluate the truthfulness of entailment between propositions it encodes as well as the consistency between concepts. We adopt this as a single, complex modelling requirement:

Requirement R15

A model of Conceptual Structure must encode information about concepts, individuals, taxonomic concept relations and relational predicates such as concept-to-concept and concept-to-individual relations. It must also provide the capability to evaluate the truthfulness of entailment between encoded propositions as well as the consistency between concepts.

Jackendoff (1996, pp. 6) specifies further that Conceptual Structure must capture all non-sensory distinctions of meaning and hence extends the list of information encoded in Conceptual Structure: pointers to representations in the sensory modalities to be able to access encodings of sensory information. These pointers are required, for example, when re-evaluating sensory input based on top-down influences. Observe that this requirement for pointers to sensory information is compatible with Spivey et al. (2004)'s argument for spatial indices presented in Section 2.4. Jackendoff further requires the expressivity of Conceptual Structure representations to comprise quantification and quantifier scope, abstract representations of actions and acting entities, social predicates and modal predicates to express semantic notions such as negation or conditionality. The requirement for the representation of social predicates as part of Conceptual Structure arises from the fact that languages like Thai or Japanese express aspects of social relation to the addressee syntactically. Jackendoff argues that – since Conceptual Structure is assumed to be universal and innate – there is no need for an additional language-specific level of representation in-between Conceptual Structure and syntax (Jackendoff, 1996, p. 8). As social predicates encode linguistically relevant information in some languages, they must form a part of Conceptual Structure for all of humanity. We hence add the above extensions to the notion of Conceptual Structure as individual modelling requirements:

Requirement R16

A model of Conceptual Structure must contain pointers to the representation of sensory information.

Requirement R17

A model of Conceptual Structure must encode quantification and quantifier scope.

Requirement R18

A model of Conceptual Structure must provide abstract representations of actions and acting entities.

Requirement R19

A model of Conceptual Structure must provide social predicates.

Requirement R20

A model of Conceptual Structure must provide modal predicates to express semantic notions such as negation or conditionality.

Given this degree of expressivity of Conceptual Structure representations, the question arises whether the need for a separate representation of linguistic semantics can still be maintained at all. If every information encoded in the representation of linguistic semantics is expressable as an informationally equivalent representation in Conceptual Structure there is no need for a distinct representation of linguistic semantics anymore. Based on the identity of the inferences drawn in visual perception and semantic representation, Jackendoff (1983, chapter 6) concludes that the representation of linguistic semantics does not encode any information that cannot also be expressed representationally in Conceptual Structure. The set of representations of linguistic semantics hence forms a subset of the set of the conceptual structures expressible in Conceptual Structure. Jackendoff argues that, for reasons of representational economy, a separate level of mental representation for linguistic semantics is superfluous. He postulates that consequently the representation of linguistic semantics is included as part of Conceptual Structure. We capture this as modelling requirement R21:

Requirement R21

A model of Conceptual Structure must encode the semantic part of linguistic representation within Conceptual Structure.

On the basis of this assumption we can further clarify how the interaction between non-linguistic modalities and language needs to be modelled: the representations produced by the non-linguistic modalities interface with Conceptual Structure. The correspondence rules in their interface with Conceptual Structure translate these representations into well-formed correspondences in Conceptual Structure. This process results in a projection of the percepts from the individual modalities into Conceptual Structure. According to the well-formedness rules of Conceptual Structure, the projected semantic representations interact with the semantic representation of language which is also encoded in Conceptual Structure. The result is a well-formed, semantically integrated Conceptual Structure representation that interacts with syntax via the representational syntax-semantics interface.
This interaction provides the answer to Question Q1 raised on page 37: non-linguistic representations engage in a direct interaction with the semantic – not the syntactic – level of linguistic representation. This account provides a plausible explanation for the observed semantic mediation in the cross-modal interaction with language as described in Sections 2.2 and 2.3. A schematic rendition of the cognitive architecture according to Jackendoff (1983) is given in Figure 3.2. Special emphasis is on the fact that there is *no* separate representation of linguistic semantics since this is fully included in Conceptual Structure.

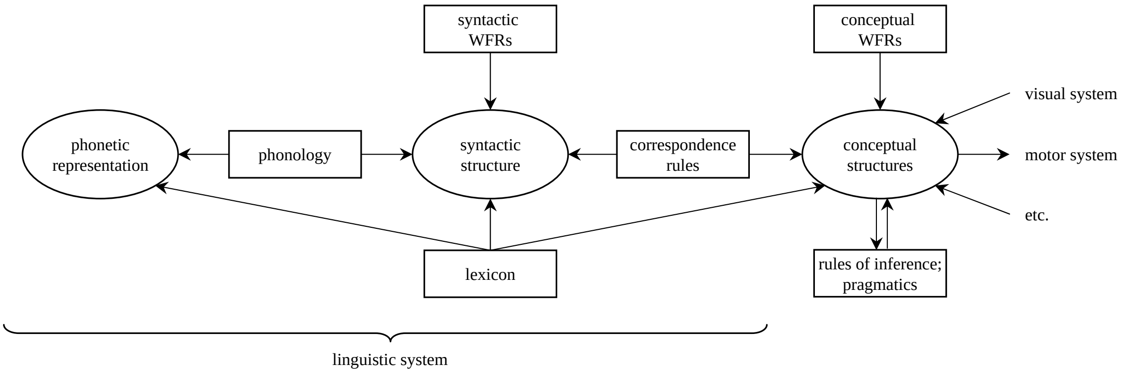


Figure 3.2: The cognitive architecture for the interaction between the linguistic system and the sensory modalities according to Jackendoff (1983).

For the purpose of modelling cross-modal interaction with language we adopt Jackendoff's Conceptual Structure hypothesis as modelling requirement R22:

> **Requirement R22**
>
> *A model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics must contain a single, uniform level of semantic representation. This level interfaces with the syntactic level of representation and constitutes the central representation of linguistic and non-linguistic semantics. Meaning-based interactions between non-linguistic modalities and language must be mediated by this level of representation.*

While in his earlier work Jackendoff describes Conceptual Structure as the single level of semantic representation into which all sensory modalities project, this view is slightly modified in his later work: (Jackendoff, 1992, Chapter 6) and Jackendoff (1996) include an interface of Conceptual Structure to *Spatial Representation*, another level of representation resulting from the higher cognitive faculties of spatial cognition. Spatial Representation acts as a similarly central representation for spatial information as Conceptual Structure is for semantic and conceptual information. A valid cognitive architecture needs to incorporate the fact that not all sensory modalities seem to project directly into Conceptual Structure. Jackendoff points out that there are indeed different kinds of auditory perception which he refers to as *general-purpose (g-p) audition* and *auditory localisation*. While general-purpose audition focuses on the perception of an auditory stimulus with the aim of interpreting it, auditory localisation is used to support spatial reasoning by means of spatial localisation of the sound source. Jackendoff argues that general-purpose audition projects into Conceptual Structure while auditory localisation projects into Spatial Representation.

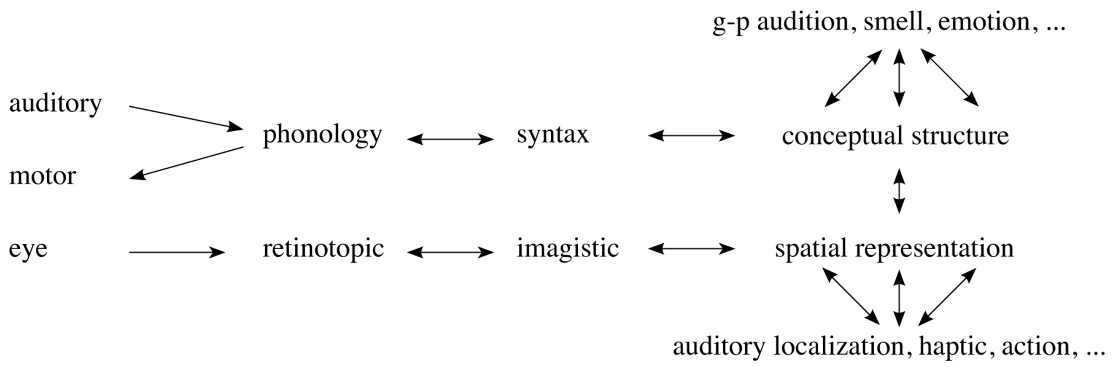


Figure 3.3: The relation between Spatial Representation and Conceptual Structure according to Jackendoff (1996).

An additional amendment to Jackendoff's cognitive architecture in Figure 3.2 is necessitated by the fact that linguistically triggered top-down effects are insufficiently represented. Cooper (1974) showed that the semantics of linguistic stimuli can indeed interact with oculomotor activity (see Section 2.2). The interaction between Conceptual Structure and non-linguistic modalities hence needs to be bidirectional rather than just unidirectional. Jackendoff includes these aspects in the 1996 version of his cognitive architecture given in Figure 3.3 (Jackendoff, 1996, p. 3).
An apparent difference between Jackendoff's architecture in Figure 3.2 and his modified architecture in Figure 3.3 worthy of discussion is the positioning of Conceptual Structure. In his earlier architecture, Conceptual Structure was explicitly marked as outside of the linguistic system. The juxtaposition of Conceptual Structure on the upper line alongside with syntax and phonology in the later version of his architecture seems to suggest that Conceptual Structure is now actually considered to be inside of the linguistic system. This difference can be reconciled by considering that, even in his later, work Jackendoff continues to argue that Conceptual Structure is a "*language-independent and universal*" level of representation (Jackendoff, 1996, p. 8). This view would clearly be incompatible with the inclusion of Conceptual Structure into the linguistic system, despite what Figure 3.3 suggests. We hence conclude that even in his later work Jackendoff continues to position Conceptual Structure outside of the linguistic system.

According to Jackendoff (1996) Conceptual Structure is the only level of representation that integrates input from sensory modalities in a direct interaction with language. It is, however, not the only level of representation which integrates sensory input from different modalities. Further, since Conceptual Structure receives input from both sensory and representational modalities, we can provide an answer to Question Q2 on page 37: Despite the different nature of the information encoded in sensory and representational modalities, their interaction with syntax proceeds along the same pathway, namely via the interface between Conceptual Structure and syntax.

3.5 Thematic Roles and Situation Representations

So far, we have mainly addressed the conceptual entities contained in a Conceptual Structure representation and have dedicated very little attention to the semantic relations assigned to hold between those entities. To build propositional semantic representations of more complex utterances we need to relate projected concept instances by semantic relations in Conceptual Structure. For verb-centred representations these semantic relations define the participants' thematic roles and allow us to specify *who did what to whom*.

Thematic roles as introduced by Gruber (1965) and later by Fillmore (1968) initially were intuitive linguistic abstractions to distinguish and classify the different, semantically unique participant functions in an utterace. While the use of thematic roles is not limited to words from a specific lexical category, thematic roles most frequently are defined for verbs where they mark the different semantic functions of each verbal argument, not just in the form of a syntactically motivated label but with a genuine semantic commitment (Dowty, 1989). Clearly, the ability to generalise over verbal argument structures is linguistically desirable.

Historically, however, the extensive and diverse debate of thematic roles in the literature has shown that a precise delineation of thematic roles is extremely difficult, if not impossible, to achieve and invariably depends on the granularity of the approach chosen.[1] While some theories only differentiate between the two basic roles PROTO-AGENT and PROTO-PATIENT (e.g., Dowty, 1989), other theories such as HPSG (e.g., Pollard and Sag, 1994) adopt a strongly lexicalised view of thematic roles, which results in a set of semantically highly differentiated, but largely verb-specific roles.[2] Despite the enormous spectrum of approaches to thematic roles, according to Löbner (2003) there appears to be a consensus on the definition of a few, central thematic roles such as AGENT, THEME/PATIENT, EXPERIENCER, INSTRUMENT, LOCATION, GOAL, and PATH. We define the set of thematic roles used in our work in Section 5.3.

Whether or not thematic roles also constitute a psycholinguistic and cognitive reality has been widely debated. Ferretti et al. (2001) conducted four single-word priming studies that provide substantial support for the hypothesis that access to verbs in the mental lexicon immediately makes available typical schema information centred around the verb. In their study Ferretti et al. tested whether the auditory presentation of verbs primed other AGENTs, PATIENTs, PATIENT features, INSTRUMENTs, or LOCATIONs that had previously been identified as typical for the stimulus verb. A priming effect of the verb-centred thematic roles would be expected if the entire situation information associated with the verb is made available immediately upon processing the verb prime.

Ferretti et al. indeed observed that priming occurred for typical fillers of the AGENT and PATIENT roles. A narrow range of typical INSTRUMENT fillers was also primed while LOCATION fillers were *not*. Associative relatedness was excluded as possible

[1] A good introduction into the different approaches to thematic roles can be found in Dowty (1989).

[2] The ability to abstract semantic relations beyond the level of single lexemes, which constituted one of the initial motivations for the introduction of thematic roles, tends to be lost in strongly lexicalised approaches.

cause for the observed priming effects. The authors conclude that thematic role information is indeed closely intertwined with the verb's definition of meaning in the mental lexicon and that activation of the verb also activates the verb's thematic roles.

Jackendoff's treatment of lexical semantics proceeds along the same lines (Jackendoff, 1990, pp. 45): A verb's entry in the mental lexicon contains argument slots, each of which carries specific requirements about the Conceptual Structure categories from which its filler candidates may be selected. Since human language permits the deliberate use of such selection criteria, e.g. in metaphoric or ironic usage, we conclude that this selection criteria are not absolute hard rules but simply express degrees of preference. We interpret this as a Conceptual-Structure equivalent of traditional sortal constraints and capture this notion as modelling requirement R23:

> **Requirement R23**
>
> *In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, a verb's lexical entry must indicate for each argument slot from which conceptual categories the argument fillers may preferably be selected.*

Jackendoff further argues that thematic roles are *"relational notions defined structurally over Conceptual Structure"*. In Jackendoff's view, a thematic role is nothing more than a specific label on a prominent semantic relation between an argument index in the verb's lexical Conceptual Structure representation and the corresponding argument slot. Jackendoff does not elaborate on how, mechanistically, the mapping between the semantic and the syntactic representations of the verb's argument slots is achieved. Since both are encoded on different levels of linguistic representation, we conclude that the mapping needs to be performed by the interface between those two levels of representation. The syntax-semantics interface must hence also contain correspondence rules that are capable of performing this specific mapping.
Kako (2006) reports consistent interpretations of nonsensical verbs and verbal arguments used in syntactically well-formed frames. These findings support the view that the mapping between syntactic structure and thematic roles is generic rather than lexically specific. This finding is also in line with our expectation that a separate mapping rule for every lexeme would be representationally highly inefficient. Irrespective of the actual mechanism via which the mapping is achieved in human language processing, we can add Requirement R24 for lexical representation:

> **Requirement R24**
>
> *In a model for the interaction between non-linguistic modalities and linguistic understanding, a verb's thematic roles must be relateable to its syntactic argument structure via correspondence rules in the syntax-semantics interface.*

Combining these verb-centred functional representations with the projections of role filling entities discussed in Section 3.4, we now have a sufficiently expressive inventory at hand to represent the semantic structure of propositions in Conceptual

$$\left[\text{KISS} \left(\left[{}^{\text{AGENT}}_{\text{HUMAN}}\,\text{BENNET} \right], \left[{}^{\text{THEME}}_{\text{HUMAN}}\,\text{LINNEA} \right] \right) \right]_{\text{EVENT}}$$

Figure 3.4: Conceptual Structure representation for the proposition *Bennet is kissing Linnea.*

Structure: They are complex function-argument representations in which thematic relations hold between an instance of a situation concept as lexicalised by a verb and instances of entity concepts that act as role fillers. An example for a representation of a simple proposition is given in Figure 3.4.

An assumption in Jackendoff (1990) with substantial cognitive and philosophical implication is that every concept instance represented in Conceptual Structure carries information about the conceptual category it instantiates, such as SITUATION, EVENT or HUMAN. This presupposes that the cognition of an entity has resulted in its categorisation as the member of a certain class prior to its projection into Conceptual Structure. We hence assume that every projected instance instantiates at least one concept from the concept hierarchy and add this as a modelling requirement for our Conceptual Structure representations:

Requirement R25

In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, every concept instance must instantiate at least one concept from the concept hierarchy.

Conceptual Semantics makes another, similarly fundamental assumption about the representation of propositional knowledge in Conceptual Structure, namely that propositional representations in Conceptual Structure are inherently verb-centric.[1] Since this view is compatible with our largely syntax-inspired view of semantic verb-argument structure, we include it as an additional modelling requirement for our model based on Conceptual Semantics.

Requirement R26

In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, Conceptual Structure representations are verb-centric.

[1] An in-depth discussion of this assumption is beyond the scope of this thesis. Since Jackendoff's Conceptual Semantics sets out to be a universal account of cognition, a rigorous, linguistically universal justification for this claim would be required. As far as we are aware, Jackendoff does not provide such an argument. For a more detailed description of the nature of Conceptual Structure representations, refer to Jackendoff (1983, Chapter 4).

3.6 Grounding

Conceptual Structure as proposed by Jackendoff is a symbolic encoding of semantic information. His theory of Conceptual Semantics clearly provides a computationalist approach to cognition. Computationalism maintains that cognition is the result of the manipulation of intrinsically meaningless symbols as exemplified by Fodor (1975): *"The mind is a symbol system and cognition is symbol manipulation."*
An opposing and largely incompatible view is provided by connectionism which holds that cognition does not arise from symbol manipulation but from dynamic patterns of activity in a multi-layered network of nodes with weighted interconnections. Activation patterns change according to the applied external stimuli and the internal network constraints. The primary appeal of connectionism lies in the superficial – and certainly not undisputed – parallelism between the structure of the artificial parallel processing networks and the structure of the human brain (Rogers and McClelland, 2004).

Today, it is widely accepted that the higher cognitive functions such as language and image understanding, spatio-temporal reasoning and mathematical thinking result from symbolic operations of the human mind.[1] However, systems that are symbolic at all levels of processing face the challenge of having to attribute intrinsic meaning to the symbols they manipulate. This is one of the fundamental challenges in the design of artificial autonomous cognitive systems — and for computational modelling in general: the *symbol grounding problem*, i.e., the challenge of relating intrinsically meaningless symbols manipulated solely on the basis of their arbitrary shapes to non-symbolic representations of their significata in the real world. Since the symbolic processes of cognition receive non-symbolic input from bottom-up sensory perception, there must be a stage in the process of cognition at which sensory — or in Harnad's words: "*iconic*" — representations connect with their symbolic correlates.

Harnad (1990) argues that in natural systems the grounding of symbols in sensory perception comprises two distinct capabilities that are executed independently of each other: *discrimination* and *identification*. Discrimination is the capability to discern whether the iconic representations of two sensory inputs differ and, if so, to what degree they differ. Discrimination, hence, is a relative judgement between two iconic representations. This capability does *not* require stimulus categorisation or interpretation and can therefore be performed at a purely sensory level prior to conceptual categorisation of the sensory stimulus. Identification, on the other hand, denotes the process of categorising a representatum in a non-symbolic representation as belonging to a given conceptual category. Identification according to Harnad is performed based on a non-symbolic *categorical representation* which is a reduction of the corresponding iconic representation down to the level of its invariant features. These invariant features permit to decide whether the represented entity is a member of a given category or not.

[1] Fodor and Pylyshyn (1988) and Harnad (1990) are prominent proponents in favour of this position.

Both tasks of discrimination and identification can be performed convincingly by a suitably trained connectionist system. What is missing in a purely connectionist system is the capability to transfer the systematic properties arising from category membership to the non-symbolic level of representation. A connectionist system may well be capable of associating a given sensor input in the form of an iconic representation with a conceptual category, e.g. HORSE.[1] A connectionist system, however, cannot assign the systematic properties of the category HORSE as inherited in a hierarchy of concepts to the iconic representation of an instantiation HORSE_01. In order to achieve such a rule-governed transfer of systematic properties, we require a symbolic representation of the conceptual category that permits rule-based operations of symbol manipulation. A purely connectionist system has no symbolic level of representation and hence offers no such symbolic operations.

This discussion of the fundamental capabilities and limitations of symbolic and connectionist systems shows that both connectionist and symbolic systems can contribute important aspects to the complex tasks of grounding and systematic symbol manipulation — but neither of them can perform discrimination, identification and symbolic manipulations all on its own.

Harnad proposes to resolve the strict dichotomy between symbolic and connectionist systems by means of a hybrid architecture that combines a symbolic and a connectionist component to complement each other such as to compensate for the individual weaknesses. The suggested framework first processes sensory input in a connectionist component such as an artificial neural network to discriminate and identify the sensory input. Systematic symbolic manipulation is then performed as a rule-based combination of grounded elementary categories. The result of these two steps is a systematically manipulable symbolic system grounded in sensory perception. Saffiotti and LeBlanc (2000), Coradeschi and Saffiotti (2001, 2003a,b), and Chella et al. (2004) report successful implementations that ground symbolic representations in the sensory representations of real-world objects based on Harnad-like hybrid architectures.

It is worth mentioning that, in the literature, the term *symbol grounding* is used primarily to denote the process of associating an abstract symbol with the sensory representation of a corresponding entity in the real world. The term *anchoring*, as formalised by Coradeschi and Saffiotti (2003a), is used in a similar fashion to denote the association of an abstract symbol with the representation of the corresponding real-world object over time. In principle, the association process between the real-world object's sensory representation and its symbolic representation can procede bottom-up, top-down or in a hybrid fashion. In this thesis, we follow Harnad (1990) in his view that *"there is really only one viable route from sense to symbols: from the ground up"*. We adopt Harnad's term *bottom-up grounding* to denote the process of grounding in which a sensory stimulus is linked to its corresponding symbolic representation by *bottom-up* processing of the sensory input.

[1] We henceforth adopt the convention of representing concepts in small capitals such EXAMPLE, and concept instances by their indexed category label such as EXAMPLE_01.

In summary, grounding representations of sensory perception requires two capabilities, namely discrimination and identification. Discrimination permits to evaluate the degree of similarity between two iconic sensory representations. Identification reduces iconic representations down to the corresponding categorical representations of distinctive and invariant features based on which the representatum can be categorised as a member of a particular conceptual category. Once identification has been accomplished, the sensory stimulus is said to ground the corresponding conceptual category. In terms of Conceptual Semantics, the stimulus can now project into Conceptual Structure as an instance of the identified conceptual category.
For our model of the cross-modal interaction between vision and language, we require the capability of discrimination. In the visual modality, discrimination allows us to distinguish between the iconic representations of visual perception. In the linguistic modality, discrimination permits us to distinguish between the different tokens in linguistic input. Our model also requires the capability to perform identification in order to achieve classification of entities in linguistic and visual input as belonging to one or more conceptual categories. We capture these aspects as modelling requirements R27 and R28:

Requirement R27

A model for the interaction between non-linguistic modalities and linguistic understanding must have the capability to discriminate individuating features of visual and linguistic input at sensory level.

Requirement R28

A model for the interaction between non-linguistic modalities and linguistic understanding must be capable of categorising sensory input in conceptual categories based on a set of individuating features.

With respect to our focus on the cross-modal interaction of representational modalities it is important to note the difference between assigning a meaning to tokens of natural language and the bottom-up grounding of a sensory stimulus representation. For language, the categorisation of the initial sensory stimulus — be it auditory, visual or haptic — results in the identification of a particular word. This word is an arbitrary linguistic symbol and has a range of associated lexical properties, one of them being the representation of word meaning as defined in the mental lexicon. According to Jackendoff, word meaning is represented in terms of semantic structures of concepts and predicates in Conceptual Structure. In contrast to the processing of a sensory stimulus, an identified word does not project into Conceptual Structure directly. Rather, the meaning of the word first needs to be retrieved as a property associated with the identified symbol. As Löbner (2003, Chapter 2) describes it, the retrieved word meaning is a conceptual expression that denotes a set of possible instantiations. It it this conceptual expression that is instantiated in Conceptual Structure. The processing of the representational modality hence includes an additional decoding step in which the meaning of the arbitrary symbol is decoded. In essence, however, both processes result in the assignment of meaning to a sensory

$$\begin{array}{c}\text{Categorical Representation}\\ \text{of a Sensory Stimulus}\end{array} \xrightarrow[\textit{Matching}]{\textit{Feature}} \begin{array}{c}\text{Identification of}\\ \text{Category Instance}\end{array}$$

$$\begin{array}{c}\text{Categorical Representation}\\ \text{of a Linguistic Stimulus}\end{array} \xrightarrow[\textit{Matching}]{\textit{Feature}} \begin{array}{c}\text{Identification of}\\ \text{Word Instance}\end{array} \xrightarrow[\textit{Meaning}]{\textit{Access to}} \begin{array}{c}\text{Identification of}\\ \text{Denoted Category}\end{array}$$

Figure 3.5: The difference between grounding concepts in sensory and linguistic stimuli.

stimulus. We therefore will refer to the process of assigning meaning to a sensory stimulus encoding linguistic symbols as *linguistic grounding* while we refer to the process of grounding in sensory stimuli as *sensory grounding*. The difference between processing a categorical representation in sensory grounding and in linguistic grounding is summarised in Figure 3.5.

3.7 Cross-Modal Matching

In the previous section we have described how sensory and linguistic input projects into a common level of mental representation: Conceptual Structure. In order for those representations to interact with each other, we require an additional step which Bushnell (1994) refers to as *cross-modal matching* or *cross-modal transfer*. In this step, the various bits of information obtained from different modalities are interconnected by establishing co-reference. Propositional information on instances from a non-linguistic modality can interact with propositional information on instances from language if the instances projected from the non-linguistic modality can be matched to one or more instances projected from the linguistic modality. Establishing referential coherence between individual propositions has been shown to be an integral part of unimodal language understanding (Kintsch and van Dijk, 1978).[1] We hypothesise that referential coherence is also established in the process of cross-modal comprehension, not only between concept instances from the linguistic modality but between *all* concept instances projected into Conceptual Structure. The modality from which an instance has projected should have no influence upon that process. Recent investigations into cross-modal graph comprehension support this view (Acartürk et al., 2008; Habel and Acartürk, 2009).
The simplest case of a cross-modal matching is token identity in which both modalities independently project the same concept instance into Conceptual Structure, e.g., when watching a visual scene and simultaneously receiving a linguistic description of a situation in that scene. Parsing the utterance *'A green car is passing by.'* while watching a green car drive past, will typically cause a human listener and speaker of English to interpret the utterance as making reference to precisely that green car

[1] Kintsch and van Dijk argue that language understanding involves the establishment of referential coherence at two levels of discourse: at the *microlevel* within the utterance being processed and at the *macrolevel* within the larger discourse unit that was encountered prior to the currently processed utterance. Due to the focus of this work on single sentence parsing we shall ignore the higher level discourse structures in this discussion.

in his or her visual field. As a result, the utterance is perceived as referring to the car that is being seen, heard and smelt at the time of utterance interpretation. To achieve this uniform and integrated percept of the situation, concept instances projected from non-linguistic modalities and language need to be projected, recognised as mutually compatible and subsequently treated as co-indexical.[1]

The cognitive benefit of cross-modal matching is an integration of initially disparate information arriving from different modalities. The consolidated information can then be linked to the same entity in the subject's perception of the real world resulting in a uniform cross-modal percept of that entity. The precise mechanisms of how the integration of information from different modalities is achieved in the human brain are still unknown and continue to attract intense research attention. While significant advances have been made in the neurophysiological investigation of multisensory integration, i.e., information fusion at sensory level (cf. Section 2.1), comparatively little is known about how information from different representational modalities is integrated by the higher cognitive functions.

An important factor in the integration of cross-modal information is the compatibility of the information obtained from the different modalities. In cases where two modalities provide compatible information, an accumulation mechanism appears plausible in which the information from both modalities is simply combined into a joined representation in Conceptual Structure. Since different modalities may exert a different degree of dominance, e.g., due to differing degrees of perceptual certainty, it is likely that the contributions of the individual modalities are weighted. Vision, e.g., has been known to dominate strongly over other modalities in humans. Congruence of information between different modalities is likely to result in a reinforcement of the representation of that information. The reinforcement of a representation by different modalities should increase the confidence in the reliability of its representata. If the mechanisms for cross-modal integration with language are analogous to those observed at sensory level[2], the independent projection of identical information from different modalities should result in superadditive responses of neural activity for the processing of cross-modally congruent representational stimuli.

Given the large amount of different – and oftentimes conflicting – stimuli we are exposed to, the processing of conflicting information across different modalities will also play an important role in our modelling. Whether or not cross-modal information is consistent and coherent can only be identified when the representations arising from different modalities are brought together in the process of cross-modal matching. We hence formulate the need for cross-modal matching as an essential

[1] Several effects of representational consolidation upon multiple tokens representing the same real-world are conceivable: (1) All tokens continue to co-exist in Conceptual Structure joined by an additional relation marking their co-indexation. (2) Alternatively, all but one of the multiple tokens are removed and the relations they engage in are redirected to the last remaining of tokens. (3) Multiple tokens are replaced by a common index which points to a single instance of that token. A discussion of how these options could be expressed in terms of Conceptual Structure representations is beyond the scope of this work. Jackendoff (1990, Chapter 3.2) provides some examples of co-indexation in Conceptual Structure representations.

[2] See the discussion of multisensory integration in Section 2.1.

part of a cross-modal interaction as modelling requirements R29 and R30.

> **Requirement R29**
> *A model for the interaction between non-linguistic modalities and linguistic understanding needs to provide a mechanism for establishing cross-modal referential links by matching entities from the linguistic modality with concept instances from the interacting non-linguistic modalities.*

> **Requirement R30**
> *A model for the interaction between non-linguistic modalities and linguistic understanding needs to provide a mechanism for establishing cross-modal referential links by matching concept instances from the non-linguistic modalities with entities from the linguistic modality.*

Bushnell (1994) argues for two different strategies to accomplish cross-modal matching. *Matching by analysis* is the approach in which salient object features detected in the first modality are tested for their correspondence in the second modality. This approach predominates for unknown items. *Matching by recognition* is an alternative approach when object perception from one modality is strong enough as to project into the second modality, e.g., when the haptic impression of an object presented under a blanket is strong enough to conjure up an image of a potato in front of the subject's inner eye. The subject then needs to compare his or her internal image against the actual visual perception of the visually presented object to judge whether they coincide. This latter strategy is employed when objects are very familiar. According to Bushnell, matching by recognition is analogous to word recognition in which an auditory linguistic stimulus activates a concept in another representational modality. The analytic and holistic methods of visual pattern perception as described by Cooper (1976) and other researchers can also be considered instances of matching by recognition.

3.8 Chapter Summary

In this chapter we have achieved a significant objective in the preparation of the detailed description of our model: we have identified Conceptual Semantics as a general theory of cognition which provides an integrated account for the cross-modal interaction between non-linguistic modalities and language. Based on the description of Conceptual Semantics we have formulated further requirements for our model.

Conceptual Semantics is based on a cognitive architecture in which non-linguistic modalities and language interact at a representational level. The interaction of non-linguistic modalities with syntax is mediated by Conceptual Structure, a single and uniform level of mental representation that encodes both conceptual knowledge and linguistic semantics. Conceptual Structure interfaces with the representations of non-linguistic modalities as well as with syntax. All of these representations are

representationally encapsulated and project into Conceptual Structure via modality-specific interfaces. The role of these interfaces is to translate between the encodings of the different representational levels by applying a finite set of correspondence rules.
We have further outlined the view of Conceptual Semantics on the significance of thematic roles. According to Jackendoff, thematic roles mark prominent argument slots in the Conceptual Structure representation of verbal concepts. Our discussion of grounding has led us to the insight that discrimination and identification are important tasks to achieve conceptual grounding in our model. Finally, cross-modal matching was introduced as the process by which cross-modal referential links between concept instances are established between concept instances from different modalities. As such, it constitutes an indispensable cognitive process for cross-modal interaction. We have argued that the compatibility of the concepts instantiated in different modalities is a key requirement for establishing cross-modal co-reference.

In the following chapter we shift our focus to language processing. We present symbolic constraint-based parsing as a suitable formalism for the integration of non-linguistic contextual constraints upon syntactic parsing and motivate an existing parser implementation as a suitable candidate for the natural language processing component in our model.

Chapter 4

Constraint-Based Analysis of Natural Language with WCDG

A model for the influence of cross-modal context upon syntactic parsing requires a parser that is capable of receiving and processing external context information in some form or another. The majority of syntax parsers today, however, are informationally encapsulated in the sense that they only accept linguistic input which they process based on their intrinsic linguistic resources. Those parsers that do permit to impose additional constraints on linguistic analysis typically employ unfication such that the additional constraints are added as hard constraints on linguistic analysis rather than as biasing preferences. The weighted-constraint dependency parser WCDG constitutes a notable exception in this respect. It comes with a generic interface that permits the inclusion of parser-external non-linguistic information into linguistic processing. WCDG is an attractive candidate for the parsing component in our model because its interface permits to influence linguistic decision making by introducing external, possibly non-linguistic information into the parsing process. WCDG is also based on weighted or graded constraints which, as we shall see, are highly suited for modelling linguistic and contextual preferences.

This chapter provides an introduction to WCDG and its approach to the analysis of natural language as a symbolic constraint satisfaction problem. While the preceding chapters focused on the development of the cognitive requirements for our model, this chapter sets out to identify further, more implementation-related requirements. The primary focus in this chapter is on the derivation of the technical requirements for the parser component in the context of our modelling framework.

Section 4.1 begins with an outline of the major differences between generation-rule-based and weighted-constraint parsers to motivate the use of WCDG in our model. Section 4.2 describes WCDG's relevant standard capabilities. Section 4.3 offers a discussion of why some of the central limitations of WCDG's standard implementation necessitate modifications to the implementation in order to meet our specific modelling objectives. Section 4.4 summarises the central points in this chapter and lists the resulting conclusions. This chapter concludes Part I of this thesis, and with it, the requirements collection process for our computational model.

4.1 Generation Rules vs. Constraints

Approaches to natural language analysis can be broadly categorised into two fundamentally different classes, depending on their method for defining the set of acceptable solution structures: generation-rule-based approaches and constraint-based approaches. The majority of the existing parser implementations follow a generation-rule-based approach. Generation-rule-based systems span open the space of well-formed sentences based on a set of generation rules.
Constraint-based systems, on the other hand, constrain the set of all possible structures by excluding ungrammatical structures, leaving only the set of desired solutions. Importantly, therefore, the set of constraint-based systems not only comprises connectionist, i.e. non-symbolic, approaches but also symbolic constraint parsers. Symbolic constraint parsers encode syntactic properties in variables and constrain the assignment of values to these variables by means of suitable constraints. In the following, we mean *symbolic constraint parsers* when we refer to 'constraint-based systems'.

4.1.1 Generation-Rule-Based Parsers

A generation-rule-based parser tries to assess whether a given input can be generated from a set of generation rules that stipulate the procedures for generating well-formed sentences. The result of the parser's analysis is a Boolean decision on the grammaticality of the input with respect to the set of generation rules. If the input is classified as grammatical, the input's syntactic structure resulting from the successful procedural application the generation rules is known as well.
Effectively, a generation-rule-based parser acts as a theorem prover attempting to prove if the input theorem can be derived from a set of axioms stated in the formal system constituted by its grammar rules and the additional information in the lexicon. If the input is generable from the rules, it is rated as grammatical, otherwise as ungrammatical.
In contrast with this binary decision on grammaticality, the human analysis of natural language also comprises the central ability to discern *preferences* of acceptability, be they syntactic, semantic or pragmatic. It is this capability that lets humans accept a given construct as perfectly grammatical in one context while rejecting it as ungrammatical in another context (Crain and Steedman, 1985).

A natural language analysis application designed with the intent to model human language processing behaviour should therefore include the capability to discern degrees of acceptability rather than just to categorise solution candidates as correct or incorrect. In a generation-rule-based system, however, the inability to derive a given input from the grammar and the lexicon cannot always be attributed to the violation of a specific grammatical axiom; the input simply cannot be deduced from the formal system constituted by the grammar and the lexicon as a whole. Consequently, a generation-rule-based system cannot provide detailed diagnostic information on specifically which property of the input was responsible for its classification as ungrammatical.

Another limitation of the generation-rule-based approach is its handling of unknown input. Even the largest of today's grammars and lexicons are inevitably limited in their modelling scope and hence do not cover the totality of natural language expressiveness. To a generation-rule-based parser *'outside of modelling scope'* is equivalent to *'ungrammatical'*. However, not every input that cannot be generated by the formal system must necessarily be ungrammatical; unrestricted natural language abounds with multi-word expressions, metaphors, creative word or expression formations whose underlying formation patterns are not always easy to predict. In rejecting input beyond the boundaries of the known as ungrammatical, generation-rule-based parsers are limited in their capability of handling unknown input robustly. Given the high productivity of natural language, the constructive handling unknown input is a key feature for the robust processing of unrestricted natural language input.

4.1.2 Symbolic Constraint-Based Parsers

Symbolic constraint-based systems approach the task of parsing as a constraint-satisfaction problem over the assignment of values to variables representing syntactic properties. The degree of complexity of the represented features depends on the formalism. In the case of WCDG, the words in the input sentence form the nodes of a constraint net whose edges correspond to the dependencies between words. Every node and every edge corresponds to a variable in the constraint system to which a value needs to be assigned. Well-formedness rules define the permissible relations between words hence act as constraints upon the values that can populate the edges in the constraint net. Edge values violating constraints are removed from the constraint net until no further restrictions can be imposed. The remaining edge values in the constraint net describe the set of structures classified as grammatical with respect to the constraint set. This approach was first described by Maruyama (1990).

In analogy to the ancient Roman legal guideline *Nulla pœna sine lege*[1] a constraint-based parser will admit every solution candidate as correct unless it violates a well-formedness rule in the grammar. The set of constraints therefore needs to be specific enough such as to admit only grammatical sentences and general enough such as not to exclude acceptable structures from the solution set. A major advantage of this approach is the system's robustness to unknown input. Every structure, including those which have not been considered by the grammar-writer, can pass as an acceptable solution as long as it does not violate a given structural constraint.
A significant difference compared with generation-rule-based systems for language analysis is that constraint-based parsers can also provide very specific feedback on which constraints in its grammar are being violated by a given input structure. Because of this, constraint-based systems are good candidates for providing diagnostic support in language analysis.

[1] 'No penalty without a [corresponding] law'

Finally, the constraints defined in the grammar all apply to a solution candidate simultaneously rather than sequentially. This aspect makes the evaluation of constraint satisfaction in constraint-based systems amenable to parallel processing. Harper and Helzermann (1995, p. 199) review a number of implementation efforts aimed at parallelising constraint-dependency parsing.

An important refinement to the constraint-based approach outlined so far is motivated by the insight that the well-formedness rules do not all contribute equally to the acceptability of the overall solution structure. The constraint-based systems described so far cannot yet express degrees of preference amongst solution candidates. Graded acceptability assessments can be incorporated by expressing the severity of a violated well-formedness rule as a numerical weight. In case of a constraint violation, rather than removing the structural candidate from the set of acceptable solutions altogether, we can retain the candidate structure as a potential solution and assign it a penalty score for each constraint that it violates. As an example, a sentence may well contain a determiner-noun incongruence and still be acceptable overall while the absence of a full verb may result in a much more severe degradation of its grammatical acceptability. A weighted constraint formalism is capable of expressing such graded acceptability ratings.

Weighted constraint-based parsers typically define a measure for the overall acceptability of a solution candidate as a function of the constraint-violation penalties it incurs. This overall measure allows the system to rank solutions and compare their acceptability against each other. The most preferred solution is the one with the best overall acceptability rating. To identify the most preferable, i.e. the least penalised, solution candidate in the potentially very large search space, we require a search algorithm that provides complete coverage of the search space. In case a complete search is infeasible due to the sheer size of the search space, we need to employ an efficient search heuristic to identify a local optimum as our preferred solution candidate.

4.2 The WCDG Parser

The weighted-constraint dependency parser WCDG is an implementation of the WCDG formalism and can be obtained in its standard release from the WCDG Download (2009). With an overall syntactic parsing accuracy of 92.5%, WCDG is the system with the highest reported accuracy for unrestricted German text to date (Foth, 2006).[1] By integrating the powerful MSTParser (McDonald, 2006; McDonald et al., 2006) as a predictor to WCDG, Khmylko (2007) even achieved a parsing accuracy of 93.9% under conditions comparable with those used by Foth.

Apart from its constraint-based processing WCDG also offers a generic interface for the integration of parser-external, non-linguistic information into the parsing process (cf. Section 4.2.5). For this reason WCDG was chosen as a suitable language processing component in our model of cross-modal integration into language. As will become apparent in the course of this chapter, a number of processing require-

[1] The evaluations of parsing results for WCDG in the literature cite accuracies. Unless otherwise stated, these are identical to the standard measures of precision, recall and the resulting f_1-measure.

```
<Lexicon Entry>            ::=    <Lexeme> ':=[' <Feature List > '];'
<Lexeme>                   ::=    { STRING | INTEGER }
<Feature List>             ::=    <Attribute-Value Pair> { ',' <Attribute-Value Pair> }
<Attribute-Value Pair>     ::=    <Attribute> ':' <Value>
<Attribute>                ::=    { STRING | INTEGER }
<Value>                    ::=    { STRING | INTEGER }
```

Figure 4.1: The form of a WCDG1 lexicon entry (EBNF).

ments necessitate implementation extensions to WCDG. In our model we therefore employ a selectively enhanced version of WCDG. We refer to the standard version of WCDG as *WCDG1* and to our extended version as *WCDG2*. By *WCDG* we henceforth refer to the common core of these implementations, i.e. the WCDG system in general, irrespective of its implementation version.

The following sections provide a brief overview over WCDG1's components and motivate the extensions realised in WCDG2. For a comprehensive description of WCDG1, see Schröder (2002).

4.2.1 Lexicon

WCDG uses a semi-automatically generated full-form lexicon with approximately 931,000 individual entries for WCDG1.[1] A lexicon entry consists of an assignment of comma-separated attribute-value pairs to a lexical full-form as shown in Figure 4.1. The set of attributes comprises part of speech (`cat`)[2], lexical base form (`base`), syntactic valence (`valence`), morphosyntactic features (`case`, `number`, `person` etc.) and others.
An example for a typical lexicon entry is shown in Figure 4.2. For economy of representation and processing, WCDG permits to assign underspecified feature values. `case = bot`, e.g., is an underspecified `case` feature that corresponds to the assignment of any possible case. Assigning the underspecified feature has the advantage of needing to represent and process only one single form rather than four individual forms as would be the case with the assignment of `case = (nom | gen | dat | acc)`.

```
kaufen:=[base:kaufen,cat:VVINF,perfect:haben,valence:'a?+d?',avz:allowed];
```

Figure 4.2: WCDG1's lexicon entry for 'kaufen'/`VVINF`

[1] See Foth (2006) for a detailed explanation of the generation of the lexicon files.

[2] For the specification of the `cat` feature, WCDG uses the standard tags from the Stuttgart-Tübingen part of speech tag set (STTS) for German as documented in STTS Tag Set (2009). We follow this practice and use this tag set in our notation of part of speech throughout this document.

Lexical Entry	`kaufen:=[base:kaufen,cat:VVINF,perfect:haben,`
	`valence:'a?+d?',avz:allowed];`

Actual Valence 1	`valence = -`
Actual Valence 2	`valence = a`
Actual Valence 3	`valence = a+d`
WCDG1 Valence	`valence = a?+d?`
Valence Expansion	`valence = ( - | a | d | a+d )`
Overgenerated Valence	`valence = d`

Figure 4.3: Syntactic valence expansion for the verb 'kaufen' *to buy* as an example for the overgeneration of syntactic valence alternatives in WCDG1.

Of particular importance to a verb's syntactic behaviour is its feature `valence` which specifies the range of syntactic dependencies that the verb may entertain as a regent. Consider the infinitive of the verb 'kaufen' *to buy* which can act as an intransitive, a transitive or a ditransitive verb and hence has several syntactic valences. While the different syntactic behaviours of the infinitives would warrant separate entries in the full-form lexicon, the verb's WCDG1-representation contracts all three verb forms into a single lexical entry with the underspecified syntactic valence `a?+d?`. This syntactic valence indicates that 'kaufen' must subcategorise either an optional accusative/direct object (`a?`) or an optional dative/indirect object (`d?`) or a combination of these (`+`).
Since the syntax for specifying syntactic valences in WCDG is limited in its power to express optionality, expanding underspecified syntactic valences may lead to overgeneration of valence alternatives (Foth, 2006, p. 172). An example for an invalid syntactic valence option generated from an underspecified valence representation is given in Figure 4.3. WCDG1's lexical entries do not contain semantic valence information such as subcategorised thematic roles. For semantic text processing with WCDG, semantic lexical information hence needs to be added to the lexicon. We henceforth use the term *semantic valence* to denote the set of thematic relations a verb can engage in. For reasons to be outlined in Section 5.7, we limit our use of the term to the mandatory thematic relations a verb needs to entertain in order to be semantically well-formed.

4.2.2 Grammar

WCDG's grammar primarily contains a collection of the constraints defined on the space of permissible parse structures. Each constraint c is weighted by a penalty score $\phi(c)$ where ϕ is a function

$$\phi : C \mapsto [0,1] \tag{4.1}$$

that assigns each constraint from the set of all constraints C its penalty.

When a solution candidate SC violates a given constraint c_i, the resulting constraint penalty $\phi(c_i)$ is aggregated into the candidate's overall score $\Phi(SC)$. In WCDG, the aggregation function Φ is multiplicative and a candidate's overall score is defined as the product of the penalties associated with the constraints it violates.[1]

$$(4.2) \qquad \Phi(SC) = \prod_i \phi(c_i)$$

The violation of a constraint with a penalty score of *0* results in the rejection of the corresponding solution candidate as unacceptable. Constraints with a penalty score of *0* are referred to as *hard constraints*. All other constraints can, in principle, be violated by an acceptable solution candidate and are referred to as *soft constraints*. Observe that the violation of constraints c_i with weight c_i = *1.0* has no effect upon the overall score of the solution candidate $\Phi(SC)$. Such constraints hence effectively do not constrain the space of acceptable solutions and should, strictly speaking, not be referred to as *constraints*. For simplicity, however, we will use the term *constraint* to denote every well-formedness rule in the grammar, irrespective of its numerical penalty score.

For each input sentence WCDG creates a directed acyclic graph of labelled dependency edges joining each dependant with its regent. With the exception of the `ROOT` node, every node in a WCDG dependency graph represents a word in the input sentence. Dependencies are assigned between nodes. In contrast to phrase-structure formalisms, dependency parsing requires that higher syntactic structures such as prepositional phrases be modelled by assigning the corresponding dependency to a particular node. With the exception of the `S`-node a dependency tree hence has no internal nodes that do not map to a word in the input sentence. This property constitutes the most significant representational difference between dependency and phrase-structure grammars.
Each dependant can only have one regent per level of analysis. For some aspects of linguistic analysis such as syntax, relative clause reference or semantics, it can be helpful to build up separate dependency structures for the same input sentence. WCDG achieves this by defining a separate level of analysis for each one of these structures. Attachment rules are defined individually for each level. In a constraint definition we therefore need to declare explicitly which level of analysis the constraint body refers to. WCDG1 provides two levels of analysis, the `SYN` level for the assignment of syntactic dependencies and the `REF` level for reference resolution of relative pronouns. WCDG1 does *not* provide any levels – and thus no constraints – for semantic analysis.

Constraints on the properties of a single edge are referred to as *unary* while constraints on the properties of two edges are referred to as *binary*. As will be discussed further in Section 5.6, the edge properties checked for in a constraint may indeed be complex. An example for a more complex edge property is its adjacency to other

[1] For a comprehensive and rigorous description of the WCDG formalism cf. Schröder (2002).

edges with specific properties. The constraint that an edge `X` be above another edge with label `PP`, say, is still considered a unary constraint — even if the reach of the constraint has effectively been extended and the properties of two edges need to be evaluated in order to assess the satisfaction of this constraint. In the view of WCDG, the adjacency condition is still considered a property of edge `X`. Observe that this view also effects a *shift of blame* for constraint violations. Since the constraint is formulated as a property of `X`, `X` is also responsible for violating the above adjacency constraint, even if the violation occurs because of an edge below `X` that does not bear the required `PP` label.

In first approximation, solving a constraint satisfaction problem in WCDG breaks down into three major steps, namely:

1. the application of unary constraints to individual dependency edges,
2. the application of binary constraints to larger dependency structures comprising more than one edge,
3. the search for the optimal solution.

Let us attempt to formulate an expression to describe the time complexity of this parsing problem. As we have just seen, dividing constraint evaluation into the application of unary and binary constraints is a simplification (cf. Section 4.2.4). Certain unary constraints also need to be evaluated on dependencies that extend over more than one edge and – analogously to binary constraints – need to be applied to larger dependency structures. Such constraints are referred to as *context-sensitive*. For the purpose of this derivation, however, we neglect the differentiation between context-sensitive and non-context-sensitive constraints. We make the simplifying assumption that every n-ary constraint contributes equally to the time complexity of n-ary constraint evaluation, regardless of whether it is context-sensitive or not.
The time complexity of unary constraint evaluation then depends on the number of constraint evaluations that need to be performed, which is equal to the product of the number of constraints and the total number of edge constellations to be evaluated for a given sentence. An upper bound for the number of unary constraint evaluations, $\mathscr{N}_{unary}$, is given by Equation (4.3). A brief derivation of Equation (4.3) is given in Appendix III.1.

$$\mathscr{N}_{unary} = |C_{unary}| \cdot n_{max}^2 \cdot s^2 \cdot \sum_i \lambda_i \tag{4.3}$$

where

C_{unary} is the set of all unary constraints in the grammar,
n_{max} is the maximum number of homonyms per slot in the sentence,
s is the number of slots in the sentence, and
λ_i is the number of dependency labels on level of analysis i.

For binary constraint evaluation, constraint application is typically combined with search such as to avoid computation of the entire hypothesis space $\mathscr{H}$. With this approach, binary constraints are not applied to solution structures that lie in pruned sections of $\mathscr{H}$. For reasons of processing efficiency – especially with longer sentences as are typically encountered in unrestricted natural language input – no WCDG search heuristic in practical use evaluates binary constraints for the entire hypothesis space. In cases where binary constraint evaluation and search are combined, time complexity for the combined step depends on the size of $\mathscr{H}$ and the algorithm employed to search it. Due to the strong dependency on the efficiency of search, no implementation-independent prediction of time complexity is possible in this case. If all binary constraints were to be evaluated prior to the search on $\mathscr{H}$, the time complexity of binary constraint evaluation would be proportional to the number of constraint evaluations. The upper bound for the number of binary constraint evaluations, $\mathscr{N}_{binary}$, is given by Equation (4.4). For a brief derivation of Equation (4.4) see Appendix III.2.

$$\mathscr{N}_{binary} = |C_{binary}| \cdot n_{max}^4 \cdot s^4 \cdot \left[\sum_i \lambda_i \right]^2 \tag{4.4}$$

In practice, the actual number of binary constraints evaluated is less than $\mathscr{N}_{binary}$ since WCDG does not evaluate binary constraints in solution candidates that contain edges which violate a hard unary constraint. The number of solution candidates violating one or more hard unary constraints, however, depends on the actual formulation of the unary constraints and thus cannot be predicted by an implementation-independent expression.

In summary we conclude that in WCDG the number of constraints in the grammar, the number of homonyms in each slot, the number of slots in the sentence and the number of labels in the grammar have a bearing on the number of constraint evaluations and thus on processing time. An understanding of the factors that influence the number of constraint evaluations will become important when trying to assess the effect of implementation changes between WCDG1 an WCDG2 upon processing time (cf. Section 5.7).

After this brief excursus into the computational complexity of the constraint satisfaction problem, let us return to the overview over WCDG's grammar. In addition to the constraint and level definitions outlined so far, the grammar also may contain some further information most of which, however, is of lesser importance in the context of modelling cross-modal interaction:

Pragmas

Pragmas permit to define macros on the basis of WCDG commands. A pragma is executed when WCDG loads the file containing the pragma into memory. The main use of pragmas is to set parameters for the current WCDG session.

Hierarchies

Named hierarchies provide a simple way of defining subsumption relations between hierarchically structured symbols in WCDG.

Word Templates

To improve WCDG's robustness to unknown input the grammar contains templates describing typical word-formation patterns in a regular expression syntax. These patterns are effectively functions that assign a set of lexical features to any matching string. They are extremely useful for recognising systematically generated words which are either too specific or too infrequent to warrant their inclusion as a full-form entry into the lexicon. One of these templates, e.g., recognises all words of the family *n-fold* such as 'zwölffach' *twelvefold.*

Datamaps

WCDG provides a limited number of functions and predicates most of which take a fixed number of input arguments. Datamaps permit to map an arbitrary number of input arguments n to a single return value. They have been employed to model subsumption hierarchies of lexical and syntactic features and as a work-around to extend the number of input arguments in WCDG functions and predicates. By including the return value of a datamap as an argument into an existing function, the number of input arguments of a function can formally be maintained while, effectively, extending its number of input arguments by *n-1.* In terms of data structures, a datamap can be considered a hash table which takes the concatenation of the input argument strings as key and returns a hash value.

4.2.3 Constraint Syntax

A WCDG constraint defines a well-formedness rule for a dependency structure or a part of it. A constraint consists of a *constraint header* and an all-quantified *logical formula* (Schulz et al., 2003, pp. 43). The logical formula is evaluated for every dependency structure that satisfies the restrictions formulated in the constraint header. A penalty score is imposed on the overall score of the structure if the truth value of the logical formula evaluates to `false`.

The majority of constraints in WCDG's standard grammar of German contain a logical formula which consists of a precondition and a postcondition joined by logical implication written as '`->`'. Constraint satisfaction can be achieved in two ways: Either the dependency structure satisfies both the precondition and the postcondition or the precondition evaluates to `false`.[1] An example of a typical WCDG constraint containing pre- and postcondition joined by logical implication is given in Figure 4.4.

```
{X!SYN} : 'PP-attachment' : stat : [ predict(X@id, PP, X^from) ] :
X.label = PP | X.label = KOM
   ->  (X@word = um | X@word = als) & X^degree = comparative
       | predict(X@id, PP, X^from) = 1;
```

Figure 4.4: A WCDG constraint with constraint header, precondition, implication, postcondition and dynamic constraint weighting.

The constraint header defines:

1. arbitrary variables to denote the edge references in the constraint's logical formula (`X` in Figure 4.4)
2. optionally, a restriction of the level of analysis on which each of the referenced edges is to be evaluated (`SYN` in Figure 4.4)
3. optionally, structural restrictions for the referenced edges. '`!`' in Figure 4.4 indicates that `X` must not attach to the `ROOT` node.
4. a unique constraint name (`'PP-attachment'` in Figure 4.4)
5. an optional indication of the constraint group to permit selective activation or deactivation of entire constraint groups (in Figure 4.4: `stat`)
6. an optional constraint weight. Constraint weights can either be declared by a static number from the closed interval between 0 and 1 or by a dynamically evaluated function. Missing weight declarations default to 0. In the example in Figure 4.4, the constraint weight is given by the expression `[ predict(X@id, PP, X^from) ]` which evaluates dynamically to the return value of the `predict()` function with the given input parameters.

The precondition in Figure 4.4 is satisfied by any edge `X` whose label is either `PP` or `KOM`.[2] The postcondition is met if the lower word of `X` is either 'um' or 'als' and the upper word of `X` is a comparative form or, alternatively, if the prediction value of the PP-attachment predictor for `X` is `1`. A comprehensive coverage of WCDG's constraint syntax is provided in Schulz et al. (2003).

[1] According to the rules of formal logic, the evaluation of an implication's precondition to `false` results in the evaluation of the implication to `true` overall (*ex falso sequitur quod libet*).

[2] The complete set of WCDG edge labels is documented in Foth (2006).

In its present form of implementation, WCDG only supports unary and binary constraints. As a result of this, a single constraint cannot relate more than two different dependency edges with each other. These limitations become particularly restrictive in the definition of constraints integrating features across levels of analysis. Syntactic extensions to constraint expressivity based on the use of additional predicates and ancillary constraints without modifications to WCDG's processing algorithms have been reported in McCrae et al. (2008). Baumgärtner (2009) reports an extension of WCDG's processing algorithms to include ternary constraints. A discussion of how expressivity challenges resulting from the limitation to unary and binary constraints have been overcome in our model is given in Section 5.6.

As an aside it might be added that a significant challenge for the WCDG grammar author is provided by the fact that – at the time of writing – there is no rigorous or consistent approach to assigning constraint weights systematically. For grammars complex enough to handle unrestricted German language input, the complex interactions between different constraints and constraint weights are virtually impossible to predict for a human grammar writer.
In a first attempt to address this issue, Schröder et al. (2001, 2002) report experiments for learning constraint weights using genetic algorithms. The approach was based on a toy grammar and a relatively small training corpus of 220 Verbmobil sentences. After mutations over a few hundred generations, small improvements in f-measure and processing time were observed for grammars with hand-selected starting values for the constraint weights. Starting with grammars containing randomly assigned constraint weights failed to produce better results than manual constraint weight selection. The approach was not pursued further due to the significant computational effort involved in the optimisation process.
Apart from this machine learning approach, no heuristics or formal approaches to assist weighted constraint grammar writers is available and manual constraint weight assignment largely and regrettably remains a matter of linguistic intuition and trial-and-error.

4.2.4 Processing Fundamentals

Knowing how constraint-based parsing has been implemented in WCDG will provide a better understanding of some of the technical limitations encountered when trying to employ WCDG for the task of integrating external, non-linguistic context. Parsing in WCDG is a multi-step process which can be decomposed into *Pre-Processing*, *Predictor Integration*, *Unary Constraint Evaluation* and *Search*. The application of the binary and context-sensitive constraints is performed in the context of Search.[1]

Pre-Processing starts with the input of the sentence to parse. A tokenizer and the T'n'T part-of-speech (POS) tagger are applied to the input to obtain individual tokens and their most probable POS tag. Each token occupies a position or *slot* in the sentence and will subsequently be referred to as *slot string*. WCDG then retrieves

[1] A description of *all* steps, which is more detailed than required for the argument put forward here, is given in Menzel and Schröder (1998).

the lexical entries that match each slot string. Due to lexical ambiguity each slot string may map to a number of unique lexical entries. These unique lexical entries will subsequently be referred to as *homonyms* of the slot string.
WCDG submits the list of slot strings together with their most probable part of speech tag according to T'n'T to all registered external predictors. Each predictor processes the input and returns its prediction results via WCDG's predictor interface. The prediction results are then accessible in the parsing process via the use of the `predict()` function in WCDG's grammar. A more detailed description of predictor integration is given in Section 4.2.5 below.

With the Unary Constraint Evaluation, WCDG starts the actual parsing process by applying the non-context-sensitive unary constraints to every single edge. Non-context-sensitive constraints are those which can be evaluated without the use of adjacency conditions, i.e.: without the use of the WCDG-predicates `is()` or `has()` which access the properties of the edge or edges above or below. Evaluated edges receive a score equal to the product of the penalties associated with the constraints they violate (see Equation (4.2)). Valid structural solutions for the parsing problem must not contain edges that violate a hard constraint. We can therefore exclude any edges that do violate a hard constraint from the set of potential constituents of a valid solution at this early stage of the process.
Having scored all edges that connect word pairs by applying the non-context-sensitive unary constraints, WCDG now needs to integrate these fragments into dependency structures spanning the entire input sentence. Only when these larger structures have been assembled, binary and context-sensitive constraints can be applied and the best, i.e. least penalised, parse structure overall will be searched for.
WCDG offers a number of different search algorithms to do so. For performance reasons, they all combine the application of binary and context-sensitive constraints with the search for an optimum such that the binary and context-sensitive constraints are not evaluated on all structures in the hypothesis space. We briefly describe the two most common search methods here: *complete search* and *frobbing*.

Ideally, search should identify the globally optimal solution structure, i.e. the solution candidate SC_{opt} with the best – and in our case: highest – score $\Phi(SC_{opt})$ overall. To ascertain global optimality, the entire solution space needs to be searched. To this end, WCDG offers the option of complete search. For the sentence lengths encountered in the parsing of unrestricted natural language, however, complete search is undesirable since its completeness comes at the cost of extremely long processing times. For sentence lengths greater than 30 words, which are encountered frequently in unrestricted German input, complete search typically incurs processing times in the order of hours for a single sentence on standard hardware. This is not surprising since the evaluation of the binary and context-sensitive constraints on all possible solution tree structures incurs an enormous computational effort. On longer sentences the size of the hypothesis space can easily reach 10^{100} candidates and more; for practical parsing purposes, the computational effort of complete search is therefore best avoided.

Moreover, even this supposedly *"complete"* search applies pruning heuristics to reduce the size of its search space. The pruning heuristics have the undesirable side effect that even the "complete" search can miss the global optimum as the result of imperfect pruning in some cases. Despite what its name suggests, WCDG's complete search therefore is not truly *complete*.

The most common search method employed in WCDG — and also the one we have used in the experiments described in Part III of this thesis — is *frobbing* (Foth et al., 2000; Foth, 2007), a highly effective transformation-based search method. Frobbing frequently — though unfortunately not always — detects the global optimum within seconds to minutes on standard hardware, depending primarily on the length of the input sentence. In its search for the best solution structure, frobbing builds *one* complete solution candidate from the edges scored in the application of non-context-sensitive unary constraints and then applies the context-sensitive and binary constraints to it. Frobbing continually makes local modifications to the initial structural hypothesis and attempts to resolve its most severe constraint violations first. Frobbing iteratively proceeds along a path of continually improving solution candidates to a local optimum until no further improvements can be effected. In this optimisation, frobbing operates on a dependency structure spanning the entire input sentence and, without further amendments, excludes the aspect of incrementality as observed in human sentence processing.[1] A detailed description of the complex algorithm underlying frobbing is given in Foth (2007, p. 36).

4.2.5 Predictor Integration

WCDG1 provides a generic interface for the integration of parser-external information prior to the commencement of the parsing process. Invariably, this external information is provided by a specialised application that delivers additional information pertaining to the input sentence. As the integrated application provides its information prior to parse time, it cannot build on knowledge about the final parse structure of the input sentence and therefore is referred to as a *predictor*.

While the interface is subject to a number of limitations to be outlined below, it constitutes a highly useful option for incorporating external, possibly non-linguistic information to influence the process of parsing. For this reason, the interface is of particular interest to our endeavour of integrating non-linguistic context information into the process of parsing. In line with WCDG's constraint-based approach, predictors can only influence dependency decisions in the parser by providing information that further constrains dependency assignments. A predictor can be employed as a veto component whose judgement is based on parser-external information, i.e., on information residing outside of WCDG's lexicon and grammar.

Predictors are being used extensively in WCDG to improve the quality of parsing further. Foth (2007) demonstrated that the integration of a combination of different predictor components in a hybrid parsing approach improved WCDG1's overall

[1] Beuck (2009) has recently reported successful modifications to WCDG's core processing algorithms to allow incremental sentence processing in WCDG. One of the reported modifications includes the repetitive invocation of frobbing on sentence fragments of increasing length.

parsing accuracy significantly. Based on these findings, the T'n'T POS tagger and a statistical predictor for PP-attachment have been included into the standard configuration of WCDG1.[1] Khmylko (2007) reports further improvements to WCDG1's parsing accuracy by integrating Ryan McDonald's MSTParser (McDonald, 2006; McDonald et al., 2006) as a predictor. The integration increased parsing accuracy to 93.9% for structural and 91.8% for labelled accuracy.

Predictors influence the parse process via the following mechanism: After tokenization and POS tagging, but still prior to parsing, WCDG1 requests predictions for dependency scores between words (*sic!*) in the input sentence. With the request WCDG submits the list of slot strings in the input sentence as well as their – based on the relative frequencies in the training corpora – most probable POS tag according to the T'n'T POS tagger to the predictor. The predictor processes its input, generates predictions and returns a line of attribute-value pairs for each slot in the input sentence. The PP-attachment predictor, for example, returns a line of attribute-value pairs for each slot string in the input sentence.[2] Each attribute-value pair contains the regent slot number as attribute and the predicted dependency score for assigning a `PP`-dependency between the two words as its value.
WCDG1 reads the predictions into memory and has access to them via the grammar's `predict()` function. With this function, integration constraints can be formulated to assign predictor-based dependency scores. Attributes for which the predictor has not returned a prediction value do not contribute information based on which a predictor-based score penalty can be imposed. WCDG hence treats missing predictor information as equivalent to receiving a permissive prediction score of `1`.

The constraint for integrating PP-attachment predictions to influence the overall scoring of a solution candidate is given in Figure 4.4. This constraint stipulates that any edge `X` on the `SYN` level which does not attach to the `ROOT` node and bears the label `PP` or `KOM` has to meet at least one of the following two conditions: Either the word on its lower node is 'um' or 'als' and the word on its upper node is a comparative or the score predicted by the PP-attachment predictor for `X` is equal to 1. Observe that this constraint will be violated by any edge for which the predictor provides a prediction value lower than 1. In that case, the constraint inflicts a dynamic penalty on the overall parse tree which is equal to the value of the PP-attachment prediction for that edge. The closer the prediction value is to 0, the harder the veto on the assignment of the `PP` label to that edge (cf. the definitions of constraint weights in Equation (4.1) and of WCDG's overall scoring function in Equation (4.2)).

[1] To be precise, the T'n'T POS tagger does not integrate via WCDG1's standard predictor interface for historical reasons. Also, the POS tagger takes a somewhat special position since it is invoked with the input sentence only. Its output is then included as part of the prediction request to all the other predictors. We can still consider the POS tagger a predictor since its output is made available to WCDG prior to the commencement of parsing — just as if it were generated by a regular predictor.

[2] Foth and Menzel (2006a) provide a detailed description of the PP-attachment predictor.

4.3 Limitations of WCDG's Standard Implementation

With regards to our modelling objectives, the implementation of predictor integration in WCDG1 is subject to limitations in two important respects:

L1 Underspecified Prediction Request

As input to the predictor prior to parsing, WCDG1 hands over the list of all slot strings in the input sentence, each of which with an assigned POS tag. The POS tag handed-over for each homonym is the most probable one for the respective slot string according to the T'n'T POS tagger. The predictor hence does not receive *all* possible POS tags for a given slot string.

In many cases, however, the different readings of a slot string exhibit fundamentally different syntactic and semantic behaviour. A capitalised sentence-initial word like 'Fragen'—which translates to either *Ask* or *Questions*—can be lexically ambiguous with a variety of readings, the most probable of which with POS classifications `NN` (regular noun), `VVINF` (verbal infinitive) or `VVFIN` (finite verb). All of these differ significantly in their syntactic and semantic behaviour. Yet, WCDG1 predictors only receive the most probable of these homonyms as input for making their predictions.

The decision of which of these homonyms is to appear in the final parse structure is only taken at parse time – and may well result in the selection of a form other than the one bearing the most probable POS tag. When generating predictions for structurally relevant features prior to parsing, we need to be able to assign different prediction values to different homonyms. This capability requires that the lexical information which homonyms are available for the given slot be made available at the time of predictor invocation.

L2 Slot-Based Prediction Encoding and Retrieval

Predictions are externally assigned additional properties of tokens in the input sentence. For WCDG, a general prediction can be represented as a quadruplet $\langle t, n, a, v\rangle$ where t is the identifier of the token for which the prediction is being made, n the predictor name, a the prediction attribute and v the prediction value.[1] An example of how such a prediction quadruplet maps onto actual parameters is given for WCDG's PP-attachment predictor in Figure 4.5.

Predictions can be encoded and retrieved homonym-specifically, if t is specific enough to reference individual homonyms. In WCDG1, however, predictions have been modelled as a slot property. The internal representation of t therefore points to a slot rather than to a homonym. Consequently, homonym-specific prediction encoding cannot be achieved. The standard implementation of prediction retrieval in WCDG1 via the `predict()` function matches this modelling view and only offers slot-based prediction retrieval.[2]

[1] For predictors that compute only a single attribute, the information encoded in n and a is redundant and can be collapsed into a single parameter. In those cases, predictions can be represented as a prediction triplet $\langle t, n, v\rangle$.

[2] In this regard, the syntax of the `predict()` function belies the implementation in WCDG1. The first

$t \longmapsto$ `X@id`, i.e., the slot identifier of `X`'s lower node,
$n \longmapsto$ PP, i.e., the predictor name,
$a \longmapsto$ `X^from`, i.e., the slot number of `X`'s upper node,
$v \longmapsto$ the actual prediction value.

Figure 4.5: Mapping a prediction quadruplet to parameters of the PP-attachment predictor.

To overcome limitation L1, we add modelling requirement R31.

> **Requirement R31**
> *A WCDG predictor for scoring meaning-related dependencies must be able to differentiate between different readings of a slot string and must be capable of generating separate, homonym-specific predictions for those readings.*

To overcome limitation L2, we add modelling requirement R32.

> **Requirement R32**
> *To enable the processing of different external predictions for the readings of a slot string, WCDG2 must provide homonym-specific encoding and retrieval of predictions.*

4.4 Chapter Summary

In this chapter we have provided an introduction into the treatment of parsing as a constraint satisfaction problem. Weighted-constraints have been presented as a particularly useful refinement to the formalism of symbolic constraint-based parsing. The main advantages of a symbolic parser operating with weighted constraints rather than generation-rules lie in its capability to a) model graded preferences rather than simply to categorise as grammatical or ungrammatical, b) provide analytic feedback, and c) react more robustly to unknown input.

We have outlined the benefits of WCDG's dependency formalism and its relational representation. Its disadvantage lies in the limitation that it expresses linguistic dependencies between individual words rather than between bracketed, more complex linguistic entities such as phrases.

We have also provided a detailed description of the system capabilities of WCDG1, an implementation of a weighted-constraint dependency parser. WCDG1 is of particular interest to our modelling challenge because of its generic predictor interface. Within clearly defined limitations, this interface permits to integrate parser-external, non-linguistic information into the parsing process. Access to the predictions is achieved via suitably formulated integration constraints in the grammar that are processed at parse time.

argument to the `predict()` function is `X@id`, which normally references a specific homonym in the input sentence unambiguously. However, at code level, the `predict()` function has been implemented such that the homonym-specific reference `X@id` is abstracted into a reference to the homonym's slot only.

Additional modelling requirements based on WCDG1's capabilities and limitations have been motivated in order to achieve the integration of cross-modal context into the parsing process. The application-focused requirements identified in this chapter complement the collection of modelling requirements from the preceding Chapters 2 and 3 and conclude the process of requirements collection in this thesis.
It needs to be noted that the list of collected requirements cannot make a claim to completeness, nor are the requirements structurally homogeneous or uniform in granularity. The main use of the requirements will be as a benchmark for the functional scope of the model we intend to argue for in Part II of this thesis. The compilation of a comprehensive, uniform and homogeneous collection of requirements for the interaction between vision and language certainly requires further rigorous investigation and warrants to undertake a separate research effort in its own right.

In the following part of this thesis, we describe in detail our model implementation which was designed with the intention to achieve maximum coverage of the identified requirements while ensuring actual implementability of the specified system. We also engage in a discussion to what extent our model meets the identified modelling requirements. For ease of revision, the list of all 32 modelling requirements is given in Appendix I.

Part II

Model Implementation

Chapter 5

The WCDG2 Parser

The implementation of our model for the interaction of non-linguistic modalities with language centres around WCDG2, a functionally enhanced version of the weighted-constraint dependency parser WCDG1 described in Chapter 4. This chapter outlines WCDG2's functional enhancements over WCDG1 and discusses the parser's interactions with the other components in our model.

As a general guideline in the design and specification of WCDG2, we have kept the number and extent of functional changes over WCDG1 to a minimum. Modifications were only made in cases where WCDG1's features made the implementation of vital aspects of our model difficult or impossible. Another guideline was to leave the syntactic processing of WCDG1 unchanged. Any additional capabilities included in WCDG2 are add-ons to WCDG1's existing functionality.

The enhancements implemented in WCDG2 comprise semantic extensions to the constraint base and the lexicon, modifications to the predictor interface and an expansion of the argument structure for prediction access in the grammar. In line with our second design guideline, no changes were made to WCDG's central constraint satisfaction algorithms nor to the heuristic search routines of frobbing (cf. Section 4.2.4). In our model implementation, we build on WCDG1's large-coverage grammar of German and leave its constraints for syntactic processing unchanged.

5.1 Architectural Overview

After loading an extended lexicon and a semantically enhanced grammar, WCDG2 receives its input sentence. In the preprocessing phase, WCDG2 integrates a plausibility predictor component (PPC) via an extended version of WCDG1's predictor interface (cf. Section 4.2.5). The purpose of the PPC is to score semantic dependencies in the input sentence based on a representation of visual context information. To this end, the PPC establishes communication with a reasoner component. The reasoner accesses a knowledge representation of visual context. The knowledge representation is made up of two components: a situation-invariant ontology containing hierarchical lexical and world knowledge (the *T-Box*) and a situation-specific representation of situation information (the *A-Box*).

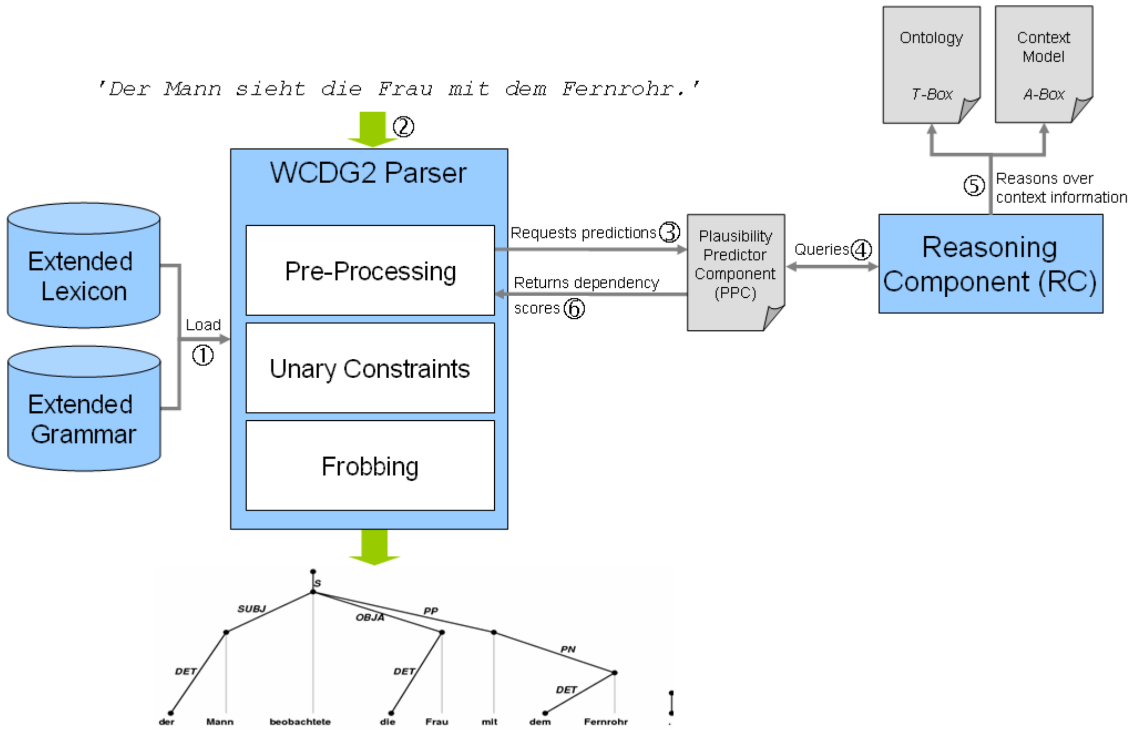


Figure 5.1: Components and their interaction in the Context Integration Architecture (CIA).

Based on this context information, the PPC computes score predictions for semantic dependencies between words in the input sentence and returns them to the parser for access at parse time. The PPC's score predictions are based on semantic context information and directly affect the assignment of the semantic dependencies in WCDG2's semantic representation of the input sentence. Syntactic analysis is affected indirectly by these predictions via the correspondence rules between semantic and syntactic representation as specified in the syntax-semantics interface in WCDG2's enhanced grammar. The syntax-semantics interface contains correspondence rules between the representations of syntax and semantics which ensure that semantic and syntactic representations align. Thus, the contextual influence upon semantic representation is propagated into syntactic analysis with semantic mediation. WCDG2 optimises the semantic and syntactic dependency structures by a heuristic search for the minimum of the severities of all constraint violations.
We refer to this entire framework as our *Context Integration Architecture* (CIA). A schematic overview over the CIA is given in Figure 5.1. The CIA components and their interaction with each other will now be outlined in detail in the following sections.

5.2 The Role-Assigning Grammar

As outlined in Section 4.2.2, WCDG1 comes with a robust grammar for syntactic parsing of unrestricted German text input. Requirement R1 demands that the interaction between non-linguistic modalities and language be mediated by a semantic representation of linguistic meaning. To meet this requirement, we include a semantic representation in WCDG2.
WCDG's implementation builds on the hard-wired[1] unique-regency constraint which stipulates that on a given level of analysis every dependant may have *precisely one* regent. A number of semantic constellations are known, however, in which a single dependant needs to depend on *more than one* semantic regent in the same sentence. An example of a sentence in which a single dependant takes two different semantic regents is given in Figure 5.2.
To be able to model such semantic constellations, the implementation of the semantic representation in WCDG2 had to be spread out over separate levels of analysis. Despite this distributed implementation, we consider our model to meet modelling Requirement R22 which demands a single, unified representation of meaning. In our model, this representation can be realised as an abstraction by projecting the results of all semantic levels of analysis in WCDG2 into a single plane. The additive projection of all semantic levels of analysis in WCDG2 constitutes an equivalent of Conceptual Structure in Conceptual Semantics. We emphasise that from a modelling perspective these different technical realisations do not constitute separate levels of semantic representation. They merely are different technical realisations of *the same* semantic representation in the sense of Conceptual Semantics (cf. Section 3.2).
For terminological clarity we refer to the levels of analysis on which WCDG2 performs semantic processing as *semantic levels of analysis*. We reserve the term *semantic representation* for the integrated, uniform representation of meaning as required by Conceptual Semantics and captured in Requirement R22.

The constraints for the semantic levels of analysis are contained in a separate grammar which we refer to as WCDG2's *role-assigning grammar*. For parsing, WCDG2 uses the *extended grammar* which consists of the union of WCDG1's syntactic grammar and the role-assigning semantic grammar. Since the two constraint sets are disjoint, WCDG2 can process their union without any conflicts.[2]

[1] By 'hard-wired' we mean that this is not a grammar-based constraint but a hard-coded restriction arising from how the constraint-dependency formalism has been implemented in WCDG.

[2] In case of multiply defined constraints by the same name, WCDG resolves the conflict by overwriting previously loaded constraint definitions with the subsequently loaded constraints by the same name. Conflicts that cannot be resolved in this way cause WCDG to reject the input grammars.

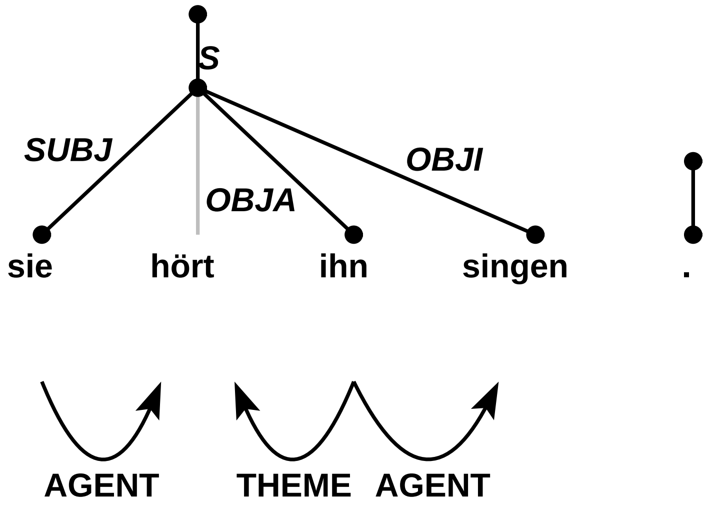


Figure 5.2: *She hears him singing.* Sentence in which the word 'ihn' *him* takes two thematic roles AGENT and THEME and hence requires a different semantic regent for each dependency.

The constraints in the role-assigning grammar can be divided into three disjoint subsets:

1. Constraints that act on semantic levels only and exclusively use information from WCDG2-internal resources such as the extended lexicon.
2. Integration constraints that act on semantic levels only and integrate the WCDG2-external predictor information.
3. Constraints that act on a combination of syntactic and semantic levels. These constraints define correspondence rules between the syntactic and the semantic representations.

Taken together, the constraints in the first two subsets define the structural properties of permissible thematic role dependency structures on the semantic levels of analysis. The constraints in the second subset propagate the contextual information from the non-linguistic representations into the semantic representation of language in WCDG2. Access to the WCDG2-external non-linguistic information is achieved via the prediction scores that the PPC returns to WCDG2. The third set of constraints governs the interaction between the semantic and the non-semantic levels of analysis in WCDG2. The non-semantic levels are the syntactic `SYN` level and the `REF` level. The constraints in this third set define WCDG2's syntax-semantics interface, a representational interface in the Jackendoffian sense (see Section 3.3). The well-formedness rules for the non-semantic levels of representation are defined in WCDG's standard grammar for German. The well-formedness for the semantic part of linguistic analysis in WCDG2 are provided by constraint subsets 1 and 2 above. We consequently consider the requirement that every level of representation have its own finite set of well-formedness rules, Requirement R10, fully implemented

in our model.[1] Also, the semantic part of linguistic representation is fully included in the single, shared level of semantic representation. We hence consider Requirement R21 fully implemented by our model.
Since each representation uses its own set of edge labels that are specific to the levels of analysis in WCDG2, we consider the syntactic and semantic representations in our model as informationally encapsulated. This constitutes a fulfilment of Requirement R11.

The following sections explain in detail which role each of these three constraint categories play in our model. Implementation specifics will only be given where required for the in-depth understanding of the model realisation.

5.3 Thematic Role Representations

We base the definition of the thematic roles supported in our model implementation on the lists of thematic roles in Dowty (1989, p. 69) and Löbner (2003, p. 174). We deliberately avoid a discussion of the granularity and appropriateness of these thematic roles by choosing comparatively general definitions for a set of roles that is widely accepted as standard (cf. also Ferretti et al., 2001). The purpose of our model implementation is *not* to demonstrate the correctness or appropriateness of specific thematic role definitions. Rather, we wish to show with our model that inherently semantic generalisations over verbal argument slots can be used as constituents of semantic representations that mediate the cross-modal interaction between non-linguistic and linguistic modalities. We acknowledge that the thematic role definitions in Table 5.1 are to some extent arbitrary and not sharply delineated – as are *all* thematic role definitions that attempt to capture semantic generalisations over verbal arguments. As long as the assumption that underlies the concept of thematic roles remains unchallenged, namely that semantic generalisations over verbal argument structures are indeed possible, the precise semantic delineation of these roles or the labels attached to them have no fundamental impact on the validity of our model.[2]

These role definitions deserve a few further comments: Our role definitions do not build on a differentiation between events, actions, states or processes. Instead, we adopt the more embracing term *situation* as used in the *situation semantics* of Barwise and Perry (1983, pp. 7) to subsume *all* of the aforementioned notions. A situation type or concept hence is taken to denote an abstract constellation by which participating individuals with certain properties are related to each other. An example of a situation concept is BARK which only involves a *barker* as participant.

[1] Strictly, the well-formedness rules for context model representations also contribute to the fulfilment of this requirement. These are described in Section 6.4.

[2] Clearly, the choice of thematic roles considered does correlate with the constraints defined in the role-assigning grammar. A more fine-grained differentiation in role definition will also require more finely differentiated role-assigning constraints in the grammar. This aspect, however, does not question the validity of our model as such.

AGENT	The participant specified as doing, causing, having, being or experiencing something in a situation. Example: *He is eating an apple.* He $\xrightarrow{is\_AGENT\_for}$ eat
THEME	The participant that something is happening to in the situation or that is immediately affected by the situation. Example: *He is eating an apple.* apple $\xrightarrow{is\_THEME\_for}$ eat
RECIPIENT	The participant that the result of the situation is directed to. Example: *She gave him a book.* he $\xrightarrow{is\_RECIPIENT\_for}$ give
INSTRUMENT	The entity enabling or facilitating the occurrence or progress of a situation. Example: *He opened the door with a key.* key $\xrightarrow{is\_INSTRUMENT\_for}$ open
OWNER	The entity extending any sort of ownership or belonging relation towards another participant. Example: *She has Kirsa's book.* Kirsa $\xrightarrow{is\_OWNER\_for}$ book
COMITATIVE	The entity that physically or figuratively accompanies another participant. Example: *He went to the cinema with her.* she $\xrightarrow{is\_COMITATIVE\_for}$ he

Table 5.1: Overview over the thematic role definitions in our model.

The situation in which a participant is *barked at* involves an enity *barked-at* in addition to the *barker* and thus constitutes a related, yet different, situation BARK.AT. Barwise and Perry consider the instances of barking observed in the real world to be instantiations of the abstract situation types BARK and BARK.AT.

By including the component of experience in the role definition of AGENT, we incorporate the aspects which Dowty (1989) lists for the separate role of EXPERIENCER into the role definition of AGENT. Our definition of THEME is in line with that of Löbner (2003) and treats the roles THEME and PATIENT as semantically equivalent.[1]
For terminological clarity, we henceforth differentiate between *entities* and *participants*. By the term 'entity' we denote anything that takes a thematic role in the context of a situation. We use the term 'participant' to refer specifically to those entities in a situation that engage in a direct and semantically mandatory thematic relation with an instance of the situation concept. We therefore denote entities taking an AGENT, RECIPIENT or THEME role as participants while entities taking an OWNER, COMITATIVE or INSTRUMENT role are considered situation entities but *not* participants.

We limit our modelling scope to these thematic roles since they are sufficient for the study of a number of interesting and notoriously difficult-to-parse syntactic phenomena such as PP-attachment or subject-object ambiguity of German plural nouns. Also, this set of thematic roles results in a manageable number and complexity of hand-written constraints in the role-assigning grammar. The role-assigning grammar used in the experimental runs reported in Chapters 8 to 11 contains about 140 individual constraints — as opposed to approximately 1050 active constraints in WCDG1's large-coverage syntactic grammar for German.

In principle, our model permits to extend or modify the list of supported thematic roles. Any extension or modification that is not simply a reduction of the set of thematic roles supported may, however, incur the need to add or change constraints in the role-assigning grammar.[2] Thematic role assignment in our model is subject to the following modelling decisions and constraints:

- Thematic dependencies originate from the role filler.
- The `AGENT` dependency is assigned on its own level of analysis.
- The `THEME` dependency is assigned on its own level of analysis.
- The dependencies `RECIPIENT`, `OWNER`, `INSTRUMENT` and `COMITATIVE` are modelled as mutually exclusive and are assigned on a separate level of analysis.
- Thematic dependencies under the same regent are unique.
- The verb-centred semantic dependencies `AGENT`, `THEME` and `RECIPIENT` can only be assigned to verb forms that have a corresponding semantic valence.

[1] Jackendoff (1990, p. 129) vehemently argues against this practice. Given the widely acknowledged fuzziness in defining thematic roles, we choose to disagree with his point of view.

[2] An arbitrarily large extension of the list clearly is not possible in our model. *Every* implementation is limited, not only by its algorithmic design, but also by the capabilities of the hardware on which it is executed. Unless stated otherwise, we assume that limitations arising from the hardware environment can be neglected in the assessment of the validity of our model.

- A verb's semantic valence must be saturated by the assignment of the corresponding semantic dependencies.

- Attachment restrictions for the semantic dependencies have been formulated based on the dependant's and regent's part of speech rather than based on genuinely semantic criteria. The role-assigning grammar imposes no selectional restrictions as a function of the role fillers' conceptual category. Our model hence fails to meet Requirement R23 which demands meaning-based selectional restrictions for thematic role fillers.

5.4 The Extended Predictor Interface

To achieve an interaction between parser-external information and linguistic decision making in the process of parsing, WCDG2 integrates the PPC as a predictor component and communicates with it prior to parsing. WCDG2's predictor interface provides functional extensions over WCDG1's interface for both out-bound and in-bound communication.
To overcome the limitation of underspecified predictor requests in WCDG1's outbound communication (see Limitation L1 on page 68), two design approaches are conceivable: either WCDG2 is modified such that it hands over more detailed information to the predictor or the predictor is enabled to procure the more detailed information independently.

In an early design phase we tested if modifications to WCDG's predictor interface could be avoided by non-invasively leaving the task of homonym collection to the external predictor. This approach turned out to be infeasible for performance reasons. To obtain the complete homonym information available to WCDG, the predictor has to perform two tasks: 1) search all lexical entries in the full-form lexicon with more than $1.01 \cdot 10^6$ entries and 2) check every slot string for matches with the templates for unknown words that have been defined in the grammar. One of the reasons for the observed performance issues is that this design replicates the task of homonym collection. Internally, WCDG already collects the complete homonym information for every slot of the input sentence.
We therefore chose to extend the predictor interface in the parser. WCDG2 now hands over the input sentence with the complete list of homonyms in the input sentence — rather than just the slot strings. For each homonym, WCDG2 provides the slot string, the slot number, an identifier that uniquely identifies the homonym in its slot and all of the homonym's lexical features. The list of lexical features is obtained either from the homonym's lexical entry or from the word templates it matches. A typical input line handed over to the PPC is shown in Figure 5.3.

```
Der 1 Der_ART_pl cat ART case gen number pl gender bot definite yes
```

Figure 5.3: A PPC input line as received from WCDG2 via the extended predictor interface.

The PPC now has access to the full lexical information for each homonym in all the sentence slots rather than just the slot string's most frequent POS tag as is the case for WCDG1 predictors (cf. Section 4.2.5). This allows the predictor to compute dependency score predictions between individual homonyms rather than just between slots. In order for homonym-specific prediction results to be processable by the parser, WCDG2's in-bound predictor communication also needed to be enhanced. Multiple homonym-specific predictions for a given dependant-regent pair can now be read in and accessed via suitable predicates in WCDG2's grammar (cf. Section 5.5).
Since the PPC typically returns multiple semantic-dependency predictions for each pair of homonyms that receive a prediction, the size of its output can be significantly larger than for WCDG1 predictors which only return a single dependency prediction between two slots. As a result, the extension of the predictor interface also requires that sufficient memory be allocated in WCDG2 to read in the entire PPC input.[1]

5.5 Context Integration

An essential aspect of our model implementation is its capability to integrate non-linguistic information into the process of parsing – and the construction of a cross-modally integrated semantic representation of sentence meaning in particular. This representation is created on the semantic levels of analysis in the parser. Contrary to natural systems in which the semantic representation of visual scene context is built up incrementally shortly before – or in some cases even in parallel to – linguistic processing, our model acquires its complete contextual information as a completed knowledge representation of visual context.[2] The complete semantic representation of visual context is analysed by the PPC which, based on the semantic context information and the input sentence, calculates its score predictions for semantic dependency edges. Integration into the process of parsing is achieved the integration constraints (see the categorisation of constraints in the role-assigning grammar on page 76). The purpose of an integration constraint is to check whether the PPC has made a score prediction for a given dependency edge and, if so, to assign the predicted score to that dependency edge. We use this mechanism to penalise dependency edges in the semantic representation based on WCDG-external information.

Technically, the propagation of the dependency score prediction into the parser is achieved by a class of dynamically weighted constraints whose constraint body declares that if the dependency edge bears a specific thematic role label, the edge must have a score prediction of 1. This constraint is violated by all dependency edges with the correct label that have a PPC prediction score less than 1.[3] In case of

[1] In one of our early experimental runs on a longer input sentence the reserved buffer was too small to contain the entire predictor input. The cause for this problem was the static buffer size allocation inherited from WCDG1. This problem was remedied in WCDG2 by implementing dynamic buffer sizing for reading in the predictor input.

[2] We discuss the assumptions underlying this modelling decision in detail in Section 6.4.

[3] The algorithm by which the PPC computes its prediction scores will be discussed in Section 7.4.

```
// ROLEhood as predicted by the PPC.
{X!LEVEL} : 'ROLE Integration': [ predict( X@id, PPC, ROLE, X^id ) ] :
X.label = ROLE
->  predict( X@id, PPC, ROLE, X^id ) = 1;
```

Figure 5.4: Generic cross-modal integration constraint for the thematic dependency ROLE.

a constraint violation, the constraint weight is set dynamically to the PPC's prediction value for the thematic dependency. The class of integration constraints is represented by the generic constraint in Figure 5.4. In that constraint, the semantic dependency `ROLE`[1] is from the set of supported thematic roles in the model implementation and `LEVEL` is the level of analysis in WCDG2 on which `ROLE` is assigned. To be able to constrain the assignment of every supported semantic dependencies, the role-assigning grammar contains one integration constraint of this form for every supported thematic role. `AGENT` dependencies, e.g., are restricted by an integration constraint checking for `X.label = ROLE` on the `AGNT` level of analysis. Owing to WCDG's scoring policy, the best contextual support a thematic dependency assignment can obtain from the PPC is a prediction score of `1` (cf. Section 4.2.5). This means that the PPC has either found positive contextual evidence for the assignment of this role or was unable to derive contextual evidence against it.

The PPC's prediction value for the semantic dependency is accessed at parse time via an extended, four-place `predict()` function. Its input arguments are the unique ID of the dependant homonym, the predictor name, the edge label and the unique ID of the regent homonym. With the fourth argument for prediction access, WCDG2 can now retrieve homonym-specific predictions for multiple dependencies between the same pair of dependant and regent homonyms. This could not be done in WCDG1. This WCDG2 capability overcomes the WCDG1 limitation of slot-based prediction encoding and retrieval (cf. Limitation L2 on p. 68) and meets Requirement R32.

It is important to stress that the context integration with WCDG2's predictor interface constrains the semantic representation of linguistic analysis based on given visual context information. As such, it incorporates a *unidirectional* influence of visual context upon linguistic processing. This fulfils Requirement R5.

A substantial limitation of the present form of our model lies in the fact that it does not offer any mechanism for propagating linguistic information into the opposite direction, i.e., from the linguistic to the non-linguistic modalities. This limitation arises from the fact that WCDG does not – as yet – provide an interface to access parser-external components *at parse time*. So far, WCDG's access at parse time is limited to its internal data structures containing information from parser-external components that was acquired prior to parse time.[2] Our model hence fails to meet Requirement R6.

[1] We henceforth use the string ROLE as a generic placeholder for an arbitrary thematic role and denote the corresponding semantic dependency that is supported in our model by `ROLE`.

[2] We wish to acknowledge that at the time of writing, a project to extend the functional scope of WCDG2 with the aim to remove this limitation is ongoing at the University of Hamburg's Department of Informatics.

In humans, the influence of language upon non-linguistic modalities, mostly mediated by attention, gives rise to effects such as the language-driven anticipatory eye movements observed by Tanenhaus et al. (see Section 2.3), visual search and active vision (Henderson, 2003). In the subsequent discussion of our model, we restrict ourselves to the unidirectional influence of non-linguistic context upon language. The implementation of a bidirectional interaction constitutes a significant challenge for modelling and implementation and is strongly encouraged as a target of future research.

5.6 The Syntax-Semantics Interface

The objective of our implementation is to achieve an interaction between non-linguistic information and syntactic parsing via a single, shared level of semantic representation. So far, our implementation description has covered how WCDG2-external prediction information is propagated into the cross-modally integrated semantic representation of sentence meaning. We now outline how WCDG2's semantic representation interacts with the syntactic representation in the course of parsing.
In WCDG, the representations for dependency structures are level-specific in the sense that each structure uses its own set of edge lables. As a result, the processing on different levels of analysis is informationally encapsulated, i.e., structural changes on a level L_1 do not affect the structures on another level L_2 unless there is an explicitly defined constraint in the grammar that requires such a correspondence. We hence consider Requirement R12 for the representational encapsulation of representations to be fulfilled by the syntactic and semantic representations in our model.
The only way that two dependency structures from L_1 and L_2 can interact with each other is via the structural correspondence rules defined in the interface between those levels of representation. In our model, such a rule typically is a binary constraint relating two edges `X` and `Y` such that `X` is from L_1 and `Y` is from L_2.

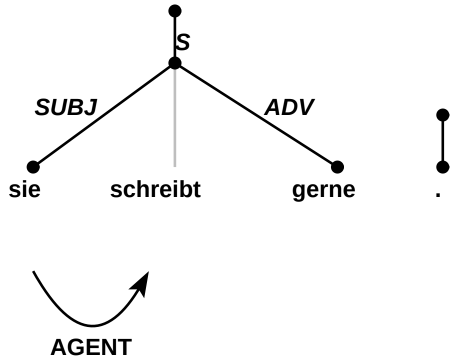
WCDG2's syntax-semantics interface establishes correspondence relations between syntactic structural constellations and their semantic correlates. The interface between syntax and semantics enables an *immediate* and *bidirectional* interaction between the two representations such that changes in semantic representation directly affect syntactic analysis and vice versa. We consider this a model feature in fulfilment of Requirement R4 for representational interfaces.
Our model makes contextual information available at the point in time of syntactic decision making at which they are needed. This way, syntactic candidate structures are assessed for their contextual compliance *at the time of their creation* rather than subsequently. As we acknowledged previously, WCDG does not provide the capability of incremental sentence processing yet (cf. Section 4.2.4, p. 83). Requirements R2 and R3 hence cannot be fulfilled in our model implementation. Still, the immediate correspondence between syntactic analysis and the contextually-informed semantic representation enables an immediate unidirectional influence of contextual information upon syntactic processing. We consider this a fulfilment of Requirement R4 which demands the online interaction between non-linguistic modalities and language at parse time.

The syntax-semantics interface propagates the influence of non-linguistic context upon semantics on to syntactic representation via the following mechanism: The integration constraints propagate non-linguistic context information into the semantic representation via context integration constraints. Simultaneously, the linguistic part of semantic analysis interacts with syntactic representation via the correspondence constraints in the syntax-semantics interface. The influence of non-linguistic context upon syntactic representation is hence mediated by the semantic representation as demanded by Requirement R1. In fulfilment of Requirement R13, this mediating effect is achieved by the correspondence rules between the syntactic and semantic representations. Since this correspondence interaction is evaluated at parse time, our model also meets Requirement R14. As there is only one level of semantic representation and this level of representation mediates the meaning-based cross-modal influence of visual context upon syntactic processing, our model meets Requirement R22 for a single and unified semantic representation as well.

With approximately 40 constraints in the syntax-semantics interface spanning more than 400 lines of constraint code[1], a comprehensive list of correspondences between the syntactic and semantic levels of analysis is beyond the scope of this work. To provide at least a qualitative impression of the kind of correspondences captured in the syntax-semantics interface, the following list gives a brief overview over some of the more important modelling rules we have implemented.

In an active-voice sentence, the verb's `AGENT` is also the subject `SUBJ` if the verb's semantic valence admits an `AGENT`. Conversely, the `SUBJ` in an active-voice sentence is also the `AGENT`.

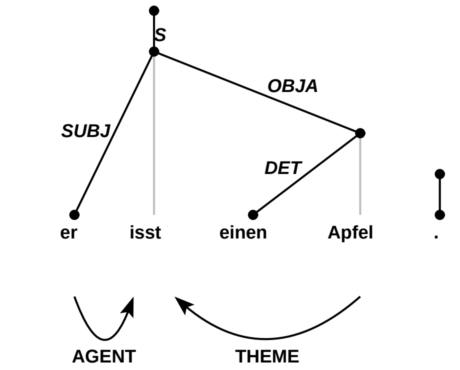
Example: 'Sie schreibt gerne'
She likes to write.



[1] This number includes all active lines of code in a constraint. In the absence of a suitable formal measure for the informational complexity of constraint dependency grammars we resort to the purely quantitative measure of expressing the size of the syntax-semantics interface in lines of code. An absolute measure of constraint-dependency grammar expressivity would be desireable for it would permit to compare the informational complexity of different grammars objectively.

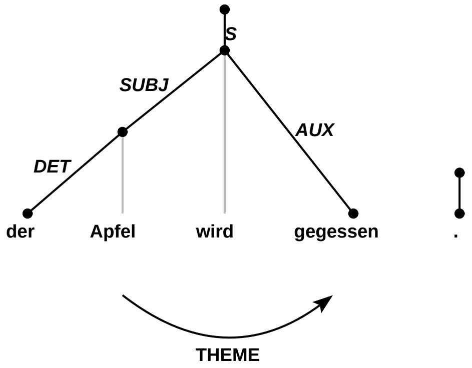
In an active-voice sentence, the `THEME` is the accusative or direct object `OBJA` if the verb's semantic valence admits a `THEME`. Conversely, in an active-voice sentence, the verb's `OBJA` is also the `THEME`.

Example:
'Er isst einen Apfel.'
He is eating an apple.



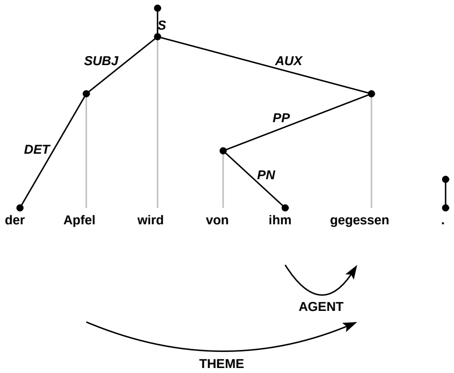
In a passive-voice sentence, the verb's `THEME` is the subject `SUBJ` if the verb's semantic valence admits a `THEME`. Conversely, in a passive-voice sentence, the `THEME` is also the `SUBJ`.

Example:
'Der Apfel wird gegessen.'
The apple is being eaten.



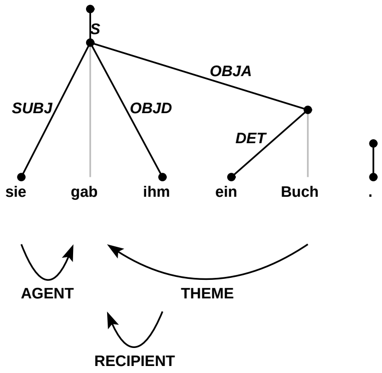
In a passive-voice sentence, the verb's `AGENT` is the prepositional complement `PN` in the prepositional phrase `PP` modifying the full-verb if the verb's semantic valence admits an `AGENT`. Conversely, in a passive-voice sentence, the `PN` of a full-verb-modifying `PP` is the `AGENT`.

Example:
'Der Apfel wird von ihm gegessen.'
The apple is being eaten by him.



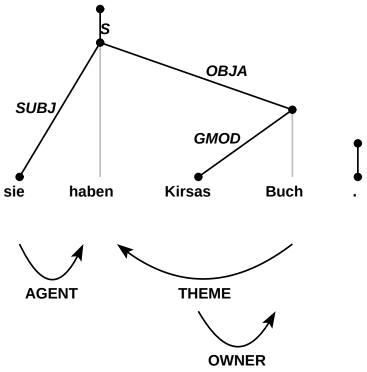
The verb's RECIPIENT is its dative or indirect object OBJD if the verb's semantic valence admits a RECIPIENT. Conversely, the OBJD is the RECIPIENT if the verb's semantic valence permits.

Example:
'Sie gab ihm ein Buch.'
She gave him a book.



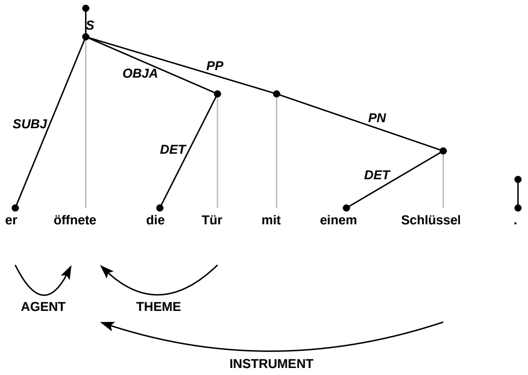
The OWNER of a syntactically modified entity is its genitive modifier GMOD. Conversely, any GMOD is also an OWNER.

Example:
'Sie haben Kirsas Buch.'
They have Kirsa's book.



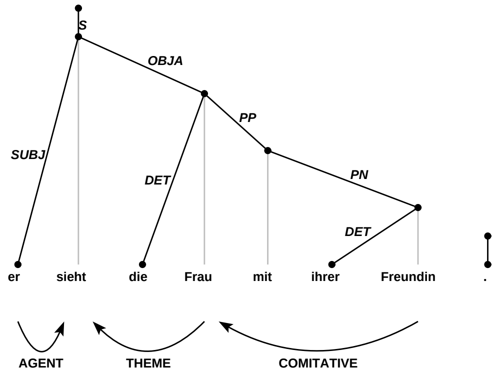
In a passive-voice sentence, the verb's INSTRUMENT is the prepositional complement PN in a full-verb modifying 'mit' *with* or 'durch' *by* prepositional phrase PP. Conversely, in a passive-voice sentence, the PN originates from the INSTRUMENT if the PN is part of a full-verb modifying 'mit' or 'durch' PP.

Example:
'Er öffnete die Tür mit einem Schlüssel.'
He opened the door with a key.



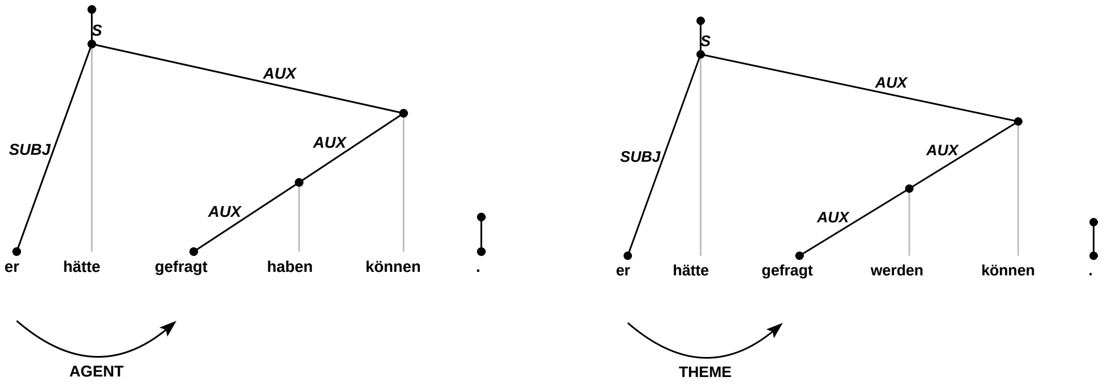
In a passive-voice sentence, the `COMITATIVE` is the prepositional complement `PN` in a 'mit' *with* prepositional phrase `PP` that modifies a non-verbal constituent. Conversely, in a passive-voice sentence, the `PN` of a 'mit' `PP` modifying a non-verbal constituent must originate from the `COMITATIVE`.

Example:
'Er sieht die Frau mit ihrer Freundin.'
He is seeing the woman with her friend.



We concede that from a semantic point of view these syntax-semantics correspondences are not unduly restrictive. Our experimental results reported in Chapters 9, 10, and 11 illustrate, however, that even with these semantically rather loosely cut correspondences very selective syntactic modulations can be effected under cross-modal context integration.
With the use of the extended grammar WCDG2 applies more constraints to its linguistic input than WCDG1. One would therefore expect that the quality of analysis in WCDG2 be higher. An aspect counteracting the benefit from the addition of further constraints is that more levels of analysis produce a larger search space. Whether or not the globally optimal dependency structure is found depends on the effectiveness of the frobbing procedure. Guided by the design principle to leave the central processing mechanisms in WCDG – including frobbing – untouched, WCDG2 operates under these two competing and counteracting influences.
In practice, we find that the majority of the cases in which WCDG2's syntactic analysis incorrectly deviates from the WCDG1 analysis is due to frobbing not finding the correct solution rather than due to incorrect grammar modelling. The cause for failure to produce the correct syntactic analysis can be tested for in WCDG: A deviant analysis can be modified manually in WCDG. Constraint violations are reevaluated after every manipulation. If WCDG2 hence scores a manually corrected structure better than the solution found by frobbing, then the obtained structural deviation is due to a search rather than a modelling error. For a more detailed discussion of the extended grammar's performance on unrestricted input and the challenge of evaluation, see Chapter 8.

From a modelling perspective, the definition of certain syntax-semantics correlations in WCDG2's grammar presented a significant challenge, in particular when resulting from expressivity limitations in WCDG's grammar. Limitations encountered were twofold: First, WCDG only permits a maximum constraint arity of two, i.e., a single constraint can only evaluate properties of up to two edges in the dependency tree.



(a) *He could have (have) asked.* (b) *He could have been asked.*

Figure 5.5: Two sentences illustrating that the check for active/passive voice may involve the evaluation of several dependency edges.

Second, the range of features that can be checked for in a constraint is primarily focused on properties of individual edges. Important supra-local properties, such as active and passive voice in a sentence, often manifest themselves on more than one edge and can even span the entire dependency tree.

One constraint in the syntax-semantics interface defines the subject of an agentive active verb in an active-voice sentence to be the verb's AGENT. In order for this constraint to be evaluated, the global sentence property of active or passive voice needs to be checked for, which may require the evaluation of multi-edge dependency constellations. Figure 5.5 illustrates an example in which the multi-edge dependency constellations of active and passive voice only differ in a single auxiliary ('haben' *have* versus 'werden' been).

In our model implementation's syntax-semantics interface we overcome this challenge by employing a novel method for capturing *global* and *supra-local* sentence properties in WCDG. Our method is based on calls to *ancillary constraints* that check for complex edge properties. A complex property is encoded as a separate constraint. If the property constraints are too complex to be expressed in a single WCDG constraint, ancillary constraints can be constructed that call further ancillary constraints that check for sub-aspects of the complex property and thus effectively decompose the complex property into less complex features that can easily be checked for with WCDG's standard predicates.

As explained in Section 4.2.4, a WCDG constraint can only evaluate the properties of up to two edges as well as their direct neighbours above and below. We extend constraint expressivity by evaluating further ancillary constraints on the neighbouring edges and their neighbours. These consecutive calls of ancillary constraints help extend the reach to neighbouring edge properties of the initial constraint by one edge with each ancillary constraint call.

A simple example is to check whether an edge has label `AUX` and the edge below it also bears the `AUX` label. This check can still be achieved in a single constraint. Once the property of the conditions on the edge below get more complex than just checking for its label, we need to encode the check as a separate ancillary constraint,

```
{X!SYN} : ’AUX above AUX’ : main : 1 :
X.label = AUX
& has( X@id, AUX )
;

{X!SYN} : ’AUX above edge above AUX’ : main : 1 :
X.label = AUX
& has( X@id, ’Is above AUX’ )
;

{X!SYN} : ’Is above AUX’ : ancillary : 1 :
has( X@id, AUX )
;
```

Figure 5.6: Constraints extending their reach to their immediate edge neighbour (`’AUX above AUX’`) and to the neighbour’s neighbour (`’AUX above edge above AUX’` in combination with `’Is above AUX’`).

e.g.: *Check that the edge label is **AUX** and the edge below it lies above another edge that lies above edge with label **AUX**.* The constraints for these checks are exemplified in Figure 5.6.

One limitation this approach has not been able to overcome is that ancillary constraints must be unary and can only be applied on the level of analysis of the calling edge. WCDG’s limitation that constraints can only be defined for edges on a maximum of two levels of analysis thus remains. In our model, this did not impose any fundamental modelling restrictions.[1]

In summary, we can evaluate global sentence properties for trains of contiguous dependency edges on the same level of analysis by consecutive calls to ancillary constraints. More details on our approach of checking global sentence properties with localised constraints are given in McCrae et al. (2008).

5.7 The Extended Lexicon

Semantic processing in WCDG2 requires additional lexical information beyond the information provided in WCDG’s standard lexicon. Most notably, a verb requires a semantic valence that expresses which verb-centred thematic relations it must engage in.

[1] The only constraint which could not be expressed even with the use of ancillary constraints was the precedence preference on participants in an active sentence (AGENT $\succ$ RECIPIENT $\succ$ THEME). Our modelling options were 1) to model this prefence pairwise for the semantic levels of analysis irrespective of active/passive voice or 2) to omit a constraint for this preference altogether. We decided for the first option which has the downside that regular word order in passive sentences violates this soft constraint. Despite the slight reduction of the overall sentence score, this effect had no adverse influence on the correctness of the overall dependency structure for the passive-voice sentences studied with our model.

WCDG's standard lexicon contains syntactic valence definitions for all verbs. Initially, it therefore appeared attractive to derive semantic valence definitions from those existing syntactic valences and simply map them onto each other by a set of correspondence rules. Doing so would bind semantics to syntax and thereby would reduce semantic representation to a mere derivative of syntactic representation. We opted against this approach in order to enable a genuinely bidirectional interaction between the semantic and syntactic levels of representation. Both levels need to build up their own representations that are based on as much independently defined information as possible. If one level of representation were simply a correspondence projection of the other, that level would not contribute any new information to the solution of the constraint satisfaction problem. In fact, the additional level of representation would be nothing more than a rule-based encoding of the original level causing a processing overhead without an actual gain of information.

An important question – both from the perspective of modelling and cognition – is which of a verb's thematic role relations are required for the definition of its core meaning and hence should be included in a lexeme's semantic valence representation. From a modelling perspective, the verb needs to entertain enough thematic relations to permit the mapping of its syntactic arguments to the thematic roles – as required by R24. From a cognitive perspective, only those thematic relations should be included that are semantically integral to the definition of the verb's core meaning. Thematic roles like LOCATION and COMITATIVE, for example, can easily be omitted without violating the semantic completeness of the verb's meaning. These roles may well be part of a situation description centred around an instantiation of the concept activated by the verb — but they are not an integral component of the verb's representation of meaning. On the other hand, the omission of roles such as AGENT or THEME is either semantically completely unacceptable or leads to a significant distortion of the original verb meaning.[1][2]
Ferretti et al. (2001) found that situation verbs prime their typical role fillers for the AGENT and THEME roles. Since these thematic roles typically find their syntactic realisation in the mandatory verbal arguments of subject and direct object, we include them in our list of required verb-centred thematic roles. While not tested for by Ferretti et al., we also treat the role RECIPIENT as essential to verb meaning for situation verbs. Encoding the participant that the result of the situation is directed to (cf. Table 5.1) is an essential aspect of a situation description, which is also reflected

[1] From our constraint-based perspective on language processing this begs the question how hard these semantic constraints actually are. One challenge in answering this question lies in the difficulty of observing such semantic constraint violations in isolation, i.e., unaccompanied by a constraint violation on another level of linguistic analysis. In natural language, likely candidates for hard semantic constraint violations, such as the omission of an AGENT on an agentive verb, always seem to be accompanied by the violation of a similarly hard syntactic constraint.

[2] Many cases of humour, metaphor, or figurative usage derive their communicative effect from the deliberate violation of *soft* semantic constraints. A constraint violation binds some of the interlocutor's cognitive resources in the effort to find an alternative utterance interpretation that results in the removal of the constraint violation or its replacement by a lesser constraint violation.

in the typical realisation as a mandatory indirect object.[1] We choose to exclude the role of INSTRUMENT from the semantic valence definitions. The reason for doing so is that the robust priming effect reported in Ferretti et al. (2001) for this thematic role only pertains to a small set verbs that describe actions which are closely associated with the corresponding INSTRUMENT. Ferretti et al.'s findings support the view that for this group of verbs the INSTRUMENT information does indeed contribute to their generalised representation of meaning. It is unclear, however, how these findings generalise to verbs that describe situation types that are not actions and do not exhibit a close semantic tie with an INSTRUMENT. Also, Ferretti et al.'s do not permit conclusions as to whether a situation verb mandatorily needs to be accompanied by a role filler for the INSTRUMENT role. Sentences containing situation verbs that prime typical INSTRUMENTs, e.g. *to stir* → *spoon* or *to paint* → *brush*, are semantically perfectly acceptable even in the absence of an explicit mention of the INSTRUMENT. Our decision to exclude the role of INSTRUMENT from semantic valence definitions is supported by earlier findings on the strength of inferences from situation verb meaning to its corresponding INSTRUMENT as discussed in Ferretti et al. (2001, pp. 524).[2]

From an implementation point of view, we now need to decide which is the most suitable way to represent semantic valences in the lexicon. As shown in Equations 4.3 and 4.4, increasing the number of homonyms in a slot also increases the overall number of unary and binary constraint evaluations for the corresponding sentence. Decisions regarding lexical representation in WCDG may hence directly affect processing time. One of these decisions is how a verb's semantic valences shall be matched against the corresponding syntactic valences.
WCDG1's underspecified syntactic valence representation collapses several valences into a single valence representation that is assigned to a single lexical entry. This condensed representation has the advantage of reducing the number of homonyms for a given slot string (cf. Section 4.2.1). The disadvantage of the WCDG1-representation of semantic valence is that it potentially overgenerates invalid syntactic valence alternatives as shown in Figure 4.3. If semantic valences were to be represented by similarly condensed – but overgenerating – representations, the number of overgenerated invalid lexical forms would increase multiplicatively (see Figure 5.7 for an example). If, on the other hand, we choose to represent every semantic valence combination explicitly, we increase the number of homonyms to be considered in parsing – without exact prior knowledge on how strongly this increase might affect processing times.

[1] We realise that quoting syntactic evidence in support of our semantic modelling decisions may expose us to Jackendoff's criticism against thematic roles as a *"thinly disguised wild card to meet the exigencies of syntax."* (Jackendoff, 1990, p. 46). We stand up to this criticism by emphasising that our model permits full control over the degree to which syntactic decisions can influence semantic role assignments. In our model semantic processing proceeds on separate levels of analysis according to independently formulated semantic WFRs. The interaction between syntactic and semantic processing is bidirectional and fully open to control via the explicitly stated constraints in the syntax-semantics interface.

[2] This modelling decision proved to be adequate for the majority of verbs studied. A notable exception was provided by the verb 'erliegen' *to succumb to* which, in German, takes a mandatory Dative object that acts as an INSTRUMENT. As a result of our modelling decision, the CIA in its present form does not support semantic processing for sentences containing this verb.

<table>
<tr><td>Lexical Entry</td><td>bezahlen:=[base:bezahlen,cat:VVINF,stress: unstressed,perfect:haben,sem_val:ag_re?_th?, valence:'a?+d?',avz:allowed];</td></tr>
<tr><td>Underspecified Syntactic Valence</td><td>valence:a?+d?</td></tr>
<tr><td>Correct Syntactic Valences</td><td>valence:- | valence:a | valence:a+d</td></tr>
<tr><td>Overgenerated Syntactic Valence</td><td>valence:d</td></tr>
<tr><td>Underspecified Semantic Valence</td><td>sem_val:ag_re?_th?</td></tr>
<tr><td>Correct Semantic Valences</td><td>sem_val:ag | sem_val:ag_th | sem_val:ag_re_th</td></tr>
<tr><td>Overgenerated Semantic Valences</td><td>sem_val:ag_re</td></tr>
<tr><td>Valid Valence Combinations</td><td>sem_val:ag & valence:- |
sem_val:ag_th & valence:a |
sem_val:ag_re_th & valence:a+d |</td></tr>
<tr><td>Overgenerated Valence Combinations</td><td>sem_val:ag & valence:a |
sem_val:ag & valence:d |
sem_val:ag & valence:a+d |
sem_val:ag_th & valence:- |
sem_val:ag_th & valence:d |
sem_val:ag_th & valence:a+d |
sem_val:ag_re_th & valence:- |
sem_val:ag_re_th & valence:a |
sem_val:ag_re_th & valence:d</td></tr>
</table>

Figure 5.7: Example for the multiplicative increase of overgenerated invalid combinations of syntactic and semantic valences for 'bezahlen' *to pay* as produced by systematic expansion of underspecified syntactic and semantic valence representations.

In our modelling effort, performance aspects play a subordinate role to the demonstration of the conceptual feasibility of the context integration. We therefore assign a higher priority to the accuracy and correctness of the lexical representation and the parses resulting from it than to an improved performance of the implementation. Based on this guideline we choose to define a separate lexical entry for each combination of syntactic and semantic valences. For each verb included in the scope of our implementation, we unambiguously specify the permissible combinations of their syntactic and a semantic valences. All other verbs continue to use their lexical representation from WCDG1.

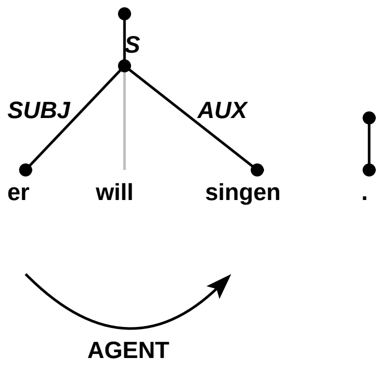


Figure 5.8: The correct semantic analysis according to our semantic modelling approach in which the auxiliary has semantic valence `null` and does not participate in a semantic dependency.

For verbs with multiple syntactic valences this approach counteracts the reduction in lexicon size achieved by the underspecified syntactic valences. At the same time, it eliminates the potentially large number of invalid syntactic and semantic valence combinations resulting from overgeneration. The example in Figure 5.7 shows that for the verb 'bezahlen' *to pay* three individual lexicon entries need to be added to the lexicon. While the condensed representation would only add one entry, it would also give rise to *nine* invalid valence combinations. This example illustrates how the addition of several lexical entries containing the separate semantic valence specifications can in fact be more economic with respect to the generation of valid homonyms than the addition of a single, underspecified entry.

Semantic valences were hand-annotated and have been included into the lexicon as values for the feature `sem_val` for a subset of 1,063 unique verbs.[1] These are the verbs for which thematic role assignments can be performed in the CIA.
Each semantic valence references a verb's mandatory thematic roles by their initial two letters in lowercase. Multiple two-letter references occur in alphabetical order and are separated by an underscore. An exception to this nomenclature is provided by the semantic valence `null` which is assigned to auxiliaries and modals that have been modelled such as not to participate in any semantic dependencies. Figure 5.8 provides the correct semantic analysis according to this modelling approach for the sentence 'Er will singen.' *He wants to sing.* In this example, 'will' *wants* has semantic valence `null` and 'singen' *sing* has semantic valence `ag`.
Exemplary size effects for extending WCDG2's lexicon according to our approach for representing semantic valence are shown in Table 5.3. The additions to the lexicon primarily result from enriching the lexical entries of verbs with the corresponding

[1] Initial attempts to automate the process of semantic valence specification were not pursued further. Semantic valences were extracted from the annotated verb frames in the SALSA Corpus (Burchardt et al., 2006; SALSA Corpus Homepage, 2009), a semantically annotated extension of the TIGER 1.0 Corpus (Brants et al., 2002; TIGER Corpus Homepage, 2009). The approach had to be abandoned due to substantial difficulties in trying to define general rules for mapping the strongly lexicalised roles in the corpus to the more generalised thematic roles in our model implementation.

Category	**Unique Entity Counted**	**WCDG1**	**WCDG2**	**Change**
Verb Infinitives	Semantic Valence `VVFIN, VAFIN, VMFIN`	8,750	9,321	+6.53%
Adjectives	Lexical Baseform `ADJA, ADJD`	9,158	9,164	+ 0.07%
Nouns	Lexical Baseform `NN`	27,473	27,485	+ 0.04%
Proper Names	Lexical Baseform `NE`	30,881	30,884	+ 0.01%
All	Lexical Entry	1,014,864	1,054,451	+ 3.90%

Table 5.3: Size comparison of lexicon components for WCDG1 and WCDG2 based on entity counts in the generated full-form lexicons.

semantic valence information. While the figures in Table 5.3 are implementation specific, they document the predicted increase in lexicon size as a result of the chosen semantic valence representation. The change in lexicon size is mainly due to the addition of new lexical entries for verbs. Observe that the addition of a semantically annotated verb forms to the full-form lexicon adds more than just the single line for the verbal infinitive since, in the process of semi-automatic lexicon generation, the infinitive is expanded into its potentially large set of corresponding inflected forms, each of which is added as a separate entry. Also note that, strictly speaking, the semantically annotated infinitives were not *added* to the existing lexicon. Rather, they *replaced* the corresponding entries that did not carry semantic valence information.

5.8 Chapter Summary

With the implementation of the functional enhancements described in this chapter, WCDG as the central component in the CIA has been prepared for cross-modal interaction. The integration of cross-modal context information as provided by the PPC proceeds via WCDG2's extended predictor interface. WCDG2 submits its homonym-specific input to the PPC and — once the computation of the contextually-informed prediction scores is complete — also receives the PPC's predictions via this interface.

To meet Requirements 1 and 22, WCDG2's interaction with cross-modal context is mediated by a semantic representation involving a set of semantic dependencies. The assignment of these dependencies is constrained by rules in the role-assigning grammar. The role-assigning grammar comprises correspondence constraints from the syntax-semantics interface and context-integration constraints. The latter constrain the assignment of semantic dependencies by propagating the cross-modal prediction scores provided by the PPC into WCDG2's semantic levels of analysis.

The enablement of semantic processing in WCDG2 further necessitated changes to the lexical representations of verb valences. The potentially overgenerating and ambiguous valence representations in WCDG1 were replaced. The improved accuracy in semantic valence representation comes at the price of an increase in the number of homonyms per verb form that need to be evaluated in parsing. Compared with WCDG1, parsing in WCDG2 involves processing more constraints on more levels of analysis with more dependencies and more verbal homonyms for the same slot string. We hence expect the total of these changes to increase the overall processing time in WCDG2.

Chapter 6

Knowledge Representation and Reasoning

The primary objective of this research project is to design and implement a cognitively motivated model in which two distinct and representationally encapsulated symbolic representations, namely those of visual understanding and syntax, are brought to interact with each other. Conceptual Semantics postulates that any cross-modal interaction with syntax is mediated by a shared, uniform level of semantic representation. This representation employs a hierarchy of concepts and their instantiations to encode both, entities represented syntactically as well as entities projected from visual understanding. It is the purpose of the present chapter to outline how the notions of *concept* and *concept instance*, which form the foundation of the model of cognition and cross-modal interaction in Conceptual Semantics, have been realised in our model.

We begin this chapter with an overview over the components required for the representation of semantic knowledge and provide a description of how they interact with each other. Section 6.2 describes how we represent situation-invariant semantic knowledge and world knowledge. Section 6.3 explains which inferences can be drawn in our model and which role these inferences play in the context of cross-modal interaction. Finally, Section 6.4 describes in detail what kind of information we include in our representations of cross-modal context and how we encode that information.

6.1 Overview

As discussed in Section 3.4, semantic representation in the view of Conceptual Semantics comprises two related but distinct aspects: a largely *situation-invariant* hierarchical representation of conceptual knowledge that, broadly speaking, corresponds to semantic memory and a *situation-dependent* representation of experience which represents episodic aspects of memory and cross-modal perception. The situation-dependent episodic representation instantiates concepts from the conceptual hierarchy as activated by perception.

Our model implements this distinction by representing cross-modal context in a *bipartite* declarative knowledge representation. This representation consists of an ontology as the knowledge base of situation-invariant semantic knowledge, the T-Box, and a situation-dependent context model, the *A-Box*, which contains indexed instantiations of T-Box concepts as well as assertions of thematic relations between those concept instantiations.

The reasoner provides inference information about the A-Box and the T-Box which may reveal additional information that has not been asserted explicitly. The PPC establishes communication with the reasoner via a generic application programming interface (API), the *Reasoner API* that permits to query for asserted and inferable information in the A-Box and the T-Box.

6.2 Representing Situation-Invariant Semantic Knowledge

The T-Box is a knowledge representation intended to contain the entire situation-invariant semantic knowledge of the system. In our model, it consists of a handcrafted OWL ontology that defines a hierarchy of *concepts* or *classes* as well as *concept instances* or *individuals*. Between these entities, a number of *relations* with well-defined semantics have been asserted.

We use the terms *'concept'* and *'class'* interchangeably in the following.[1] The subsequent sections describe these ontology constituents in further detail.

6.2.1 The Concept Hierarchy

The concept hierarchy in the T-Box is established by successive assertions of subsumption relations which we denote by *is_a*. The assertions of subsumption relations also comprises the assertion of disjoint classes.[2] The result of defining all concepts, concept instances and relations in the T-Box is an ontology proper with the typical ontological properties such as inheritance between a class and its superclass(es). Apart from its significance for the mechanisms of inference, inheritance has the advantage of representational economy: inherited information only needs to be asserted once, namely for the most general superclass, in order to apply to all subclasses that inherit from it.

To avoid confusion between the name of a concept and its lexicalisations, the latter being invariably German strings in our model, we assign every entity concept an English name. In cases where the English concept name does not reflect the unambiguous gender marking of the concept's German lexicalisation, the gender marking suffix '.M' for 'male' and '.F' for 'female' is appended to the concept name for disambiguation. The corresponding super-concept with underspecified gender definition has been defined in most cases and bears the gender marking suffix '.M.F'.

As for the naming of situation concepts, it turned out to be virtually impossible to find precise translation matches for all German situation verbs in English. We

[1] While the term 'concept' is used with preference in the domain of description logic, 'class' is predominantly used in the OWL community with focus on ontology-based implementations. The same analogy applies for the terms 'concept instance' and 'individual'.

[2] Asserting that two classes A and B are disjoint can also be expressed equivalently in terms of the subsumption relations $A \sqsubseteq (\top \sqcap \neg B)$ and $B \sqsubseteq (\top \sqcap \neg A)$.

Lexicalisation	**Concept Type**	**Situation Arity**	**Concept Name**
‘Mann’	Lexicalised Entity Concept	–	MAN
‘Student’	Lexicalised Entity Concept	–	STUDENT.M
‘Studentin’	Lexicalised Entity Concept	–	STUDENT.F
–	Non-Lexicalised Structure Concept	–	STUDENT.M.F
‘geben’	Lexicalised Situation Concept	Unary	NULL.GEBEN
‘geben’	Lexicalised Situation Concept	Binary	ETW.GEBEN
‘geben’	Lexicalised Situation Concept	Ternary	JMD.ETW.GEBEN

Figure 6.1: Concept naming exemplified for selected entity and situation concepts in the T-Box.

therefore adopt a different approach in the naming of situation concepts: The name of a situation concept consists of the infinitive form of the German verb that activates the concept. The infinitive is preceded by placeholders for the verb’s syntactic arguments. Examples for this concept naming convention are shown in Figure 6.1.

The T-Box contains four types of non-trivial concepts[1]: *entity concepts*, *structure concepts*, *helper concepts* and *situation concepts*. Entity concepts determine concrete or abstract entities in the real world and have a specific lexicalisation in the German language, e.g. MANN or FRAU. Structure concepts have been introduced to improve the representational structure and transparency of the ontology’s concept hierarchy. In most cases, structure concepts do not have a concrete lexicalisation in German. The four concept types just listed have been implemented as structure concepts in the ontology: ENTITY.CONCEPT, STRUCTURE.CONCEPT, HELPER.CONCEPT and SITUATION.CONCEPT. Helper concepts have been introduced in the ontology for purely technical reasons, e.g., to contain a well-defined subset of classes from the T-Box after classification by the reasoner. The content of classes representing these helper concepts in the ontology can conveniently be queried by the PPC via the Reasoner API. Helper concepts can be considered selective filters over the totality of concepts in the ontology. As an example for a helper concept, consider the class LEXICALISED.CONCEPT which contains all concepts for which a lexicalisation has been asserted in the T-Box.

[1] We refer to $\top$ and $\bot$ as *trivial concepts* of an ontology. All other concepts are considered *non-trivial.*

6.2.2 Relations

We define a set of 14 relations in the T-Box. Some of these relations hold between concept individuals, e.g. *is_AGENT_for*, and some between concepts and individuals, e.g. *has_Lexicalisation*. The complete list of modelled relations is given in Figure 6.2. In addition, the specification of the concept hierarchy and the individuals associated with it require the relations *is_a*, which holds between concepts, and *is_instance_of*, which holds between an individual and the concept it instantiates. The latter two relations are provided by the OWL formalism. The social predicates demanded by Requirement R19 have not been included as they were not needed for our modelling purposes.
We have adopted the modelling convention that *has_X* relations are used for assertions in the T-Box and *is_Y* relations are used for assertions in the A-Box. A-Box assertions are discussed in Section 6.4.

Forward Relation	**Function**	**Inverse Relation**
has_AGENT	Relates a situation concept instance to its AGENT.	*is_AGENT_for*
has_COMITATIVE	Relates an accompanied entity to its COMITATIVE.	*is_COMITATIVE_for*
has_INSTRUMENT	Relates a situation concept instance to to its INSTRUMENT.	*is_INSTRUMENT_for*
has_LEXICALISATION	Relates a concept to its lexicalisation.	*is_Lexicalisation_for*
has_OWNER	Relates an owned entity to its OWNER.	*is_OWNER_for*
has_RECIPIENT	Relates a situation concept instance to its RECIPIENT.	*is_RECIPIENT_for*
has_THEME	Relates a situation concept instance to its THEME.	*is_THEME_for*

Figure 6.2: The set of relations defined in the T-Box.

A concept C in the ontology can be assigned a lexicalisation λ_i by asserting the relation *has_Lexicalisation* between C and an instance of the concept LEXICALISATION. The relation can be asserted as the triplet (*has_Lexicalisation*, C, λ_i). Our model permits to assert multiple lexicalisations for a concept: (*has_Lexicalisation*, C, λ_1), (*has_Lexicalisation*, C, λ_2), . . . (*has_Lexicalisation*, C, λ_n). Like this, we can model synonymy, homonymy and polysemy:

Synonymy
$(\mathit{has\_Lexicalisation}, C, \lambda_1) \wedge (\mathit{has\_Lexicalisation}, C, \lambda_2) \wedge \lambda_1 \neq \lambda_2$

Homonymy and Polysemy
$(\mathit{has\_Lexicalisation}, C, \lambda_1) \wedge (\mathit{has\_Lexicalisation}, D, \lambda_1) \wedge C \neq D$

For clarity and economy of representation, the T-Box contains structure classes that group concepts which share certain properties or engage in the same type of relations. As a modelling convenience we assign these shared properties to the superordinate structure class. In our model, the structure classes of situation concepts, for instance, all bear cardinality restrictions on the thematic relations they must engage in.[1] These restrictions are passed on to all the members of the structure class by inheritance. For the structure class TAKES.AGENT.THEME, e.g., a subclass of BINARY.SITUATION (see Appendix II, page 230), the following cardinality restrictions are imposed: Members of this class must not engage in a *has_RECIPIENT* relation and must have exactly one *has_AGENT* relation to an ENTITY.CONCEPT and exactly one *has_THEME* relation to an ENTITY.CONCEPT. Analogous cardinality restrictions apply to other situation concepts in the T-Box. In the subsequent course of this thesis, we frequently refer to two important properties of situation concepts: *situation valence* and *situation arity*. The thematic relation restrictions imposed upon a class determine its situation valence. In our model, situation valence is labelled in complete analogy to the semantic valence of verbs (cf. Section 5.7). The class TAKES.AGENT.THEME, hence, is assigned the situation valence `ag_th`.
In contrast, we take the more general term 'situation arity' to denote the number of mandatory thematic relations that instances of a certain concept must engage in. The class BINARY.SITUATION, e.g., contains all situation concepts of binary situation arity, i.e., all situation concepts that engage in precisely two mandatory thematic relations.

6.2.3 Modelling Domain and Domain Modelling

The modelling domain of the T-Box is based on the concepts activated by the content words in the set of structurally ambiguous sentences for which the influence of cross-modal context upon linguistic processing has been studied. A detailed description of the sentences and the sources from which they have been extracted is provided in Section 8.2.

[1] We adopt this procedure as a general design guideline for restricting the use of relations in the T-Box: Rather than to impose a global domain or range restriction on a relation, we restrict the use of the relation via class-specific properties and hence localise the effect of the restriction. In this manner, we can formulate class properties that, for the members of this class, have the same effect as a global domain restriction on the corresponding relation. For instance, in asserting the cardinality restriction (*has exactly 1*, *property*, CONCEPT) as a class property, we achieve that members of this class must engage in exactly one *property* relation with a member from the class CONCEPT. As a result, members of the restricted class cannot enter a *property* relation with a member of any other class – just as if a global range restriction had been imposed upon the *property* relation.

VK-011	'Er wusste, dass die Magd der Bäuerin den Korb suchte.'
'Er'	gives rise to the concept HUMAN.M.
'wusste'	gives rise to the concept ETW.WISSEN$_{\texttt{ag\_th}}$.
'Magd'	gives rise to the concept MAID.
'Bäuerin'	gives rise to the concept FARMER.F.
'Korb'	gives rise to the concept BASKET.
'suchte'	gives rise to the concepts NULL.SUCHEN$_{\texttt{ag}}$, ETW.SUCHEN$_{\texttt{ag\_th}}$, and JMD.ETW.SUCHEN$_{\texttt{ag\_re\_th}}$.

Figure 6.3: The selection of content words for conceptualisation in the T-Box (underlined) from one of the studied globally ambiguous sentences.

We have conceptualised content words such as verbs and nouns as well as function words such as personal pronouns in the input sentences by the corresponding situation and entity concepts in the T-Box. An illustration of this process is given in Figure 6.3. Instantiations of these T-Box concepts are modelled in the situation-specific A-Boxes to represent a disambiguating cross-modal context. A detailed description of situation modelling is provided in Section 6.4. At the time of writing, the T-Box contains 427 classes, 310 individuals, and 14 relations. A representation of the asserted concept hierarchy in the T-Box is given in Appendix II.

On the first hierarchy level, the T-Box contains four structure concepts that categorise entity concepts, helper concepts, meta data and situation concepts. The class ENTITY.CONCEPT subsumes all concepts that can act as an argument to the relation is_ROLE_for.[1] The class HELPER.CONCEPT subsumes LEXICALISED.CONCEPT which, in turn, subsumes all concepts that have been assigned a lexicalisation in the T-Box.

The class SITUATION.CONCEPT is disjoint from the class ENTITY.CONCEPT and subdivides into classes containing the unary, binary and ternary situation concepts of our model implementation. Each of these subclasses further subdivides into classes that contain situation concepts of the same situation valence only.

META.DATA subsumes the abstract concepts GRAMMATICAL.NUMBER and GENDER. The latter have been introduced to facilitate the adequate modelling of syntactically relevant information. The inclusion of these concepts permits a more accurate semantic representation of visual scenes and thereby increases the specificity of reference formation during cross-modal matching. Without the concepts SINGULAR and PLURAL as subsumed by GRAMMATICAL.NUMBER, concept instantiations would always be underspecified with respect to grammatical number as illustrated in Figure 6.4 (a).

The interpretation of 6.4 (a) shows that the omission of grammatical number from the representation of visual scene context results in a rather crude approximation of the visual scene contents. Cognitively, it is virtually impossible to conceive a scenario in which the accuracy of visual perception is so strongly degraded that information

[1] In $A \xrightarrow{relation} B$ we refer to A as the *relation argument* and B as the *relation value*.

(a) $\text{MAN\_01} \xrightarrow{\textit{is\_instance\_of}} \text{MAN}$

(b) $\text{MAN\_02} \xrightarrow{\textit{is\_instance\_of}} \text{MAN} \sqcap \text{SINGULAR}$

(c) $\text{MAN\_03} \xrightarrow{\textit{is\_instance\_of}} \text{MAN} \sqcap \text{PLURAL}$

Figure 6.4: Concept instantiations in our model representing (a) an unspecified positive number of men, (b) precisely one man and (c) several men.

about the grammatical – not the actual – number of concept instances observed cannot be extracted from the visual modality.[1] Such unusual conditions may, perhaps, be encountered in the presence of extremely poor lighting or in extreme physical distance to the observed scene. At any rate, these are fringe phenomena of marginal importance to a general model for the interaction between visual understanding and linguistic processing. Even if we admit such percepts as possible – which in our model, we do – without incurring undesirable consequences for the more specific representations of visual scene context, it remains questionable how accurate the classification of individuals could be under such limited visibility conditions. As discrimination temporally precedes classification in perceptual bottom-up grounding (cf. Section 3.6), it is plausible that the grammatical number of participants can be assessed prior to their conceptual categorisation.
The reverse, i.e., the classification of participants without a precise knowledge of their grammatical number, seems improbable since grammatical number could, in principle, always be inferred from the number of concept instantiations that have projected into Conceptual Structure during classification. The representations in Figure 6.4 (b) and (c) illustrate how the concepts SINGULAR and PLURAL can be employed to specify concept instances of well-defined grammatical number.

In our model, the expression of quantification and quantifier scope for concept definition is determined by the expressivity of the OWL language. For our modelling purposes, we express quantification as conjuncts of entity concepts with concepts from the class GRAMMATICAL.NUMBER. Presently, this class only containts the subclasses SINGULAR and PLURAL. Other types of quantification cannot be expressed conceptually in the current version of our model. We consider Requirement R17, which demands the capability to express quantification and quantifier scope, partially implemented in our model. A comprehensive coverage of all facets of quantification is likely to require an extensive elaboration of the model.

Whilst the list of concepts subsumed by META.DATA clearly is incomplete (cf. Footnote 4 on page 117), the inclusion of these concepts into the T-Box forms an important first step towards a more precise representation of referentially relevant

[1] Due to the very restricted range of values that grammatical number can adopt in most languages, the perception of the grammatical number of concept instances clearly is considerably easier than the perception of the actual number. In German, the perception of grammatical number only requires the cognitive discrimination between *none*, *one*, and *many*.

information in models of visual context. GRAMMATICAL.NUMBER is an obvious candidate for inclusion in the T-Box since its manifestation is overtly detectable by sensory perception. As far as cross-modal reference to people is concerned, this argument can also be extended to GENDER in most cases. The experimental findings for context integration with sentence SO-9681 in Experiment 3.4 (to be discussed in Section 10.5) also support the view that the inclusion of meta-data such as GENDER can further improve the specificity of bottom-up grounding and cross-modal matching. An analogous argument can be developed for GRAMMATICAL.NUMBER.[1]

6.3 Reasoning and Inferences

The use of declarative knowledge representations without the capability to reason over those representations would be severely limited in a number of respects. Most significantly, without reasoning every piece of knowledge needs to be asserted explicitly. Even for relatively small knowledge bases, the explicit assertion of *all* relevant relations that do and that do not hold true between the represented concepts and individuals would be an arduous and time-consuming task. Worse still, with increasing ontology size, the complexity of the endeavour grows dramatically and soon exceeds the limits of the manageable.

From a representational point of view, it is therefore much more economical to assert a smaller number of independent relations and then to infer further implicit knowledge from reasoning over the representation. Given the well-defined semantics of the relations in an OWL ontology, we can draw inferences based on the systematic behaviour of these relations. The *is_a* relation, e.g., is transitive and hence permits us to infer $(\,is\_a, A, C\,)$ from $(\,is\_a, A, B\,)$ and $(\,is\_a, B, C\,)$. In our model, we use a reasoner that draws inferences over the T-Box and the A-Box and communicates with the PPC via the Reasoner API. The PPC utilises this API to query the ontology.

6.3.1 The Reasoner

In the implementation of our model we use the FaCT++ description logic reasoner (FaCT-PlusPlus Download Page, 2009) to draw inferences over the T-Box and the A-Box. From the many description logic reasoners available, FaCT++ was chosen as most suitable for our needs because it offers a convenient and stable API, supports the decidable OWL description logic dialect OWL DL and also supports the forthcoming description logic standard OWL 2.[2] The latter standard also includes the assertion of cardinality restrictions, which we also employ in a number of class definitions in our T-Box, e.g. for the definition of the class TAKES.AGENT.THEME (cf. Section 6.2.2).

[1] We would go as far as to speculate that the cognitive prominence of these features accounts for the fact that 'number' and 'gender' are grammatical features which undergo marking in most – if not all – natural languages of the world.

[2] We report the use of a different reasoner with an earlier version of our model in McCrae (2007). Reasoning in this earlier version was subject to a number of technical and representational limitations, all of which could be overcome with the inclusion of the FaCT++ reasoner.

6.3.2 Inferences

In our model, the reasoner computes two important types of inference on the asserted concept hierarchy of the T-Box: *concept subsumption* and *concept satisfiability*. The check for subsumption permits to detect class-subclass relationships in the ontology that have not been asserted explicitly. This is particularly useful in large and complex ontologies where this kind of relation may not be apparent from inspection of the class hierarchy any more. The test for subsumption can also be used to obtain the set of all concepts subsumed by a given superconcept. We use this method to retrieve the contents of suitably defined helper classes.

The test for concept satisfiability evaluates whether the addition of a given concept to the knowledge base maintains consistency of the knowledge representation.[1] Of particular importance in the context of our model is that the check for satisfiability also allows us to check for concept compatibility. A pair of concepts A and B from the T-Box is mutually compatible if and only if $A \sqcap B$ is satisfiable in the T-Box. The importance of concept compatibility for the cross-modal influence of visual context upon linguistic processing in our model will be discussed in Section 7.3.

6.3.3 The Reasoner API

The PPC sends queries to the reasoner and obtains reasoning results from it via the Java OWL API (OWL API Homepage, 2009). This API provides a large number of classes and methods that permit to query for almost all ontology properties of relevance to our modelling objective. The API also permits to trigger some of the reasoning operations such as the classification of the asserted hierarchy to afford the inferred hierarchy.

The only case in which we needed to extend the querying capabilities provided by the Reasoner API was for the query about all individuals in the ontology and the classes they instantiate. We use this query to obtain the set of all concept instances asserted in the A-Box. Due to the lack of a convenience method for this information in the Reasoner API, we have added the corresponding query capabilities in the PPC. It now includes a method that parses the assertions of all OWL individuals in a given ontology and extracts the information about which class each of them instantiates.

6.4 Representing Situation-Dependent Visual Context

Following the fundamental tenets of Conceptual Semantics, we assume in our model that situation-dependent percepts are encoded in terms of concept instances joined by thematic relations. We consider the description of the cognitive processes that lead to the generation of these mental representations out of scope of our modelling effort.

[1] An inconsistent ontology may produce inconsistent inferences (*'Ex contradictione sequitur quodlibet'*). The overall consistency of the ontology is therefore a prerequisite for correct and consistent reasoning.

We hence represent the entities that are perceived as participating in the visual scene as well as the thematic relations between them in a context model. We further assume that the cross-modal influence of the visual modality upon linguistic processing occurs due to the represented visual scene being co-present with the linguistic stimulus in the input sentence. Such a co-occurrence is given, for instance, when an individual is exposed to a visual scene context and simultaneously is exposed to the input sentence as an auditory linguistic stimulus. It is the resulting cross-modal influence of visual context upon linguistic processing in such scenarios that we are trying to approximate with our model implementation. In terms of Conceptual Semantics, the A-Box can be seen to contain the projections of entities and thematic relations as identified in the process of visual understanding.

6.4.1 The Contents of Visual Scene Representations

The amount of information that can potentially be extracted from a visual scene is enormous. A number of cognitive top-down processes such as visual attention and context-based expectation, however, help to reduce this large amount of information down to a cognitively manageable set of salient features that are extracted from the visual scene for further processing. Strohner et al. (2000) report a number of experiments that illustrate the strong influence of attentional focus upon cross-modal reference formation in ambiguous cross-modal matching situations. As we deliberately exclude the complexity of these top-down processes from consideration in our model, we need to make appropriate assumptions about the effect that these processes have upon the representation of visual percepts. For our model, we assume that top-down cognitive processes have already effected a pre-selection of entities from visual context. Precisely these entities will be represented and shall be modelled to interact with linguistic processing. In our context models we encode exactly this selected visual scene information. The selection processes that lead to the decision which part of visual context shall be focussed on are considered outside of the present scope of our model.

Visual scenes offering several situations for extraction can be studied with our model by designing a separate context model for each of those situations. Each situation then gives rise to a separate cross-modal interaction with language. Each cross-modal interaction requires a separate parse run with a distinct context model. Our model can therefore only approximate the effect of multistable visual percepts (see Section 3.1) upon linguistic analysis. As the context representations in our model are inherently static, each of the multistable states needs to be represented as a distinct context model that gives rise to a separate cross-modal interaction with linguistic processing.

In order for linguistic processing to be influenced by visual understanding, the representation of visual understanding must contain the linguistically relevant entities and relations. Our context models intend to represent the output of the process of visual understanding. As such, they primarily represent the entities observed in a visual scene, the situation that binds these entities together and the thematic relations that relate the entities to each other. We also include information beyond the visually perceivable when this information is likely to be known or inferable from prior knowledge or world knowledge. An example for this is the visual perception of entities that are identified by their relation to other entities which themselves are not part of the visual scene. Consider the context representation resulting from the visual perception of *Dominik's son*. Prior knowledge identifies the visually perceived person as *Dominik's son*, even if *Dominik* is not part of the visual scene. Our context model of this visual scene will hence include a representation of both entities, SON_01 and DOMINIK_01.

Some thematic roles may be easier to observe visually than others. AGENT and THEME, for example, are generally quite easy to extract from a visual scene, especially, if dynamic rather than static visual scene information is available. The role OWNER, on the other hand, is an example of a role that is more difficult – if not even impossible in some cases – to extract from inspection of a visual scene. For our model, we assume that additional knowledge about the entities perceived in the situation is also incorporated in the process of visual understanding. It is additional knowledge in the form of prior context and world knowledge that permits the assignment of the visually less accessible thematic role OWNER.
As an example, consider Figure 6.5, where one participant has been identified as a PhD student.[1] If we assume that this participant is already known as '*the researcher's PhD student*', the recognition of the entity PHD.STUDENT.F_01 in the visual scene permits the inclusion of the additional, visually inaccessible thematic relation in the output representation of visual understanding. An implication of this argument is that the output representation of visual understanding can, in some cases, include entities that are not even physically present in the observed visual scene. In the example in Figure 6.5, prior knowledge about PHD.STUDENT.F_01 can warrant the inclusion of the *is_OWNER_for* relation with RESEARCHER.F_01 in the context model, even if the latter entity physically is not present in or detectable from the scene. This argument is in line with our approach to use the visual context model as a representation of all linguistically relevant entities and relations identified in the process of visual understanding.

Refinements to the semantic representation of visual scene context such as *modal aspects* or *negation* are demanded by Requirement R20. These aspects have not been incorporated into our representation of visual context to limit the modelling complexity in the interaction with linguistic analysis. We hypothesise that modals and negations differ in their effect on linguistic analysis from factual assertions and hence require different modelling with regards to their effect on the assignment of semantic dependencies in the linguistic analysis. An appropriate modelling approach for these contextual aspects may presumably require different types of inferences and, possibly, even a different logic. We recommend that a systematic investigation into these phenomena be undertaken in the context of future research.

[1] We omit the indication of German gender marking in the English translation of German sentence material, unless it is essential for the argument.

Binary Visual Scene Context:

VK-274 '. . . , dass die (Doktorandin der Forscherin) den Beweis lieferte.'

. . . that the researcher's PhD student delivered the evidence.

Class Assertions in Cross-Modal Context:

$\text{PHD.STUDENT.F\_01} \xrightarrow{is\_instance\_of} \text{PHD.STUDENT.F} \sqcap \text{SINGULAR}$

$\text{RESEARCHER.F\_01} \xrightarrow{is\_instance\_of} \text{RESEARCHER.F} \sqcap \text{SINGULAR}$

$\text{EVIDENCE\_01} \xrightarrow{is\_instance\_of} \text{EVIDENCE} \sqcap \text{SINGULAR}$

$\text{ETW.LIEFERN\_01} \xrightarrow{is\_instance\_of} \text{ETW.LIEFERN}_{\texttt{ag\_th}}$

Object Property Assertions in Cross-Modal Context:

$\text{PHD.STUDENT.F\_01} \xrightarrow{is\_AGENT\_for} \text{ETW.LIEFERN\_01}$

$\text{EVIDENCE\_01} \xrightarrow{is\_THEME\_for} \text{ETW.LIEFERN\_01}$

$\text{RESEARCHER.F\_01} \xrightarrow{is\_OWNER\_for} \text{PHD.STUDENT.F\_01}$

Figure 6.5: The inclusion of the thematic role OWNER into the representation of visual context to reflect the contribution of contextual and world knowledge.

According to Jackendoff's Conceptual Semantics, only the entities that have projected into Conceptual Structure subsequently have the potential to interact with syntactic representation. An entity needs to have been cognised, or — in terms of Conceptual Semantics — must have projected into Conceptual Structure in order to be able to affect linguistic processing (Jackendoff, 1983, p. 35). We consequently require that only those entities may exert an influence upon linguistic processing that have been represented in the context model. Effectively, this assumption provides a closure on the default *open-world assumption* of OWL reasoning. As our model centres around a constraint-based linguistic processor, we need to make this *closed-world assumption* to be able to derive constraints on linguistic analyses that do not receive the support of positive evidence in visual context. A purely OWL-based formalism does not provide closed-world inference mechanisms. We hence implement these inferences in the PPC at a process stage posterior to communication with the OWL reasoner. A detailed description of the inferences resulting from this closure is provided in the description of the PPC's scoring algorithm in Section 7.4.

A mental representation of cross-modal context according to Conceptual Semantics is a representation of cognised entities, encoded as concept instances and thematic relations between them. The creation of such a representation presupposes the identification of perceived entities and hence can only be populated by the process of

Visual scene context:

A man is giving a woman a book.

Context Model Class Assertions:

MAN_01 $\xrightarrow{is\_instance\_of}$ MAN $\sqcap$ SINGULAR

BOOK_01 $\xrightarrow{is\_instance\_of}$ BOOK $\sqcap$ SINGULAR

WOMAN_01 $\xrightarrow{is\_instance\_of}$ WOMAN $\sqcap$ SINGULAR

JMD.ETW.GEBEN_01 $\xrightarrow{is\_instance\_of}$ JMD.ETW.GEBEN$_{\texttt{ag\_re\_th}}$

Context Model Property Assertions:

MAN_01 $\xrightarrow{is\_AGENT\_for}$ JMD.ETW.GEBEN_01

BOOK_01 $\xrightarrow{is\_THEME\_for}$ JMD.ETW.GEBEN_01

WOMAN_01 $\xrightarrow{is\_RECIPIENT\_for}$ JMD.ETW.GEBEN_01

Figure 6.6: Typical assertions contained in a context model.

visual understanding. We also expect a representation of the output of visual understanding to comprise additional cognitively relevant information such as spatial, temporal or causal relations. In the current form of the model, however, this type of information is not captured in our representation of visual context. Due to the absence of spatial information in our context models, our representation of visual context in its current form also fails to meet Requirement R16 for pointers to sensory representations.
The contents of a context model comprise instances of acting entities or, more generally, situation entities, instances of actions or, more generally, situation concepts and thematic relation assertions between those instances. As such, our knowledge representation of visual context satisfies Requirement R18 for the abstract representation of actions and acting entities. Typical context model assertions are exemplified in Figure 6.6.

To achieve a further reduction of the modelling complexity in the interaction between visual understanding and linguistic processing, we make another assumption: The representation of cross-modal context as provided by the A-Box remains valid throughout the course of sentence processing. We hence assume that the visual scene information provided at the onset of linguistic processing is static and remains unchanged by the interim and final results of linguistic processing. This assumption hence excludes the possibility that linguistic processing influences the process of visual understanding. In natural systems, the latter kind of interaction is observed frequently, for example, in cases where local syntactic or referential ambiguities are resolved by means of visual information in the course of incremental sentence processing. In those cases, the disambiguated linguistic information is found to have a directing effect on the course of eye fixations in a co-present visual scene (Tanenhaus et al., 1995, Ambiguous conditions of Experiments 1 and 2). A full model of the cross-modal interaction between vision and language — as opposed to a model of the cross-modal influence of vision upon language — will need to incorporate such *bidirectional* cross-modal interactions. Their exclusion from the scope of our model results from a technical limitation of the parser's predictor interface: the predictor interface only permits the unidirectional integration of non-linguistic information. A bidirectional interaction at parse time is currently not possible.[1]

Assuming that the representation of visual context is static over parse time justifies the use of a predictor for the computation of thematic relation scores prior to parse time. If the contextual representation remains unaffected by the course of linguistic processing it makes no difference whether we query that representation prior to or during the process of parsing.

A criticism that has been raised against our format of context representation is that the influence of the visual context upon linguistic processing is not *cross-modal* in nature anymore.[2] We argue against this view for two reasons: First, our model does indeed influence the processing of one representational modality based on information contained in another representational modality: the semantic representation of visual context modulates linguistic analysis on the syntactic level of representation. Importantly, both of these representational modalities are independent of each other and are representationally encapsulated in the sense of Jackendoff (1996). Second, the information encoded in the representation of visual context is an adequate approximation of some of the non-linguistic information that is readily available from inspection of a visual scene. As none of the entities or relations used to encode visual context information have linguistic properties, the nature of this representation as well as the information it encodes are genuinely non-linguistic.

6.4.2 Representing Situation Entities and Participants

Every entity in visual context is encoded as an instantiation of a *single* concept that has been *asserted* in the T-Box. For technical reasons, instantiations of multiple or anonymous classes[3] are not possible in the current implementation: While OWL and the reasoner permit the assertion of concept instances of anonymous classes in principle, the axioms resulting from this kind of instantiation cannot be parsed by the current version of the PPC. This technical limitation can be circumvented simply by asserting an additional class in the T-Box that has the properties of the desired anonymous class. This new class can then conveniently be instantiated in the context model.

[1] It is currently planned to overcome this limitation in the context of a separate research project dedicated to the extension of the parser's capabilities to enable online access to contextual information at parse time.

[2] This criticism was voiced by an anonymous reviewer of McCrae (2009).

[3] An anonymous class is a class that has not been asserted explicitly but can be expressed by a description logic expression of asserted classes which is satisfiable in the T-Box.

Since every concept instance in the A-Box grounds precisely one asserted concept in the T-Box, Requirement R25 is met. This requirement demands that every concept instance must instantiate at least one concept from the concept hierarchy. In combination with the technical limitation of the current PPC version, this requirement has the representational consequences we have just outlined. It is important to stress, however, that this requirement does not constitute a limitation on which concept instances can be expressed in a context model. In principle, any concept can be instantiated in an A-Box provided it has been asserted in the T-Box beforehand. Missing concepts can simply be added to the T-Box, as long as they are expressible in the OWL description logic.

An analogous approach can be taken to model conceptual uncertainty in visual perception. In cases where visual perception is ambiguous, the visual percept can be defined as an instantiation of a class with sufficiently vague class definition. For example, if a situation participant was observed and it could not be discerned whether the person was a man or a woman, this person could be represented in the context model as an instantiation of the asserted T-Box concept PERSON defined in 6.1:

$$\text{PERSON} \equiv \text{MAN} \sqcup \text{WOMAN} \tag{6.1}$$

Analogously, the ambiguous visual percept of an entity categorisable as instantiating either of the two mutually exclusive concepts A or B can be represented as instantiating concept C defined according to 6.2.[1]

$$\text{C} \equiv (\,\text{A} \sqcap \neg\text{B}\,) \sqcup (\,\neg\text{A} \sqcap \text{B}\,) \tag{6.2}$$

Another representational implication arising from the implementation is that for every individual it should be asserted which class it instantiates. While OWL and the reasoner in principle permit the assertion of OWL individuals without an explicit class assertion, the reasoner treats such individuals as belonging to the class OWL:THING by default.

6.4.3 Representing Thematic Roles

A fundamental assumption underlying our modelling approach is that the thematic relations between situation entities in a visual scene can be extracted from a visual scene and are a part of the output of visual understanding. We hence assume that an observer arrives at an interpretation of a visual scene in which each of the observed entities is assigned a thematic role.

Thematic role assignments to entities in visual context are represented as assertions of an is_ROLE_for thematic relation between two concept instances. Depending on the type of thematic relation asserted, the relation can hold between instances of two entity concepts or between instances of a participant and a situation concept. The set of thematic relations supported in our model is given in Figure 6.2. In

[1] Note that the assertion of C in 6.2 defines a conceptual underspecification that is *diachronically invariant*. This type of ambiguity is different from the type of concept assertion that would be required to express a bistable visual percept that oscillates over time between instantiations of concepts CONCEPT.A and CONCEPT.B.

fulfilment of Requirement R26 our situation representations are always verb-centric in the sense that they centre around an instance of a situation concept which, in German, typically are lexicalised by a verb. As the assignment of thematic relations to entities in visual context is situation specific, such assertions are represented in the A-Box.

In our representations of visual context, only one thematic relation assertion can be processed for a given pair of individuals. The model does not support uncertain or ambiguous role assignments. To model the effect of a visual scene in which it is unclear whether an individual A_01 engages in an *is_AGENT_for* or an *is_THEME_for* relation with an instance of situation concept SITUATION.CONCEPT_01, we need to model each visual context in a separate context model.

6.5 Chapter Summary

In this chapter we have argued for a bipartite knowledge representation to encode the linguistically relevant aspects from the outcome of visual understanding. The representational division into a situation-invariant T-Box and a situation-dependent A-Box reflects the representational requirements of Conceptual Semantics. In our model, the T-Box corresponds to semantic memory and encodes temporally invariant lexical knowledge in a hierarchy of concepts, a hierarchy of thematic relations and a set of situation-invariant concept instances. The A-Box encodes episodic aspects of visual understanding and contains situation-specific instantiations of T-Box concepts joined by thematic relations from the T-Box.

The hierarchy of concepts results from multiple assertions of *is_a* relations between concepts. Thematic relations are asserted betwen individuals from the classes in the conceptual hierarchy. In addition, the T-Box specifies relations between individuals and concepts, such as the *is_instance_of* relation, and between concepts and individuals such as the *has_Lexicalisation* relation. The inclusion of a reasoner permits to check whether a given proposition is consistent with the asserted hierarchy and the set of defined relations and individuals. As such, the form of knowlege representation and reasoning we choose for our model meets Requirement R15 for such reasoning capabilities.

We argue for the adoption of a closed-world assumption for the cross-modal influence of visual context upon linguistic processing. Our modelling approach is based on the premise that both positive and negative evidence from a visual scene context can influence linguistic processing. We further stress the importance of reasoning for the efficient extraction of implicit knowledge from a knowledge representation as well as for the economy in the representation of semantic knowledge in general.

The next chapter discusses in detail how the PPC uses the information in the T-Box and the A-Box in combination with the closed-world assumption to compute dependency score predictions for the semantic dependencies between homonyms in the input sentence. We illustrate in detail how our model utilises the reasoning operations of concept subsumption and concept satisfiability that were introduced in this chapter to compute dependency score predictions for linguistic processing.

Chapter 7

The PPC — A Cross-Modal Predictor Component

The PPC is the central component in our model. It computes the semantic dependency scores based on which parser-external non-linguistic context information influences linguistic processing in WCDG2. With its scores, the PPC determines which effect a semantic relation asserted in the representation of visual context shall have upon linguistic processing.

This chapter describes in detail the steps performed by the PPC, from processing the initial sentence input to the eventual return of a homonym-specific list of contextually-informed dependency scores to WCDG2. The detailed understanding of the PPC's decision processes in computing semantic dependency scores will also be essential for an appreciation of the experimental results discussed in Chapters 9, 10 and 11.

Section 7.1 describes how the predictor receives its input information from WCDG2. Section 7.2 explains how we achieve linguistic grounding of the words in the input sentence in our model. Section 7.3 addresses the decision process implemented in the PPC for matching words from the input sentence to assertions of concept instances in the context model. Section 7.4 provides a detailed discussion of the PPC's scoring algorithm. In Section 7.5 we outline how different effects of perceptual uncertainty have been incorporated in our model. Section 7.6 describes how the PPC communicates its score predictions back to WCDG2 and makes them accessible at parse time.

7.1 Predictor Invocation

The integration of non-linguistic context information begins with the invocation of the PPC predictor by WCDG2. Resulting from the extensions to the predictor interface (cf. Section 5.4), WCDG2 provides the predictor with the list of *all* homonyms in the input sentence as well as their full lexicon information. The limitation of unspecific predictor request, Limitation L1 in Section 4.2.5, has thus been overcome.

According to Requirement R22, the cross-modal interaction between non-linguistic context and linguistic processing must be semantic in nature. At the point in time when the PPC receives the input from WCDG2 via the extended predictor interface, the homonyms have not been assigned a conceptualisation yet. In terms of linguistic grounding, their linguistic symbol has been identified but the symbol's meaning has not been assigned yet. The PPC converts the inherently meaningless representation of lexically specified homonyms into symbolic representations, which are linked to an intrinsic representation of meaning (cf. Section 5.4). In order for a homonym's symbolic representation to be linked to a representation of meaning, it must have one or more conceptualisations. The next important processing step in our model therefore is to establish the link between the homonyms as arbitrary linguistic symbols and concepts in the ontology. We describe this process in detail in the following section.

7.2 Linguistic Grounding

The grounding of conceptual categories in sensory perception in natural systems is a complex, bidirectional process during which bottom-up and top-down processes converge to produce a link between one or more conceptual categories and the sensory representation of the input stimulus. As discussed in Section 3.6, the processes for grounding in sensory and representational modalities differ. In representational modalities, the sensory stimulus, e.g. the sound wave pattern conveying a spoken word, encodes a symbol rather than a real world entity or situation. The word itself cannot be grounded in the sensory stimulus since the word is an arbitrary symbol and, as such, intrinsically meaningless. To assign a meaning to the sensory stimulus an additional step is required: the word's meaning needs to be accessed in the mental lexicon (see Figure 3.5).

A model for the interaction between visual and linguistic understanding must perform the grounding process in both modalities. In our model we assume that the instantiation of concepts in the context model already has occurred; the percepts in the visual modality are therefore already grounded. In the linguistic modality, WCDG2 identifies different homonyms that have been matched against their corresponding lexical entries. These homonyms result from successful discrimination and identification of the linguistic input in Harnad's sense. Since our model does not include analogous processes for the visual modality, Requirements R27 and R28 for the discrimination and identification of sensory input are fulfilled by the linguistic modality only. We consequently rate these requirements as partially fulfilled by our model. Complete fulfilment of these requirements can only be achieved by a model with full sensory processing in both modalities.

According to our definition in Section 3.6, linguistic grounding is complete when the identified symbols in the linguistic modality have been assigned a meaning. In our model, we approximate the assignment of meaning to a word by a *bottom-up* process based on lexical properties. This modelling decision implements Harnad's bottom-up grounding discussed in Section 3.6 and arises from the fact that our model does not incorporate top-down effects such as the influence of context-based expectations or world knowledge in the assignment of concepts to words. For a given word, we refer to a concept representing the word's meaning as its *conceptualisation*. The set of concepts available in our model is represented in the T-Box.[1] In our model, a word is mapped to its set of conceptualisations based on the three lexical features *normalisation*, *semantic valence* and *grammatical number*. The internal representation of these lexical features form a categorical representation of the linguistic input in the sense of Harnad (1990).

The normalisation is computed for every homonym in the input sentence. It can be thought of as a generalised lexical base form that is available for *all* homonyms — in contrast to the extant feature `lexical baseform` in WCDG which remains undefined for all input words that are not contained in the lexicon, have not matched an existing word template or bear the POS tag `NE` for proper names. Depending on which lexical template[2] an unknown word matches, its `lexical baseform` takes the uninformative value `'unknown'` or `'-'`. For a homonym that has an informative lexical baseform listed in its lexicon entry, the normalisation is set to its lexical baseform. For a homonym that has no or no informative lexical baseform listed, the normalisation is set to the its surface string. To eliminate encoding issues in the course of this process, German special characters (umlauts and 'ß') are transliterated into their two-letter equivalents during normalisation.[3] In our model, a word maps to a concept if and only if the following three conditions are met:[4]

1. The conceptualisation has the homonym's normalisation as one of its asserted lexicalisations.[5]
2. The conceptualisation has a situation valence equal to the homonym's semantic valence (if defined).
3. The conceptualisation is compatible with the homonym's grammatical number (if defined).

The effect of applying these criteria in linguistic bottom-up grounding is illustrated in Figure 7.1 for different homonyms of the same transitive verb. As can be seen, the application of these grounding criteria ensures that every homonym is mapped to another conceptual expression. In some cases, a homonym in the input sentence of course may also be mapped to an empty set of conceptualisations by the end of that process. This happens if no concept in the T-Box meets all three conditions for that homonym. The experimental findings reported in Part III of this thesis show,

[1] A detailed description of the T-Box is provided in Section 6.2. The full list of concepts asserted in the T-Box is given in Appendix II.

[2] WCDG's lexical templates were introduced in the list of grammar elements on page 62.

[3] The conventional mappings apply: 'ä' ⟼ 'ae', 'ö' ⟼ 'oe', 'ü' ⟼ 'ue' and 'ß' ⟼ 'ss'.

[4] From the perspective of cognitive linguistics, the list of lexical constraints imposed is clearly incomplete. We expect lexical and semantic features such as gender, person or animacy to influence the process of word understanding as well. However, the results reported in Part III will demonstrate, that the three selected lexical features above permit to obtain convincing results.

[5] In word interpretation, the PPC only considers a concept's asserted lexicalisations. The fact that a concept may also inherit lexicalisations from its superclasses in the ontology remains inconsequential in our implementation of bottom-up grounding.

that a contextual influence can be exerted upon syntactic analysis even if conceptualisations are not available for all homonyms in the input sentence. Note that the failure to assign a homonym its conceptualisation also propagates into cross-modal matching for that homonym (cf. Section 7.3): a homonym which does not map to a concept from the T-Box also cannot be matched with a concept instance in the representation of visual context.

A salient feature of our model is that a single homonym can map to a whole set of conceptualisations rather than just a single conceptualisation. This set-based *one-to-n* mapping permits to model linguistic phenomena such as homophony and lexical ambiguity robustly (see Section 6.2.2). The unweighted representation of word meaning as a set of conceptualisations suggests that all conceptualisations contribute to the meaning of a homonym to an equal extent. Such a uniform semantic representation of word meaning is in discord with notion of multi-facetted word meanings as deducible from human preferences in lexical semantics. Moreover, human preferences are dynamic and can be influenced by factors such as discourse context or world knowledge (e.g., Crain and Steedman, 1985). In lexical disambiguation, for instance, different readings of an utterance are adopted with different degrees of preference. Stronger preferences have precedence but may be dropped in favour of alternative readings once other, even stronger semantic factors enforce an alternative analysis. To achieve a more adequate representation of observable semantic saliencies in word meaning, the ability to represent gradients of semantic preference in the mapping of homonyms to conceptualisations should be included in future extensions of our model.[1]

With the successful mapping of a homonym to a set of conceptualisations from the T-Box the linguistic bottom-up grounding process for that homonym is complete. This step has attributed one or more conceptualisations to the linguistic symbol of a homonym in the input sentence. These conceptualisations are given by concepts in the T-Box. In order for the homonym's semantic representation to interact with cross-modal context in the A-Box, it must now be matched to the representational entities in the model of cross-modal context. This mapping is achieved in the process of cross-modal matching which will be discussed in the following section.

7.3 Cross-Modal Matching

Cross-modal matching in natural systems as introduced in Section 3.7 refers to the establishment of co-reference between representational entities from different modalities. In our model, we face the challenge of matching homonyms, whose meaning is expressed in terms of concepts in the concept hierarchy, with sets of concept instances in the representation of visual context. Effectively, this process results in the creation of cross-modal referential links between entities in the linguistic modality

[1] In first approximation, semantic preferences could be modelled by the inclusion of normalised weights that reflect the contribution of each conceptualisation. Ideally, these weights should be context-sensitive rather than static.

Homonyms

H_1 `schenkt:=[base:schenken,cat:VVFIN,...,person:third, number:sg,...,sem_val:ag_re_th,valence:'a+d',...];`

H_2 `schenkt:=[base:schenken,cat:VVFIN,...,person:second, number:pl,...,sem_val:ag_re_th,valence:'a+d',...];`

H_3 `schenkt:=[base:schenken,cat:VVFIN,...,person:third, number:sg,...,sem_val:ag_th,valence:a,...];`

H_4 `schenkt:=[base:schenken,cat:VVFIN,...,person:second, number:pl,...,sem_val:ag_th,valence:a,...];`

Concepts

C_1 ETW.SCHENKEN ⊓ SINGULAR
`has_lexicalisation:schenken, situation valence:ag_th,`

C_2 JMD.ETW.SCHENKEN ⊓ SINGULAR
`has_lexicalisation:schenken, situation valence:ag_re_th`

C_3 ETW.SCHENKEN ⊓ PLURAL
`has_lexicalisation:schenken, situation valence:ag_th,`

C_4 JMD.ETW.SCHENKEN ⊓ PLURAL
`has_lexicalisation:schenken, situation valence:ag_re_th`

Cross-Modal Matching

$$\begin{tikzcd}
H_1 \arrow[d, "\textit{normalises to}"] & H_2 \arrow[d, "\textit{normalises to}"] & H_3 \arrow[d, "\textit{normalises to}"] & H_4 \arrow[d, "\textit{normalises to}"] \\
\begin{array}{c}\texttt{schenken}\\ \{C_1, C_2, C_3, C_4\}\end{array} \arrow[d, "\textit{compatible valence}"] & \begin{array}{c}\texttt{schenken}\\ \{C_1, C_2, C_3, C_4\}\end{array} \arrow[d, "\textit{compatible valence}"] & \begin{array}{c}\texttt{schenken}\\ \{C_1, C_2, C_3, C_4\}\end{array} \arrow[d, "\textit{compatible valence}"] & \begin{array}{c}\texttt{schenken}\\ \{C_1, C_2, C_3, C_4\}\end{array} \arrow[d, "\textit{compatible valence}"] \\
\begin{array}{c}\texttt{ag\_re\_th}\\ \{C_2, C_4\}\end{array} \arrow[d, "\textit{compatible number}"] & \begin{array}{c}\texttt{ag\_re\_th}\\ \{C_2, C_4\}\end{array} \arrow[d, "\textit{compatible number}"] & \begin{array}{c}\texttt{ag\_th}\\ \{C_1, C_3\}\end{array} \arrow[d, "\textit{compatible number}"] & \begin{array}{c}\texttt{ag\_th}\\ \{C_1, C_3\}\end{array} \arrow[d, "\textit{compatible number}"] \\
\begin{array}{c}\textsc{singular}\\ \{C_2\}\end{array} & \begin{array}{c}\textsc{plural}\\ \{C_4\}\end{array} & \begin{array}{c}\textsc{singular}\\ \{C_1\}\end{array} & \begin{array}{c}\textsc{plural}\\ \{C_3\}\end{array}
\end{tikzcd}$$

Figure 7.1: The effect of the three implemented criteria in linguistic bottom-up grounding for the present tense indicative `VVFIN` homonyms of 'schenkt' *give(s).*

to entities in the visual modality. The underlying idea of our model architecture is that the process of assigning semantic dependencies in linguistic processing shall be influenced by cross-modal context information if the homonyms to be scored in the linguistic analysis have a cross-modal match in visual context.

Similarly to grounding, cross-modal matching in natural systems is a bidirectional process susceptible to bottom-up and top-down influences. Top-down influences such as expectations arising from a percept in one modality will influence likely matching candidates in the other modality. For example, hearing the noise of a loud diesel engine approach when attempting to cross a road will trigger visual search for the corresponding large vehicle; the perceived auditory stimulus will *not* be attributed to the bicycle seen passing at the same time. Conversely, noticing a heavy truck approach without hearing the corresponding motor noise would give rise to an extremely bewildering percept.[1]

In our model implementation we reduce the complexity of the cross-modal matching process down to a single criterion: *concept compatibility*. A linguistic entity is modelled to be co-referent with an entity in visual scene context if its conceptualisation is compatible with the concept instantiated by the visually observed entity or entities. This, clearly, is a simplification in several respects. First of all, we assume that the given utterance is about the visual scene thus and makes reference to the entities or situations in the visual scene. Roy and Mukherjee (2005) refer to this approach as the *assumption of immediate reference*. We hence assume that the natural language utterance refers to the immediate visual scene context represented in the context model. This assumption may not hold for all cross-modal interactions between vision and language since not all situated utterances actually make reference to entities in the scene in which they are being uttered.

Secondly, we assume that cross-modal reference is established between entities that activate concepts which are semantically consistent or compatible with each other. While, in first approximation, this assumption is plausible for descriptive utterances, there are a number of linguistic devices such as *irony* or *sarcasm* which may not obey this rule.

Thirdly, conceptual compatibility is a weaker criterion than actual co-reference. Concept compatibility is a necessary but not a sufficient criterion for co-reference. An illustration of this point is given in Figure 7.2: The presupposition arising from the use of the definite article in 'der Mann' *the man* results in a strong preference for the interpretation that 'der Mann' and 'der Schauspieler' *the actor* do *not* refer to the same male individual, despite the fact that these words have conceptually compatible conceptualisations.

[1] The deliberate violation of such cross-modal top-down expectations based on world-knowledge has occasionally been used for artistic effect, e.g. in the deliberately bewildering cinematographic art of the French *Nouvelle Vague* director Alain Resnais (*1922).

Input Sentence:

'Der Mann sieht den Schauspieler in einem Kinofilm.'
The man is seeing the actor in a movie.

From the T-Box:

$$(\textit{is\_satisfiable}, \text{MAN} \sqcap \text{ACTOR}, \text{T-Box}) = \textit{true}$$

Preferred Interpretation:

$$\begin{array}{llll} & \text{`der Mann'} & \xrightarrow{\textit{refers to}} & \text{MAN\_01} \\ \wedge & \text{`der Schauspieler'} & \xrightarrow{\textit{refers to}} & \text{ACTOR\_01} \\ \wedge & \text{MAN\_01} \neq \text{ACTOR\_01} & & \end{array}$$

Figure 7.2: In the majority of cases, concept compatibility is a necessary – but not a sufficient – criterion for co-reference.

A homonym may have several meanings, each of which can be compatible with a different concept instance in the representation of visual context. As a result, a homonym can match an entire set of entities in visual context. The mapping from homonym to concept instances hence need not be injective (one-to-one) or surjective (onto-mapping), let alone bijective (both one-to-one and onto-mapping). All of the matched entities, however, must instantiate a concept compatible with at least one of homonym's conceptualisations.

The cross-modal matching example in Figure 7.3 illustrates a case in which no homonym matches more than one instance in visual context, which actually is a special case. Due to the comparative looseness of the applied cross-modal matching criterion of concept compatibility the mapping turns out to be non-injective in the majority of cases. In particular, semantically underspecified word classes such as pronouns tend to map to several entities in visual context. In analogy to the uni-modal linguistic bottom-up grounding of homonyms, our set-based approach permits the robust handling of the influence of lexical ambiguity and homophony upon cross-modal matching as well. Words that have several distinct meanings also have the potential to refer to different entities in a visual scene context.
In fulfilment of Requirement R1 our model implements the process of cross-modal matching as mediated by a representation of word meaning. The experimental findings by Cooper (1974) and Huettig and Altmann (2005) presented in Section 2.2 further support this modelling decision. With our model's focus on the influence of visual context upon linguistic processing this realisation of cross-modal matching maps entities from the linguistic input to concept instantiations in the representation of visual context. It therefore fulfils Requirement R29 for the formation of cross-modal referential links from the linguistic to the non-linguistic modalities. As we have excluded the cross-modal interaction in the reverse direction from the modelling scope, our model does not meet Requirement R30 for establishing cross-modal referential links in the reverse direction.

Input Sentence:

‘Er hört die Männer singen.’
He hears the men sing.

Class Assertions in Cross-Modal Context:

MAN_01	$\xrightarrow{is\_instance\_of}$	MAN $\sqcap$ SINGULAR
MAN_02	$\xrightarrow{is\_instance\_of}$	MAN $\sqcap$ PLURAL
ETW.HOEREN_01	$\xrightarrow{is\_instance\_of}$	$\text{ETW.HOEREN}_{\texttt{ag\_th}}$
NULL.SINGEN_01	$\xrightarrow{is\_instance\_of}$	$\text{NULL.SINGEN}_{\texttt{ag}}$

Object Property Assertions in Cross-Modal Context:

MAN_01	$\xrightarrow{is\_AGENT\_for}$	ETW.HOEREN_01
MAN_02	$\xrightarrow{is\_THEME\_for}$	ETW.HOEREN_01
MAN_02	$\xrightarrow{is\_AGENT\_for}$	NULL.SINGEN_01

Bottom-Up Linguistic Grounding:

‘Er’ *He*	$\xrightarrow{is\_conceptualised\_by}$	{MALE $\sqcap$ SINGULAR}
‘hört’ *hears*	$\xrightarrow{is\_conceptualised\_by}$	$\{\text{ETW.HOEREN}_{\texttt{ag\_th}}\}$
‘die’ *the*	$\xrightarrow{is\_conceptualised\_by}$	{}
‘Männer’ *men*	$\xrightarrow{is\_conceptualised\_by}$	{MAN $\sqcap$ PLURAL}
‘singen’ *sing*	$\xrightarrow{is\_conceptualised\_by}$	$\{\text{NULL.SINGEN}_{\texttt{ag}}\}$

Cross-Modal Matching:

‘Er’ *He*	$\xrightarrow{matches}$	{MAN_01}
‘hört’ *hears*	$\xrightarrow{matches}$	$\{\text{ETW.HOEREN}_{\texttt{ag\_th}}\text{\_01}\}$
‘die’ *the*	$\xrightarrow{matches}$	{}
‘Männer’ *men*	$\xrightarrow{matches}$	{MAN_02}
‘singen’ *sing*	$\xrightarrow{matches}$	$\{\text{NULL.SINGEN}_{\texttt{ag}}\text{\_01}\}$

Figure 7.3: An example of cross-modal matching based on concept compatibility.

Concept compatibility depends on the semantic properties asserted for that concept in the T-Box. The concept’s position in the T-Box conceptual hierarchy as well as its list of disjoint classes determine which other concepts it is compatible with. An illustration of cross-modal matching based on concept compatibility is given in Figure 7.3.

The implicit assumption that all members of the set of cross-modal matches are equally likely cross-modal referents of a given homonym clearly constitutes a simplification. In humans, the preference for one lexical reading over another one is

expected to propagate from linguistic grounding into cross-modal matching (cf. Section 7.2): a preference for one specific conceptualisation of a homonym should also result in a preference of the homonym's cross-modal match. Future extension of our model should incorporate the ability to propagate semantic saliencies from homonym grounding into the process of cross-modal matching.
As outlined above, cross-modal matching on the basis of conceptual compatibility may result in cross-modal matching ambiguity in cases in which the mapping from homonym to concept instance is not injective. Since this seems to be the norm rather than the exception, natural systems apply additional criteria to reduce ambiguity in cross-modal matching. One factor to establish cross-modal matching preferences is the degree of conceptual fit: a homonym that matches several context entities will preferentially match that entity which exhibits the largest conceptual overlap with the homonym's preferred conceptualisation. An implementation of this notion in our model would require a gradable measure of conceptual overlap in addition to a weighted representation of word meaning. At the level of implementation described in this work, neither of these has been included.

We finally need to explicate the minimal conditions under which our model can produce a cross-modal influence of visual context upon linguistic processing. Our model is based on the notion that only those semantic relations in visual context can give rise to dependency score predictions which have been asserted between entities that match homonyms in the input sentence. Our model thus requires at least two homonyms from different slots to have different cross-modal matches in order for the context model to be able to affect linguistic processing. Otherwise, the PPC cannot make a context-based prediction and all context-based predictions for semantic dependencies will default to unity. We now give a complete description of the PPC's scoring algorithm in the following section.

7.4 Relation Scoring

The scoring of thematic role dependencies between homonyms is the central mechanism via which the PPC influences the subsequent parsing process in WCDG2. The fundamental idea is that the assertion of a thematic relation in the representation of visual context can affect semantic dependency assignments in linguistic processing. A thematic relation assertion between two entities E_1 and E_2 in the context model can have an influence upon linguistic processing if and only if the homonyms between which a semantic dependency is to be assigned have E_1 and E_2 as cross-modal matches. This is to say: a semantic dependency between a dependant and a regent will be affected by visual context if the visual context contains entities that are identified as cross-modal referents of the dependant and the regent, respectively. If this is the case, the PPC admits the semantic dependency that corresponds to the thematic relation asserted in visual context by assigning it a score of `1`. We refer to this admission based on thematic relation assertions as an influence by *positive evidence*.

As discussed in Section 4.2.5, a predictor influences the dependency assignment in the parser by penalising certain dependencies. Assigning an acceptance score of `1` as such does not have a constraining effect on the assignment of dependencies in the parser. To be able to derive constraints from visual context information, we need to adopt the closed-world assumption as introduced in Section 6.4.1. Imposing a closure on information that has not been asserted in the representation of visual context enables us to derive constraints from *positive and negative evidence* in the context model. As an example, consider two homonyms that match contextual entities between which no thematic relations have been asserted. The open-world assumption that normally applies in OWL reasoning does not permit to draw any constraining conclusions as long as no explicit negative evidence is available. Under the closed-world assumption, we can now infer that all semantic dependencies must be penalised between those homonyms. The closed-world assumption implies that the admission of a semantic dependency in linguistic processing is only possible based on the explicit assertion of the corresponding thematic relation in the context model. Conversely, we can infer from the absence of such an assertion that the thematic relation is not detected in the visual scene. The absence in the visual scene consequently motivates the veto of the corresponding semantic dependency in linguistic processing.

Situations evolve over time, and so does the perception of them. It is therefore a common occurrence that our knowledge of a visual scene increases or changes over time. In this context the question arises how the closed-world assumption is compatible with incremental changes to the representation of visual context. There is no immediately apparent reason why a visual scene containing a number of participants — say a person buying a book, which involves an AGENT and a THEME for the binary buying situation ETW.KAUFEN — should not be expandable by another participant such as a RECIPIENT, at a later stage of observation. The later inclusion of an additional participant into the situation changes the the situation arity from binary to ternary. We fully acknowledge the cognitive reality of such incremental expansions of visual scene interpretations and their corresponding mental representations. The point is that the addition of further participants to the interpretation of the visual scene gives rise to a semantically different scene interpretation and hence a different scene representation in its own right. In our model, a binary situation concept is treated as genuinely distinct from the corresponding ternary situation concept, even if both of them are lexicalised by the same verb.[1] Conceptually, they differ because they describe situations of different situation valence, i.e., situations involving different sets of participants with a different number of mandatory arguments in their syntactic and semantic representations. This semantic difference is reflected in the concepts' different syntactic realisations: Higher situation arity is typically realised syntactically in the subcategorisation of additional arguments on the situation verb. We expect that the instantiation of these different situation concepts in visual context should also give rise to different cross-modal interactions with linguistic processing.

[1] The lexicographic discussion of whether the transitive and the ditransitive form of a verb are indeed manifestations of the *same* verb or of two *distinct* verbs is outside of the scope of this thesis.

The CIA permits to model such extensions of situation valence over time, e.g., from `ag_th` to `ag_re_th` as in the example above. However, it is not possible to model this extension as a single interaction between visual scene context and linguistic processing. Instead, we need to render the two different visual contexts as discrete and distinct representations. Each context model can then engage in a separate – and potentially different – cross-modal interaction with linguistic processing in the course of a separate parse run. Admittedly, this is a work-around since the change in contextual information does not effect a re-analysis of an existing parse in our model as would be the case in a natural system; rather, the effect of the temporally posterior context model upon linguistic processing is modelled as a completely new cross-modal interaction in an independent *ab initio* parse run. Chapter 11 discusses our empirical investigations of the effect of situation arity in the visual scene upon the interaction between context and linguistic processing.

In conclusion, our model is based on the assumption that context models are *informationally complete* representations of the output of visual scene comprehension. By 'informationally complete' we mean that all semantic information required for the cross-modal interaction with linguistic processing is encoded in the representation of visual context. Clearly, this is a modelling idealisation since real-world perception may be subject to uncertainty.
Entities and relations that have been represented can participate in a cross-modal interaction with linguistic processing while those that have not been represented cannot. Imposing the closed-world assumption on the representation of visual context permits us to formulate constraints on linguistic analysis based on both positive and negative evidence from visual context. Based on the closed-world assumption we can exclude any information outside of the context model from consideration in the cross-modal interaction with linguistic processing.

Given this general overview over the PPC's use of positive and negative contextual evidence, we now provide a more detailed discussion of the complete set of inferences and underlying assumptions used in the PPC's computation of score predictions between two homonyms. We then integrate the declarative description of the scoring process into a procedural context and describe the PPC's scoring algorithm in its entirety. In our description we use the following symbols:

S_i	the i-th slot in the sentence,
$H_{i.j}$	the j-th homonym in S_i,
t	an arbitrary thematic relation in the modelling scope of WCDG2's context models,
ϵ	the empty, non-thematic relation which is assumed to hold between two entities in the context model for which no thematic relation t has been asserted,
$\mathbb{T}$	the set of all thematic relations t in WCDG2,
$\mathbb{T}_\epsilon$	the set of all thematic relations t in WCDG2's modelling scope extended by the empty relation ϵ,
$\delta(t)$	the semantic dependency in WCDG2 that corresponds to the thematic relation t in the context model,
$p(H_{i.j}, H_{m.n}, \delta(t))$	the PPC's score prediction for the dependency $\delta(t)$ between dependant $H_{i.j}$ and regent $H_{m.n}$,
$v(\delta(t))$	the penalty score assigned in WCDG2 for the semantic dependency $\delta(t)$,
$\mathbb{M}(H_{i.j})$	the set of all concept instances in the context model that match $H_{i.j}$ cross-modally,
$M(H_{i.j})$	an arbitrary element of $\mathbb{M}(H_{i.j})$, and
$\theta(M(H_{i.j}), M(H_{m.n}))$	a thematic or empty relation that holds between $M(H_{i.j})$ and $M(H_{m.n})$ in the context model.

As discussed in Section 3.4, a major postulate of Conceptual Semantics – and a direct consequence of Jackendoff's Conceptual Structure Hypothesis (cf. p. 38) – is that representationally $\delta(t)$ and t are the same type of relation, namely semantic relations in Conceptual Structure assigned on the single and uniform level of semantic representation for both linguistic and non-linguistic input. In the description of our model we deliberately list them as separate entities since the two relation types are assigned in different technical components of the model: t-relations are asserted in the A-Box while $\delta(t)$-relations are assigned on the semantic levels of analysis in the parser. The assertion in different technical components, however, does not permit the conclusion that the assigned relations represent conceptually different relation types in our model. The fact that they are assigned in different components is the result of purely technical constraints in the implementation. The identity of these relations is encoded in our model; without this identity a cross-modal interaction between visual context and linguistic processing could not be achieved. We exploit the identity of these relations when assigning $\delta(t)$-dependencies in the parser: The assignment is made based on PPC score predictions that reflect the result of queries to a context model containing only t-type relations. Consulting t-relation scores for the assignment of δ-relations is a meaningful procedure if and only if we assume t- and δ-relations to be of the same nature. This is the technical realisation of the central tennet of Conceptual Semantics, namely that cross-modal context and the semantic part of linguistic analysis project into the same semantic representation and consequently also make use of the same type of semantic relations between projected entities.

Given $\theta \in \mathbb{T}_\epsilon$ such that $\theta = \theta(M(H_{i.j}), M(H_{m.n}))$, the PPC draws the following inferences to compute its score predictions:

Inference 1. Veto all unscored semantic dependencies for this dependant-regent pair.

These dependency vetoes are based on the fact that only those semantic dependencies shall be admitted for which positive evidence has been asserted in visual context. If a given thematic relation from $\mathbb{T}$ has not been asserted in the context model, we veto its corresponding semantic dependency in the linguistic analysis, provided that dependency has not been scored yet. This inference applies regardless of whether θ is a thematic or the empty relation. If visual context provides positive evidence for a thematic relation, this pre-assigned veto will be overwritten in Inference 2. If no positive evidence is found to overrule this pre-assigned veto, the veto persists.

$$\forall\, t \in \mathbb{T}: \quad p(H_{i.j}, H_{m.n}, \delta(t)) = \texttt{NULL}$$
$$\implies \quad p(H_{i.j}, H_{m.n}, \delta(t)) \longleftarrow v(\delta(t))$$

If a non-empty thematic relation has been asserted in the context model between $M(H_{i.j})$ and $M(H_{m.n})$, continue with the following inferences:

Inference 2. Admit the semantic dependency that corresponds to the contextually asserted thematic relation.

The central idea of the PPC's scoring policy is to admit those semantic dependencies for which positive evidence in the form of an asserted thematic relation from $\mathbb{T}$ could be found in visual context. Note that the admission of positive evidence in this step is performed regardless of whether any scores have been assigned to $\delta(\theta)$ before. Previously assigned vetoes will thus be overwritten.

$$\theta \in \mathbb{T}$$
$$\implies \quad p(H_{i.j}, H_{m.n}, \delta(\theta)) \longleftarrow 1$$

Inference 3. Veto the reverse direction of the admitted semantic dependency, provided it has not been scored yet.

This dependency veto is based on the fact that our semantic dependencies are unique in a given situation, i.e., a dependency can only be admitted once per situation. Admitting the semantic dependency in the forward direction in Inference 2 therefore permits us to exclude the admittance of the same dependency in the reverse direction.

$$\theta \in \mathbb{T} \quad \wedge \quad p(H_{m.n}, H_{i.j}, \delta(\theta)) = \texttt{NULL}$$
$$\implies \quad p(H_{m.n}, H_{i.j}, \delta(\theta)) \longleftarrow v(\delta(\theta))$$

Inference 4. Veto *all* unscored semantic dependencies for dependants from the same dependant slot.

Only those semantic dependencies for the dependant slot can be admitted for which positive evidence is found in the context model. All other semantic dependencies for the dependant slot are vetoed.

$$\theta \in \mathbb{T}, \forall\, t \neq \theta, \forall\, k \neq j, \forall\, o, \forall\, p : \quad p(H_{i.k}, H_{o.p}, \delta(t)) = \texttt{NULL}$$
$$\Longrightarrow \quad p(H_{i.k}, H_{o.p}, \delta(t)) \longleftarrow v(\delta(t))$$

Inference 5. Veto *all* unscored semantic dependencies for regents from the same regent slot.

Only those semantic dependencies with the regent slot can be admitted for which positive evidence is found in the context model. All other semantic dependencies with the regent slot are vetoed.

$$\theta \in \mathbb{T}, \forall\, t \neq \theta, \forall\, k \neq n, \forall\, o, \forall\, p : \quad p(H_{o.p}, H_{m.k}, \delta(t)) = \texttt{NULL}$$
$$\Longrightarrow \quad p(H_{o.p}, H_{m.k}, \delta(t)) \longleftarrow v(\delta(t))$$

The veto scores from Inference 2 to Inference 5 are inferred whenever a pair of homonyms has cross-modal matches between which a thematic relation has been asserted in the context model. The complete PPC scoring algorithm is given as pseudocode in Algorithm 1. Note that the inferred vetoes on the semantic dependencies can only be imposed because of the closed-world assumption. Essentially, we are using the positive evidence for θ to infer a whole range of other semantic dependency scores. With these inferences in place, our model meets Requirement R8 which demands that the cross-modal interaction between visual context and linguistic processing be based on a representation of the visual context information.

A point worth discussing is the scope of the vetoing we apply. Inference 4 and Inference 5 impose specific vetoes that allow the context model to have a powerful yet selective influence upon linguistic processing. Concretely, these two inferences leave $p(H_{i.j}, H_{m.n}, \delta(t))$ untouched and veto all other semantic dependencies that originate from a dependant homonym in S_i or that are directed towards a regent homonym in S_m.

An alternative approach would have been to extend vetoing to *all* homonyms in the entire sentence such that only those semantic dependencies would be admitted for which a corresponding thematic relation has been asserted in visual context. A crucial effect of this approach is that semantic dependencies are vetoed between homonyms which refer to entities entirely unrelated to the situation encoded in the context model. This constitutes a significant challenge when multiple situations are expressed in a single sentence, as frequently is the case in unrestricted natural language. A simple example is shown in Figure 7.4. If vetoes were applied to all semantic dependencies across the entire sentence, a thematic relation asserted between SHE_01 and NULL.TANZEN_01 would have an effect upon the dependency assignment between ‘Er’ *he* and ‘beobachten’ *observe*.

Algorithm 1 The PPC scoring algorithm for semantic dependency scores.

```
Require: Sentence
for $i = 1$ to number of slots do
  for $j = 1$ to number of homonyms in dependant slot $S_i$ do
    for $m = 1, m \neq i$ to number of slots do
      for $n = 1$ to number of homonyms in regent slot $S_m$ do
        if $\mathbb{M}(H_{i.j}) \neq \{\}$ and $\mathbb{M}(H_{m.n}) \neq \{\}$ then
          for all $t \in \mathbb{T}$ do
            if $p(H_{i.j}, H_{m.n}, \delta(t)) = \texttt{NULL}$ then
              $p(H_{i.j}, H_{m.n}, \delta(t)) \longleftarrow v(\delta(t))$          // Inference 1
            end if
          end for
          for all $D \in \mathbb{M}(H_{i.j}), R \in \mathbb{M}(H_{m.n})$ do
            $\theta \longleftarrow \theta(D, R)$
            if $\theta \in \mathbb{T}$ then
              $p(H_{i.j}, H_{m.n}, \delta(\theta)) \longleftarrow 1$          // Inference 2
              if $p(H_{m.n}, H_{i.j}, \delta(\theta)) = \texttt{NULL}$ then
                $p(H_{m.n}, H_{i.j}, \delta(\theta)) \longleftarrow v(\delta(\theta))$          // Inference 3
              end if
              for all $H_{o.p}$ in the sentence do
                for all $t \in \mathbb{T}$ do
                  for all $H_{i.k} \neq H_{i.j}$ in $S_i$ do
                    if $p(H_{i.k}, H_{o.p}, \delta(t)) = \texttt{NULL}$ then
                      $p(H_{i.k}, H_{o.p}, \delta(t)) \longleftarrow v(\delta(t))$          // Inference 4
                    end if
                  end for
                  for all $H_{m.k} \neq H_{m.n}$ in $S_m$ do
                    if $p(H_{o.p}, H_{m.k}, \delta(t)) = \texttt{NULL}$ then
                      $p(H_{o.p}, H_{m.k}, \delta(t)) \longleftarrow v(\delta(t))$          // Inference 5
                    end if
                  end for
                end for
              end for
            end if
          end for
        end if
      end for
    end for
  end for
end for
```

In principle there are two ways to address this challenge: We can limit the scope of the vetoes applied or we can choose to represent all situations expressed linguistically in the context model. The latter is undesirable for two reasons: 1) A sentence may express situations that are inaccessible to visual perception or that do not refer to the co-present visual context. 2) We do not wish to impose constraints on the amount of visual context to be represented. In some cases, the visual information available at the time of sentence processing may be limited to one specific situation; in other cases, visual context may provide a plethora of visually accessible information with a multitude of observed thematic relations. In either case, visual information should only affect semantic dependencies between those homonyms that are directly or indirectly related to the entities between which the thematic relation is being observed. By *direct relation* we mean a cross-modal reference relation based on concept compatibility, by *indirect relation* we refer to a connection via the inference mechanisms just outlined.

For our model this means that the effect of cross-modal context must remain neutral with respect to linguistic processing unless the asserted thematic relation θ holds between two concept instances D and R, respectively, such that $D \in \mathbb{M}(H_{i.j})$ and $R \in \mathbb{M}(H_{m.n})$. The example in Figure 7.4 illustrates that there is no reason why the vetoes resulting from the AGENT-relation between SHE_01 and NULL.TANZEN_01 as asserted in the context model should give rise to a veto on the AGENT-dependency between 'Er' *He* and 'beobachtet' *observes* in the introductory main clause. The restriction of vetoing scope is hence a modelling decision of particular relevance to the processing of longer sentences. The latter are frequently encountered in unrestricted natural language input and typically contain several verb forms, each of which require independent semantic dependency assignment.

7.5 Perceptual Uncertainty

Every sensory perception is afflicted with a degree of perceptual error. This inherent uncertainty of sensory perception propagates into the representational level as well. If the identifying features of an instance – be it an object instance or a word token – are perceived with an accuracy of 70%, say, we cannot expect the categorisation performed on the basis of these discerning features to be free of error. The challenges resulting from uncertainty affect representation and processing. We expect that the representations arising from sensory input also include information about the degree of perceptual certainty of its representata. Further, processing of uncertain information needs to comprise the systematic propagation of uncertainties. Most importantly, perceptual uncertainty is not a unidimensional phenomenon but can affect various aspects of perception. Consider the processes of sensory perception and subsequent understanding of a visual scene. Uncertainty may affect the perception of spatial and temporal relations. Uncertainty may also affect other dimensions of visual perception such as the conceptual categorisation of the situation instance, the identity of the entities participating in it or the thematic relations that are perceived to hold between those entities. The integration of all facets of uncertainty into a model of cross-modal interaction thus introduces additional levels of complexity, both in representation and processing.

In an attempt to reduce modelling complexity, we largely exclude the effects of uncertainty from consideration in our model. We make the simplifying assumption that the discrimination and categorisation of word tokens in text input is achieved with absolute certainty. We further assume that the class assertions for concept instances in the context model are free of error. We do, however, incorporate categorisation

Input Sentence:

'Er beobachtet, wie sie tanzt.'
He observes how she is dancing.

Class Assertions:

$$\text{SHE\_01} \xrightarrow{is\_instance\_of} \text{SINGULAR} \sqcap \text{FEMALE}$$

$$\text{NULL.TANZEN\_01} \xrightarrow{is\_instance\_of} \text{NULL.TANZEN}$$

Object Property Assertions in Cross-Modal Context:

$$\text{SHE\_01} \xrightarrow{is\_AGENT\_for} \text{NULL.TANZEN\_01}$$

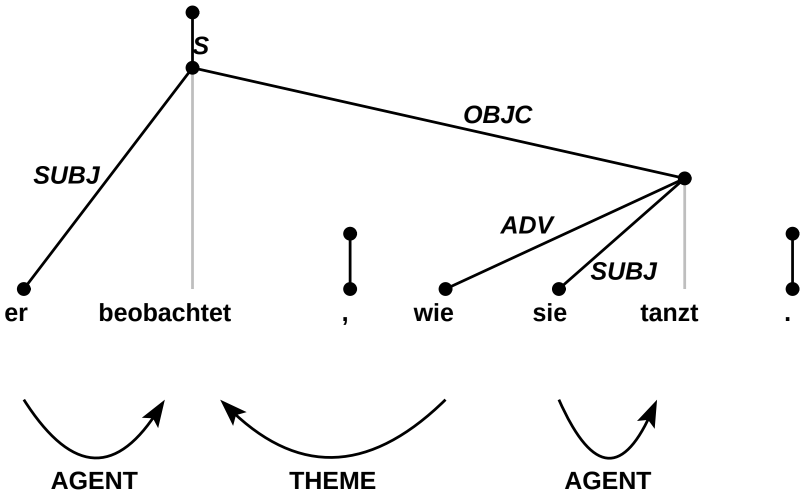


Figure 7.4: The importance of veto scope limitation as illustrated by the assignment of AGENT-dependencies in multiple-situation sentences.

uncertainty and ambiguity of visually perceived entities. These types of uncertainty are represented by underspecified entity and situation concepts as illustrated by the expressions in Equations 6.1 and 6.2 on page 111. We provide a validation of our model's ability to process conceptually underspecified concept instances in Experiment 4 reported in Chapter 11.

We further expect that in a natural system the perceptual certainty of one modality has an influence upon the strength with which it can influence other modalities in cross-modal interaction. The more reliable the percept in one modality is, the more difficult it should be to overrule the information it provides with conflicting information from another modality. This compensatory effect of cross-modal interactions has been shown for numerous modalities. A frequently cited example is the supporting effect of lip reading on phoneme categorisation in audio-visual speech recognition under different signal-to-noise ratios in the auditory channel (see Potamianos et al., 2001, 2004, e.g., for detailed comparisons of the performance of natural and artificial systems).

In our model, the dominance of the visual modality can be achieved by adjusting the numeric value of the veto score assigned to $\delta(t)$. The model permits to specify a separate value of $v(\delta(t))$ for each semantic dependency. $v(\delta(t))$ is the WCDG2 prediction score assigned in the absence of positive evidence for t in the context model and hence also the numerical value of the score penalty incurred by a dependency edge in the linguistic analysis that violates the integration constraint for $\delta(t)$ dependencies. The harder the integration constraint, the more strongly the compliance of $\delta(t)$ dependencies is enforced with the assertions of t relations in the context model. We chose to model $v(\delta(t))$ in a linear relation with *context compliance* γ, the extent to which a semantic dependency assignment is enforced to align with the assertion of the corresponding thematic relation t in the context model. The parameter γ thus effectively models the strength of the influence of visual context upon linguistic processing. The implemented relation between γ and $v(\delta(t))$ in our model is given in Equation 7.2.

$$\gamma \in [0, 1] \tag{7.1}$$

$$v(\delta(t)) = 1 - \gamma \tag{7.2}$$

7.6 Result Communication

The PPC returns to WCDG2 the list of homonyms from the input sentence. For each homonym, the PPC quotes its slot string, the slot number, the identifier that uniquely identifies the homonym in its slot and a set of attribute-value pairs. Each attribute-value pair consists of an attribute identifier that uniquely denotes the scored semantic relation in the sentence and a prediction score as the corresponding value. The semantic dependency identifier consists of a semicolon-separated concatenation of the semantic dependency label, the regent slot number and the regent-homonym's identifier. The beginning of a typical line of PPC output returned to WCDG2 is shown in Figure 7.5.

```
Beide 1 beide_PIDAT_acc RECIPIENT;3;drängten_VVFIN_first_past_- 0.1 ...
```

Figure 7.5: The beginning of a single line of PPC output as received via WCDG2's extended predictor interface.

The unique dependency identifier in combination with the slot number permits the unambiguous encoding of all homonym-specific dependency predictions in a sentence. This encoding of homonym-specific predictions fulfils Requirement R31 for the homonym-specific generation of predictions. WCDG2's four-place `predict()`-function (cf. Section 5.5) maps its input parameters to this identifier and can thus retrieve the corresponding homonym-specific prediction score for all homonyms in the input sentence.

7.7 Chapter Summary

This chapter concludes the description of the individual components making up the CIA. We have provided a comprehensive description of how the PPC, as the central component in the CIA, uses non-linguistic information from the semantic representation of cross-modal context to compute its contextually-informed dependency score predictions.

PPC processing starts with the pre-processing of the linguistic input received from WCDG2. Based on the three lexical features *normalisation*, *semantic valence* and *number*, homonyms from the input sentence are assigned a set of conceptualisations from the T-Box. In the subsequent process of cross-modal matching, the PPC maps each homonym in the input sentence to a set of concept instances asserted in the context model. A cross-modal match is established if and only if at least one of the homonym's conceptualisations is conceptually compatible with the concept instantiated in the context model. The overall flow of cross-modal matching in the PPC is summarised diagrammatically in Figure 7.6.

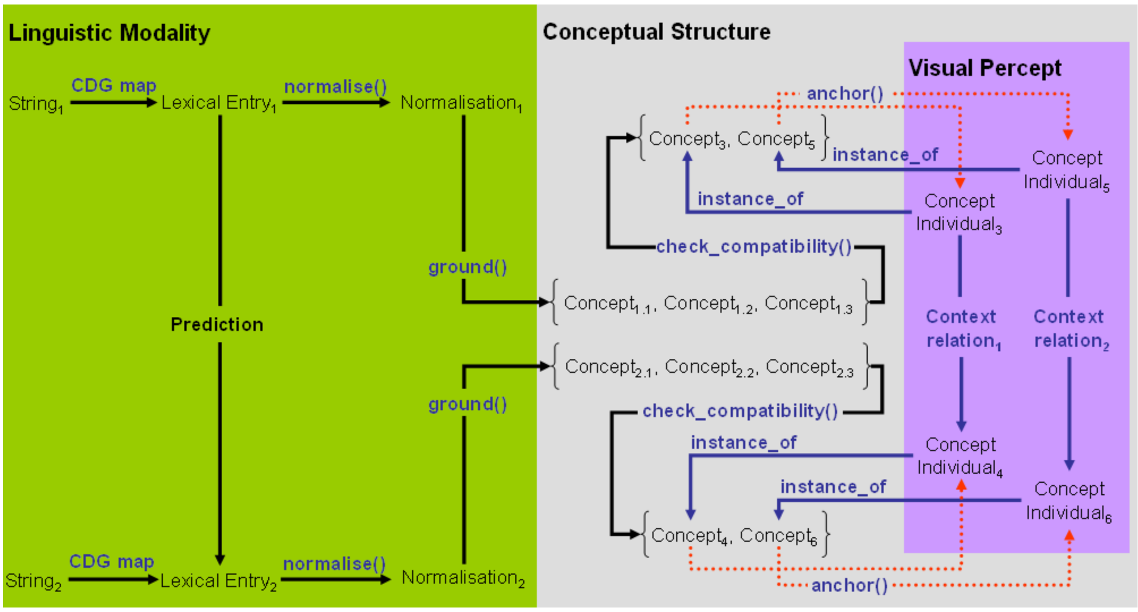


Figure 7.6: Overall process flow in the PPC.

Prediction scoring is triggered if both dependant and regent homonyms have cross-modal matches. For homonym pairs in which both the dependant and the regent have a cross-modal match, the PPC admits the semantic dependency that corresponds to the thematic relation asserted between the cross-modal matches in the context model. Based on the closed-world assumption, all other semantic dependencies originating from the same dependant slot or directed towards the same regent slot are vetoed. When the dependency score predictions have been computed for all eligible homonym pairs, the PPC returns these predictions to WCDG2 where they can be accessed homonym-specifically by the integration constraints in the role-assigning grammar.

A significant strength of our model is that the steps in the process of establishing cross-modal referential links between words and concept instances in visual context are essentially language-independent. Language-independence is also a central claim that Jackendoff makes for Conceptual Structure. While the individual features used in linguistic grounding may be language specific in our model, the overall process which connects input words to conceptualisations via linguistic bottom-up grounding and checks these concepts for compatibility with the concepts instantiated in visual context, generalises to languages other than German.

Part III

Model Validation and Conclusions

The argument for our model of the cross-modal influence of visual scene context upon linguistic processing presented in this thesis is structured into three main parts: Part I was dedicated to the identification and formulation of modelling requirements. Part II served the purpose of providing a detailed specification of the model as well as an in-depth discussion of the extent to which the requirements from Part I have been met by our model implementation. The third part of the thesis now addresses the empirical investigation and validation of the implemented model and discusses the model's behaviour under different experimental conditions.

Each of the subsequent chapters has a specific experimental focus. Chapters 9, 10 and 11 build upon each other in that each subsequent chapter releases one simplifying assumption that was maintined in the preceding chapter or chapters. Chapter 8 describes a pre-experiment to the actual study of our model's behaviour. As context integration is mediated by the semantic level of analysis, the model's integration success crucially depends on the quality of semantic analysis. We therefore evaluate the effect that adding a semantic level of analysis has upon the quality of syntactic analysis in WCDG in the absence of any contextual information. Chapter 9 reports the first actual integration experiment with the CIA: We demonstrate the technical feasibility of context-driven syntactic modulations by enforcing an absolute dominance of visual context over linguistic analysis. This is achieved by integrating visual context information into linguistic analysis via hard integration constraints. Chapter 10 discusses the effect of relaxing the integration constraints while leaving all other experimental conditions unchanged. In Chapter 11 we remove the simplifying assumption that visual and linguistic representations be of the same level of conceptual specificity. The chapter examines how conceptually underspecified representations of visual percepts can still contribute to syntactic disambiguation in the linguistic analysis. We also discuss an experimental investigation into how concept instantiations of different conceptual specificity vary in their ability to induce syntactic modulations under context integration. Chapter 12, finally, concludes this thesis with a summary of the central claims, a collection of the conclusions we draw and an outlook to future directions of research that arise from the work presented here.

Chapter 8

Semantic Grammar Evaluation

This chapter describes two pre-experiments in preparation of the actual study of our model's context integration behaviour. We evaluate the effect that the addition of semantic levels of analysis in WCDG2's extended grammar has upon the quality of syntactic parsing. We also evaluate the extended grammar on two other corpora and select sentences for the subsequent study of context integration phenomena. As regards the overall line of argument for our model, this chapter takes a preparatory function to motivate the selection of the studied linguistic material in the forthcoming context integration experiments.

This chapter is structured into two main sections that correspond to the evaluations of the extended grammar we conducted: Section 8.1 describes the extended grammar evaluation on 1,000 sentences from the NEGRA corpus. Section 8.2 reports the evaluation of the extended grammar on three smaller sets of globally ambiguous sentences that were extracted from a psycholinguistic examination and the SALSA corpus.

8.1 Evaluation on the NEGRA Corpus

8.1.1 Experimental Motivation

In the preceding chapters, we have argued extensively for a context-integration model based on the propagation of non-linguistic context information into syntactic analysis via a shared semantic representation. The interaction between the semantic and the syntactic representations in this model is enabled by correspondence rules in the syntax-semantics interface. Ideally, this interface should propagate referentially relevant context information into syntactic representation while remaining neutral with respect to syntactic analysis in case of referentially unrelated contextual assertions. These modelling aspects have previously been captured as Requirements R5 and R7. Before we set out to study the effect of non-empty context models on syntactic analysis in the following chapters, we need to understand whether the addition of semantic processing has an influence on syntactic analysis in the absence of a contextual bias.

8.1.2 Approach

To see whether the addition of the semantic levels of analysis in WCDG2 has an influence on syntactic analysis, we compare the syntactic parsing accuracy of WCDG2 under integration of an empty context model with the results obtained for WCDG1 on the same corpus. An empty context model contains no assertions of concept instances or thematic relations.[1] We refer to the WCDG2 parse runs as Experiment 1.1. For convenience of expression, we use the general term *accuracy* to denote parsing quality which, more accurately, is quoted in terms of the standard measures *precision*, *recall* and their resulting f_1*-measure*. For our evaluations, we parse sentences from the NEGRA corpus (Skut et al., 1997; NEGRA Homepage, 2006). The NEGRA corpus is a standard corpus of German which has been used extensively for parsing evaluations of WCDG1 on previous occasions. We compare our results against the accuracies reported by Foth and Menzel (2006b) and Khmylko et al. (2009) for WCDG1 evaluations on the same set of sentences.

8.1.3 Setup

We parse sentences 18,602 to 19,601 from the NEGRA corpus. This set of sentences was also used in the reference evaluations of WCDG1 by Foth and Menzel and Khmylko et al. The sentences are parsed with WCDG2's extended grammar under integration of an empty context model. Evaluations are performed against a manually corrected version of the gold standard annotations. Manual correction removed some known orthographic mistakes and amended a few obvious annotation inconsistencies. Following the practice adopted in the cited prior work, we report the *structural* and the *labelled* measures precision, recall and f_1-measure. The structural measures refer to edges that have been structurally correctly attached, irrespective of whether they have been labelled correctly. The labelled measures refer to edges that have been correctly attached *and* correctly labelled. We evaluate parsing accuracy on all sentences with and without punctuation marks to ensure the comparability of our results with prior work. While Foth and Menzel reports parsing accuracy for all edges including those originating from punctuation marks, Khmylko et al. exclude those edges from their evaluation. The latter approach is becoming standard evaluation practice nowadays.

8.1.4 Results

Of the 1,000 sentences in the parsed corpus subset, WCDG2 was found to process only 865 sentences to completion. For the remaining 135 sentences, the parser aborted processing for technical reasons prior to completion. An analysis of the number of the tokens per sentence reveals a clear trend: with an average of 34.8 tokens

[1] Note that the effect of integrating an empty context model with respect to the context-driven modulation of syntactic dependencies is equivalent to parsing with the extended grammar without invoking the PPC at all. An empty context model contains no potential referents for the homonyms in the input sentence. Consequently, an empty context model offers no cross-modal match candidates and hence does not give rise to any constraining PPC predictions.

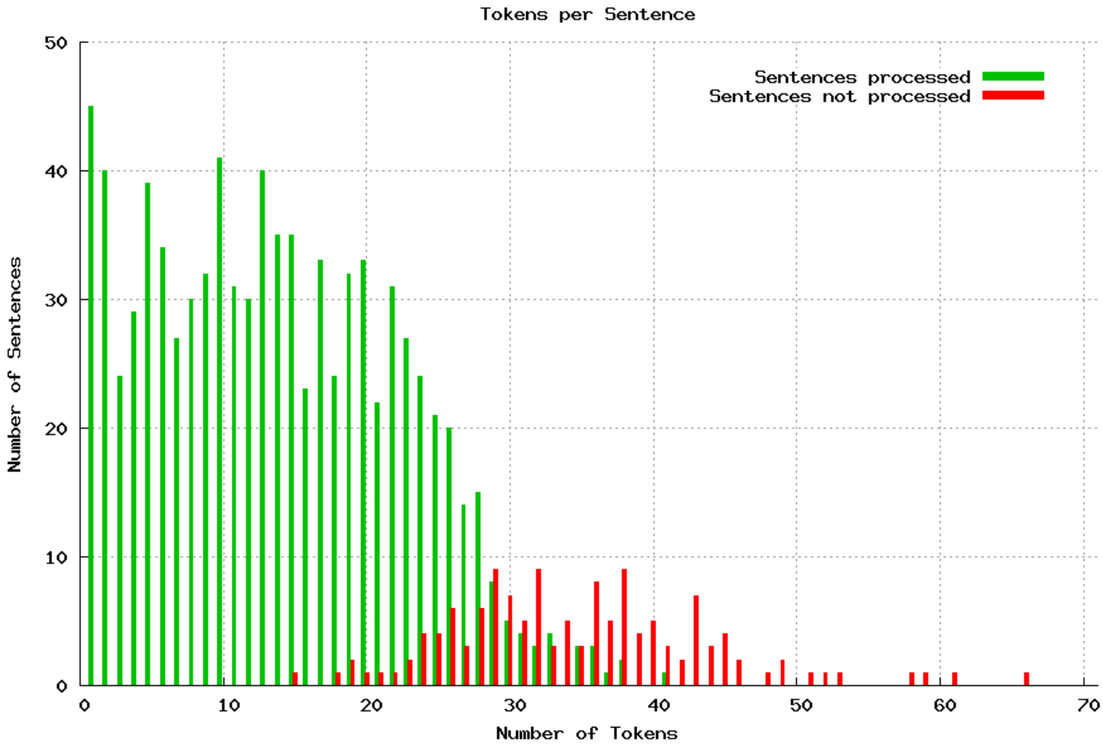


Figure 8.1: A plot of sentences processed against the number of tokens per sentence for the studied 1,000 NEGRA sentences under empty context integration.

per sentence[1], the sentences that were not processed to completion were considerably longer than the average sentence in the studied corpus subset with 16.4 tokens. The average length of the 865 sentences that did process to completion was 13.9 tokens. The plot in Figure 8.1 clearly illustrates the increasing tendency of WCDG2 to fail for sentences longer than approximately 20 tokens. The graph plots the sentence counts against the number of tokens for the 1,000 sentences processed. Colour-coding distinguishes between sentences that were processed to completion (green) and sentences that were not processed to completion (red).

We suspect that processing for the latter sentences requires more working memory than was available on the standard hardware used. Another possibility might be that the implementation of WCDG2 contains a memory leak whose adverse effect remains unnoticed for sentences requiring moderate processing effort but becomes noticeable in more complex analyses. The system errors received for the sentences that did not process to completion did not permit to determine the exact cause. Further investigation is warranted here to determine the exact cause of WCDG2's failure to process these sentences to completion.

We report precision, recall and f_1-measure for the 1,000 NEGRA sentences in Table 8.1. The WCDG1 evaluation results including punctuation marks are quoted from Foth and Menzel (2006b), evaluation results excluding punctuation marks are quoted from Khmylko et al. (2009).

[1] All measures quoted here include punctuation marks as tokens.

	WCDG1				WCDG2			
	Punctuation +		Punctuation –		Punctuation +		Punctuation –	
	str	lbl	str	lbl	str	lbl	str	lbl
Recall [%]	92.5	91.1	91.3	90.0	65.0	63.1	63.8	61.6
Precision [%]	92.5	91.1	91.3	90.0	90.4	87.8	88.8	85.9
f_1-Measure	92.5	91.1	91.3	90.0	75.6	73.5	74.2	71.7

Table 8.1: The structural (str) and labelled (lbl) results for 1,000 Negra sentences with WCDG1's standard grammar and WCDG2's extended grammar. Evaluation results including and excluding punctuation marks are listed separately (Punctuation + and Punctuation –, respectively).

8.1.5 Discussion

Compared with WCDG1's syntax-only analysis, the extended grammar results in an overall degradation of syntactic parsing quality both with regards to precision and recall. The drop in recall to a value substantially lower than for the standard grammar in WCDG1 is drastic but not surprising in view of the fact that WCDG2 did not complete processing for 135 longer-than-average sentences. A comparison of the parsing precisions for syntactic analysis in Table 8.1 shows that the addition of semantic to the syntactic analysis only reduces precision by 2.1% to 4.1%.
Considering that no attempt has been made in this research project to optimise the role-assigning grammar for full coverage of unrestricted input, we consider these precision values on unrestricted text encouraging, even if they fail to meet the overall expectation of matching or superceding the challenging baseline set by WCDG1. It should be kept in mind that the role-assigning grammar has been developed with the objective to ensure correct syntactic and semantic analysis for a small set of specific sentences, typically considerably less complex than most of the Negra sentences studied in this evaluation. WCDG1's standard grammar, in contrast, has been improved continually over a period of years with the express goal of achieving a substantial coverage of German. The large differences between the good precision values and the disappointing recalls are an accurate reflection of the extended grammar's history: the grammar achieves good grammatical precision but suffers from limited coverage.

In conclusion, this evaluation has shown that the addition of the semantic levels of analysis in WCDG2 results in an overall degradation of syntactic analysis quality compared with WCDG1. With good to very good precisions that almost reach the level of the standard grammar, and a significantly lower recall, the primary issue to address in our model's grammar for full compliance with Requirement R7 is the extended grammar's coverage. To achieve the required improvements, significant further grammar modelling effort is needed. We estimate the additional modelling effort to be in the order of magnitude of one to three person years.
To achieve robust and wide coverage of German at a level comparable to that of the syntactic analysis of WCDG1, any effort to improve the grammar also needs to include a systematic validation of WCDG2's implementation integrity. Specifically,

it needs to be ensured that the inability to complete the 135 sentences, most of which longer than the average in the corpus subset, was not caused by a memory leak as this could nullify the benefits expected from further grammar development.

With respect to the selection of linguistic stimuli for the further study of our model's context-integration behaviour, we conclude that the selection of arbitrary linguistic stimuli from a corpus of unrestricted natural language is not a viable option with the present version of the role-assigning grammar. To be able to predict and analyse our model's context integration behaviour systematically, we hence need to study context integration on sentences for which correct syntactic and semantic analysis has been ensured prior to context integration. The following section discusses the selection of suitable linguistic input for our context integration investigations in the subsequent chapters.

8.2 Evaluation on Three Sets of Ambiguous Sentences

8.2.1 Experimental Motivation

In order for a systematic prediction and analysis of our model's context integration behaviour to be possible, we need to ensure that context integration is studied on sentences that are analysed correctly, both syntactically and semantically, by the extended grammar prior to context integration. The evaluation on the NEGRA corpus reported in Section 8.1 illustrates that the extended grammar has not yet reached a maturity level — both in coverage and precision — to afford results superior to the baseline established by WCDG1's syntax-only analysis. For the investigation of our model's context integration behaviour we therefore need to select globally ambiguous sentences that are analysed correctly by the current version of WCDG2's extended grammar in the absence of contextual information.

8.2.2 Approach

We extract three types of globally ambiguous sentences from two sources. The syntactic ambiguities selected for extraction are genitive-dative ambiguity, subject-object ambiguity and PP-attachment ambiguity in German. Examples for each these ambiguities are given in Figure 8.2. Extraction was performed from the following two sources:

1. the SALSA corpus (Burchardt et al., 2006; SALSA Corpus Homepage, 2009), a semantically annotated subset of the TIGER corpus (Brants et al., 2002; TIGER Corpus Homepage, 2009). We performed two extractions, one of sentences containing subject-object ambiguities and one of sentences containing 'mit'-PPs with an INSTRUMENT-COMITATIVE ambiguity.

2. the psycholinguistic study van Kampen (2001) which focuses on ambiguity effects invoked by genitive-dative-ambiguous feminine nouns in German subclauses. From that work, we extracted sentences containing an introductory

main clause and a subclause. In a subset of the extracted sentences the subclause contained a global genitive-dative ambiguity. With our focus on context-induced resolution of syntactic ambiguity, we normalised the introductory main clauses to be the same in all sentences. Normalisation of the introductory main clauses resulted in a reduction of the total number of unique sentences after the removal of duplicate sentences.

VK-274 ‘Er wusste, dass die Doktorandin der Forscherin den Beweis lieferte.’

$$\textit{He knew that} \left\{ \begin{array}{l} \textit{the researcher's PhD student delivered the evidence.} \\ \textit{the PhD student delivered the evidence to the researcher.} \end{array} \right\}$$

Genitive Reading Forscherin $\xrightarrow{\texttt{SYN:GMOD, INST:OWNER}}$ Doktorandin

Dative Reading Forscherin $\xrightarrow{\texttt{SYN:OBJD, INST:RECIPIENT}}$ lieferte

SO-9792 ‘Sie vertritt die Gesellschaft, und ihr obliegt die Geschäftsführung.’

$$\left\{ \begin{array}{l} \textit{She represents the association,} \\ \textit{It is her that the association represents,} \end{array} \right\} \textit{and management is her responsibility.}$$

Subject-Object Reading sie $\xrightarrow{\texttt{SYN:SUBJ, AGNT:AGENT}}$ vertritt

Gesellschaft $\xrightarrow{\texttt{SYN:OBJA, THME:THEME}}$ vertritt

Object-Subject Reading sie $\xrightarrow{\texttt{SYN:OBJA, THME:THEME}}$ vertritt

Gesellschaft $\xrightarrow{\texttt{SYN:SUBJ, AGNT:AGENT}}$ vertritt

PP-7177 ‘Insgesamt werden Braunkohlemeiler mit zusammen 8500 Megawatt (MW) abgeschaltet.’

$$\textit{Overall, lignite-fired plants} \left\{ \begin{array}{l} \textit{with a total of 8,500 megawatts (MW) will be switched off.} \\ \textit{will be switched off by a total of 8,500 megawatts (MW).} \end{array} \right\}$$

INSTRUMENT Reading Megawatt $\xrightarrow{\texttt{SYN:PN}}$ mit $\xrightarrow{\texttt{SYN:PP}}$ abgeschaltet

Megawatt $\xrightarrow{\texttt{INST:INSTRUMENT}}$ abgeschaltet

COMITATIVE Reading Megawatt $\xrightarrow{\texttt{SYN:PN}}$ mit $\xrightarrow{\texttt{SYN:PP}}$ Braunkohlemeiler

Megawatt $\xrightarrow{\texttt{INST:COMITATIVE}}$ Braunkohlemeiler

Figure 8.2: Examples for the ambiguity types selected for study under context integration.

The selected sentences are parsed with WCDG2's extended grammar under integration of an empty context model (genitive-dative ambiguity in Experiment 1.2, subject-object ambiguity in Experiment 1.3, and PP-attachment ambiguity in Experiment 1.4). We test which of the extracted sentences are assigned a *correct* syntactic and semantic analysis by WCDG2's extended grammar. 'Correct' in this context does not necessarily mean that the default analysis also represents the preferred reading that human linguistic intuition would favour. Rather, 'correct' in this case expresses that the analysis of the sentence permits to construct a context in which the analysis represents a plausible reading of the sentence.

From the set of correctly analysed sentences we select three subsets of 10 sentences each for use in the subsequent context integration experiments. The selection criterion for the sentences in the subsets is that at least one of the two readings of the syntactic ambiguity, preferably even both, should correspond to a visually perceivable situation. Due to the extensive use of figurative language, especially in the newspaper articles of the SALSA corpus, this criterion turned out to be surprisingly difficult to fulfil. The three sets of selected sentences are listed in Appendices IV.1, IV.2, and IV.3, respectively. We henceforth refer to the parses obtained under integration of an empty context model as *default parses*.

8.2.3 Results

Ten subject-object-ambiguous sentences were selected from a set of 1,813 sentences extracted from the SALSA corpus. The ten sentences with PP-attachment ambiguity were picked from an extract of 152 sentences that contained a 'mit'-PP. From the cited psycholinguistic investigation we extracted 427 sentences. Normalisation of the introductory main clauses and subsequent removal of the resulting duplicate sentences reduced the number of sentences down to 337. A subset of these sentences exhibited global genitive-dative ambiguity from which we randomly selected ten sentences.

As a consequence of the selection criteria, all of the sentences in the three subsets had a syntactically and semantically 'correct' analysis in the sense laid out in Section 8.2.2. The genitive-dative ambiguous sentences all received the same structural analysis which corresponds to the dative-reading and involves the ternary verb form. The corresponding generic tree structure is shown in Figure 8.3. The full list of parse trees for the genitive-dative-ambiguous sentences in the absence of a contextual bias is given in Appendix VI.1.1. The analyses for the sentences containing subject-object ambiguity and PP-attachment ambiguity do not exhibit a uniform preference pattern within each set. The parse trees for subject-object and PP-attachment ambiguities are listed in Appendices VI.2.1 and VI.3.1, respectively.

8.2.4 Discussion

The selected globally ambiguous sentences largely afford analyses in WCDG2 that also represent the reading favoured by human linguistic intuition. However, some of the solutions, though formally correct, represent a reading that differs from human linguistic intuition in the absence of a biasing context. As an example, consider

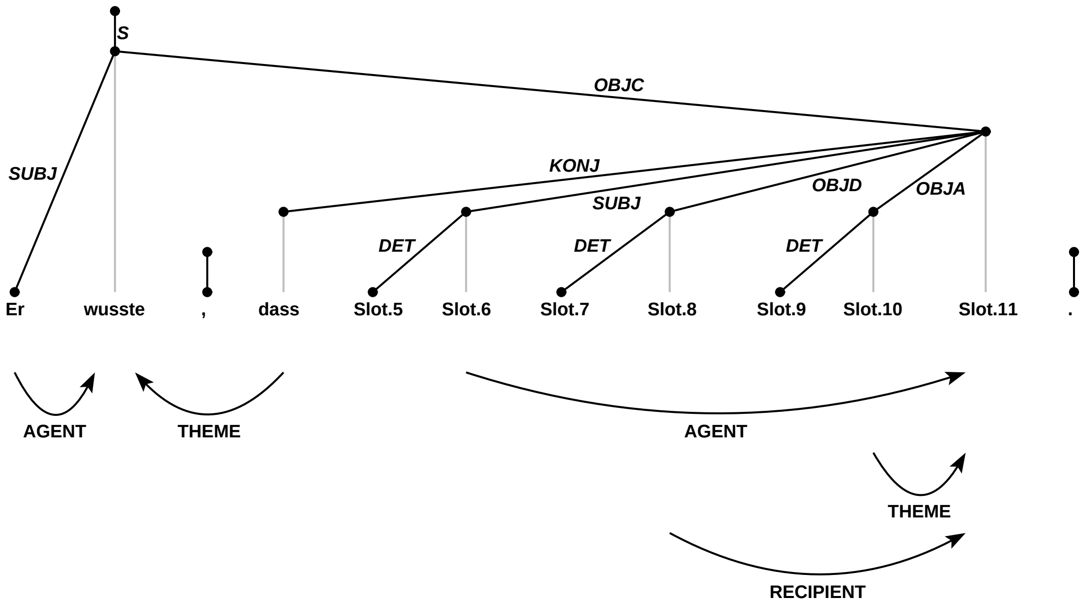


Figure 8.3: Generic parse tree structure for the extracted genitive-dative-ambiguous sentences under integration of an empty visual context model (default analysis).

sentence SO-10744 'Beide Kriegsparteien drängten sie, an den Verhandlungstisch zurückzukehren.' *It was both war parties that they urged to return to the negotiating table.* This sentence parses as the subject-object analysis by default: *Both war parties urged them to return to the negotiating table.* While this analysis is clearly possible, it is certainly the less likely reading in the absence of a biasing context. Based on world-knowledge we know that negotiations can be an alternative means of conflict resolution for warring parties; we hence would assume that it was the war parties that were urged to return to the negotiating table. Since the analyses of the sentences in this experiment have been obtained under integration of an empty context model, i.e., in the absence of a visual context bias, they are a direct reflection of the linguistic preferences encoded in the extended grammar. In this case, the preference of the subject-before-object word order dominated the entire analysis.

The evaluation of the semantic grammar on the 1,000 NEGRA sentences in Experiment 1.1 showed that the extended grammar fails to meet Requirement R7 for the neutrality of referentially unrelated visual context on unrestricted input. The selection of the sentences in Experiments 1.2 to 1.4 was made to ensure that our model meets Requirement R7 at least for the selected sentences. We hence rate Requirement R7 as partially fulfilled by our model. As illustrated by the low recall values obtained in the validation of the extended grammar on unrestricted input in Experiment 1.1, the full satisfaction of this requirement necessitates a substantial extension of the grammar's coverage.

8.3 Conclusions

With precisions of 88.8% and 85.9% (structural and labelled, respectively) for the syntactic analysis of unrestricted input, WCDG2's extended grammar achieves results within reach of the basline established by WCDG1's standard grammar. We have demonstrated that the extended grammar performs well, both syntactically and semantically, on sentences of moderate length spanning up to 20 or 30 words. For longer and syntactically more complex sentences, as are frequently encountered in unrestricted German language input, the parsing quality drops off significantly or fails to process to completion altogether. The extended grammar's present scope limitations in the analysis of larger and more complex sentences impose restrictions on the generalisability of our model to arbitrary input of German.

We re-emphasise that the primary focus of the research described in this thesis is on the motivation, development and validation of a feasible model for the influence of visual context upon syntactic processing — rather than on the scaling of such a model to large or full coverage of German. With this focus in mind, we have selected three subsets of sentences containing 10 sentences each for further study. For all of the selected sentences we ensured that the extended grammar affords a correct syntactic and semantic analysis in the absence of a contextual bias. We will use these sentences in the following chapters for the systematic investigation of the CIA's capacity to effect context-driven syntactic disambiguation.

Chapter 9

Syntactic Attachment Modulation by Hard Integration

In the preceding chapter we have established the neutrality of the semantic grammar with respect to parsing accuracy under integration of an empty visual context for three corpus subsets of 10 sentences each. We will use these subsets throughout the experimental part of this thesis to study the behaviour of the CIA in further detail. This chapter contains a discussion of Experiment 2, the first experiment in which non-empty cross-modal context information is integrated into the process of syntactic parsing. In first approximation to the effects of cross-modal integration in natural systems, we investigate the case of cross-modal context integration via hard integration constraints. The discussion of the experimental observations in this chapter includes a detailed analysis of how the model achieves contextually modulated syntactic analyses. We study hard context integration on the set of genitive-dative-ambiguous sentences.

9.1 Experimental Motivation

While the detailed functional specifications of our model have been presented in Part II of this thesis, the empirical evidence is still pending that the implemented features indeed suffice to drive the parser's syntactic attachments towards a contextually modulated syntactic analysis that is consistent with the integrated visual context information. A prediction of the CIA's behaviour based on the feature descriptions in Part II is complicated by two factors: the complexity of how the different constraints will interact for a given input sentence and the complexity of WCDG's heuristic search algorithm *frobbing* that we employ to locate the optimal solution.

Experiment 2 reported in this chapter illustrates how the integration of cross-modal context information can be enforced by making the integration constraints hard constraints. As a result of this, the parser will only consider solutions whose semantic representation is compatible with the semantic representation of the integrated visual context. We consider two semantic representations R_1 and R_2 compatible with each other if and only if the semantic relations they assert between coreferential entities agree with each other.[1] As an example consider the semantic analysis in the parser R_1 and the semantic representation in a context model R_2: The two representations are incompatible with each other if a level of semantic analysis in the parser R_1 asserts a semantic dependency $\delta(t)$ between $H_{i,j}$ and $H_{m,n}$ while the context model R_2 asserts a thematic relation θ between $M(H_{i,j})$ and $M(H_{m,n})$ such that $t \neq \theta$.

9.2 Approach

Experiment 2 comprises two parts. In Experiment 2.1 we study 10 genitive-dative-ambiguous sentences under integration of context models that describe the corresponding binary situation with three entities. The context models include an instance of a binary situation concept with two participants relating to the situation concept instance via thematic is_AGENT_for and is_THEME_for relations, respectively. Additionally, the context models include a third entity in an is_OWNER_for relation with the situation's AGENT.

We can interpret the represented binary visual scenes as situations in which two participants, the AGENT and the THEME, interact with each other while the third entity, the OWNER, need not necessarily be physically present in the scene. Consider sentence VK-011 as an example: 'Er wusste, dass die Magd der Bäuerin den Korb suchte.' *He knew that* $\{$*the farmer's maid was looking for the basket* | *the maid was looking for a basket for the farmer*$\}$. Based on the linguistic representation alone, we cannot make a conclusive statement about whether the farmer is actually co-present in the described scene or not. The same holds true for the representation of visual context: We consider the OWNER relation asserted in the context model the result of the process of visual understanding which associates the visually perceived AGENT with another, potentially not co-present, entity. The cases in which the OWNER is not co-present in the visually perceived scene are situations in which visual understanding has recognised the AGENT to be a specific AGENT, namely the AGENT which entertains an is_OWNER_for with the OWNER in that context.

In Experiment 2.2 we repeat the conditions of Experiment 2.1 with a different set of context models. Here, we use visual context models for the corresponding ternary situations involving three participants, i.e., a context model containing an instance of a verb-specific ternary situation concept with the participants AGENT, THEME, and RECIPIENT. The analyses obtained under hard integration of the binary and ternary visual contexts are compared with the parses obtained under integration of an empty visual context (see Experiment 1.2 discussed in Chapter 8).

[1] An exception to this notion of compatibility is provided by the case that no relation has been asserted between two contextual entities. Note that in our model we do *not* have an explicit semantic NULL relation to assert that no semantic relation exists between to entities of a semantic representation. Consequently, our model cannot differentiate between the express assertion of an absence of semantic relations and the case in which simply no assertion about the relation between two entities has been made. In our model, these two cases are equivalent and all relations are considered compatible in case no relation has been asserted.

For all of the parses in both parts of the experiment we record three measures: the average processing time required for frobbing to find the optimal solution, the number of structural candidates in the hypothesis space prior to frobbing as reported by WCDG2, and the number of unary and binary constraint evaluations performed in the course of parsing. The average processing times reported are average values based on 10 individual measurements for each sentence. All processing times were recorded on the same machine with no other applications running.
We record the number of structural candidates *prior to* frobbing. This is necessary in order to obtain the actual size of the hypothesis space after the removal of structures that violate hard, unary, non-context-sensitive constraints. The number of structural candidates quoted by WCDG *after* frobbing typically is smaller than this value. This is because in the course of frobbing, WCDG performs additional pruning operations to eliminate further candidates from the hypothesis space. As our focus is on the effect of the unary integration constraints upon the size of the hypothesis space, we report the number of structural candidates prior to frobbing such as to eliminate the effect of pruning during frobbing.

9.3 Setup

The parse runs are performed with the parameter settings shown in Table 9.1. The quoted slot indices for the homonyms in that table refer to the word slots in the normalised sentences with genitive-dative ambiguity. All of those sentences follow the pattern illustrated at the top of the table. The asserted binary and ternary situation context models for those sentences are given in Appendices V.1.1 and V.1.2, respectively.

9.4 Results

The parses obtained under hard integration of a binary visual context containing three entities all conform to the generic structure shown in Figure 9.1. For convenience, we refer to these parses as *binary situation parses* in the subsequent discussion of this experiment. The complete list of the binary situation parse trees is given in Appendix VI.1.2.

The binary situation parses differ from their corresponding default parses, which are represented by the generic tree structure in Figure 8.3, both on the syntactic and the semantic levels of analysis. The structural differences between the binary situation parses in this experiment and the default parses are summarised in Table 9.2.

The parses obtained under hard integration of a ternary visual context with three participants all instantiate the generic tree structure shown in Figure 9.2. For the complete list of parse trees obtained refer to Appendix VI.1.3. The only observed structural difference between the ternary situation parses and their corresponding default parses is the absence of context integration.

Pattern	Er	wusste	,	dass	ART	NN	ART	NN	ART	NN	VVFIN	.
	|	|	|	|	|	|	|	|	|	|	|	|
Slot	1	2	3	4	5	6	7	8	9	10	11	12
Example	Er	wusste	,	dass	die	Magd	der	Bäuerin	den	Korb	suchte	.

		Experiment 2.1		
Context Compliance		Context Model Scheme		
AGENT	1.0	$M(H_{6,j})$	$\xrightarrow{is\_AGENT\_for}$	$M(H_{11,n})$
OWNER	1.0	$M(H_{8,j})$	$\xrightarrow{is\_OWNER\_for}$	$M(H_{6,n})$
RECIPIENT	1.0			
THEME	1.0	$M(H_{10,j})$	$\xrightarrow{is\_THEME\_for}$	$M(H_{11,n})$
INSTRUMENT	1.0			
COMITATIVE	1.0			
		Experiment 2.2		
Context Compliance		Context Model Scheme		
AGENT	1.0	$M(H_{6,j})$	$\xrightarrow{is\_AGENT\_for}$	$M(H_{11,n})$
OWNER	1.0			
RECIPIENT	1.0	$M(H_{8,j})$	$\xrightarrow{is\_RECIPIENT\_for}$	$M(H_{11,n})$
THEME	1.0	$M(H_{10,j})$	$\xrightarrow{is\_THEME\_for}$	$M(H_{11,n})$
INSTRUMENT	1.0			
COMITATIVE	1.0			

Table 9.1: Parameter settings for Experiments 2.1 and 2.2.

Level of Analysis	Dependant	Default Parse		Binary Parse	
		Regent	Label	Regent	Label
SYN	Slot.8	Slot.11	OBJD	Slot.6	GMOD
INST	Slot.8	Slot.11	RECIPIENT	Slot.6	OWNER
AGNT	Slot.1	Slot.2	AGENT	ROOT	*none*

Table 9.2: The structural differences between the generic parse trees for the default and the binary analyses of sentences with genitive-dative ambiguity.

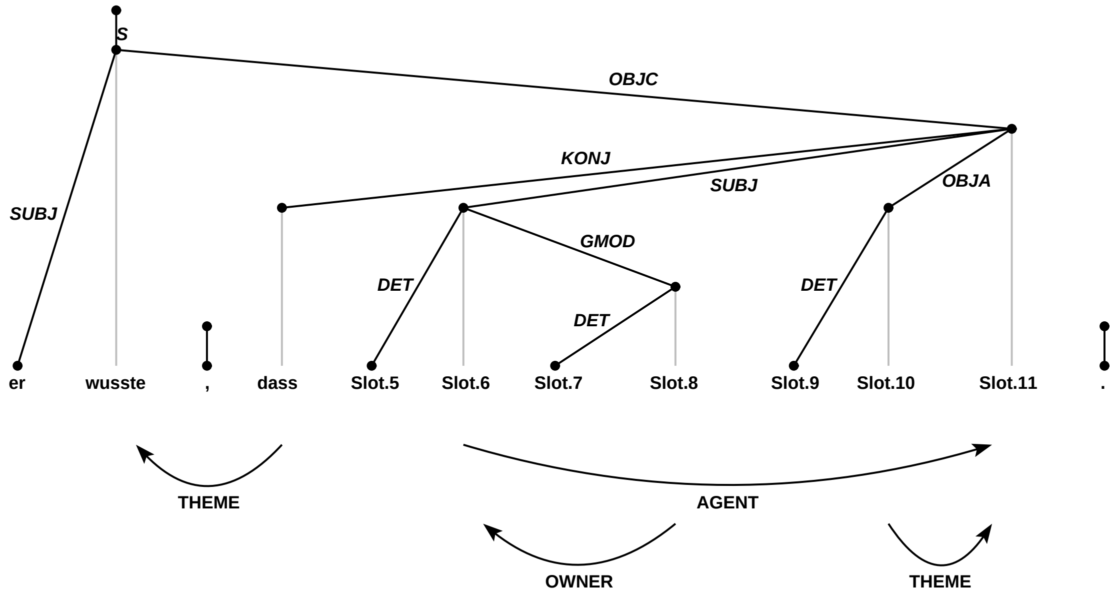


Figure 9.1: Generic parse tree structure for the hard integration of a binary visual context containing three entities, two of which participants.

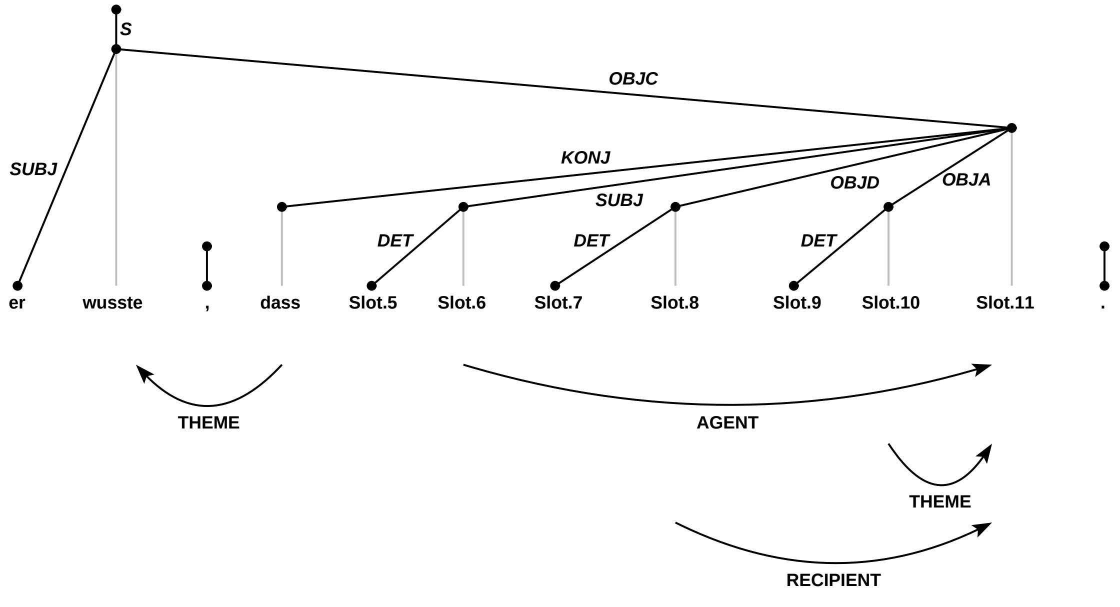


Figure 9.2: Generic parse tree structure for the hard integration of a ternary visual context containing three entities, all of which participants.

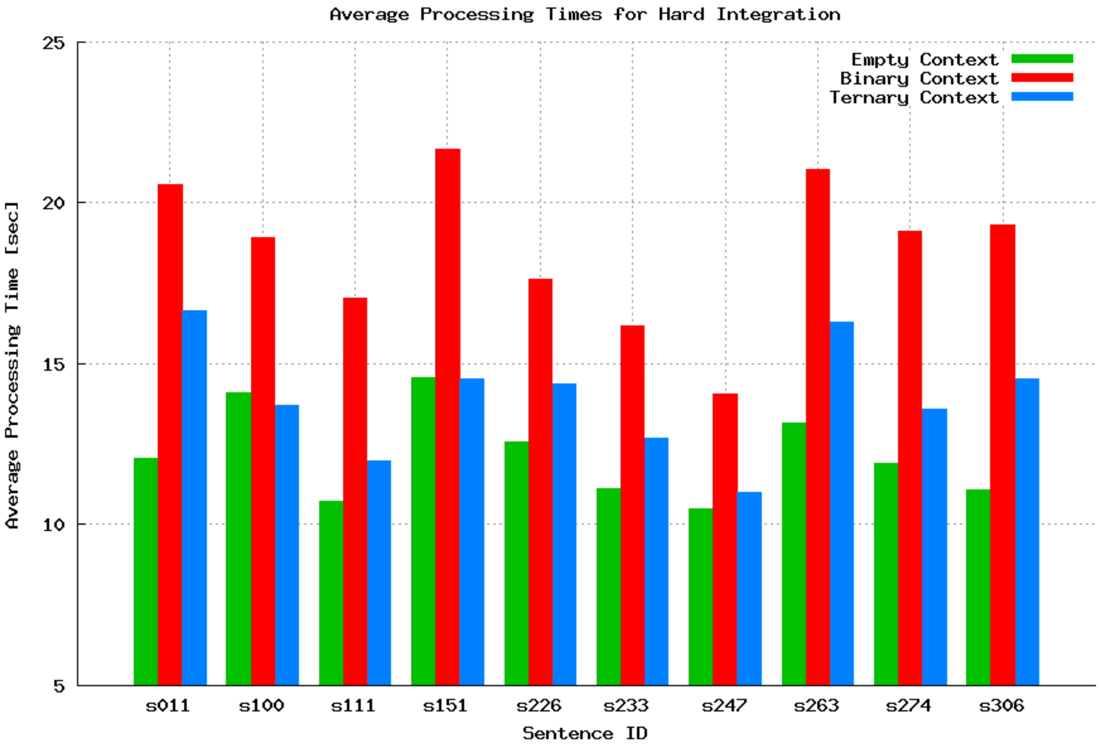


Figure 9.3: The average processing time per sentence for hard context integration.

A plot of the average processing times for the sentences is shown in Figure 9.3. The corresponding numerical values are listed in Table 12, Appendix VII.1. In both cases, the average processing time for hard integration of a non-empty context exceeded – or at best matched – the corresponding processing time under default conditions. The graph also shows that processing under binary context integration in this experiment took systematically longer than under ternary context integration.

The measures for the number of structural candidates reported by WCDG2 for each of the parses are plotted in Figure 9.4. The exact values are given in Table 13, Appendix VII.1. From the plot two observations are immediately obvious: First, the number of structural candidates under hard integration of a non-empty context is about 10 orders of magnitude smaller than the number of candidates for the default parse. Second, the difference between the number of structural candidates for the binary and the ternary situation parses under hard integration is comparatively small. Inspection of Table 13 reveals that the number of candidates for binary context integration marginally exceeds that for ternary context integration.
The plot in Figure 9.5 contrasts the number of constraint evaluations performed in the course of Experiment 2 with the values for the default parses. The detailed values are listed in Table 14, Appendix VII.1. Constraint evaluations performed during the same parse are shown in the same colour. The graph shows that binary context integration results in the largest number of unary and binary constraint evaluations on every sentence in this experiment.

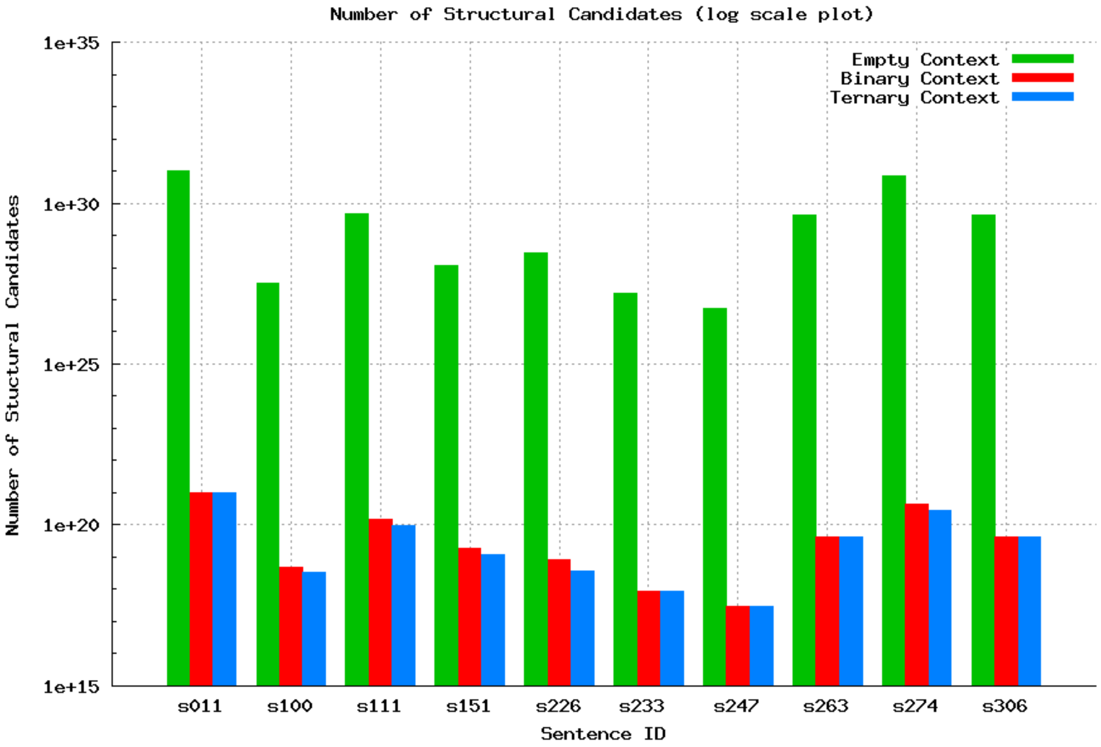


Figure 9.4: Log scale plot of the number of structural candidates prior to frobbing under hard integration as reported by WCDG2.

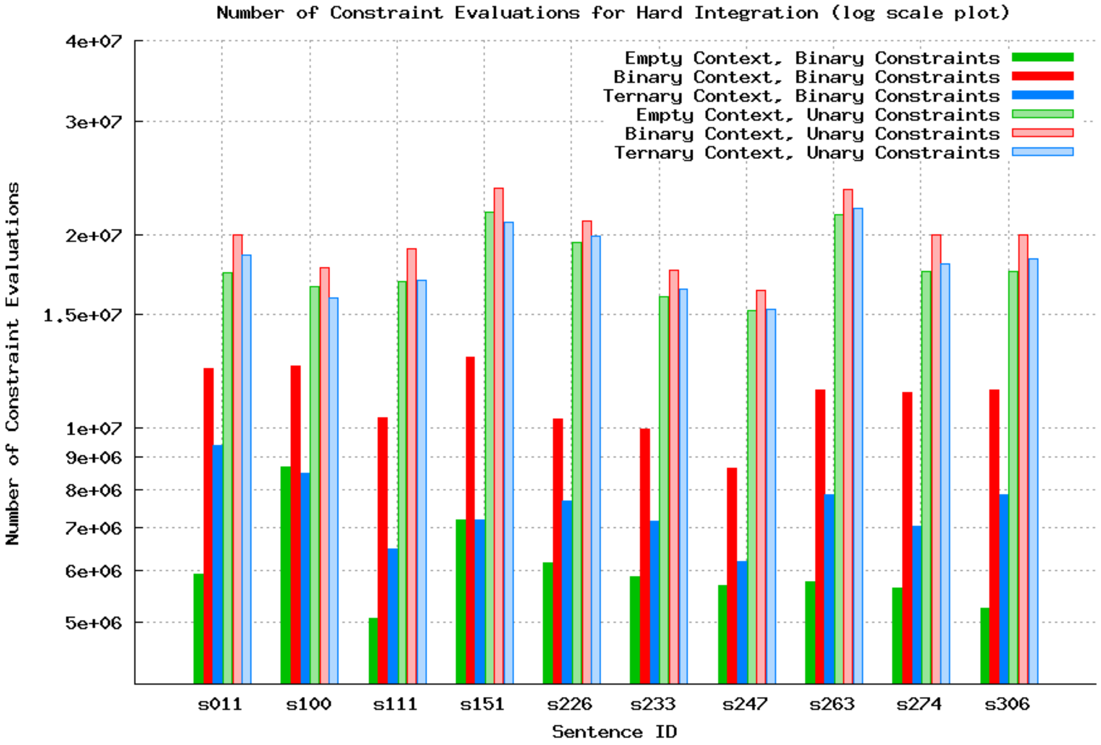


Figure 9.5: The number of unary and binary constraint evaluations under hard integration plotted for each sentence.

9.5 Discussion

As the parse structures for the hard integration of binary and ternary visual contexts illustrate, the CIA has succeeded in propagating the propositional semantic information from the context model into syntactic structure via WCDG2's syntax-semantics interface. According to Figure 8.3, the default parses for the genitive-dative-ambiguous sentences coincide with the results for ternary context integration — at least with regards to the structurally ambiguous subclause. This shows that context integration indeed results in a confirmation of the structural default analysis when visual context is compatible with the default reading. This setup models the real world scenario in which a sentence is uttered and its preferred reading is endorsed by the co-present visual context.

More interesting from the point of view of context integration is the case in which the default reading is *incongruent* with the information provided by visual context. In humans, it is by no means obvious which source, linguistic analysis or context information, will dominate the final utterance interpretation. The different degrees of reliability of the modalities involved, e.g., due to adverse visibility or error-afflicted linguistic input, motivate the hypothesis that in natural systems the weight with which each modality is integrated is dynamically adjusted based on the cognitive and communicative conditions of the given situation.

Under the conditions of hard context integration in this experiment, we can be sure that any structural hypothesis whose semantic representation is incompatible with the semantic representation of visual context will cause the violation of a hard integration constraint. As a result, the parser will reject such a solution candidate as invalid. With visual context information providing a uni-directional hard constraint on syntactic analysis in our model, we hence expect to observe a precedence of visual context information over linguistic default preferences.

This expectation is confirmed by experimental observation: The hard integration of a binary visual context succeeds in overriding the syntactic default analysis of a three-argument verb subcategorising a dative object (`OBJD`). Instead, the parser favours an analysis with a two-argument verb and an additional genitive modifier (`GMOD`) to the verb's subject (`SUBJ`). These syntactic modulations are driven by the corresponding modulations on the semantic level of analysis, concretely the change from `RECIPIENT` to `OWNER` dependencies between Slot.8 and Slot.6.

Under binary context integration, the parser selects the transitive verb form over the ditransitive form because visual context instantiates a situation concept from the class TAKES.AGENT.THEME. The asserted thematic relations consequently authorise the semantic dependencies to the transitive verb of semantic valence `ag_th` while all other dependencies, including those to the ditransitive verb forms, are all vetoed.

In terms of ambiguity resolution, we can thus confirm that, under hard context integration, the CIA achieves disambiguation in line with the information provided in the visual context model. These modulations are precisely the effect we had hoped to observe: the thematic relations and situation arity asserted for the visual scene constrain the dependencies assigned in the parser's semantic analysis. The syntax-semantics interface then propagates these semantic dependencies into the syntactic level of analysis via correspondence rules. In summary, this experiment has shown for the CIA that propositional semantic context information effects syntactic modulations mediated by a shared level of semantic representation.

Apart from these structural modulations affecting verb valence, Table 9.2 lists another structural difference as a result of context integration: the systematic absence of the `AGENT` dependency between Slot.1 and Slot.2 in the context-integrated structures. The reason for this is also a direct consequence of context integration via hard constraints, albeit a somewhat less obvious — and, in terms of the integrity of the semantic analysis, a less desirable one.

The observation shows that context integration has an adverse effect on the semantic analysis of the introductory main clause for which no information had been included in the context model. The absence of the incoming `AGENT` dependency on the verb with semantic valence `ag_th` may appear surprising at first glance. Firstly, it seems easy to fix, namely by just assigning the missing `AGENT` dependency between Slot.1 and Slot.2. Secondly, the absence incurs a comparatively hard constraint violation penalty of `0.1` which persists unremedied throughout all binary and ternary situation parses.

Still, the parser curiously prefers not to assign the missing `AGENT` dependency. The only plausible explanation for this observation is that the assignment of the dependency would give rise to an even more severe constraint violation. Indeed, the observed preference arises from the fact that the inclusion of an `AGENT` dependency would cause a hard constraint violation on the `AGENT` integration constraint. This hard constraint violation results from the PPC veto on this specific relation which, in turn, has been assigned on the basis of information in the context model. The mechanism via which this veto is imposed is as follows: The word 'er' *he* in Slot.1 grounds the concept HE which has been modelled as a rather general concept in the ontology:

$$\text{HE} \equiv \text{PERSONAL.PRONOUN} \sqcap \text{MALE} \sqcap \text{SINGULAR}$$

For most concepts in our ontology no assignment of natural gender or number has been made. Nor have any superclasses corresponding to gender or number been defined for the classes instantiated in the context models (see Appendix V.1 for the detailed context models used in this experiment). Since personal pronouns can refer to any type of entity, be it concrete or abstract, animate or inanimate, we have not defined any disjoint classes in the ontology for the class PERSONAL.PRONOUN. This conceptual underspecification is responsible for the fact that in most sentences personal pronouns have several cross-modal matches in the context model, some of which less obvious than others. Once a thematic relation has been asserted for one of those matches, the PPC imposes a veto on all other thematic relations.[1] This is a general property of our model: *A dependant in the input sentence can only engage in those semantic dependencies which are equivalent to the thematic relations that have been asserted for its cross-modal matches.* All other semantic dependencies for that dependant are vetoed by the PPC. The fact that the `AGENT` dependency

[1] See Sections 7.3 and 7.4 for details on the PPC's algorithm for cross-modal matching and relation scoring, respectively.

between Slot.1 and Slot.2 is missing throughout the context-integrated parses is a direct consequence of the fact that a cross-modal match for ‘er’ was found in every context model. We have confirmed the successful cross-modal matching of ‘er’ *he* by diagnostic output from the PPC.

As regards the number of structural candidates in WCDG2’s hypothesis space, Figure 9.4 shows that the PPC’s introduction of hard penalties on a number of semantic dependencies effects a drastic reduction of the size of the hypothesis space. This is in line with expectation for hard integration since WCDG does not include candidate structures in the hypothesis space that give rise to a violatation of hard integration constraints. The hypothesis space for empty context integration, which we list for comparison, reflects the size of the hypothesis space in the absence of contextual influences.

We further observe that both binary and ternary contexts give rise to a similar number of structural candidates. In our view this is due to the relatively large similarity of the context models integrated. In our model, the number of structural candidates that are eliminated from the hypothesis space as a result of hard context integration depends on the following factors:

1. The number of cross-modal matches for each word
2. The number of words with cross-modal matches in the sentence
3. The number of thematic relations asserted in the context model

Despite the drastic size reduction of the hypothesis space, Figure 9.3 shows that the average processing times under hard context integration were longer than under default conditions — which, at first sight, may seem counterintuitive. It may seem more reasonable to expect that a smaller hypothesis space should also make it easier, i.e.: faster, to locate the optimal solution.

Since both the default and the context-integrated parses are evaluated on the same constraint set, a difference in the constraint base upon which the evaluation is performed in the different conditions can be ruled out as a possible cause for the observation. Figure 9.4 shows that – in line with longer processing times – the number of constraint evaluations also increased under context integration, i.e., WCDG2 had to evaluate more structural candidates in order to arrive at the global optimum.

With solution candidates removed from the hypothesis space due to their violating a hard constraint, transformation pathways to the optimal solution can become obstructed — or in some cases even blocked completely. As outlined in Section 4.2.4 frobbing gradually modifies the best known solution in its search for the global optimum. In the course of this process, frobbing only attempts those transformations that do not incur excessively severe constraint violations. Frobbing will therefore not be able to progress to the global optimum directly when the best known solution candidate is separated from that optimum by interim transformation structures that are *unacceptably bad*. While the global optimum may still be reached via other round-about pathways through the hypothesis space, longer processing times are required to compute the additional interim structures along those alternative pathways. In some cases frobbing may even fail to find the global optimum altogether.

Moreover, there is another way in which hard integration affects the progress of frobbing: Frobbing always attempts to remove the most severe containt violation in a solution candidate first (see Section 4.2.4). The list of constraint violations therefore provides important guiding information for the direction that the frobbing process takes through the hypothesis space. As WCDG rejects structures that violate hard constraints as invalid, none of the interim structures in frobbing will contain hard constraint violations. Consequently, frobbing under hard integration faces the challenge that none of the interim structures may violate an integration constraint. Under hard integration frobbing hence has to proceed without the guiding information of which integration constraints were violated and thus may take longer to locate the local optimum.

The presented experimental evidence supports the view that hard context integration forces frobbing to perform additional structural transformations – and hence constraint evaluations – in order to find the global optimum. As a result, processing times increase under hard integration, despite the reduction in size of the hypothesis space. The default analysis reflects preferences arising from the entire constraint base. In order to arrive at the non-default analysis, some of these preferences need to be overridden by the integration constraints.

The reason for why binary context integration takes longer to process than ternary context integration is because the default context for the studied sentences is structurally almost identical with the parse obtained from ternary context integration. It is therefore likely that the interim structures evaluated during ternary context integration are very similar to those for the default analysis. Binary context integration, in contrast, produces a parse output that is structurally significantly different from the default parse such that different interim structural candidates need to be evaluated by frobbing.

To wrap up this discussion, let us briefly address the degree of conceptual specificity with which the context models in the experiment have been designed. We acknowledge that it is a significant idealisation to assume that the output of the process of visual understanding will be a representation that contains instances of concepts which precisely correspond to the concepts activated by the linguistic input. In our view it is indeed highly unlikely in most cases that visual understanding can provide representations that are conceptually so fine-grained as to differentiate between very similar situation instances in the same way that language can. This holds true in particular for cases in which there are no top-down expectations regarding the classification of the observed visual scene. As an example, consider the ternary visual context models for sentences VK-151 and VK-306 in Figure 9.6.

These context models contain instantiations of the concepts JMD.ETW.SCHICKEN and JMD.ETW.SENDEN, respectively. In the ontology, these concepts are modelled as disjoint. Semantically, however, these concepts are so closely related that they are even rendered by the same verb in the English translations. It is highly unlikely in any case that a visual observer would be able to tell a JMD.ETW.SCHICKEN situation from a JMD.ETW.SENDEN situation by visual inspection alone.

VK-151 'Er wusste, dass die Bergsteiger der Referentin die Warnung schickten.'
He knew that the mountaineers sent the speaker the warning.

MOUNTAINEER.M_01 $\xrightarrow{is\_AGENT\_for}$ JMD.ETW.SCHICKEN_01
SPEAKER.F_01 $\xrightarrow{is\_RECIPIENT\_for}$ JMD.ETW.SCHICKEN_01
WARNING_01 $\xrightarrow{is\_THEME\_for}$ JMD.ETW.SCHICKEN_01

VK-306 'Er wusste, dass die Managerin der Unternehmerin den Vertreter sendete.'
He knew that the manager sent the entrepreneur the sales rep.

MANAGER.F_01 $\xrightarrow{is\_AGENT\_for}$ JMD.ETW.SENDEN_01
ENTREPRENEUR.F_01 $\xrightarrow{is\_RECIPIENT\_for}$ JMD.ETW.SENDEN_01
SALES.REP.M_01 $\xrightarrow{is\_THEME\_for}$ JMD.ETW.SENDEN_01

Figure 9.6: Ternary context models representing the scenes described in the sentences VK-151 and VK-306, respectively.

A more realistic approach to modelling the representations from visual understanding, in our view, must accommodate concept generalisation and perceptual uncertainty. Our model permits to approach this modelling challenge by instantiating conceptually underspecified concepts as illustrated in Section 6.4.2. An experimental validation of this approach is given in Chapter 11 which addresses the influence of grounding and conceptual specificity upon our model's capability to achieve syntactic disambiguation. Suffice it to say for now that our model is indeed capable of exploiting the ontological properties of the concepts involved such that context-modulated syntactic disambiguation can be achieved, even under integration of conceptually underspecified context models. We will see in due course that the type of syntactic ambiguity to resolve determines the degree of permissible conceptual generalisation that we may adopt in the representation of visual context. For the contextual resolution of ambiguities affecting verb valence, such as the genitive-dative ambiguity, situation arity is vital information to be extracted from visual context. Syntactic disambiguation can be achieved as long as this information is provided.

As a final remark we need to comment on the cognitive plausibility of hard context integration in this experiment. As outlined above, a context compliance of `1.0` in this experiment enforces that any solution acceptable to the parser must have a semantic representation which is compatible with the context model. This is another way of saying that a context compliance of `1.0` enforces an absolute dominance of visual context information upon the semantic analysis in the linguistic modality.
We can, of course, easily conceive a number of situations in which visual context information should be subordinate to linguistic interpretation. Typical examples would be conditions of limited visibility, the presence of unknown or unidentifiable

entities in the visual scene or cases of visual ambiguity, such as in a snapshot of a dynamic, potentially bi-directional event which makes it impossible to tell in which direction the scene is evolving. It would be cognitively highly ineffective if humans *always* integrated visual information with the same strength at all times. More to the point, the degree to which humans rely on visual information to support their linguistic processing is dynamic and adjusts situation-specifically. With the introduction of the modelling parameter *context compliance*, we have incorporated precisely this aspect as an important feature in our model.

9.6 Conclusions

This experiment has shown that the CIA successfully integrates propositional semantic information from ontology-based representations of visual scene context into the process of syntactic parsing. The outcome of hard context integration is a linguistic analysis whose semantic representation is enforced to be compatible with the parser-external representation of visual scene context. This contextually compatible semantic analysis drives the corresponding syntactic structure via the model's syntax-semantics interface. A minor corruption to the semantic analysis of the context-integrated structures has highlighted the challenge arising from the conceptual underspecification of personal pronoun concepts in the ontology.

We have further shown that, compared to the default parse, the successful integration of non-linguistic context information comes at the price of longer processing times. Our analysis revealed that despite a drastic reduction in the number of structural candidates an increase in the number of constraint evaluations caused an overall increase in processing time. The experimental data support our hypothesis that the observed increases in processing time under context integration are due to a reduced accessibility of transition structures in the process of transformational search as well as missing guidance information from violated integration constraints.

Having established the feasibility of context integration with our model in this experiment we now continue to study the model's behaviour with respect to central issues of cross-modal integration in natural systems. The following chapter is dedicated to the discussion of constraint relaxation in the integration constraints and discusses the benefits of softer visual context integration in general. It describes the model's response to conflicting visual and linguistic information and discusses the benefits of softer context integration. Chapter 11 then elaborates on the model's robustness to conceptual underspecification in the representations of visual context as may arise from perceptual uncertainty.

Chapter 10

Syntactic Attachment Modulation by Soft Integration

Experiment 2 in the preceding chapter showed that the CIA successfully performs the integration of propositional semantic context information into the process of syntactic parsing. The hard integration scenario discussed effectively models an absolute dominance of the visual modality over linguistic processing. In most real-world situations, however, visual understanding is subject to challenges such as uncertainty, conflicting information or perceptual ambiguity. It is therefore implausible to assume that the integration of visual information into linguistic processing is always performed with the same strength. More realistically, humans dynamically adjust the strength with which they integrate the semantic information from visual context into linguistic processing. In some cases visual information will have a strong effect upon linguistic processing while in other cases it will remain inconsequential. The ability to perform dynamic adjustments of integration strength can suitably be modelled in the CIA based on the WCDG's capability to process weighted constraints. In this chapter, we investigate the results of context integration via soft constraints as a viable alternative to the previously studied hard integration. Our experimental findings show that soft integration provides a number of benefits over hard integration such as context integration without a damage to contextually unrelated syntactic and semantic dependencies, the accommodation of conflicting visual and linguistic information in a uniform linguistic representation as well as diagnostic capabilities to highlight semantic dependencies in discord with the modelled contextual information. The capability of a cognitive system to perform diagnosis of which parts of a given input violate context-driven expectations is highly important in contextualised cognition. Rather than just to say that a given sentence is inconsistent with contextual expectations diagnosis permits to say which aspects of linguistic analysis are in conflict with context-based expectations. In natural systems, the ability to perform such diagnostics enables a more specific and effective response to and interaction with the environment.

10.1 Experimental Motivation

The detailed analysis of the experimental findings in Experiment 2 reported in Section 9.5 revealed that hard context integration can have the undesirable side effect of modulating semantic dependencies in referentially unrelated parts of the input sentence that should remain unaffected by the give visual context information. Concretely, we observed that the semantic analysis of the main clause, for which no context representation was available, was modulated by the integration of visual context information related to the subclause.

While the system of constraint weights is robust enough to leave the syntactic analysis unaffected, the hard integration constraints on the semantic levels of analysis enforce partial defects on the semantic analysis of the introductory main clause. As a result, all hard-integration structures were missing a specific semantic dependency that should have been present.

It would be desirable to achieve context integration that affects only those areas of the sentence that the visual context actually refers to. Ideally, context integration should be selective such as to leave *all* other aspects of linguistic analysis unchanged. For this reason, soft integration is an attractive option: It permits to adjust the strength with which contextual influences are enforced upon linguistic analysis.

In Experiment 3 it is our aim to find an appropriate weighting for the strength of context integration: On the one hand, visual context integration should be strong enough to drive linguistic analysis in line with the visual scene information; on the other hand, context integration must be soft enough such as not to enforce linguistic structures that violate any of the harder constraints in the grammar. An example for one of these constraints is the semantic valence constraint with a constraint weight of `0.1` which was violated under hard integration by the absence of the `AGENT` dependency in the context-integrated structures (see Section 9.5).

With soft context integration we expect the hypothesis space to take the same size as for empty context integration. This is because visual context does not impose any hard constraints on the linguistic analysis anymore. The violation of the integration constraints now does not result in the exclusion of a structural candidate from the hypothesis space anymore. The hypothesis space hence contains more structural candidates and it should also be easier for frobbing to progress towards the optimal solution. We therefore expect processing times and the number of constraint evaluations to go down for the integration of non-empty context models.

10.2 Approach

In Experiment 3 we repeat the parses from Experiment 2 with a smaller, i.e., weaker or softer, value of context compliance. Binary context integration is investigated in Experiment 3.1, ternary context integration in Experiment 3.2. We compare the outcome of soft context integration with the parse results obtained under hard integration for both of these experiments. We also record the average processing

times, number of structural candidates and number of constraint evaluations for comparison. As in Experiment 2, the number of structural candidates recorded is the size of the hypothesis space prior to frobbing.
Experiments 3.3 and 3.4 are aimed at demonstrating the generalisability of our soft-integration results. We set out to verify our previous observations on a small subset of subject-object-ambiguous sentences taken from the SALSA Corpus (see Section 8.2.2 for details). In Experiment 3.3 we parse the subject-object-ambiguous sentences with a visual context that describes a scene in line with the syntactic `SUBJ-OBJA` analysis. In Experiment 3.4, we integrate a visual context that is consistent with the syntactic `OBJA-SUBJ` analysis.

10.3 Setup

The parameter settings for the parse runs are listed in Table 10.1. The context models for Experiments 3.1 and 3.2 are the same as those used in Experiment 2 (see Appendices V.1.1 and V.1.2, respectively). The context models for Experiments 3.3 and 3.4 are given in Appendices V.2.1 and V.2.2, respectively. The latter two experiments are also conducted with a context compliance of `0.8`. We chose this value because context integration with a hardness of `0.2` (see Equation (7.2)) was found to strike a good balance between contextual alignment and grammar-driven linguistic analysis. `0.2` is a good value because dependencies arising from the harder constraints in the syntax-semantic interface do not get overwritten. Typically, the harder constraint in the grammar bear weights that fall in the interval between `0` and `0.2`. As these are harder than the integration constraint, WCDG2 will rather violate the integration constraint than one of these constraints. At the same time, a weight of `0.2` still makes the integration constraint hard enough to drive the overall sentence structure towards a compliance with visual context information.

10.4 Results

The structures obtained from soft context integration for genitive-dative ambiguity follow the structural paradigms in Figure 10.1 for the binary and Figure 10.2 for the ternary contexts. The complete list of parse trees obtained is given in Appendices VI.1.4 and VI.1.5, respectively. In contrast to the generic structures obtained under hard integration (see Figures 9.1 and 9.2), all of these trees contain the `AGENT`-dependency from Slot.1 to Slot.2 that should be contained in the correct semantic analysis of the introductory main clause. What cannot be seen from the parse trees is that WCDG2 also reports a violation of the `AGENT` integration constraint by the dependency between Slot.1 and Slot.2 for all of these parses. Despite the penalty of `0.2` incurred by these structures, frobbing identifies them as optimal.
The recorded average processing times, the number of structural candidates and the number of constraint evaluations are given in Figures 10.3, 10.4 and 10.5, respectively. The numerical values are listed in Tables 15 through 18, Appendix VII.2. A comparison between soft and hard integration shows that processing times are systematically shorter for soft integration, both under binary and ternary context

Pattern	Er	wusste	,	dass	`ART`	`NN`	`ART`	`NN`	`ART`	`NN`	`VVFIN`	`.`
	|	|	|	|	|	|	|	|	|	|	|	|
Slot	1	2	3	4	5	6	7	8	9	10	11	12
Example	Er	wusste	,	dass	die	Magd	der	Bäuerin	den	Korb	suchte	.

Experiment 3.1

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	$M(H_{11,n})$
OWNER	`0.8`	$M(H_{8,j})$	$\xrightarrow{\textit{is\_OWNER\_for}}$	$M(H_{6,n})$
RECIPIENT	`0.8`			
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	$M(H_{11,n})$
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Experiment 3.2

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	$M(H_{11,n})$
OWNER	`0.8`			
RECIPIENT	`0.8`	$M(H_{8,j})$	$\xrightarrow{\textit{is\_RECIPIENT\_for}}$	$M(H_{11,n})$
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	$M(H_{11,n})$
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Table 10.1: Parameter settings for Experiments 3.1 and 3.2.

integration (see Table 10.2). Interestingly, the processing times for the integration of a non-empty ternary context are also shorter than for the integration of an empty context. Figure 10.3 also reveals that – just as for hard integration – the processing times for soft binary context integration are the longest on all runs.

Figure 10.4 shows that for soft integration of the non-empty contexts, the number of structural candidates is the same as for the integration of an empty context. The number of unary and binary constraint evaluations under soft integration of binary and ternary contexts is found to be smaller than for hard integration (see Figure 10.5).

Soft context integration on unrestricted language input also resulted in the successful syntactic modulation based on contextual information. With the exception of sentence SO-9681, all sentences studied exhibited the desired syntactic modulation to integrate the visual context information. A detailed analysis of the reasons why context integration failed for sentence SO-9681 is provided in the following section. The parse trees obtained under soft integration in Experiments 3.3 and 3.4 are given in Appendices VI.2.2 and VI.2.3, respectively.

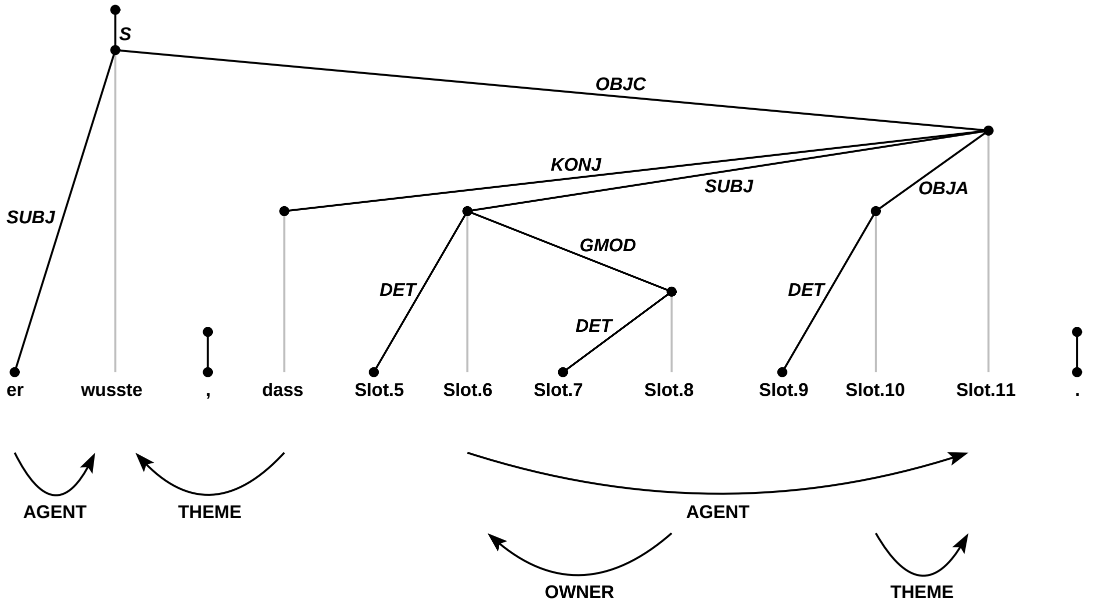


Figure 10.1: Generic parse tree structure for the soft integration of a binary visual context containing three entities, two of which participants (context compliance = `0.8`).

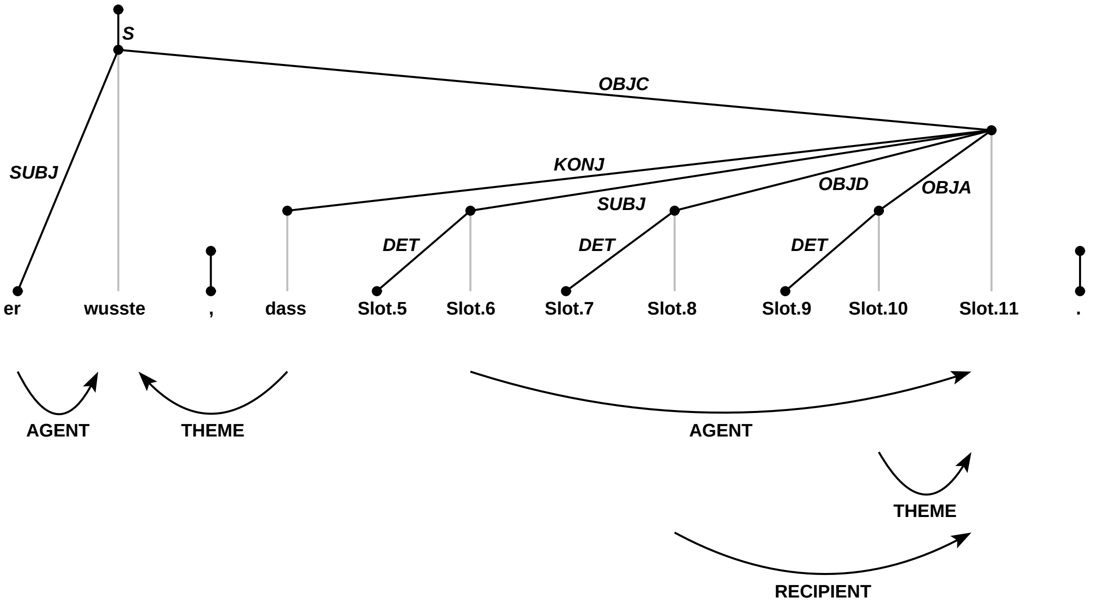


Figure 10.2: Generic parse tree structure for the soft integration of a ternary visual context containing three participants (context compliance = `0.8`).

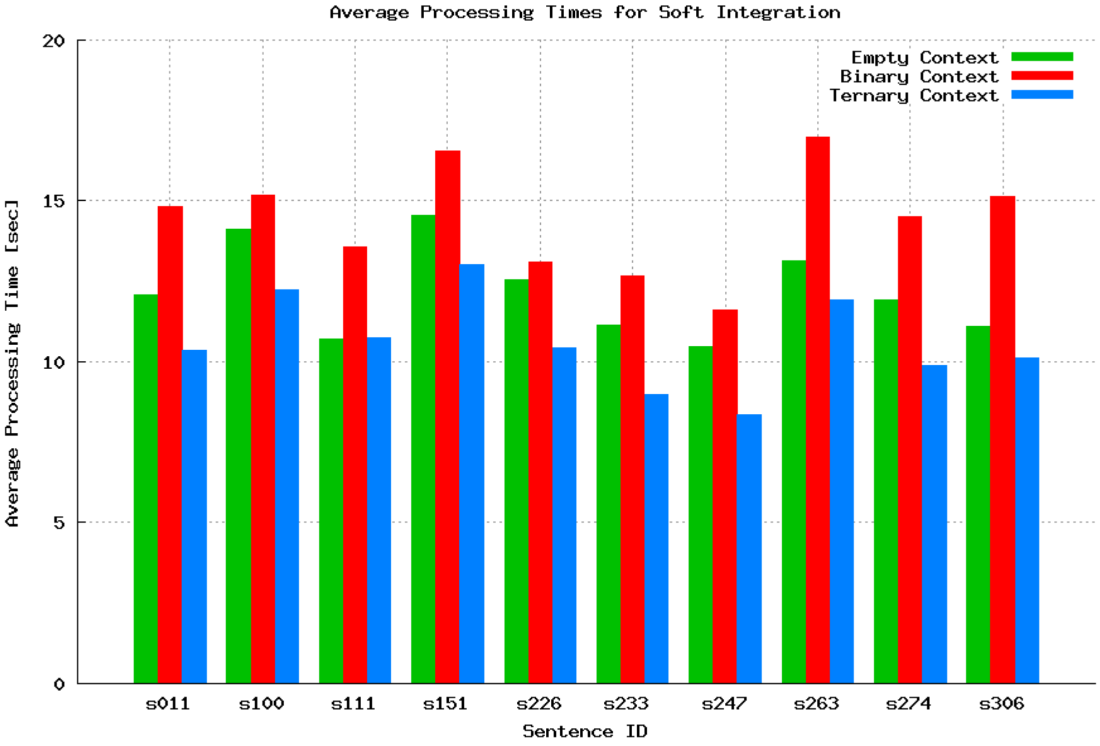


Figure 10.3: The average processing time per sentence for soft integration with a context compliance of `0.8`.

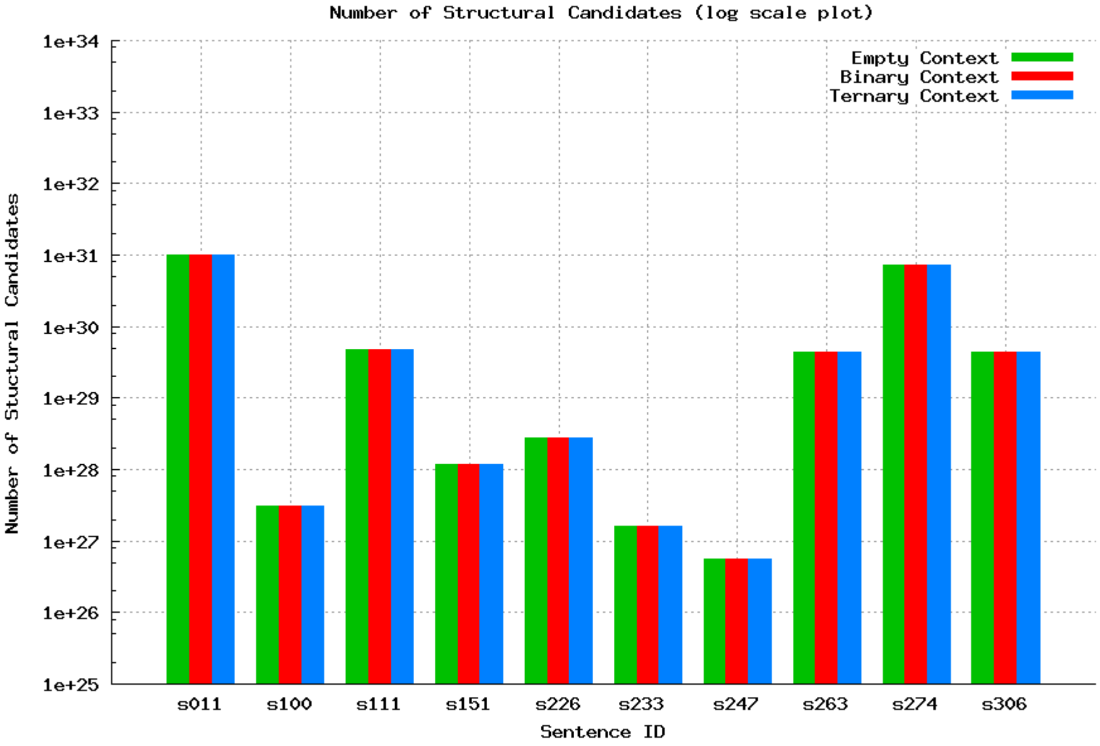


Figure 10.4: Log scale plot of the number of structural candidates under soft integration as reported by WCDG2 (context compliance = `0.8`).

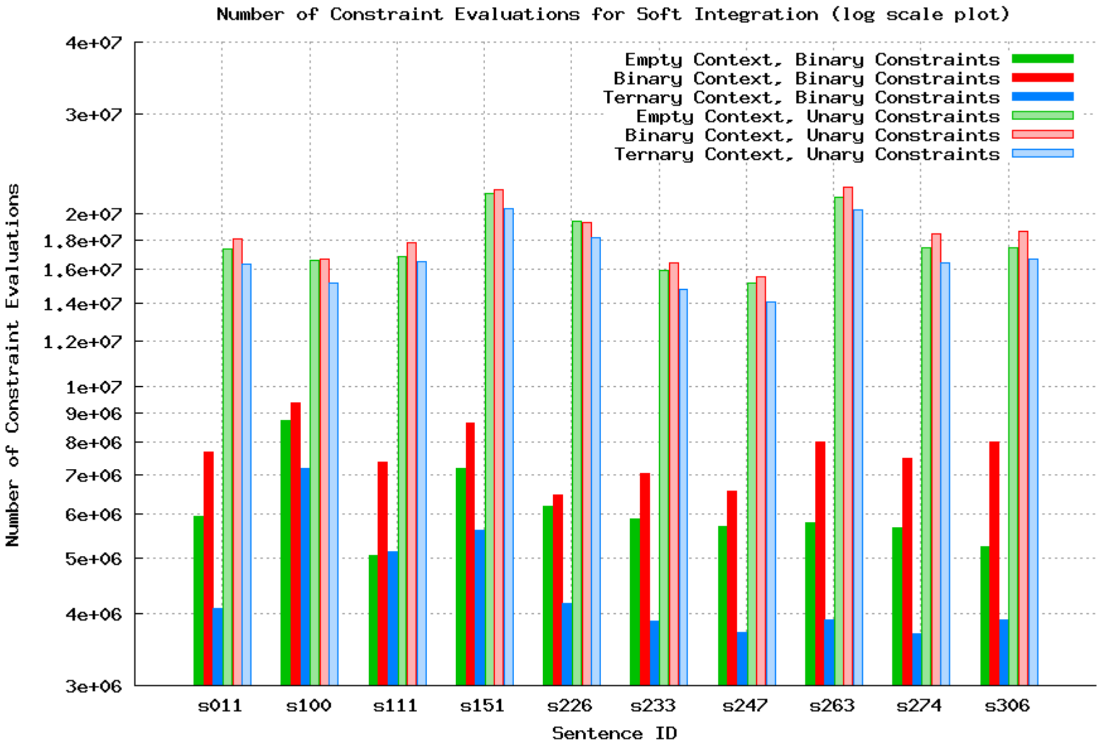


Figure 10.5: The number of unary and binary constraint evaluations under soft integration (context compliance = `0.8`).

10.5 Discussion

The main purpose of soft context integration in Experiment 3 was to incorporate visual context information into syntactic analysis in a way that permits to drive contextually relevant parts of linguistic analysis while leaving contextually unrelated parts unaffected. In soft integration we achieve this goal by reducing the hardness of the integration constraints such that their violation becomes more acceptable than the violation of any of the harder structural constraints in WCDG2's grammar.

One of the advantages of using a weighted constraint-based parser for linguistic analysis is that — given a suitably adjusted set of constraint weights — it can identify a solution as optimal even if that solution violates one or more of the less severe constraints. The concept of *correctness* of a linguistic analysis is hence relativised from an absolute judgement of `true` or `false` to a relative one, which permits to express and compare degrees of acceptability. By reducing the weight of the integration constraints to `0.2` in this set of experiments, we penalise a semantic misalignment with context less weakly than a large number of important structural properties in WCDG2's grammar.

The parse trees obtained for the sentences with genitive-dative ambiguity in Experiments 3.1 and 3.2 confirm the success of this approach: All of the soft-integration structures are semantically and syntactically well-formed according to our model's grammar. They also all contain the `AGENT` dependency from Slot.1 to Slot.2 which

	Ratios of Average Processing Times			
	$\frac{Soft}{Hard}$		$\frac{Soft}{Empty}$	
Sentence ID	Binary	Ternary	Binary	Ternary
VK-011	0.720	0.622	1.229	0.857
VK-100	0.802	0.893	1.076	0.868
VK-111	0.797	0.897	1.265	1.001
VK-151	0.764	0.896	1.137	0.893
VK-226	0.743	0.726	1.043	0.831
VK-233	0.784	0.710	1.139	0.808
VK-247	0.826	0.762	1.109	0.799
VK-263	0.808	0.732	1.292	0.907
VK-274	0.758	0.728	1.216	0.830
VK-306	0.784	0.695	1.365	0.911

Table 10.2: The ratios of average processing time for soft over hard and empty context integration of binary and ternary contexts (context compliance = `0.8`).

was missing in the hard-integration structures obtained in Experiment 2. The `AGENT` dependency is assigned although there is no positive evidence for it in visual context. WCDG2 assigns it based on the structural well-formedness rules in the syntax-semantics interface. The dependency hence is linguistically driven rather than contextually. Precisely for this reason, the observed structures cause a constraint violation on the `AGENT` integration constraint. The violation arises from the fact that a semantic dependency has been assigned in the linguistic analysis which is not expressly endorsed by a corresponding assertion in the context model. Still, WCDG2 identifies the resulting parse tree as the preferred overall solution.

It constitutes a significant strength of our model that a solution candidate can become the preferred structure overall, even if it violates an integration constraint. The softer integration constraint weights used in this experiment permit to accommodate conflicting linguistic and contextual preferences in a single solution structure: while the context model expresses a preference for the removal of the `AGENT` dependency at the cost of a `0.2` penalty, the syntax-semantics interface expresses an even stronger preference in favour of retaining that `AGENT` dependency, thus avoiding to incur an even harder penalty of `0.1`.[1]

[1] Recall that the weight of a constraint is incurred as a penalty for violating the constraint. The penalty affects the overall score of the solution structure multiplicatively. Hence, harder constraints have constraint

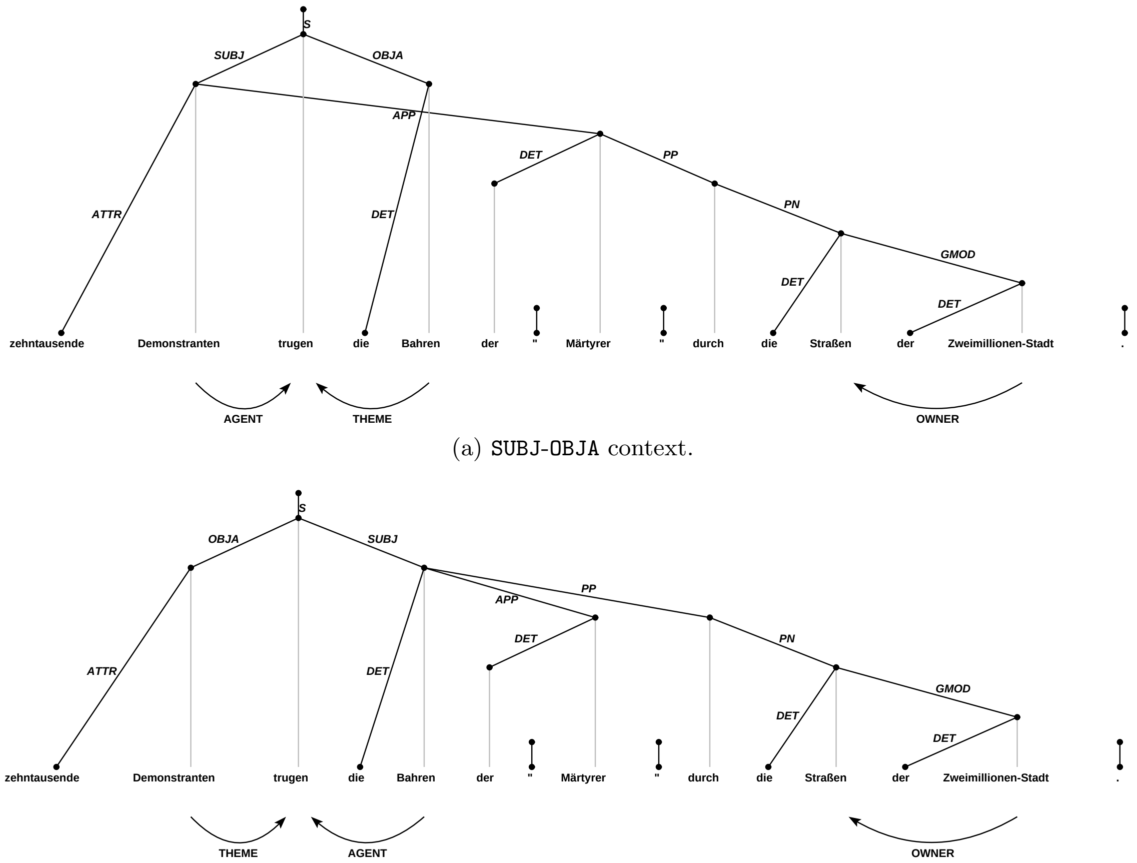
The observed systematic reduction in processing times and the number of constraint evaluations compared to hard integration is consistent with our argument regarding the accessibility of interim transformation structures in the hypothesis space (see Section 9.5): soft integration does not cause the removal of structural candidates from the hypothesis space. Hence, frobbing can traverse the hypothesis space more directly towards the optimal solution, which results in shorter processing times. Another reason for the reduced processing times is that soft integration leads to the retention of structural candidates in the hypothesis space which violate integration constraints. Their constraint violations provide valuable guiding information for the direction that the structural transformations in frobbing will take through the hypothesis space.

Of particular interest with regards to cross-modal integration is the effect on processing times observed for soft integration of a visual context that confirms the linguistic default structure: processing under these conditions is found to be faster than for the integration of an empty visual context model. The constraining information provided by visual context hence improves the effectiveness of localising the optimal structure in the hypothesis space. In analogy to the reduction in processing times observed for temporally and spatially aligned sensory stimuli, we interpret this observation as an instance of *cross-modal facilitation*, i.e., a measurable processing improvement in one modality based on information provided by another modality. Due to the structural identity of the sentences studied, caution needs to be applied not to generalise this observeration without further scrutiny. To make the general claim that our model reproduces cross-modal facilitation under integration of visual contexts that confirm the linguistic default analysis, a systematic investigation of processing times across a large number of structurally diverse sentences is required. The pattern observed for the number of structural candidates is in line with expectation: since visual context is integrated via soft constraints, no structural candidates are removed from the hypothesis space. As a result, the number of pre-frobbing structural candidates observed should be the same for all integrated contexts, be they empty or non-empty. This is precisely what we find in Figure 10.4.

Looking at the outcome of Experiments 3.3 and 3.4 we can say that the integration of contextual information successfully achieves the desired syntactic modulations. However, two aspects deserve further elaboration: The syntactic modulation under context integration for sentence SO-360 requires specific context modelling to yield the correct linguistic analysis. Furthermore, for reasons not immediately apparent, syntactic modulation is not observed for sentence SO-9681 under the given experimental conditions. We will now discuss both of these points in detail.

Let us address the analysis of sentence SO-360 first. The list of context models in Appendices V.2.1 and V.2.2 shows that – in contrast to the other context models – the context models defined for sentence SO-360 contain more information than just the assertion of the is_AGENT_for and is_THEME_for relations. An additional

weights with smaller numerical values. See Section 4.2.2.



(a) `SUBJ-OBJA` context.

(b) `OBJA-SUBJ` context.

Figure 10.6: Incorrect analyses obtained for sentence SO-360 under soft integration of a context model asserting `AGENT` and `THEME` dependencies only.

is_OWNER_for relation is asserted because otherwise the parser's linguistic preferences do not result in the assignment of `SYN:GMOD` and `INST:OWNER` dependencies for this sentence. Instead, WCDG2's grammar defaults into the assignment of an incorrect apposition dependency `APP` on the syntactic level with empty semantic dependencies pointing to `ROOT` on the `INST` level.[1] The effect of these structural assignments is such that frobbing fails to locate the absolute optimum and returns the incorrect structures in Figures 10.6 (a) and (b).

These structures illustrate that for this sentence the assertion of an is_AGENT_for and an is_THEME_for relation in visual context leads to a situation in which the system's linguistic and contextual preferences conflict. In this case, linguistic preferences dominate – but, alas, yield an incorrect overall analysis.

The problem can be fixed in two ways: either the linguistic preferences are adjusted or the constraining effect of visual context is increased. We chose to provide a more constraining visual context to override the underlying linguistic preferences. Our

[1] We have observed the tendency of WCDG's standard grammar for German to assign `APP` labels too readily on a number of unrelated occasions. We recommend a systematic review of the grammar's apposition-handling constraints to correct a potential overgeneration of `APP`-labels.

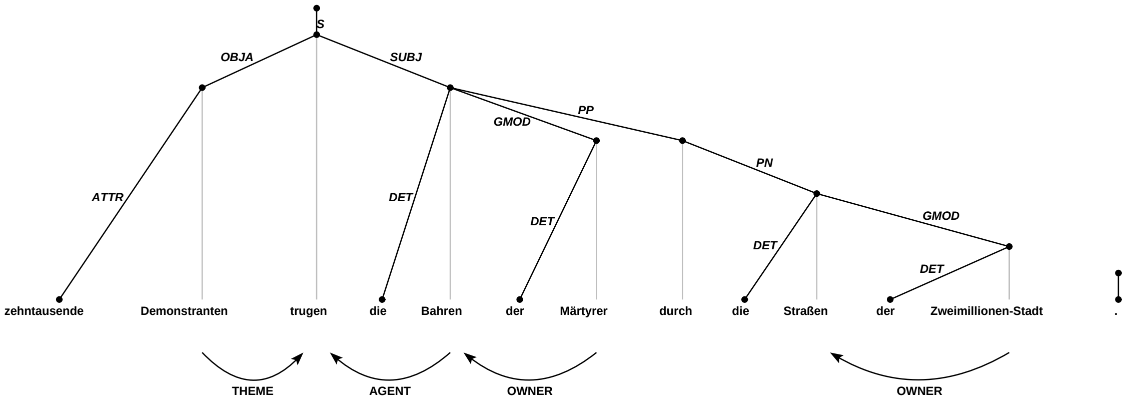


Figure 10.7: The syntactically and semantically correct non-default analysis of sentence SO-360 obtained by integrating a `OBJA-SUBJ` context model that also includes an is_OWNER_for assertion.

decision is motivated by the consideration that humans tend to re-examine visual context for additional information rather than to question their linguistic preferences in cases where the integration of visual context yields an unsuitable analysis. For sentence SO-360, the assertion of the additional is_OWNER_for relation in the context models disfavoured the `APP` assignment originating from Slot.8 and drove the correct assignments of the `SYN:GMOD` and `INST:OWNER` dependencies, instead. This yielded the correct overall analyses, of which the non-default analysis is shown in Figure 10.7.

Summarisingly we can say that this context modelling exception is a direct consequence of the linguistic preferences in the grammar. This example has shown that soft context integration can give rise to conflicting linguistic and contextual preferences in some input sentences. The balance between these preferences can be shifted by modification of the visual context information or the adjustment of the linguistic preferences.

We now turn to the discussion of sentence SO-9681. For this sentence we obtain the `SUBJ-OBJA`-analysis shown in Figure 10.8 for both contexts. This analysis is afforded despite the bias provided by the `OBJA-SUBJ` context model; in that context model we instantiate the concept HUMAN.F rather than the more general concept HUMAN.M.F (see Appendix II). The latter would actually be a more adequate categorisation of the visually perceivable entities referred to by the gender-underspecified personal pronoun ‘sie’, *she* or *they*, in the linguistic input. The incorrect analysis is still obtained for the `OBJA-SUBJ` context, despite the integration of a context representation that is more restrictive than the level of detail provided by the linguistic input. We shall now illucidate why this is the case.

Let us investigate the influence that the context models with different instantiations of HUMAN.M.F and its subconcepts will have: First, consider the context model with the instantiation of the more general concept HUMAN.M.F:

$$\begin{array}{llll} \text{SO-9681} & \text{ADVERTISER.M\_01} & \xrightarrow{is\_AGENT\_for} & \text{ETW.SCHICKEN\_01} \\ & \text{HUMAN.M.F\_01} & \xrightarrow{is\_THEME\_for} & \text{ETW.SCHICKEN\_01} \end{array}$$

The first step in cross-modal matching is the grounding of concepts from the ontology in words of the input sentence. This step is independent of the integrated context model. The critical word in this sentence is the word 'Werber' *advertiser(s)* in Slot.5 whose homonyms all are assigned the conceptualisation ADVERTISER.M as the following excerpt from diagnostic PPC output shows:

$$\begin{array}{lll} \texttt{Werber\_NN\_pl} & \xrightarrow{denotes} & \text{ADVERTISER.M} \\ \texttt{Werber\_ADJA} & \xrightarrow{denotes} & \text{ADVERTISER.M} \\ \texttt{Werber\_FM} & \xrightarrow{denotes} & \text{ADVERTISER.M} \\ \texttt{Werber\_NE} & \xrightarrow{denotes} & \text{ADVERTISER.M} \\ \texttt{Werber\_NN\_sg} & \xrightarrow{denotes} & \text{ADVERTISER.M} \end{array}$$

Due to its underspecification with respect to number and gender, the concept HUMAN.M.F exhibits gender and number compatibility with *all* entity concepts in the ontology. In particular, ADVERTISER.M is compatible with HUMAN.M.F which also has an instantiation in the context model. As a result of this compatibility, the PPC assigns the word 'Werber' two cross-modal matches, namely HUMAN.M.F_01 *and* ADVERTISER.M_01.

In our model, the permissible semantic dependencies of a homonym are determined by the semantic relations asserted for the homonym's cross-modal matches. The context model asserts an *is_THEME_for* relation for ADVERTISER.M_01 and an *is_AGENT_for* relation for HUMAN.M.F_01. The corresponding semantic dependencies `AGENT` and `THEME` therefore both are permissible dependencies for the homonyms of 'Werber'. By the same argument, the two relevant homonyms of 'sie', namely `sie_PPER_pl_acc` and `sie_PPER_pl_nom`, also map to the contextually asserted individuals ADVERTISER.M_01 and HUMAN.M.F_01 such that the words 'Werber' and 'sie' can engage in an `AGENT` or a `THEME` dependency with 'schicken'. Effectively, a visual context containing HUMAN.M.F_01 as an `AGENT` or `THEME` for ETW.SCHICKEN_01 hence has no constraining effect on the linguistic analysis of SO-9681, which is why `OBJA-SUBJ` context integration defaults back to the `SUBJ-OBJA` reading for that sentence.

Our goal was to show that the model for the integration of semantic context information into syntactic parsing also works for unrestricted German language input. To achieve this goal for SO-9681, we investigated how that sentence's context representation needed to be modified such that the desired syntactic modulation would occur. To block the default `SUBJ-OBJA` reading, we need a visual context that can effect a veto on the `AGENT` dependency from 'Werber' to 'schicken' or on the `THEME` dependency from 'sie' to 'schicken' — or both.

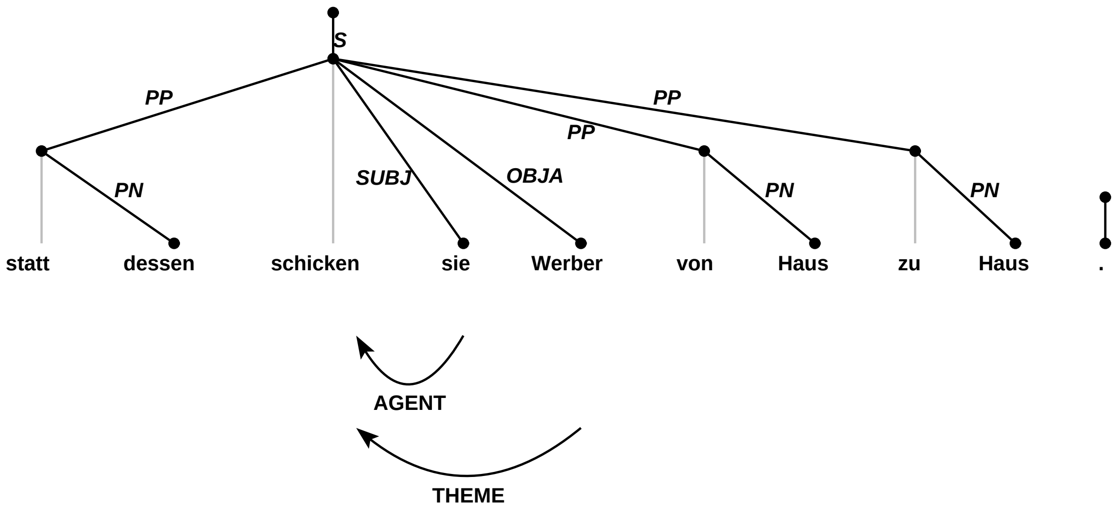


Figure 10.8: The SUBJ-OBJA analysis obtained for sentence SO-9681 in both contexts: under integration of the SUBJ-OBJA and of the OBJA-SUBJ context.

It was our hope to achieve this by enforcing an incompatibility between the concept ADVERTISER.M and the concept instantiated by the instance in visual context. One way to achieve this was to interpret the personal pronoun ‘sie’ as a reference to HUMAN.F_01, i.e. an unspecified number of female persons.[1]. The introduction of the feminine gender specification then results in an in incompatibility with ADVERTISER.M defined as ADVERTISER.M $\equiv$ ADVERTISER.M.F $\sqcap$ MALE.

The resulting context model

$$\begin{array}{llcl} \text{SO-9681} & \text{ADVERTISER.M\_01} & \xrightarrow{is\_AGENT\_for} & \text{ETW.SCHICKEN\_01} \\ & \text{HUMAN.M.F\_01} & \xrightarrow{is\_THEME\_for} & \text{ETW.SCHICKEN\_01} \end{array}$$

was the one used in the parses of Experiment 3. Integrating this context model should suffice to induce the OBJA-SUBJ reading on SO-9681: due to a gender constraint on ‘Werber’, this word now only has one cross-modal match, namely the individual ADVERTISER.M_01. Following from this, THEME dependencies are vetoed for the dependant ‘Werber’ such that the parser assigns it an AGENT dependency, leaving ‘sie’ with the THEME dependency. But why, then, is this behaviour not observed for this context model and SO-9681 under the conditions of Experiment 3? The answer to this question lies in the realisation that, so far, we have only considered concept compatibilities in our argument — and according to those, the described context model should indeed have effected the OBJA-SUBJ analysis. However, we have not yet questioned whether the vetoes resulting from context integration with a context compliance of 0.8 are indeed strong enough to override the linguistic preferences.

[1] Note that in line with the other context models we have used so far, we omit the modelling of number. While the CIA supports the inclusion of number, so far, the number of entities instantiated in the context model was not needed as relevant information in cross-modal matching.

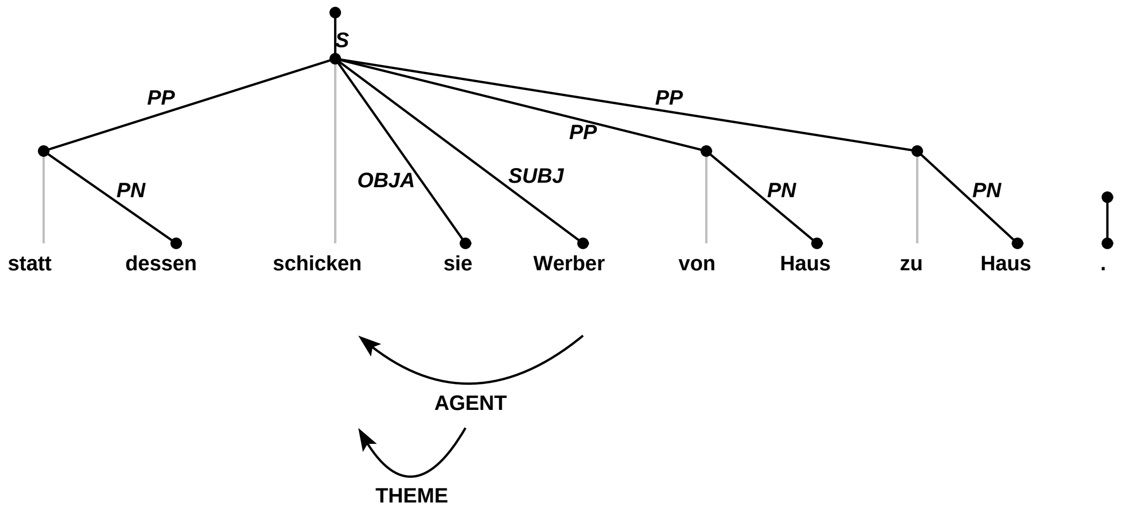


Figure 10.9: Raising context compliance to `0.9` effects the correct linguistic analysis for sentence SO-9681 with the non-default `OBJA-SUBJ` context model.

To pursue this point further, we investigated for SO-9681 how strong visual context integration needs to be enforced in order to override the default analysis, i.e., we studied to which value of context compliance the non-default reading is obtained when integrating the non-default context. To do so, we employed 15 iterations of simple interval bisection on $[0, 1]$, the interval of possible context compliance values. The switch value for SO-9681 was found to be `0.89`, i.e.: WCDG2 returns the default parse for context compliance values below `0.89`. Since in Experiment 3 context integration was performed with a context compliance of `0.8`, we now understand why the default parse was received despite expecting the contrary based on concept compatibility considerations: The weight of the integration constraints of `0.2` is simply not hard enough to enforce the non-default reading against the pressure created by the linguistic preferences. When we re-parse SO-9681 with the HUMAN.F-based context model and a stronger context compliance of `0.9`, indeed the desired non-default `OBJA-SUBJ` reading in Figure 10.9 is received.

We can summarise the discussion of Experiment 3.4 with the central insight that the successful context-based modulation of linguistic dependencies requires a careful balancing of linguistic and contextual preferences. The effect of the visual contextual preferences upon linguistic analysis largely depends on two factors, namely on the concept compatibilities between the concepts activated in the linguistic and non-linguistic modalities and the hardness with which visual context is integrated into linguistic processing.

10.6 Conclusions

The experimental results presented in this chapter illustrate the effect of soft context integration upon linguistic processing in our model. By integrating contextual information via soft constraints, visual context loses the absolute dominance over linguistic analysis that it had under the experimental conditions of hard integration. Soft integration allows for contextual preferences to be overruled by linguistic preferences if and when the latter are stronger. Major benefits of our model are its capability to incorporate conflicting linguistic and contextual information in a uniform linguistic representation, the ability to diagnose which linguistic and contextual preferences are in conflict with a given structural analysis, and the possibility to adjust the strength with which contextual information is integrated.

Experiments 3.1 and 3.2 showed that soft integration permits to incorporate contextual information with the potential of avoiding an adverse effect on the linguistic analysis of contextually unrelated sentence parts. Whether or not context integration indeed maintains the structural integrity of unrelated sentence sections depends on the delicate balance between linguistic and contextual constraints for the given sentence. Soft context integration can only override linguistic constraints that are softer than the integration constraint. Conversely, a linguistic constraint needs to be harder than the integration constraint in order for the linguistic preference to override the contextual preference.
In comparison with hard integration, we observed a decrease in processing times and the number of constraint evaluations as well as an increase in the number of structural candidates for soft integration of non-empty context models. We have advanced an argument to account for this observation based on the notion that the retention of candidate structures in the hypothesis space has two beneficial effects with regards to processing time: it results in easier structural transitions between the interim transformation structures and provides frobbing with better directional guidance for the structural transformations to attempt next.
The results of Experiments 3.3 and 3.4 demonstrate that our model of context integration can also be applied successfully to unrestricted language input. Our observations support the view that the preferred linguistic analysis is afforded by a careful balancing act between linguistic and contextual preferences. We have highlighted the importance of concept compatibility in the process of cross-modal matching and discuss some of the challenges that result for context modelling.

In the following chapter we will expand further on the effect of grounding upon context integration in our model. Concretely, we will discuss how different degrees of concept specificity in grounding affect context integration. We explain in detail how our model exploits ontological properties of the concepts instantiated in a context model and illustrate the power of the resulting inferences.

Chapter 11

The Effect of Grounding on Cross-Modal Matching

In the experiments discussed so far, we have modelled situations in which the visual modality instantiated precisely those concepts that were also activated in the linguistic modality. While this might be an acceptable approximation in a closed domain or under strong top-down expectations, it is unrealistic to assume that bottom-up visual processing will always be able to identify instantiations of object or situation concepts unambiguously. To grasp the gist of a situation, it is not necessary to perform a complete classification of the participating entities and the situation involved.

In this chapter we hence investigate how conceptual underspecification in the representation of visual context affects linguistic processing in our model. As we will see, our model is capable of accommodating uncertainty arising from the grounding of underspecified concepts. In exploiting the ontological properties of underspecified concepts instantiated in the visual modality, cross-modal integration in our model exhibits robustness against grounding uncertainties of visual perception.

11.1 Experimental Motivation

A critical review of the context models employed in the experiments reported in the preceding two chapters confirms that the tokens in the visual modality instantiate precisely those concepts that are also activated in the linguistic modality. As an example, consider the binary and ternary context models for sentence VK-011 that were integrated in Experiments 2 and 3 and are shown in Figure 11.1.

The concept instance BASKET_01 is modelled to represent the visual percept of an object referred to as ‘Korb’ *basket* in the linguistic input. Since the lexicalisation of the concept BASKET is ‘Korb’, this concept is also activated by the word ‘Korb’ in the linguistic modality. The setup in these experiments hence models a scenario in which the objects and situation concepts in the visual scene are perceived to instantiate precisely the same categories that are also activated in the linguistic modality. Since our model does not include a bidirectional interaction between vision and language, we cannot justify the instantiation of exactly the same concepts in both modalities by top-down expectations induced from the linguistic modality. While

VK-011 ‘Er wusste, dass die Magd der Bäuerin den Korb suchte.’

He knew that the farmer’s maid was looking for the basket.

$$\text{MAID\_01} \xrightarrow{is\_AGENT\_for} \text{ETW.SUCHEN\_01}$$

$$\text{FARMER.F\_01} \xrightarrow{is\_OWNER\_for} \text{MAID\_01}$$

$$\text{BASKET\_01} \xrightarrow{is\_THEME\_for} \text{ETW.SUCHEN\_01}$$

He knew that the maid was looking for the basket for the farmer.

$$\text{MAID\_01} \xrightarrow{is\_AGENT\_for} \text{JMD.ETW.SUCHEN\_01}$$

$$\text{FARMER.F\_01} \xrightarrow{is\_RECIPIENT\_for} \text{JMD.ETW.SUCHEN\_01}$$

$$\text{BASKET\_01} \xrightarrow{is\_THEME\_for} \text{JMD.ETW.SUCHEN\_01}$$

Figure 11.1: The binary and ternary context models for sentence VK-011 as used in Experiments 2 and 3.

top-down expectations do exist in natural systems, they cannot be modelled with the current level of implementation in our Context Integration Architecture. Instead, we need to challenge the tacit assumption in our context modelling so far that visual and linguistic modalities activate precisely the same concepts.

The instantiation of exactly the same concept in the visual modality as the result of pure bottom-up processing is unlikely for three reasons: First, visual perception is typically subject to uncertainty as arises from factors such as insufficient lighting, full or partial occlusion, visual ambiguity and others. Nonetheless, humans integrate such underspecified visual scene information into linguistic processing without any difficulty. Second, natural language exhibits synonymy, i.e., different words can be used to denote the same or a very similar concept. The integration of visual context information yields the same result, irrespective of which of these synonyms has been chosen in the linguistic description of the visual scene. Third, it is impossible to make a general prediction as to which modality is going to provide the conceptually more specific information, vision or language. In principle, either modality could be conceptually more specific than the other one such that in some cases the visual modality may add specific information to the linguistic situation description while in other cases the linguistic input provides a more specific description of the situation. Of course, mixed cases may also arise in which one modality is more specific than the other with respect to one referent, but less specific than the other modality with regards to another referent.

Important for our argument is that we assume the underlying cognitive processes that perform the integration of visual context information into linguistic processing to be the same, irrespective of which modality is conceptually more specific. Based on this assumption, these cognitive processes must also comprise the capability to match up more general information from one modality with conceptually more specific information from the other modality.

In addition, we consider the capability to process instantiations of generalised concepts a way to model categorisation uncertainty and categorisation ambiguity in visual perception. Uncertainty in the categorisation of perceived entities can be modelled via the instantiation of concepts general enough to include all of the categorisations that are consistent with the visual percept. Categorisation ambiguity, for example, can be modelled by instantiating the union of the distinct concepts that represent possible categorisations of the perceived entity or situation.

11.2 Approach

Experiment 4 reported in this chapter examines the effect of visual context information that is less specific than the given linguistic information. Concretely, we integrate context models instantiating concepts that are higher up in the T-Box's conceptual hierarchy – and hence are more general – than the concepts activated by the linguistic input. Concept generalisations have been selected based on the following guidelines: Concepts denoting concrete entities are generalised to the next higher visually perceivable superclass, e.g., MAID is generalised to HUMAN.F and SON to HUMAN.M. Abstract concepts such as MOOD or ADDRESS are represented by their next higher superclass in the ontology, ABSTRACT in this case. Verb-specific concepts are generalised to the most general concept of the same situation arity. At this level, verb-specific properties such as lexicalisation and situation valence are lost. Hence, instances of all binary situation concepts are taken to instantiate the concept BINARY.SITUATION, and ternary situation concepts are abstracted to instantiate TERNARY.SITUATION.

In Experiments 4.1 and 4.2, we integrate generalisations of the context models used in Experiments 3.1 and 3.2 based on the guidelines just outlined. All instantiated situation concepts are generalisations of the verb-specific situation concepts integrated in the previous experiments.
Critical inspection of the binary context models in Experiment 4.1 raises the question to what extent the information represented in the context models is really attainable from inspection of a visual scene. In Section 9.2 we argued that a binary context model represents a visual scene in which the physical presence of the OWNER is not mandatory. While the is_OWNER_for relation as such is not visually perceivable, it can still be the part of the representation resulting from the process of visual understanding. Entity recognition in combination with world knowledge can produce a mental representation that includes an is_OWNER_for relation between two entities. In the example of VK-011, it is not just an arbitrary maid that has been identified but a very specific maid, namely *the farmer's* maid.

Entity recognition presupposes that the object in question has been uniquely identified. Our context models, however, only contain instances of more general classes, which makes unique identification in an open domain impossible. For this reason, a context representation such as the following would be cognitively implausible for VK-011:

$$\begin{array}{lll} \text{WOMAN\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{WOMAN\_02} & \xrightarrow{is\_OWNER\_for} & \text{WOMAN\_01} \\ \text{PHYSICAL.OBJECT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

If the visual information is so uncertain that is does not permit the identification of WOMAN_01 as MAID_01, then the world-knowledge-based association with FARMER.F_01 or the more general concept instance HUMAN.F_01 via an is_OWNER_for relation cannot plausibly occur, either.
Based on this argument, we conduct Experiment 4.3 in which we repeat the parse runs of Experiment 4.1 with modified binary context models. We now use the binary context representations from which the OWNER assertion has been removed. These contexts are cognitively more plausible because they only contain information that can be extracted from the visual scene under the assumed level of perceptual uncertainty. We investigate whether the information provided in this reduced context model is still sufficient to constrain the parser to the non-default binary analysis. Note that there is no need to perform an analogous modification to the ternary context models from Experiment 4.2 since all relations asserted in those contexts denote a situation partipant and hence should, in principle, be visually perceivable.

In Experiment 4.4 we examine how strongly the situation information of a visual scene can be generalised in order to still afford the non-default linguistic analysis. We study if visual contexts instantiating SITUATION.CONCEPT, the most general situation concept possible, are still restrictive enough to drive the syntactic modulations required for the non-default binary analysis. Note that the situation information integrated in this experiment is so general that *all* verb-specific information of the observed visual scene, including lexicalisation, situation valence and situation arity, is lost.
Our expectation is that with the loss of situation arity information visual context can no longer constrain the parser's selection of the correct transitive or ditransitive verb form. We therefore expect to see no more context-induced modulation towards the binary analysis for contexts instantiating SITUATION.CONCEPT.

Experiments 4.5 and 4.6, finally, validate our model's handling of conceptually generalised context models on a set of PP-ambiguous sentences from a corpus of unrestricted natural language. For some of the visual contexts it is difficult to argue how the information they represent can be extracted from a visual scene. However, each of the context models integrated represents a distinct contextual state of affairs – be it visually perceivable or not – which corresponds to a unique syntactic analysis of the PP-attachment ambiguity. As such, we expect the context model to be able to bias the parser towards one or the other attachment constellation.

11.3 Setup

In Experiments 4.1 and 4.2 we re-parse the sentences from Experiments 3.1 and 3.2 with generalised three-participant context models centring around instances of BINARY.SITUATION and TERNARY.SITUATION, respectively. The detailed context models for these parse runs are given in Appendices V.1.3 and V.1.4.
In Experiment 4.3, the sentences with genitive-dative ambiguity are re-parsed under soft integration of two-entity contexts centred around an instantiation of the concept BINARY.SITUATION. These context models only contain the assertions of the *is_AGENT_for* and the *is_THEME_for* relations. The complete list of context assertions is given in Appendix V.1.5.
Experiment 4.4 is performed under integration of the three-entity context models centring around an instance of SITUATION.CONCEPT. The complete list of context models is provided in Appendix V.1.6. The experimental parameters for the genitive-dative parses of Experiments 4.1 through 4.4 are summarised in Table 11.1.
Experiments 4.5 and 4.6, finally, are performed on the PP-ambiguous sentences under integration of the generalised COMITATIVE and INSTRUMENT contexts as given in Appendices V.3.1 and V.3.2, respectively.

11.4 Results

The parse trees obtained in Experiment 4.1 all comply with the structural scheme in Figure 11.2. The complete list of parse trees is given in Appendix VI.1.6. Structurally, the trees for integration of the generalised contexts are identical with those obtained under soft integration of the conceptually specific contexts in Experiment 3.1 (see Section 10.4).

Structural identity with the parse results for the corresponding conceptually specific contexts under soft integration (see Experiment 3.2 in Section 10.4) is also observed for the generalised ternary contexts in Experiment 4.2: all sentences comply with the structural paradigm in Figure 11.3. The complete list of parses is given in Appendix VI.1.7.

The reduction of the context models in Experiment 4.3 was found to have no adverse effect on the induction of the non-default binary analysis: all afforded parse trees were compliant with the structural scheme in Figure 11.2. For completeness of documentation, the full list of parses is given in Appendix VI.1.8.

The analyses obtained for Experiment 4.4 exhibit a pattern, the cause for which will be discussed in the following section. The majority of the parse trees follow the structural scheme in Figure 11.3; however, three of the sentences, namely VK-100, VK-111 and VK-151, follow the structural scheme in Figure 11.2. Using WCDG's capability to score manually modified parse trees, we were able to exclude search errors as a possible cause for the difference in analyses: for sentences VK-100, VK-111 and VK-151, the binary analysis does indeed receive a better overall score than the ternary analysis. We list the individual parse trees obtained in Appendix VI.1.9.

Pattern	Er	wusste	,	dass	`ART`	`NN`	`ART`	`NN`	`ART`	`NN`	`VVFIN`	`.`
Slot	1	2	3	4	5	6	7	8	9	10	11	12
Example	Er	wusste	,	dass	die	Magd	der	Bäuerin	den	Korb	suchte	.

Experiment 4.1

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	BINARY.SITUATION_01
OWNER	`0.8`	$M(H_{8,j})$	$\xrightarrow{\textit{is\_OWNER\_for}}$	$M(H_{6,n})$
RECIPIENT	`0.8`			
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	BINARY.SITUATION_01
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Experiment 4.2

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	TERNARY.SITUATION_01
OWNER	`0.8`			
RECIPIENT	`0.8`	$M(H_{8,j})$	$\xrightarrow{\textit{is\_RECIPIENT\_for}}$	TERNARY.SITUATION_01
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	TERNARY.SITUATION_01
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Experiment 4.3

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	BINARY.SITUATION_01
RECIPIENT	`0.8`			
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	BINARY.SITUATION_01
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Experiment 4.4

Context Compliance		Context Model Scheme		
AGENT	`0.8`	$M(H_{6,j})$	$\xrightarrow{\textit{is\_AGENT\_for}}$	SITUATION.CONCEPT_01
RECIPIENT	`0.8`			
THEME	`0.8`	$M(H_{10,j})$	$\xrightarrow{\textit{is\_THEME\_for}}$	SITUATION.CONCEPT_01
INSTRUMENT	`0.8`			
COMITATIVE	`0.8`			

Table 11.1: Parameter settings for Experiments 4.1, 4.2, 4.3, and 4.4.

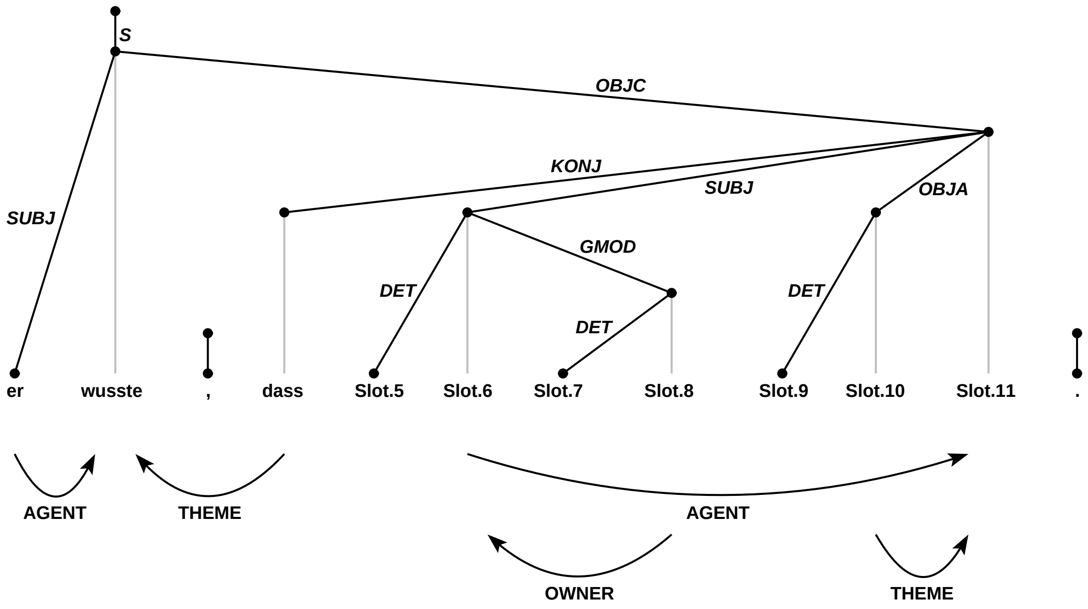


Figure 11.2: Generic parse tree structure for the soft integration of a visual context containing three generalised entities, two of which participants (context compliance = `0.8`).

The verification on unrestricted natural language input yielded the following results: Integration of `PP`-directing contexts containing instances of generalised concepts effected the desired syntactic modulations in almost all cases. All sentences with the exception of PP-3839 in the `COMITATIVE` case exhibited the desired syntactic PP-attachment modulation.

Of the sentences that displayed successful context integration some yielded a syntactically or semantically incorrect overall analysis. This was the case for sentence PP-17512 in the COMITATIVE case and for sentences PP-3025, PP-17512 and PP-31611 in the INSTRUMENT case. An analysis of the causes is provided towards the end of the following section. The resulting parse trees are listed in Appendices VI.3.2 and VI.3.3, respectively.

11.5 Discussion

The results of Experiments 4.1 and 4.2 clearly show that the CIA successfully integrates conceptually underspecified context models to achieve syntactic disambiguation in the course of linguistic processing. These experiments model the scenario in which neither the situation entities nor the kind of interaction in which they engage with each other can be categorised precisely. The only situation information that the visual scene context can provide in these cases is the arity of the interaction between the observed entities. Under integration of the binary visual context, this results in the dismissal of the ternary verb forms as possible readings. Accordingly, integration of the ternary context penalises the selection of the binary verb forms.

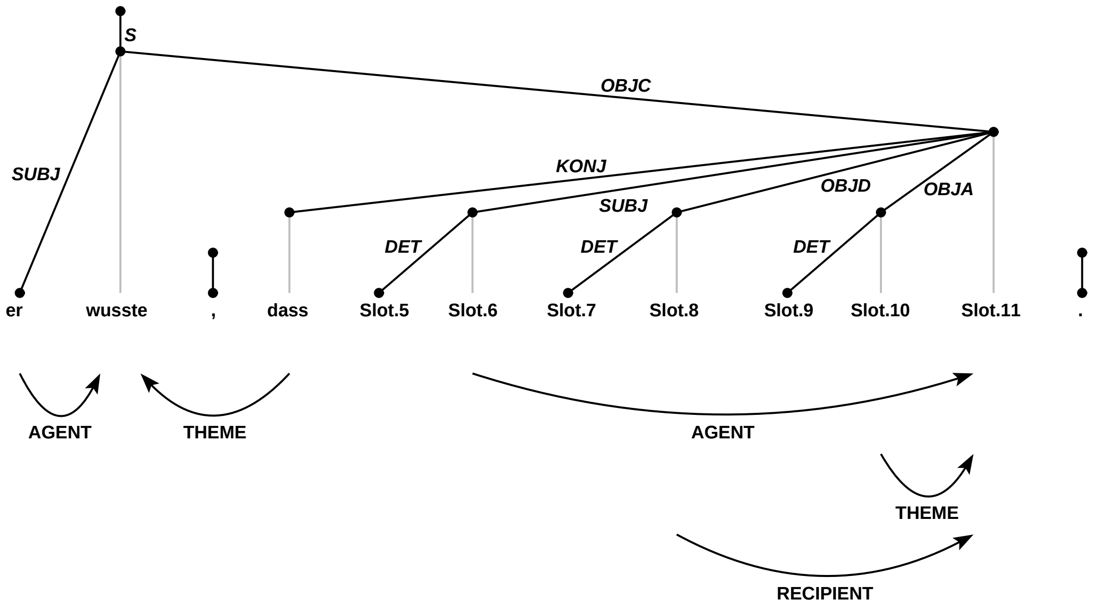


Figure 11.3: Generic parse tree structure for the soft integration of a visual context containing three generalised entities, all of which participants (context compliance = 0.8).

In contrast to the other syntactic ambiguities studied in this thesis, genitive-dative ambiguity has an effect on verb valence: the GMOD/OWNER reading requires the binary verb form while the OBJD/RECIPIENT reading needs the ternary verb form. We expect visual context to lose its constraining effect upon the resolution of genitive-dative ambiguities when the instantiated situation concepts become so general that their situation arity information is lost. They will then fail to restrict the selection of homonyms with the appropriate valence in the parser.
Increasing the generality of concepts instantiated in visual context typically has two effects: both the number of cross-modal matches per homonym and the number of homonyms receiving a cross-modal match increase. The less specific a modelled visual percept is conceptually, the less constraining its effect upon linguistic processing will be.

Table 11.2 juxtaposes the cross-modal matches assigned by the PPC for sentence VK-011 under soft integration of visual contexts that instantiate concepts of increasing generality. In that table, concept generality increases across columns from left to right. It is plain to see that the instantiation of more general concepts results in a tendency towards more cross-modal matches per homonym and towards more homonyms receiving cross-modal matches.
The cross-modal matches in Table 11.2 also illustrate why the integration of a context model centred around an instance of BINARY.SITUATION succeeds in enforcing the non-default analysis in the same way as the conceptually specific context: both contexts assign cross-modal matches to the binary verb forms. All other homonyms

from the same slot are left without cross-modal matches. Those homonyms consequently receive no predictions and are subject to PPC vetoing, which leads the parser to prefer the homonyms that have a cross-modal match. As can be seen from the table, cross-modal matching assigns the same cross-modal matches, irrespective of whether we integrate a conceptually specific or an underspecified situation context. For this reason, both the situation-specific and the conceptually underspecified situation context have the same effect on linguistic analysis. Experimentally, this was confirmed in Experiment 4.1.

The results of Experiment 4.2 can be argued for on the same grounds, the only difference being that the instantiation of TERNARY.SITUATION favours the ternary verb forms rather than the binary ones and imposes vetoes on all other verb homonyms of Slot.11 in that sentence.

The explanation for the outcome of Experiment 4.3 follows the same rationale: The contextual influence of BINARY.SITUATION_01 favours the selection of a binary verb form in the parser. A ternary analysis would incur penalties from the integration constraints of all three participant roles as the context model only asserts an instance of BINARY.SITUATION. The binary analysis, on the other hand, incurs only one contextual penalty, namely for the OWNER dependency. Since the latter has not been asserted in the reduced context model, its assignment in the linguistic analysis gives rise to an integration constraint violation. With just one integration constraint violation as opposed to three – as would result from the selection of a ternary verb form in linguistic analysis –, the parser favours the binary analysis under integration of the two-entity binary context.

Extending this line of argument, we expect the disambiguating effect of visual context to break down once the context model instantiates SITUATION.CONCEPT, the most general situation concept available. In contrast to all other situation concepts in the ontology, this concept carries *no* situation arity information.
As expected, the integration of the two-entity contexts that centre around the instance SITUATION.CONCEPT_01 fails to induce the binary analysis consistently. Most of the structures afforded follow the structural paradigm of the ternary situation analysis and thus comply with the linguistic default preferences.
The question remains: If the hypothesis is correct that contexts instantiating instances of SITUATION.CONCEPT cannot drive the binary analysis, why then do not *all* sentences in Experiment 4.4 afford the ternary analysis? The reason for this becomes apparent when we consider the integration constraints that are violated by the individual solution structures selected by WCDG2. Their constraint violations are listed in Table 11.3.

	Cross-Modal Matches in Visual Context		
Homonym	ETW.SUCHEN_01, MAID_01, FARMER.F_01, BASKET_01	BINARY.SITUATION_01, HUMAN.F_01, HUMAN.F_02, PHYSICAL.OBJECT_01	SITUATION.CONCEPT_01, HUMAN.F_01, HUMAN.F_02, PHYSICAL.OBJECT_01
er_PPER	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
er_FM	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
er_NE	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
wusste_VVFIN_first	{}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
wusste_VVFIN_third	{}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
Magd_NN	{MAID_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Magd_FM	{MAID_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Magd_NE	{MAID_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Bäuerin_NN	{FARMER.F_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Bäuerin_FM	{FARMER.F_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Bäuerin_NE	{FARMER.F_01}	{HUMAN.F_01, HUMAN.F_02}	{HUMAN.F_01, HUMAN.F_02}
Korb_NN	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
Korb_FM	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
Korb_NE	{BASKET_01}	{PHYSICAL.OBJECT_01}	{PHYSICAL.OBJECT_01}
suchte_VVFIN_first_past_-	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_first_past_aip	{ETW.SUCHEN_01}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
suchte_VVFIN_first_past_aip+d	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_first_present_-	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_first_present_aip	{ETW.SUCHEN_01}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
suchte_VVFIN_first_present_aip+d	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_past_-	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_past_aip	{ETW.SUCHEN_01}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_past_aip+d	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_present_-	{}	{}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_present_aip	{ETW.SUCHEN_01}	{BINARY.SITUATION_01}	{SITUATION.CONCEPT_01}
suchte_VVFIN_third_present_aip+d	{}	{}	{SITUATION.CONCEPT_01}

Table 11.2: The cross-modal matches assigned for context models instantiating concepts of different degrees of specificity.

Sentence ID	Dependencies that violate an integration constraint	Dependant	Regent
VK-011	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-226	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-233	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-247	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-263	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-274	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-306	AGENT	Slot.1	Slot.2
	THEME	Slot.4	Slot.2
	RECIPIENT	Slot.8	Slot.11
VK-100	THEME	Slot.4	Slot.2
VK-111	THEME	Slot.4	Slot.2
VK-151	THEME	Slot.4	Slot.2

Table 11.3: Integration constraint violations for the parse trees in Experiment 4.4.

In contrast to the majority of the sentences in Experiment 4.4, the context models for VK-100, VK-111 and VK-151 assert the entity HUMAN.M_01 as an AGENT. This entity instantiates a concept that is incompatible with any concept subsumed by FEMALE. Consequently, HUMAN.M_01 becomes the cross-modal match for the homonyms in Slot.1 and Slot.6. Furthermore, SITUATION.CONCEPT_01 is identified as the cross-modal match for the verb form in Slot.2. With the contextual assertion of

$$\text{HUMAN.M\_01} \xrightarrow{is\_AGENT\_for} \text{SITUATION.CONCEPT\_01}$$

the PPC admits the AGENT dependency from Slot.1 to Slot.2. The parse trees for VK-100, VK-111 and VK-151 therefore do not violate the integration constraint for the AGENT dependency. For these sentences, the parser resolves the genitive-dative ambiguity in favour of the binary analysis because the OWNER dependency can be assigned without causing the violation of an integration constraint: The context models for VK-100, VK-111 and VK-151 contain no assertion of an entity that would be compatible with the concept activated by the dependant of the OWNER dependency in Slot.8. In VK-100, e.g., 'Schauspielerin' *actress* in Slot.8 activates the concept ACTRESS which is conceptually incompatible with all other entities asserted in the context model. As a result, none of the homonyms in this slot have cross-modal matches, no predictions are made and no vetos are imposed for dependencies with Slot.8 as their dependant.

We emphasise that the parser favours the binary analysis for VK-100, VK-111 and VK-111 not because of a constraining influence of the asserted situation concept's arity in the context model. This was the mechanism by which the genitive-dative ambiguities were resolved in the previous cases where situation arity information was available for the contextually instantiated situation concept. In the case of Experiment 4.4, the binary analysis is afforded because it represents the more plausible interpretation of the input sentence given the high level of generality of the concepts instantiated in visual context.

A final comment is owed to the cognitive plausibility of visual contexts centring around instances of SITUATION.CONCEPT. Effectively, these context models represent percepts of visual contexts in which the information contained is so general that neither the nature of the interaction between the observed entities nor the arity of the interaction are known. In our view it is highly questionable whether the instantiation of such concepts can serve a cognitive purpose — and hence whether such percepts can arise in natural systems at all. We may rephrase this doubt as the question *whether* SITUATION.CONCEPT *is encoded in the human cognitive system at all.* A substantial amount of further investigation in the area of cognitive psychology and cognitive science will be needed to answer this question conclusively.

For our model, we conclude that the generalisation of the central situation concept to a degree at which situation arity information is lost, results in the breakdown of its power to effect the systematic disambiguation in verb-related syntactic ambiguities such as genitive-dative ambiguity.

Sentence ID	Context	WCDG2 Score	
		Frobbing	Manually Corrected
PP-3839	COMITATIVE	$3.984 \cdot 10^{-3}$	$4.121 \cdot 10^{-3}$
PP-17512	COMITATIVE	$6.882 \cdot 10^{-1}$	$7.061 \cdot 10^{-2}$
PP-3025	INSTRUMENT	$1.638 \cdot 10^{-3}$	$5.460 \cdot 10^{-3}$
PP-17512	INSTRUMENT	$2.750 \cdot 10^{-2}$	$2.821 \cdot 10^{-3}$
PP-31611	INSTRUMENT	$4.245 \cdot 10^{-4}$	$1.024 \cdot 10^{-3}$

Table 11.4: Comparison of scores for the best scored – but incorrect – candidate as found by WCDG2's frobbing and the parse tree obtained from manual correction of that solution.

We now turn to the discussion of Experiments 4.5 and 4.6. The integration of generalised COMITATIVE and INSTRUMENT contexts to direct PP-attachment were successful in the majority of cases. We provide a causal analysis of those cases where an incorrect overall analysis was obtained.

The reason for the incorrect overall analysis of sentences PP-3839, PP-3025, and PP-31611 is a search error in WCDG2. Using WCDG's manual tree manipulation, we were able to establish that the best solution scores found by WCDG2 were below the scores for the structurally correct trees. WCDG2 hence failed to locate the correct tree structures as optimal even though the modelled grammar scores them higher than the best solutions found by frobbing. The comparison of the numerical scores found by WCDG2 and those obtained after manual correction of the parse trees in WCDG2 is given in Table 11.4.

The analyses obtained for sentence PP-17512 under COMITATIVE and INSTRUMENT context integration are incorrect in a part of the sentence that is unrelated to the integration of visual context information. While the integration of the PP-directing contexts as such was successful for this sentence, an incorrect overall analysis was obtained. The score comparison in Table 11.4 shows that the incorrect analysis was not the result of a search error in frobbing for this sentence.
The semantically incorrect assignment of the AGENT and THEME dependencies in PP-17512, combined with the resulting incorrect SUBJ and OBJA dependency assignments on the syntactic level, result from the standard grammar's strong preference for subject-object word order. Syntactically, 'Renditen' *returns* and 'Agenturen' *agencies* in sentence PP-17512 can be labelled as either SUBJ or OBJA. The structure favoured by WCDG2 is therefore syntactically acceptable — but must be rejected as semantically unacceptable based on world knowledge.

The score comparison in Table 11.4 shows that the present form of the grammar disfavours object-first word order for `SUBJ-OBJA` ambiguous sentences in the absence of an additional semantic bias. Such a semantic bias can, of course, be introduced in the form of additional visual context information. Integrating the representations

$$\begin{array}{lll}
\text{GROUP\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\
\text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \\
\text{HUMAN.M.F\_01} & \xrightarrow{is\_COMITATIVE\_for} & \text{ABSTRACT\_01}
\end{array}$$

and

$$\begin{array}{lll}
\text{GROUP\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\
\text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \\
\text{HUMAN.M.F\_01} & \xrightarrow{is\_INSTRUMENT\_for} & \text{BINARY.SITUATION\_01}
\end{array}$$

as augmented context modes for PP-17512, we do indeed obtain the syntactically *and* semantically correct analyses shown in Appendix VI.3.4. These structures exhibit all the desired context-driven syntactic modulations for the `SUBJ-OBJA` and the PP-attachment ambiguities.

In summary we can say that the integration of generalised visual contexts successfully achieved the desired syntactic modulations for the large majority of PP-attachment sentences of unrestricted natural language input. The observed errors in structural analysis were caused by search errors in WCDG2 or by the unavailability of contextual preferences for the resolution of global ambiguities. While these problems are caused by factors outside of our model, they do result in a degradation of the overall quality of analysis under context integration. We conclude that further grammar modelling effort is needed to improve the detectability of the actual global optimum by WCDG2. The present state of the role-assigning grammar is capable of scoring the correct structures higher than those suggested by WCDG2 as global optima but fails to enable WCDG2 to detect these global optima.

11.6 Conclusions

As Experiment 4 has shown the question how strongly we can generalise the concepts instantiated in visual context in order to still achieve the desired syntactic modulations cannot be answered in full generality. The degree of permissible concept generalisation depends on the type of syntactic ambiguity in the input sentence as well as on the concept properties modelled in the T-Box. For an ambiguity affecting verb valence, we saw that reliable syntactic disambiguation requires the induction of situation arity information from visual context. The resolution of syntactic ambiguities that do not affect verb valence, such as PP-attachment, can be achieved with visual contexts that are specific enough to yield different attachment predictions for the constituents in question. A visual context that is so general as to effect the same predictions for all words in the sentence, e.g. a context instantiating only instances of THING, i.e. $\top$, loses all of its potential to constrain linguistic analysis.

While this chapter has extensively investigated the effect of concept underspecification on context integration in our model, the underspecification of the relations between the contextual entities has been left untouched so far. It may well be the case that the perception of a visual scene results in the instantiation of entities joined by thematic relations that are more general or more ambiguous than the relations considered so far. We encourage the investigation of this field as part of future work, both from the perspective of cognitive science and as a potential extension to our framework's modelling capabilities.

Chapter 12

Conclusions

In the preceding 193 pages we have argued extensively for a computational model of the influence of cross-modal context upon syntactic parsing. Starting with the model motivation, we collected a set of 32 requirements for the model implementation in Part I of this thesis. Part II provided detailed model specifications and an in-depth description of the model implementation. Part III focused on the empirical validation of the implementation.
This chapter rounds off the thesis with a summary of the overall line of argument and the central conclusions that result from it. The thesis summary in Section 12.1 is given in the form of an annotated list of the central tenets in this document. We then draw our final conclusions in Section 12.2. Section 12.3, eventually, is motivated by the realisation that much more work remains to be done; it closes with an outlook to future directions of research that arise from and relate to the work presented in this thesis.

12.1 Thesis Summary

Tenet 1: *Ambiguity is an inherent property and a ubiquitous feature of natural language.*

While linguists have a tendency to view ambiguity as a defective or undesirable feature of natural language, the study of unrestricted natural language *in vivo* leads to the insight that natural language abounds with ambiguity. The causes of ambiguity are manifold and comprise lexical, syntactic, referential and scope ambiguities as most frequent representatives. In most cases of human communication, ambiguity is as abundant as it is harmless in the sense that it does not cause critical misunderstandings. Most linguistic ambiguities pass unnoticed or only become apparent upon subsequent review, mainly for two reasons: humans either perform automatic and unconscious disambiguation based on prior knowledge or access to additional sources of external information; or they simply leave ambiguities unresolved of their resolution is not strictly required in the given communicative context. The latter cognitive strategy has been discussed in the literature under the label ‘Good-Enough Approach’. An important source of external information to support automatic and unconscious disambiguation is visual scene context.

Tenet 2: ***One of the human cognitive strategies to support the resolution of inherent linguistic ambiguity is to access non-linguistic sources of information such as visual scene context.***

In contrast to the majority of artificial systems, humans are capable of processing language input contextually. This capability allows the enrichment of the purely linguistic input by additional, possibly non-linguistic information to support disambiguation. Humans automatically access and integrate their knowledge of the world, the discourse context, the speaker or author, the domain and many other factors in order to constrain the set of possible interpretations of a given linguistic utterance. In the absence of other contextualisations, a listener will try to establish cross-modal referential links between the linguistic input and entities in the co-present visual scene.

Tanenhaus et al. used eye-tracking to study the effect of visual scene information on the processing of syntactically ambiguous sentences. They observed that fixations in the visual scene closely align with the linguistic stimulus unfolding over time. Their conclusion was that visual and linguistic processing operate in close temporal interlock. More importantly, they found that different visual scenes resulted in different patterns of anticipatory eye movements for the same structurally ambiguous sentence. These findings provide compelling empirical evidence that linguistic processing and visual context information interact with each other from the earliest processing stages onwards. Section 2.3 addresses the implications of these experiments in detail.

Tenet 3: ***The interaction between visual scene understanding and syntactic processing constitutes a cross-modal interaction of representational modalities; it is therefore fundamentally different from cross-modal interactions at sensory level.***

Research into multisensory integration has produced a number of remarkable findings over the last three decades. Most notably, the neural substrate for certain types of cross-modal processing has been identified. It was shown to exhibit superadditive responses in neuronal activity when processing temporally concurrent bimodal stimuli. For these cases of cross-modal processing a direct correspondence between the multimodal stimulus and the neuronal activity has been established. Importantly, the incoming stimuli are strictly sensory in nature, i.e., they do not convey a meaning beyond the sensory stimulation they invoke. Processing these stimuli, typically light flashes and sound beeps, does not require higher cognitive functions such as the analysis of linguistic meaning or symbolic reasoning.

When we consider the interaction between visual scene perception and language understanding, we are concerned with cognitively higher and — in terms of processing — temporally later processes. For a linguistic stimulus to give rise to a semantic cross-modal interaction with visual understanding, it must first have been perceived sensorially and then have been decoded semantically. The same holds true for the interpretation of a visual scene: understanding a visual scene requires the sensory discrimination of the scene, the categorisation of the entities and situations perceived as well as an extraction of the relations identified between those cognised entities. A semantic cross-modal interaction between vision and language takes place during these later processing stages. In contrast to sensory stimuli, visual scenes and language do carry a meaning beyond the level of the mere sensory stimulation they invoke. Representationalists argue that they give rise to higher-level symbolic rather than sensory mental representations. These representations are informationally encapsulated in that they have a representational encoding of their own. It is for these reasons that we consider visual scene perception and language understanding to be representational modalities. We extensively argue this point in Section 2.1

Tenet 4: *The interaction between vision and language is mediated by a shared level of semantic representation.*

Sections 2.1 to 2.4 discuss a range of interaction phenomena between visual and linguistic processing. Both the Stroop effect (Stroop, 1935) and the findings of Cooper (1974) provide strong empirical support for the claim that the interaction between visual and linguistic processing is mediated by semantics. In the Stroop effect, the meaning of a word exhibits a clear modulating effect on the response times for naming the colour in which the word is printed. Cooper, on the other hand, found that object fixations in humans are significantly influenced by the meaning of auditory linguistic stimuli presented during fixations. Subjects were much more likely to fixate those objects that were semantically related to the meaning of the auditorily presented linguistic stimulus. These findings were corroborated by Huettig et al. (2006) who showed that the influence of the linguistic stimulus results from conceptual rather than associative relatedness.

These empirical observations integrate well into the framework of Jackendoff's Conceptual Semantics. Specifically, they are consistent with Jackendoff's Conceptual Structure Hypothesis (Jackendoff, 1983) which postulates a single, uniform level of semantic representation as the mediating representation for all cross-modal interactions involving language. We adopt Jackendoff's cognitive architecture resulting from Conceptual Semantics as the basis of our computational model. A detailed discussion of Conceptual Semantics is provided in Chapter 3.

Tenet 5: *There are no extant computational models for the influence of visual scene context upon linguistic processing that mediate the contextual influence upon linguistic processing via a shared level of semantic representation.*

Over the last two to three decades a number of computational models have been reported that successfully model the interaction between vision and language to some extent. We discuss the more recent ones in Section 2.5.

All of these models apply to a limited domain and are subject to limitations in their potential to scale up, both with regards to their linguistic and their visual processing scope. Applications arising in the robotics domain exhibit a strong focus on language in the form of speech and are used primarily for object identification and spatial reasoning. The more linguistically motivated implementation efforts such as Mayberry et al. (2005a,b, 2006) are connectionist approaches that suffer from an evident lack of linguistic upward scalability. None of the models reported so far have been derived from an established, comprehensive theory of human cognition. What was missing, hence, is a model for the influence of visual context upon linguistic processing that has large or even unrestricted linguistic scope and is not confined mechanistically to a particular domain. Ideally, this model should be cognitively motivated with an architecture that can be argued for in the context of an established cognitive framework.

Tenet 6: ***Our computational model implements the semantic mediation for the influence of visual context upon linguistic processing as hypothesised in Jackendoff's Conceptual Semantics.***

According to Jackendoff's Conceptual Semantics (Jackendoff, 1983), all sensory modalities project into Conceptual Structure as the central level of semantic representation. The modalities interact with this level of semantic representation via interfaces that map between the informationally encapsulated representations. Conceptual Structure, in turn, interacts with the syntactic level of representation via correspondence rules in the syntax-semantics interface. Sensory modalities such as vision can thus exert an influence upon syntactic processing via Conceptual Structure as the mediating level of representation.
In our model, we implement precisely this flow of information: Starting from the assumption that a projection of the visual percept into Conceptual Structure has already occurred, we propagate the asserted visual context information into syntactic processing. The distinct and informationally encapsulated levels of syntactic and semantic representation interact via constraints in the syntax-semantics interface. These constraints act as correspondence rules between the syntactic and the semantic representations.

Tenet 7: ***The mechanism of assigning cross-modal referential links between linguistic and contextual entities can be approximated by matching linguistic and contextual entities based on the compatibility of the concepts they instantiate.***

Our model performs the process of cross-modal matching, i.e., the assignment of cross-modal referential links, based on the compatibility of the concepts activated in the linguistic modality and the concepts instantiated in the situation-specific representation of visual context. As the decision process is based on intrinsic properties of the activated concepts, it generalises to languages other than German. The Boolean decision of concept compatibility permits the decision of which entities in visual context are potential referents of a given word in the input sentence and which ones are not. The detailed description of how homonyms in the input sentence are mapped to a set of concept instances in visual context is given in Chapter 7.

With conceptual compatibility as the deciding criterion, our model is even capable of establishing cross-modal referential links in cases for which the concepts activated in the two modalities are of different degrees of conceptual specificity. An experimental validation of this capability is reported in Chapter 11.
A short-coming of the implemented Boolean decision criterion is that referential preferences based on different degrees of conceptual overlap cannot be expressed. A more realistic model of cross-modal matching should include the capability of expressing weighted representations of lexical meaning and a graded measure of conceptual similarity.

Tenet 8: ***The proposed model achieves selective syntactic modulations based on the integration of non-linguistic propositional information representing entities and their thematic relations in a visual scene context.***

We use non-linguistic propositional representations to model the entities observed in a visual scene context. Entities are related to each other by thematic relations. This contextual information is propagated into syntactic analysis via a shared level of semantic representation equivalent of Jackendoff's Conceptual Structure. The PPC assigns penalties for certain semantic dependencies based on the context model: dependencies compliant with the assertions in visual context are admitted while other dependencies are penalised. Details of the algorithm are described in Section 7.4.
The integration constraints in the role-assigning grammar use these penalties to constrain the shared semantic representation. The asserted visual context can thus exert a direct influence on semantic representation. Correspondence rules in the syntax-semantics interface propagate the imposed semantic constraints into syntactic representation by ensuring that the syntactic representation of the sentence always is consistent with the shared semantic representation.
In summary, the representation of visual context information gives rise to constraints on the shared level of semantic representation. The semantic representation modulates attachments in the syntactic representation via the correspondence rules in the syntax-semantics interface. Overall, the syntactic modulations based on visual scene context are achieved with semantic mediation.

Tenet 9: ***The proposed model for the influence of visual context upon syntactic processing is capable of diagnosing which aspects of linguistic analysis conflict with contextual preferences.***

Human cognition is very well adapted to the processing of a multitude of diverse cross-modal stimuli. In particular, human cognition can still arrive at a consistent interpretation of a linguistic input even if the visual scene context is in semantic conflict with that analysis. Moreover, human cognition is not only able to identify that a conflict between linguistic analysis and contextual expectation exists, human cognition can also specify concretely which aspects of linguistic analysis conflict with contextual expectations. Noteably, our model also provides this capability.

With hard integration, conflicts between linguistic and contextual preferences cause hard constraint violations and hence result in the dismissal of the corresponding structure from the hypothesis space. Hard context integration, therefore, does not allow any conflicts between linguistic and contextual preferences to arise. To exploit WCDG2's diagnostic capabilities for conflicting contextual and linguistic preferences we need to study soft context integration.
Whether or not a given conflict is resolved in favour of the contextual or the linguistic preference depends on the hardness of the preferences involved. WCDG2 will always attempt to satisfy the harder constraint. In cases where the linguistic preference is harder than the contextual one, WCDG2 satisfies the linguistic constraint at the cost of incurring a violation of the corresponding integration constraint. In Experiment 3 we observed this behaviour for the semantic dependency assignments that were not endorsed by visual context. The violated integration constraint then appears in the list of constraints violated by the given solution candidate to highlight a mismatch between the assigned linguistic analysis and the contextual preferences. WCDG2's list of violated constraints can hence be used as a diagnostic tool to detect conflicts between the linguistic analysis and the contextual assertions in the context model.

12.2 Conclusions

In this thesis we have provided a comprehensive description of a fully operational computational model for the influence of cross-modal context upon syntactic parsing. Our model integrates propositional, non-linguistic information from a representation of visual scene context into the parsing process of a syntax parser. The model is the result of an interdisciplinary research effort in the fields of informatics, cognitive science and linguistics.

The proposed Context Integration Architecture comprises the following components:

1. WCDG2, a weighted-constraint dependency parser, for linguistic processing,

2. the T-Box as an ontology to represent situation-invariant semantic knowledge,

3. an A-Box or context model to represent situation-specific semantic information,

4. FaCT++ as an OWL description logic reasoner over A-Box and T-Box, and

5. the PPC as a scoring component that computes and communicates dependency score predictions to the parser based on A-Box and T-Box queries with the reasoner.

The model uses WCDG2's extended predictor interface to establish a communication between the parser and the PPC. The PPC computes homonym-specific dependency score predictions prior to parse time. These score predictions express the acceptability of a given semantic dependency in the linguistic analysis in view of the modelled visual context information. The dependency score predictions are accessed at parse time and constrain the shared semantic representation. The semantic representation further constrains the syntactic representation by means of the correspondence rules in the syntax-semantics interface. These rules stipulate correspondences between semantic and syntactic dependency constellations.

To compute score predictions, the model establishes cross-modal referential links in a sequence of two steps. In contrast to other existing implementations, especially connectionist approaches, these steps are mechanistically transparent and predictable: First, bottom-up grounding in the linguistic modality assigns each word a set of concepts from the T-Box that denote its meaning. This set-based approach allows the robust modelling of lexical polysemy. In the second step referred to as cross-modal matching, the words in the input sentence are assigned to referents in the visual scene context. A context entity that instantiates a concept which is compatible with at least one of the concepts activated by a given word will be assigned as a cross-modal match. The model attempts to assign cross-modal matches for all words that have a concept-based meaning representation, including verb forms that denote situation rather than object concepts. Most extant models, especially from the robotic domain, restrict their cross-modal matching to linguistic entities denoting physical objects in visual context. We argue that assigning cross-modal matches based on the criterion of conceptual compatibility is a language-independent process that supports the generalisation of our model to languages other than German.
With the use of the weighted constraints in WCDG2, our context integration architecture is capable of expressing degrees of linguistic acceptability. We have shown how soft context integration permits the integration of contextual information that conflicts with the linguistic default preferences. The list of violated constraints provides valuable diagnostic feedback about the identified mismatches between linguistic and contextual preferences in those cases.

We highlight that our model architecture implements central aspects of Jackendoff's Conceptual Semantics; all cross-modal interactions with the syntactic level of representation are mediated by Conceptual Structure as the single, unified level of semantic representation. Conceptual Structure interfaces with both the visual modality and the representation of syntax. Interfaces containing modality-specific correspondence rules map between these informationally encapsulated levels of representation. In our model, we make the assumption that the visual modality already has projected its percept into Conceptual Structure. Our representations of visual context hence reflect the outcome of the process of visual understanding. As such, context models contain visually perceived information that may be enriched by world-knowledge and the results of elementary symbolic reasoning.

We have successfully employed our model for the context-driven disambiguation of three types of global syntactic ambiguities in German: genitive-dative ambiguity in feminine singular nouns, subject-object ambiguity and PP-attachment. Following the evaluation of the extended grammar and the selection of linguistic material to

study (Experiment 1), we investigated the model's behaviour with regards to hard integration (Experiment 2), soft integration (Experiment 3) and conceptual underspecification (Experiment 4). The grammar evaluation revealed its good to very good precisions for the syntactic analysis of unrestricted text in combination with strong limitations in coverage. Hard context integration demonstrated the general technical feasibility of propagating contextual information into linguistic analysis. Hard integration successfully enforces full compliance of the linguistic analysis with visual context. This strong compliance comes at the price of potentially adverse effects on the quality of the semantic analysis in referentially unrelated parts of the input sentence.
Soft context integration resulted in an overall improvement of the model's integration behaviour. Under soft integration, contextual information influences linguistic analysis as a graded preference that can be overridden by sufficiently strong linguistic preferences in the parser's grammar. The result is a more realistic model of vision-language interaction in which contextual preferences are strong enough to influence the process of linguistic analysis but still weak enough to be overruled by harder linguistic preferences. This balance produces a semantically and syntactically well-formed linguistic analysis that is compliant with the asserted context information. In case of conflicting preferences, the list of violated constraints provides valuable diagnostic information about the semantic dependencies that have been assigned in the linguistic analysis but are inconsistent with visual context.

When assigning cross-modal referential links, the presented model can exploit the world-knowledge encoded in the T-Box's conceptual hierarchy. In doing so, cross-modal matches can even be assigned in cases in which the conceptual specificity between visual and linguistic modality differ. This behaviour is in line with our expectation that cross-modal interactions should result in a support of the perceptually less specific modality by the perceptually more specific modality.
We have emphasised that concept generalisation can be utilised to model uncertainty and ambiguity in the categorisation of visual percepts. Our findings show that – despite the system's robust handling of conceptual underspecification – visual context does need to maintain a certain degree of conceptual specificity in order to maintain its disambiguating power. As would be expected, the concepts instantiated in visual context have to exhibit a sufficient degree of conceptual specificity to be able to exert a biasing effect upon linguistic analysis.

While our model is reasonably successful at integrating visual context information into syntactic parsing, it is also subject to a number of important limitations. Despite the ample empirical evidence for a bidirectional interaction between vision and language, our model is technically confined to a unidirectional influence of vision upon linguistic processing. Effects such as language-driven visual behaviour as in visual attention shifting or visual search are outside the scope of the present implementation of the model. The model is primarily constrained by the technical limitations of the predictor architecture. The use of this architecture is based on the assumption that the visual context remains unchanged during parse time. This constitutes a significant simplification, especially in highly dynamic visual environments where visual contexts change rapidly as the linguistic stimulus unfolds. Our model furthermore employs a simplified approach to the representation of word meaning: a word is mapped to a set of concepts, each of which is taken to contribute equally to the word's meaning. This approach does not permit to express referential preferences in case two words can refer to the same entity in visual context. Another limitation of the model arises from the limited linguistic scope of its role-assigning grammar. As a result of the semantic mediation of the contextual influence on syntactic parsing, context integration can only succeed for sentences within the linguistic scope of the role-assigning grammar. The corpus studies performed on unrestricted German language input suggest that further modelling effort is required in order for WCDG2's grammar to reach full coverage of German.

We see the model's primary contribution to the scientific community in the following three aspects:

1. The Context Integration Architecture is the first working implementation of visual context integration centred around a symbolic constraint-based parser.
2. The model is scalable, has a substantial linguistic scope and can be applied to arbitrary domains.
3. The model architecture is cognitively motivated and realises central aspects of Conceptual Semantics, an established and comprehensive theory of human cognition.

12.3 Directions for Future Research

The model presented in this thesis achieves the integration of propositional semantic context into the process of syntactic parsing in a weighted-constraint dependency parser. Contextual information is propagated into the syntactic level of analysis via the syntax-semantics interface which specifies correspondence rules between the syntactic and semantic representations. Visual contexts are asserted as concept instantiations joined by thematic relations. The model's capabilities to integrate contextual information under a number of experimental conditions such as hard and soft integration have been discussed *in extenso*.

We have also highlighted that the model is subject to a number of significant limitations. These limitations offer perspectives for further scientific enquiry. The following list addresses some of the central issues associated with our model and points out directions for further study in the computational modelling of vision-language interactions.

1. **Graded Conceptual Representations**

   While context integration can be performed successfully with the model in its current form, the underlying mechanisms are based on a number of simplifications that may be remedied in the context of future investigations. To begin with, the concept grounding of words in the input sentence is presently set-based; an input word can either activate a given concept or not. Clearly, equal activation of all concepts contained in that set is a somewhat crude approximation of word meaning. More realistic would be a weighted representation of concept activation such that more salient aspects of meaning could be emphasised over less important ones. A suitable internal representation of the different degrees of concept activation may be in the form of fuzzy sets.

   The inability to express conceptual preferences in our model carries over into the process of cross-modal matching. Whether or not a word is assigned a cross-modal match is currently a purely Boolean decision based on concept compatibility. For words with several cross-modal matches, our model offers no way to express a preference for one of those cross-modal matches based on word meaning. Similarly, two words in the input sentence may indeed be assigned the same cross-modal match. It would therefore be desirable to be able to express degrees of preference with which entities in the input sentence establish reference to the entities in cross-modal context. Incorporating these nuances into the process of score prediction would result in a more differentiated and cognitively more plausible influence of visual scene context upon linguistic processing.

2. **Automated Context Model Generation**

   In its present form, our model integrates context models that are generated manually using an ontology editor. This approach defies scalability and prevents the processing of larger text segments with visual context integration. It would be desirable to be able to generate context models automatically, e.g., based on input from computer vision. This would require techniques for image and visual scene understanding. While both of these are in the focus of extensive research activity, the scientific challenges involved are substantial, especially in open domains.

   Alternatively, our model may be extended to integrate contexts other than cross-modal context. In principle, the context representations can also be adapted to represent the semantic analysis of preceding discourse. A high-level sketch of how incremental context model generation from discourse could be approached is given in McCrae and Menzel (2007, Section *Future Work*) where we propose to build up a representation of discourse context based on the incremental addition of semantic analyses of sentences in the discourse. We see the main challenge of this approach in devising a suitable heuristic for the integration and fusion of new discourse information into an existing context representation.

3. **Perceptual Uncertainty at Relation Level**

   The inclusion of the modelling parameter *context compliance* was a first step towards accommodating the fact that visual perception — and in fact *all* sensory perception — is subject to degrees of uncertainty. Our model is capable of handling conceptual underspecification as the result of perceptual uncertainty or ambiguity. So far, however, our model does not permit the modelling of uncertainty in the perception of thematic relations between perceived entities. It may well be the case that two entities A and B are perceived but that the perceiver is uncertain as to which thematic relation these entities entertain with each other. Possible routes towards the processing of uncertain thematic relations may be the introduction of new roles to represent certain types of uncertainty or the support for asserting several thematic roles between two entities. In both cases an extension to the model's scoring and vetoing algorithm as well as the role-assigning grammar would be required.

4. **Bidirectional Vision-Language Interaction**

   All studies of vision-language interactions at sentence level in humans suggest that the interaction between the two modalities is bidirectional. As a direct consequence of the predictor architecture, our model presently implements a unidirectional influence of vision upon language processing. We further assume that each context model only represents a single situation and thus reflects the perceiver's focus of attention in the visual scene. This excludes a number of important vision-language interaction phenomena from the modelling scope. Shifts in visual attention, the perception of multi-situation contexts, the integration of visual contexts that change at parse time, the inclusion of language-initiated top-down effects that give rise to expectation-driven behaviour, as in active vision or visual search, are all outside of our model's scope. For a cognitively more plausible modelling of vision-language interactions, these factors would need to be included.

   A direct implication of attempting to model a bidirectional vision-language interaction must be the dismissal of the predictor architecture in favour of an online-access to contextual information. If the state of linguistic processing at a given point in time affects the information represented in the context model, an online communication between the parser and the context model must be established at that point.

   With an online communication between parser and context model in place, another limitation of the model could be improved, namely the cross-modal matching of words as isolated units. This limitation is a direct consequence of the predictor architecture which is forced to generate score predictions at a point in time when no syntactic information is yet available. Given an online access, syntactic information accumulated at some point in parse time could be utilised for more specific access to contextual information. As an example, consider a sentence about two cars, a green one (*car 1*) and a red one (*car 2*). When syntactic processing has identified the adjectives as dependants of the corresponding nouns, entity-specific context information could be extracted for each one of them, i.e., it would be possible to assign different contextual integration scores for the semantic dependencies of car 1 and car 2. This differentiation cannot be achieved with cross-modal matching at word level in the present implementation of the model.

5. **Incremental Processing**

   An important feature of human cognition in general and linguistic processing in particular is that they evolve over time. Linguistic processing proceeds in close temporal alignment with the unfolding of the linguistic stimulus. WCDG2 as used in our implementation, however, processes complete sentences in its search for the optimal solution candidate. With the functional extension of Beuck (2009), a WCDG derivative is available that can also parse input sentences incrementally. Adding one slot of the input sentence per time increment, this version of WCDG evaluates sentence fragments of growing size until the entire sentence has been parsed.

   In our view, the inclusion of incrementality into linguistic processing in combination with an implementation of a bidirectional interaction between vision and language offers the most promising perspective for future research that arises from our model. Modelling predictions that include incremental and bidirectional language-vision interactions can readily be evaluated against actual behavioural results such as eye-tracking evidence from psycholinguistic experiments or EEG-data from neurophysiological measurements.

6. **Model Coverage**

   While the extension of model coverage is not primarily a scientific question, it poses challenges in language engineering. Seeing that the implementation of this model also involved a significant part of natural language engineering, we consider it valid to include this aspect in the list of future directions.

   Scaling the model – preferably to the level that unrestricted natural language input can be processed without the need for further modifications to the lexicon, the grammar or the T-Box – requires further extensions to the semantic language processing capabilities of the system. At present, the latter are primarily limited by the scope of the semantically annotated lexicon, the T-Box and the role-assigning grammar. While manual extensions to these resources are conceivable, the overall effort involved in doing so may be substantial. We estimate the additional effort to achieve this goal between one and three person years.

   In view of this substantial grammar modelling effort it may be of interest to consider extension approaches based on existing linguistic resources such as semantic lexical networks, large-scale general purpose ontologies etc. Alternative approaches to this challenge could comprise the automated extraction of the required information from semantically annotated corpora.

At the time of writing, we are aware of intense research activity regarding some of these aspects. It is our sincere hope that some of the research questions put forward in this section might help to inspire present or future research endeavours related to this work.

* * *

References

Acartürk, C., Habel, C., and Cagiltay, K. (2008). Multimodal comprehension of graphics with textual annotations: The role of graphical means relating annotations and graph lines. In Howse, J., Lee, J., and Stapleton, G., editors, *Diagrammatic Representation and Inference*, number 5223 in Lecture Notes in Computer Science, pages 335–343. Berlin: Springer.

Ali, A. N. (2007). Exploring semantic cueing effects using McGurk fusion. In Vroomen, J., Swerts, M., and Krahmer, E., editors, *Auditory-Visual Speech Processing 2007 (AVSP2007)*. Tilburg: Universiteit van Tilburg.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world. *Cognition*, 93:B79–B87.

André, E., Herzog, G., and Rist, T. (1988). On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer. In *Proceedings of the 8^{th} European Conference on Artificial Intelligence (ECAI-88)*, pages 449–545.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. Cambridge, MA: MIT Press.

Bauckhage, C., Fritsch, J., Rohlfing, K., Wachsmuth, S., and Sagerer, G. (2002). Evaluating integrated speech- and image understanding. In *Proceedings of the IEEE International Conference on Multimodal Interfaces (ICMI 02)*, pages 9–14.

Baumgärtner, C. (2009). Parsing mit dreistelligen Constraints. Diploma Thesis, Department of Informatics, University of Hamburg, Germany.

Bertelson, P. and Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review*, 5(3):482–489.

Beuck, N. (2009). Inkrementelle Analyse mit Constraint-Dependency-Grammatiken. Diploma Thesis, Department of Informatics, University of Hamburg, Germany.

Brandt, S. A. and Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1):27–38.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The tiger treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories (Sozopol, Bulgaria)*, pages 24–41.

Brick, T. and Scheutz, M. (2007). Incremental natural language processing for hri. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 263–270.

Brown, M. K., Buntschuh, B. M., and Wilpon, J. G. (1992). Sam: A perceptive spoken language understanding robot. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:1390–1402.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of the Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, pages 969–974.

Bushnell, E. W. (1994). *The development of intersensory perception: comparative perspectives*, chapter A Dual-Processing Approach to Cross-Modal Matching: Implications for Development, pages 19–38. Lawrence Erlbaum Associates.

Chella, A., Coradeschi, S., Frixione, M., and Saffiotti, A. (2004). Perceptual anchoring via conceptual spaces. In *Proceedings of the AAAI-04 Workshop on Anchoring Symbols to Sensor Data (San Jose, CA)*. AAAI Press.

Christianson, K., Williams, C. C., Zacks, R. T., and Ferreira, F. (2006). Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*, 42(2):205–238.

Cooper, L. A. (1976). Individual differences in visual comparison process. *Perception and Psychophysics*, 19:433–444.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6:84–107.

Coradeschi, S. and Saffiotti, A. (2001). Perceptual anchoring of symbols for action. In *Proceedings of the 17th IJCAI Conference (Seattle, WA)*, pages 407–412.

Coradeschi, S. and Saffiotti, A. (2003a). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43((2–3): Special issue on perceptual anchoring):85–96.

Coradeschi, S. and Saffiotti, A. (2003b). Perceptual anchoring with indefinite descriptions. In *Proceedings of the First Joint SAIS-SSLS Workshop (Örebro, Sweden)*.

Crain, S. and Steedman, M. (1985). *Natural language parsing: Psychological, computational, and theoretical perspectives*, chapter 10: On not being led up the garden path: the use of context by the psychological syntax processor, pages 320–358. Cambridge: Cambridge University Press.

Dalrymple-Alford, E. C. (1972). Associative facilitation and interference in the Stroop color-word task. *Perception and Psychophysics*, 11:274–276.

Dowty, D. R. (1989). *Properties, Types and Meaning*, chapter On the Semantic Content of the Notion of the 'Thematic Role', pages 69–129. Dordrecht: Kluwer Academic Publishers.

Dunbar, K. N. and MacLeod, C. M. (1984). A horse race of a different color: Stroop interference patterns with transformed words. *Journal of Experimental Psychology: Human Perception and Performance*, 10:622–639.

FaCT-PlusPlus Download Page (2009). http://code.google.com/p/factplusplus/. Link verified: 20 April 2009.

Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15.

Ferreira, F. and Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1/1-2:71–83.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44:516–547.

Fillmore, C. J. (1968). *Universals in Linguistic Theory*, chapter The Case for Case, pages 1–90. New York: Holt, Rinehart and Winston.

Fodor, J. A. (1975). *The Language of Thought*. New York: Thomas Y. Crowell.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical appraisal. *Cognition*, 28:3–71.

Foth, K. A. (2006). Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Handbook.

Foth, K. A. (2007). *Hybrid Methods of Natural Language Analysis*. PhD thesis, Department of Informatics, Hamburg University, Germany.

Foth, K. A. and Menzel, W. (2006a). The benefit of stochastic PP-attachment to a rule-based parser. In *Proceedings of the 21st International Conference on Computational Linguistics (Sydney, Coling-ACL-2006)*.

Foth, K. A. and Menzel, W. (2006b). Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (Sydney, Australia)*, pages 321–328. Morristown, NJ: Association for Computational Linguistics.

Foth, K. A., Menzel, W., and Schröder, I. (2000). A transformation-based parsing technique with anytime properties. In *4th International Workshop on Parsing Technologies (IWPT-2000)*, pages 89–100.

Gee, J. P. (2001). Reading as situated language: A sociocognitive perspective. *Journal of Adolescent & Adult Literacy*, 44(8):714–725.

Gruber, J. S. (1965). *Studies in Lexical Relations. Indiana University Linguistics Club, Bloomington, Indiana. Reprinted 1976 as part of Lexical Structures in Syntax and Semantics. North-Holland, Amsterdam.* PhD thesis, MIT, Cambridge, MA.

Habel, C. and Acartürk, C. (2009). Eye-tracking evidence for multimodal language-graphics comprehension: The role of integrated conceptual representations. In Navarretta, C., Paggio, P., Allwood, J., Alsén, E., and Katagiri, Y., editors, *Proceedings of the NODALIDA 2009 Workshop on Multimodal Communication - from Human Behaviour to Computational Models*, volume 6, pages 9–14. Odense, Denmark: Northern European Association for Language Technology (NEALT).

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Harper, M. P. and Helzermann, R. A. (1995). Extensions to constraint dependency parsing for spoken language processing. *Computer Speech and Language*, 9(3):187–234.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504.

Huettig, F. and Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition*, 96:B23–B32.

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. T. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121:65–80.

Jackendoff, R. (1983). *Semantics and Cognition.* Cambridge, MA: MIT Press.

Jackendoff, R. (1990). *Semantic Structures.* Cambridge, MA: MIT Press.

Jackendoff, R. (1992). *Languages of the Mind.* Cambridge, MA: MIT Press.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F., editors, *Language and Space*, chapter 1, pages 1–30. Cambridge, MA: MIT Press.

Kako, E. (2006). The semantics of syntactic frames. *Language and Cognitive Processes*, 21(5):562–575.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, Vol. 39, No. 2(2):170–210.

Khmylko, L. (2007). Hybrid parsing with a maximum spanning tree predictor. Master's thesis, Hamburg University of Technology, Germany.

Khmylko, L., Foth, K. A., and Menzel, W. (2009). Co-parsing with competitive models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP '09), Borovets (Bulgaria)*, pages 173–179.

Kintsch, W. and van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, 85, Number 5:364–394.

Koller, D., Heinze, N., and Nagel, H.-H. (1991). Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 90–95.

Löbner, S. (2003). *Understanding Semantics.* London: Arnold.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2):163–203.

Maruyama, H. (1990). Structural disambiguation with constraint propagation. In *Proceedings of the 28th Annual Meeting of the ACL (ACL-90)*, pages 31–38, Pittsburgh, PA.

Massaro, D. W. and Stork, D. G. (1998). Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, 86:236–244.

Mast, F. W. and Kosslyn, S. M. (2002). Eye movements during visual mental imagery. *Trends in Cognitive Sciences*, 6(7):271–272.

Matin, E., Shao, K. C., and Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, 53(4):372–380.

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. S. (2005a). A connectionist model of anticipation in visual worlds. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing, (IJCNLP-05, Jeju, Korea).*

Mayberry, M. R., Crocker, M. W., and Knoeferle, P. S. (2005b). A connectionist model of sentence comprehension in visual worlds. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society (COGSCI-05, Stresa, Italy).* Mahwah, NJ: Erlbaum.

Mayberry, M. R., Knöferle, P. S., and Crocker, M. W. (2006). A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 481–529, Vancouver, Canada.

McConkie, G. W. and Currie, C. B. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 22:563–581.

McCrae, P. (2007). Integrating cross-modal context for PP attachment disambiguation. In *Proceedings of the 3rd International Conference on Natural Computation (ICNC 2007, Haikou, China)*, volume 3, pages 292–296. Los Alamitos, CA: IEEE.

McCrae, P. (2009). A model for the cross-modal influence of visual context upon language processing. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 09, Borovets, Bulgaria)*, pages 230–235. Shoumen: INCOMA.

McCrae, P., Foth, K., and Menzel, W. (2008). Modelling global phenomena with extended local constraints. In Villadsen, J. and Christiansen, H., editors, *Proceedings of the 5th International Workshop on Constraints and Language Processing (CSLP 2008, Hamburg, Germany)*, pages 48–60. Roskilde University.

McCrae, P. and Menzel, W. (2007). Towards a system architecture for integrating cross-modal context in syntactic disambiguation. In Sharp, B. and Zock, M., editors, *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007, Funchal, Portugal)*, pages 228–237. INSTICC Press.

McDonald, R. (2006). *Discriminative Learning and Spanning Tree Algorithms for Dependency Parsing*. PhD thesis, University of Pennsylvania.

McDonald, R., Lerman, K., and Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 746–748:246.

McGurk, H. and MacDonald, J. (1978). Visual influences on speech perception process. *Perception and Psychophysics*, 24:253–257.

Menzel, W. and Schröder, I. (1998). Decision procedures for dependency parsing using graded constraints. In Kahane, S. and Polguère, A., editors, *Proceedings of the Coling-ACL Workshop on Processing of Dependency-based Grammars*, pages 78–87, Montreal, Canada.

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *The Journal of Neuroscience,*, 7(10):3215–3229.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroedera, C. E., and Foxea, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, 14:115–128.

NEGRA Homepage (2006). http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html. Link verified: 03 March 2010.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). Word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/Intro.htm. Link verified: 10 June 2009.

O'Regan, J. K. (1992). Solving the 'real' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Experimental Psychology*, 46(3):461–488.

O'Regan, J. K., Deubel, H., Clark, J. J., and Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, 7:191–211.

OWL API Homepage (2009). http://owlapi.sourceforge.net/. Link verified: 07 August 2009.

Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Potamianos, G., Neti, C., Iyengar, G., and Helmuth, E. (2001). Large-vocabulary audio-visual speech recognition by machines and humans. In *Proceedings of the John Hopkins Summer 2000 Workshop on Signal Processing*, pages 619–624.

Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. In Bailly, G., Vatikiotis-Bateson, E., and Perrier, P., editors, *Issues in Visual and Audio-visual Speech Processing*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, 4(5):197–207.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects and, situated vision. *Cognition*, 80:127–158.

Rensink, R. A., O'Regan, J. K., and Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373.

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, 110(1):88–125.

Rogers, T. T. and McClelland, J. L. (2004). *Semantic Cognition*. Cambridge, MA: MIT Press.

Roy, D. and Mukherjee, N. (2005). Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248.

Saffiotti, A. and LeBlanc, K. (2000). Active perceptual anchoring of robot behavior in a dynamic environment. In *Proceedings of the IEEE Internaltional Conference on Robotics and Automation (ICRA), San Francisco, CA*, pages 3796–3802.

SALSA Corpus Homepage (2009). http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index. Link verified: 29 June 2009.

Schröder, I. (2002). *Natural Language Parsing with Graded Constraints.* PhD thesis, Department of Informatics, Hamburg University, Germany.

Schröder, I., Pop, H. F., Menzel, W., and Foth, K. (2001). Learning grammar weights using genetic algorithms. In *Proceedings of the Euroconference on Recent Advances in Natural Language Processing (Tsigov Chark, Bulgaria)*, pages 235–239.

Schröder, I., Pop, H. F., Menzel, W., and Foth, K. A. (2002). Learning weights for a natural language grammar using genetic algorithms. In Giannakoglou, K. C., Tsahalis, D. T., Periaux, J., Papaillou, K. D., and Fogarty, T., editors, *Evolutionary Methods for Design, Optimisation and Control*, pages 243–247. Barcelona: CIMNE.

Schulz, M., Hamerich, S., Schröder, I., Foth, K., and By, T. (2003). *[X]CDG User Guide, Version 1.3.* Natural Language Systems Group, Department of Informatics, Hamburg University, Germany.

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, 14:147–152.

Simons, D. J. and Ambinder, M. S. (2005). Change blindness: Theory and consequences. *Current Directions in Psychological Science*, 14:44–48.

Simons, D. J. and Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1):16–20.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In Jacobs, P., editor, *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 88–95. Morgan Kaufmann Publishers.

Socher, G. (1997). *Qualitative Scene Descriptions from Images for Integrated Speech and Image Understanding.* PhD thesis, Technical Faculty, University of Bielefeld, Germany.

Socher, G., Sagerer, G., Kummert, F., and Fuhr, T. (1996). Talking about 3d scenes: Integration of image and speech understanding in a hybrid distributed system. In *Proceedings of the International Conference on Image Processing (ICIP-96), Lausanne*, page 18A2.

Socher, G., Sagerer, G., and Perona, P. (2000). Bayesian reasoning on qualitative descriptions from images and speech. *Image And Vision Computing*, 18(2):155–172.

Spivey, M. J. and Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological Research*, 65:235–241.

Spivey, M. J., Richardson, D. C., and Fitneva, S. A. (2004). *The Interface of Vision, Language, and Action*, chapter 5 Thinking outside the brain: Spatial indices to linguistic and visual information, pages 161–189. New York: Psychology Press.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., and Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, 45:447–481.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., and Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science*, 12(4):282–286.

Srihari, R. K. (1995). Computational models for integrating linguistic and visual information: A survey. *Artificial Intelligence Review*, 8:349–369.

Srihari, R. K. and Burhans, D. T. (1994). Visual semantics: Extracting visual information from text accompanying pictures. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pages 793–798.

Strohner, H., Sichelschmidt, L., Duwe, I., and Kessler, K. (2000). Discourse focus and conceptual relations in resolving referential ambiguity. *Journal of Psycholinguistic Research*, 29(5):497–516.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–662.

STTS Tag Set (2009). http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts.html. Link verified: 24 July 2009.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *SCIENCE*, 268:1632–1634.

TIGER Corpus Homepage (2009). http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/. Link verified: 29 June 2009.

van Kampen, A. (2001). *Syntaktische und semantische Verarbeitungsprozesse bei der Analyse strukturell mehrdeutiger Verbfinalsätze im Deutschen: Eine empirische Untersuchung*. PhD thesis, Freie Universität Berlin, Germany.

Wachsmuth, S., Brandt-Pook, H., Socher, G., Kummert, F., and Sagerer, G. (1999). Multilevel integration of vision and speech understanding using bayesian networks. In Christensen, H. I., editor, *Proceedings of the International Conference on Vision Systems, Gran Canaria, Spain*, number 1542 in Lecture Notes in Computer Science (LNCS), pages 231–254.

Wallace, M. T., Meredith, M. A., and Stein, B. E. (1998). Multisensory integration in the superior colliculus of the alert cat. *Journal of Neurophysiology*, 80:1006–1010.

WCDG Download (2009). http://nats-www.informatik.uni-hamburg.de/view/CDG/DownloadPage. Link verified: 20 April 2009.

Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language*, 50(2):212–230.

Winograd, T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. PhD thesis, Massachusetts Institute of Technology.

Winograd, T. (1972). *Understanding Natural Language*. Academic Press.

Wittgenstein, L. (1953). *Bemerkungen über die Philosophie der Psychologie: Letzte Schriften über die Philosophie der Psychologie*, volume Werkausgabe 7. Frankfurt am Main: Suhrkamp, 2005 edition.

Yantis, S. and Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance*, 16:121–134.

Appendix

I List of Requirements

The following table lists all the requirements identified in Part I of this thesis. For each requirement the table lists the requirement ID, the requirement body, the page on which the requirement has been identified, the requirement's implementation status in the model and the page on which the implementation-related information is provided in this document.

ID	Modelling Requirement	Page	Status	Feature
R1	In a model for the interaction between visual context and linguistic understanding, the cross-modal interaction must be mediated by a representation of linguistic meaning.	p. 17	Fully implemented.	p. 84
R2	In a model for the interaction between visual context and linguistic understanding, linguistic processing must be incremental.	p. 19	Not implemented. Fundamental extension.	p. 83
R3	A model for the interaction between visual scene context and linguistic processing must be based on temporally synchronised interactions between the visual modality and linguistic processing.	p. 20	Not implemented. Fundamental extension.	p. 83
R4	A model for the interaction between visual scene context and linguistic processing must be based on continual interactions between non-linguistic information and linguistic processing.	p. 20	Fully implemented.	p. 83

R5	A model for the interaction between visual scene context and linguistic processing must include the influence of visual understanding upon linguistic processing.	p. 20	Fully implemented.	p. 82
R6	A model for the interaction between visual scene context and linguistic processing must include the influence of linguistic processing upon visual understanding.	p. 20	Not implemented. Fundamental extension.	p. 82
R7	In a model for the interaction between visual scene context and linguistic processing, referentially unrelated visual context information must leave linguistic processing unaffected.	p. 21	Partially implemented. Completion constitutes substantial extension.	p. 146
R8	In a model for the interaction between visual scene context and linguistic processing, linguistic processing interacts with a representation of the visual scene context.	p. 22	Fully implemented.	p. 128
R9	A model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics must contain distinct levels of representation for syntax and semantics.	p. 35	Fully implemented.	p. 75
R10	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the set of permissible representations on a given level of representation must be defined by a finite set of well-formedness rules.	p. 36	Fully implemented.	p. 77
R11	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the encoding of each representational level is domain-specific.	p. 37	Fully implemented.	p. 77

R12	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the processing on each level of representation is representationally encapsulated.	p. 37	Fully implemented.	p. 83
R13	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the mapping between representations is achieved by correspondence rules.	p. 37	Fully implemented.	p. 84
R14	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, the interaction of levels of representation via representational interfaces occurs online, i.e., at the time of linguistic processing.	p. 37	Fully implemented.	p. 84
R15	A model of Conceptual Structure must encode information about concepts, individuals, taxonomic concept relations and relational predicates such as concept-to-concept and concept-to-individual relations. It must also provide the capability to evaluate the truthfulness of entailment between encoded propositions as well as the consistency between concepts.	p. 39	Fully implemented.	p. 112
R16	A model of Conceptual Structure must contain pointers to the representation of sensory information.	p. 39	Not implemented. Fundamental extension.	p. 109
R17	A model of Conceptual Structure must encode quantification and quantifier scope.	p. 39	Partially implemented. Fundamental extension.	p. 103
R18	A model of Conceptual Structure must provide abstract representations of actions and acting entities.	p. 39	Fully implemented.	p. 109

R19	A model of Conceptual Structure must provide social predicates.	p. 39	Not implemented. Minor extension.	p. 100
R20	A model of Conceptual Structure must provide modal predicates to express semantic notions such as negation or conditionality.	p. 40	Not implemented. Major extension.	p. 107
R21	A model of Conceptual Structure must encode the semantic part of linguistic representation within Conceptual Structure.	p. 40	Fully implemented.	p. 77
R22	A model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics must contain a single, uniform level of semantic representation. This level interfaces with the syntactic level of representation and constitutes the central representation of linguistic and non-linguistic semantics. Meaning-based interactions between non-linguistic modalities and language must be mediated by this level of representation.	p. 41	Fully implemented.	p. 84
R23	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, a verb's lexical entry must indicate for each argument slot from which conceptual categories the argument fillers may preferably be selected.	p. 44	Not implemented. Major extension.	p. 80
R24	In a model for the interaction between non-linguistic modalities and linguistic understanding, a verb's thematic roles must be relateable to its syntactic argument structure via correspondence rules in the syntax-semantics interface.	p. 44	Not implemented. Minor extension.	p. 80

R25	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, every concept instance must instantiate at least one concept from the concept hierarchy.	p. 45	Fully implemented.	p. 111
R26	In a model for the interaction between non-linguistic modalities and linguistic understanding based on Conceptual Semantics, Conceptual Structure representations are verb-centric.	p. 45	Fully implemented.	p. 112
R27	A model for the interaction between non-linguistic modalities and linguistic understanding must have the capability to discriminate individuating features of visual and linguistic input at sensory level.	p. 48	Partially implemented. Fundamental extension.	p. 116
R28	A model for the interaction between non-linguistic modalities and linguistic understanding must be capable of categorising sensory input in conceptual categories based on a set of individuating features.	p. 48	Partially implemented. Fundamental extension.	p. 116
R29	A model for the interaction between non-linguistic modalities and linguistic understanding needs to provide a mechanism for establishing cross-modal referential links by matching entities from the linguistic modality with concept instances from the interacting non-linguistic modalities.	p. 51	Fully implemented.	p. 121
R30	A model for the interaction between non-linguistic modalities and linguistic understanding needs to provide a mechanism for establishing cross-modal referential links by matching concept instances from the non-linguistic modalities with entities from the linguistic modality.	p. 51	Not implemented. Major extension.	p. 121

R31	A WCDG predictor for scoring meaning-related dependencies must be able to differentiate between different readings of a slot string and must be capable of generating separate, homonym-specific predictions for those readings.	p. 51	Fully implemented.	p. 132
R32	To enable the processing of different external predictions for the readings of a slot string, WCDG2 must provide homonym-specific encoding and retrieval of predictions.	p. 68	Fully implemented.	p. 82

II The Asserted T-Box Class Hierarchy

```
Thing
  → Entity.Concept
      → Abstract
          → Address
          → Application
          → Article
          → Clarity
          → Cogeneration
          → Comment
          → Commitment
          → Comparison
          → Compensation
          → Competition
          → Damage
          → Danger
          → Debts
          → Decision
          → Demand
          → Electricity
          → Every.Day.Life
          → Excursion
          → Extrapolation
          → Force
              → Hit
              → Kick
          → Geographic.Region
              → Mecklenburg-Western_Pomerania
              → Schleswig.Holstein.Concept
          → Group
              → Agency
              → Commission
              → Company
              → Europe
              → Family
              → Government
              → Imperial.Armed.Forces
              → Management
              → Nippon
              → Peoples.Party
              → Pool
              → Prosecutor
              → Tenthousand
              → Trade.Union
              → War.Party
          → Hour
          → Idiom
          → Incident
          → Language
              → English
              → French
              → German
          → Law
              → Tax.Law
          → Leadership
          → Lexicalisation
          → Location
          → Market
          → Megawatt
          → Month
          → Mood
          → Negotiation
          → Night
          → Patience
          → Payment
          → Peace
```

  - → Place
  - → Police
  - → Price
  - → Promise
  - → Proof
  - → Representative
  - → Restructure
  - → Return
  - → Saving
  - → Sleep
  - → Song
  - → Source
  - → Statement
  - → Storm
  - → Strike
  - → Technology
  - → Time
  - → Warning
  - → Weekday
    - → Friday
    - → Monday
    - → Saturday
    - → Sunday
    - → Thursday
    - → Tuesday
    - → Wednesday

- → Concrete
  - → Human.m.f
    - → Adult
      - → Accuser.m.f
        - → Accuser.f
        - → Accuser.m
      - → Admirer.m.f
        - → Admirer.f
        - → Admirer.m
      - → Aged.m.f
        - → Aged.f
        - → Aged.m
      - → Boss
      - → Consumer.m.f
        - → Consumer.f
        - → Consumer.m
      - → Holidaymaker.m.f
        - → Holidaymaker.f
        - → Holidaymaker.m
      - → Man
        - → Man.pl
        - → Man.sg
          - → Mueller.Concept
          - → Rabin.Concept
          - → Schulz.Concept
          - → Soares.Concept
      - → Neighbour.m.f
        - → Neighbour.f
        - → Neighbour.m
      - → Parent.m.f
        - → Father
        - → Mother
      - → Professional.m.f
        - → Actor.m.f
          - → Actor
          - → Actress

- → Advertiser.m.f
  - → Advertiser.f
  - → Advertiser.m
- → Author.m.f
  - → Author.f
  - → Author.m
- → Baker.m.f
  - → Baker.f
  - → Baker.m
- → Beggar.m.f
  - → Beggar.f
  - → Beggar.m
- → Employee.m.f
  - → Employee.f
  - → Employee.m
- → Entrepreneur.m.f
  - → Entrepreneur.f
  - → Entrepreneur.m
- → Farmer.m.f
  - → Farmer.f
  - → Farmer.m
- → Gymnast.m.f
  - → Gymnast.f
  - → Gymnast.m
- → Hair.Dresser.m.f
  - → Hair.Dresser.f
  - → Hair.Dresser.m
- → Headmaster
- → Headmistress
- → Maid
- → Manager.m.f
  - → Manager.f
  - → Manager.m
- → Medical.Doctor.m.f
  - → Medical.Doctor.f
  - → Medical.Doctor.m
- → Mountaineer.m.f
  - → Mountaineer.f
  - → Mountaineer.m
- → Nurse.m.f
  - → Nurse.f
  - → Nurse.m
- → Painter.m.f
  - → Painter.f
  - → Painter.m
- → Parson.m.f
  - → Parson.f
  - → Parson.m
- → Pharmacist.m.f
  - → Pharmacist.f
  - → Pharmacist.m
- → Police.Officer.m.f
  - → Police.Man
  - → Police.Woman
- → President.m.f

```
                    → President.f
                    → President.m
                → Professor.m.f
                    → Professor.f
                    → Professor.m
                → Publisher.m.f
                    → Publisher.f
                    → Publisher.m
                → Researcher.m.f
                    → Researcher.f
                    → Researcher.m
                → Sales.Assistent.m.f
                    → Sales.Assistent.f
                    → Sales.Assistent.m
                → Sales.Rep.m.f
                    → Sales.Rep.f
                    → Sales.Rep.m
                → Sociologist.m.f
                    → Sociologist.f
                    → Sociologist.m
                → Speaker.m.f
                    → Speaker.f
                    → Speaker.m
            → Protester.m.f
                → Protester.f
                → Protester.m
            → Smoker.m.f
                → Smoker.f
                → Smoker.m
            → Student.m.f
                → Student.f
                    → PhD.Student.f
                → Student.m
                    → PhD.Student.m
            → Sufferer.m.f
                → Sufferer.f
                → Sufferer.m
            → Terrorist.m.f
                → Terrorist.f
                → Terrorist.m
            → Woman
                → Movie.Diva
        → Cousin.m.f
            → Cousin.f
            → Cousin.m
        → Guest
        → Human.f
        → Human.m
        → Inhabitant.m.f
            → Inhabitant.f
            → Inhabitant.m
        → Martyr.m.f
            → Martyr.f
            → Martyr.m
            → Member
            → Offspring
                → Child
                    → Boy
                    → Girl
                → Daughter
                → Son
            → Patient.m.f
                → Patient.f
                → Patient.m
            → Ruffian.m.f
                → Ruffian.f
                → Ruffian.m
            → Visitor.m.f
                → Visitor.f
                → Visitor.m
        → Physical.Object
            → Aeroplane
            → Airport
            → Award
            → Basket
            → Bier
            → Binocular
            → Book
                → Diary
            → Bouquet
            → City
            → Clock
            → Coast
            → Diagnosis
            → Gas
            → Gun
            → Highwater
            → House
            → Letter
            → Newspaper
            → Plant
            → Prescription
            → Stick
            → Street
            → Sun
            → Table
            → Water
            → Wind
    → Entity.Feature
        → Age
            → Old
            → Young
        → Personal.Pronoun
            → He
            → It
            → She
            → They
                → They.f
                → They.m
                → They.mixed
→ Helper.Concept
    → Lexicalised.Concept
    → Participant
        → AGENT
        → RECIPIENT
        → THEME
        → THEME_THEME
        → Situation

→ Meta.Data
        → Natural.Gender
                → Female
                → Male
                → Mixed
                → Neuter
        → Number
              → Plural
              → Singular
→ Situation.Concept
      → Binary.Situation
              → Takes.AGENT.RECIPIENT
                      → Jmd.Trauen

              → Takes.AGENT.THEME
                    → Etw.Abschalten
                    → Etw.Abwehren
                    → Etw.Aendern
                    → Etw.Anbieten
                    → Etw.Anrichten
                    → Etw.Antreten
                    → Etw.Argumentieren
                    → Etw.Aufspueren
                    → Etw.Aufsuchen
                    → Etw.Bedienen
                    → Etw.Belasten
                    → Etw.Beobachten
                    → Etw.Bitten
                    → Etw.Daempfen
                    → Etw.Draengen
                    → Etw.Erwarten
                    → Etw.Erwerben
                    → Etw.Erwirtschaften
                    → Etw.Fordern
                    → Etw.Fragen
                    → Etw.Greifen
                    → Etw.Halten
                    → Etw.Herausgreifen
                    → Etw.Kaufen
                    → Etw.Landen
                    → Etw.Liefern
                    → Etw.Malen
                    → Etw.Nennen
                    → Etw.Praesentieren
                    → Etw.Richten
                    → Etw.Schaffen
                    → Etw.Schenken
                    → Etw.Schicken
                    → Etw.Schildern
                    → Etw.Sehen
                    → Etw.Sein
                    → Etw.Senden
                    → Etw.Sprechen
                    → Etw.Spueren
                    → Etw.Suchen
                    → Etw.Tragen
                    → Etw.Treffen
                    → Etw.Treten
                    → Etw.Trinken
                    → Etw.Uebergeben
                    → Etw.Uebermitteln
                    → Etw.Verbieten
                    → Etw.Verbringen
                    → Etw.Verderben
                    → Etw.Verkaufen
                    → Etw.Verlangen
                    → Etw.Versorgen
                    → Etw.Vertreten
                    → Etw.Vorsingen
                    → Etw.Wissen
```

- → Etw.Zeigen
- → Etw.Zurueckweisen
- → Fuer.Etw.Sorgen
- → Jmd.Auffordern
- → Jmd.Beschuldigen
- → Jmd.Bitten
- → Jmd.Richten
- → Jmd.Verdaechtigen

- → Ternary.Situation
  - → Takes.AGENT.RECIPIENT.THEME
    - → Jmd.Etw.Anbieten
    - → Jmd.Etw.Geben
    - → Jmd.Etw.Greifen
    - → Jmd.Etw.Herausgreifen
    - → Jmd.Etw.Kaufen
    - → Jmd.Etw.Liefern
    - → Jmd.Etw.Nennen
    - → Jmd.Etw.Praesentieren
    - → Jmd.Etw.Schenken
    - → Jmd.Etw.Schicken
    - → Jmd.Etw.Schildern
    - → Jmd.Etw.Schulden
    - → Jmd.Etw.Sein
    - → Jmd.Etw.Senden
    - → Jmd.Etw.Stehlen
    - → Jmd.Etw.Suchen
    - → Jmd.Etw.Tragen
    - → Jmd.Etw.Uebergeben
    - → Jmd.Etw.Verbieten
    - → Jmd.Etw.Verderben
    - → Jmd.Etw.Verkaufen
    - → Jmd.Etw.Vertreten
    - → Jmd.Etw.Vorsingen
    - → Jmd.Etw.Zeigen
  - → Takes.AGENT.THEME.THEME
    - → Jmd.Etw.Fragen
- → Unary.Situation
  - → Takes.AGENT
    - → Null.Abschalten
    - → Null.Arbeiten
    - → Null.Argumentieren
    - → Null.Belasten
    - → Null.Daempfen
    - → Null.Fragen
    - → Null.Landen
    - → Null.Praesentieren
    - → Null.Richten
    - → Null.Schlafen
    - → Null.Senden
    - → Null.Sprechen
    - → Null.Suchen
    - → Null.Treffen
    - → Null.Uebergeben
    - → Null.Verderben
    - → Null.Vorsingen
    - → Null.Zurueckkehren
    - → Null.wissen

III Derivations

Be

$\mathscr{N}_{n\text{-}ary}$ the upper bound for the number of n-ary constraint evaluations,

C_{unary} the set of unary constraints in the grammar,

$\mathscr{P}_m$ the upper bound for the number of attachment possibilities in a constellation of m labelled dependencies,

n_{max} the maximum number of homonyms per slot,

s the number of slots in the sentence, and

λ_i the number of labels on level of analysis i.

III.1 The Absolute Upper Bound for the Number of Unary Constraint Evaluations as given in Equation (4.3)

Consider the number of possible combinations to attach a dependant homonym to a corresponding regent homonym by an unlabelled dependency. Since no slot in the input sentence contains more than n_{max} homonyms, the upper bound for the number of homonyms in the sentence is given by $s \cdot n_{max}$.

Since every dependant homonym connects to exactly one regent homonym, and we include `ROOT` as a potential regent, the number of homonym-to-homonym attachments in the sentence is bounded above by $s^2 \cdot n_{max}^2$.

Furthermore, a given edge can take $\sum_i \lambda_i$ labels across all levels of analysis i. The number of labelled attachment possibilities for a single edge constellation ($m = 1$) in the sentence $\mathscr{P}_1$ is therefore given by

$$\mathscr{P}_1 = n_{max}^2 \cdot s^2 \cdot \sum_i \lambda_i$$

The grammar contains $|C_{unary}|$ unary constraints, all of which need to be evaluated on all possible edge and labelled attachment combinations. The upper bound for the number of unary constraint evaluations $\mathscr{N}_{unary}$ is therefore given by

$$\mathscr{N}_{unary} = |C_{unary}| \cdot n_{max}^2 \cdot s^2 \cdot \sum_i \lambda_i \ ,$$

which is Equation (4.3), as was to be shown.

III.2 The Absolute Upper Bound for the Number of Binary Constraint Evaluations as given in Equation (4.4)

The derivation for the upper bound on the the number of binary constraint evaluations is based on some of the results from the derivation of the upper bound for the number of unary constraint evaluations in Appendix III.1.

For a constellation of two labelled dependency edges the number of possible labelled attachments is the product of the number of possible labelled constellations $\mathscr{P}_1$ for the participating edges. $\mathscr{P}_2$ is hence given by

$$\mathscr{P}_2 = n_{max}^4 \cdot s^4 \cdot \left[\sum_i \lambda_i \right]^2$$

The grammar contains $|C_{binary}|$ binary constraints, all of which need to be evaluated on all possible labelled attachment combinations. The upper bound for the number of binary constraint evaluations $\mathscr{N}_{binary}$ is therefore given by

$$\mathscr{N}_{binary} = |C_{binary}| \cdot n_{max}^4 \cdot s^4 \cdot \left[\sum_i \lambda_i \right]^2,$$

which is Equation (4.4), as was to be shown.

IV List of Studied Sentences

IV.1 Unified Sentences with Genitive-Dative Ambiguity

ID	Sentence
VK-011	Er wusste, dass die Magd der Bäuerin den Korb suchte.
VK-100	Er wusste, dass der Verehrer der Schauspielerin den Blumenstrauß schenkte.
VK-111	Er wusste, dass der Sohn der Raucherin die Laune verdarb.
VK-151	Er wusste, dass die Bergsteiger der Referentin die Warnung schickten.
VK-226	Er wusste, dass die Nachbarin der Rektorin die Adresse nannte.
VK-233	Er wusste, dass die Cousine der Besucherin den Vorfall schilderte.
VK-247	Er wusste, dass die Pflegerin der Greisin den Ausflug verbot.
VK-263	Er wusste, dass die Verlegerin der Autorin den Artikel verkaufte.
VK-274	Er wusste, dass die Doktorandin der Forscherin den Beweis lieferte.
VK-306	Er wusste, dass die Managerin der Unternehmerin den Vertreter sendete.

IV.2 SALSA-Sentences with Subject-Object Ambiguity

ID	Sentence
SO-360	Zehntausende Demonstranten trugen die Bahren der “Märtyrer” durch die Straßen der Zweimillionen-Stadt.
SO-706	Markt & Technik fordert Geduld
SO-841	Bis auf wenige Stunden Schlaf arbeiten diese Frauen rund um die Uhr, weil sie zu Hause Mann und Kinder versorgen.
SO-1090	Japan bittet Europa um Geduld
SO-4493	Erhebliche Schäden richteten Stürme und Hochwasser am Samstag in mehreren Städten an der Küste von Schleswig-Holstein und Mecklenburg-Vorpommern an.
SO-6179	Die mitregierende Volkspartei (ÖVP) wies seine Forderung als “obszön” zurück:
SO-9681	Statt dessen schicken sie Werber von Haus zu Haus.
SO-9792	Sie vertritt die Gesellschaft, und ihr obliegt die Geschäftsführung.
SO-10744	Beide Kriegsparteien drängten sie, an den Verhandlungstisch zurückzukehren.
SO-40722	Die Kommission fordert die Bundesregierung nun auf, binnen eines Monats für Klarheit zu sorgen.

IV.3 SALSA-Sentences with PP-Attachment Ambiguity

ID	Sentence
PP-3025	Dort griff die Polizei unter Gewaltanwendung einzelne Demonstranten heraus, wobei Tritte und Schläge mit dem Knüppel von Polizisten beobachtet wurden.
PP-3277	Nach Darstellung der Nippon-Firma hatten Gewerkschaftsvertreter Bezahlung für die Zeit verlangt, die sie während früherer Streiks in Verhandlungen mit der Unternehmensleitung verbrachten.
PP-3839	Staatschef Soares argumentiert, daß die Regierung nicht einfach Gesetze mit früheren Laufbahnzusagen kurzfristig ändern könne.
PP-7177	Insgesamt werden Braunkohlemeiler mit zusammen 8500 Megawatt (MW) abgeschaltet.
PP-7650	Die ganze Nacht über landeten auf dem internationalen Ben-Gurion-Flughafen Flugzeuge mit Trauergästen.
PP-17512	Die höchsten Renditen erwirtschafteten Agenturen in Hauptgeschäftslagen von Orten mit 10000 bis 500000 Einwohnern.
PP-19569	Um Familien mit Kindern nicht zusätzlich zu belasten, wird eine Neuordnung des Familienlastenausgleichs und des Steuerrechts erwartet.
PP-23135	Vorrangig erwirbt der Pool Strom zu höheren Preisen (Einspeisevergütung) aus regenerativen Quellen (Wind, Wasser, Sonne, Biogas) und Anlagen mit Kraft-Wärme-Kopplung.
PP-28600	“Wir richten uns nicht nach Müller und Schulz”, wehrt er Vergleiche mit der Konkurrenz ab.
PP-31611	Die Region ist offiziell zweisprachig, im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom.

V Context Models

V.1 Sentences with Genitive-Dative Ambiguity

V.1.1 Verb-Specific Binary Situation Contexts with Three Entities

VK-011	MAID_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SUCHEN_01
	FARMER.F_01	$\xrightarrow{is\_OWNER\_for}$	MAID_01
	BASKET_01	$\xrightarrow{is\_THEME\_for}$	ETW.SUCHEN_01
VK-100	ADMIRER.M_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SCHENKEN_01
	ACTRESS_01	$\xrightarrow{is\_OWNER\_for}$	ADMIRER.M_01
	BOUQUET_01	$\xrightarrow{is\_THEME\_for}$	ETW.SCHENKEN_01
VK-111	SON_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERDERBEN_01
	SMOKER.F_01	$\xrightarrow{is\_OWNER\_for}$	SON_01
	MOOD_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERDERBEN_01
VK-151	MOUNTAINEER.M_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SCHICKEN_01
	SPEAKER.F_01	$\xrightarrow{is\_OWNER\_for}$	MOUNTAINEER.M_01
	WARNING_01	$\xrightarrow{is\_THEME\_for}$	ETW.SCHICKEN_01
VK-226	NEIGHBOUR.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.NENNEN_01
	HEADMISTRESS_01	$\xrightarrow{is\_OWNER\_for}$	NEIGHBOUR.F_01
	ADDRESS_01	$\xrightarrow{is\_THEME\_for}$	ETW.NENNEN_01
VK-233	COUSIN.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SCHILDERN_01
	VISITOR.F_01	$\xrightarrow{is\_OWNER\_for}$	COUSIN_01
	INCIDENT_01	$\xrightarrow{is\_THEME\_for}$	ETW.SUCHEN_01
VK-247	NURSE.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERBIETEN_01
	AGED.F_01	$\xrightarrow{is\_OWNER\_for}$	NURSE.F_01
	EXCURSION_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERBIETEN_01
VK-263	PUBLISHER.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERKAUFEN_01
	AUTHOR.F_01	$\xrightarrow{is\_OWNER\_for}$	PUBLISHER.F_01
	ARTICLE_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERKAUFEN_01

VK-274	PHD.STUDENT.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.LIEFERN_01
	RESEARCHER.F_01	$\xrightarrow{is\_OWNER\_for}$	PHD.STUDENT.F_01
	PROOF_01	$\xrightarrow{is\_THEME\_for}$	ETW.LIEFERN_01
VK-306	MANAGER.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SENDEN_01
	ENTREPRENEUR.F_01	$\xrightarrow{is\_OWNER\_for}$	MANAGER.F_01
	SALES.REP.M_01	$\xrightarrow{is\_THEME\_for}$	ETW.SENDEN_01

V.1.2 Verb-Specific Ternary Situation Contexts with Three Participants

VK-011	MAID_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.SUCHEN_01
	FARMER.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.SUCHEN_01
	BASKET_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.SUCHEN_01
VK-100	ADMIRER.M_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.SCHENKEN_01
	ACTRESS_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.SCHENKEN_01
	BOUQUET_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.SCHENKEN_01
VK-111	SON_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.VERDERBEN_01
	SMOKER.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.VERDERBEN_01
	MOOD_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.VERDERBEN_01
VK-151	MOUNTAINEER.M_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.SCHICKEN_01
	SPEAKER.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.SCHICKEN_01
	WARNING_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.SCHICKEN_01
VK-226	NEIGHBOUR.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.NENNEN_01
	HEADMISTRESS_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.NENNEN_01
	ADDRESS_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.NENNEN_01
VK-233	COUSIN.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.SCHILDERN_01
	VISITOR.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.SCHILDERN_01
	INCIDENT_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.SCHILDERN_01
VK-247	NURSE.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.VERBIETEN_01
	AGED.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.VERBIETEN_01
	EXCURSION_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.VERBIETEN_01

VK-263	PUBLISHER.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.VERKAUFEN_01
	AUTHOR.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.VERKAUFEN_01
	ARTICLE_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.VERKAUFEN_01
VK-274	PHD.STUDENT.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.LIEFERN_01
	RESEARCHER.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.LIEFERN_01
	PROOF_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.LIEFERN_01
VK-306	MANAGER.F_01	$\xrightarrow{is\_AGENT\_for}$	JMD.ETW.SENDEN_01
	ENTREPRENEUR.F_01	$\xrightarrow{is\_RECIPIENT\_for}$	JMD.ETW.SENDEN_01
	SALES.REP.M_01	$\xrightarrow{is\_THEME\_for}$	JMD.ETW.SENDEN_01

V.1.3 Generalised Binary Situation Contexts with Three Entities

VK-011	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_OWNER\_for}$	HUMAN.F_01
	PHYSICAL.OBJECT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-100	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	HUMAN.F_01	$\xrightarrow{is\_OWNER\_for}$	HUMAN.M_01
	PHYSICAL.OBJECT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-111	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_OWNER\_for}$	HUMAN.F_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-151	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_OWNER\_for}$	HUMAN.F_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-226	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_OWNER\_for}$	HUMAN.F_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01

VK-233
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ BINARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_OWNER\_for}$ HUMAN.F_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ BINARY.SITUATION_01

VK-247
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ BINARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_OWNER\_for}$ HUMAN.F_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ BINARY.SITUATION_01

VK-263
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ BINARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_OWNER\_for}$ HUMAN.F_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ BINARY.SITUATION_01

VK-274
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ BINARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_OWNER\_for}$ HUMAN.F_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ BINARY.SITUATION_01

VK-306
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ BINARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_OWNER\_for}$ HUMAN.F_01
HUMAN.M_01 $\xrightarrow{is\_THEME\_for}$ BINARY.SITUATION_01

V.1.4 Generalised Ternary Situation Contexts with Three Participants

VK-011
HUMAN.F_01 $\xrightarrow{is\_AGENT\_for}$ TERNARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_RECIPIENT\_for}$ TERNARY.SITUATION_01
PHYSICAL.OBJECT_01 $\xrightarrow{is\_THEME\_for}$ TERNARY.SITUATION_01

VK-100
HUMAN.M_01 $\xrightarrow{is\_AGENT\_for}$ TERNARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_RECIPIENT\_for}$ TERNARY.SITUATION_01
PHYSICAL.OBJECT_01 $\xrightarrow{is\_THEME\_for}$ TERNARY.SITUATION_01

VK-111
HUMAN.M_01 $\xrightarrow{is\_AGENT\_for}$ TERNARY.SITUATION_01
HUMAN.F_02 $\xrightarrow{is\_RECIPIENT\_for}$ TERNARY.SITUATION_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ TERNARY.SITUATION_01

VK-151
HUMAN.M_01 $\xrightarrow{is\_AGENT\_for}$ TERNARY.SITUATION_01
HUMAN.F_01 $\xrightarrow{is\_RECIPIENT\_for}$ TERNARY.SITUATION_01
ABSTRACT_01 $\xrightarrow{is\_THEME\_for}$ TERNARY.SITUATION_01

VK-226	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01
VK-233	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01
VK-247	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01
VK-263	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01
VK-274	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01
VK-306	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	TERNARY.SITUATION_01
	HUMAN.F_02	$\xrightarrow{is\_RECIPIENT\_for}$	TERNARY.SITUATION_01
	HUMAN.M_01	$\xrightarrow{is\_THEME\_for}$	TERNARY.SITUATION_01

V.1.5 Generalised Binary Situation Contexts with Two Entities

VK-011	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	PHYSICAL.OBJECT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-100	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	PHYSICAL.OBJECT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-111	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01
VK-151	HUMAN.M_01	$\xrightarrow{is\_AGENT\_for}$	BINARY.SITUATION_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	BINARY.SITUATION_01

VK-226
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

VK-233
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

VK-247
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

VK-263
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

VK-274
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

VK-306
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{BINARY.SITUATION\_01} \\ \text{HUMAN.M\_01} & \xrightarrow{is\_THEME\_for} & \text{BINARY.SITUATION\_01} \end{array}$$

V.1.6 Generalised Situation Contexts with Two Entities, all of which Participants

VK-011
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{PHYSICAL.OBJECT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-100
$$\begin{array}{rcl} \text{HUMAN.M\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{PHYSICAL.OBJECT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-111
$$\begin{array}{rcl} \text{HUMAN.M\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-151
$$\begin{array}{rcl} \text{HUMAN.M\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-226
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-233
$$\begin{array}{rcl} \text{HUMAN.F\_01} & \xrightarrow{is\_AGENT\_for} & \text{SITUATION.CONCEPT\_01} \\ \text{ABSTRACT\_01} & \xrightarrow{is\_THEME\_for} & \text{SITUATION.CONCEPT\_01} \end{array}$$

VK-247	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	SITUATION.CONCEPT_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	SITUATION.CONCEPT_01
VK-263	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	SITUATION.CONCEPT_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	SITUATION.CONCEPT_01
VK-274	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	SITUATION.CONCEPT_01
	ABSTRACT_01	$\xrightarrow{is\_THEME\_for}$	SITUATION.CONCEPT_01
VK-306	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	SITUATION.CONCEPT_01
	HUMAN.M_01	$\xrightarrow{is\_THEME\_for}$	SITUATION.CONCEPT_01

V.2 Sentences with Subject-Object Ambiguity

V.2.1 Verb-Specific SUBJ-OBJA Contexts

SO-360	PROTESTER.M_01	$\xrightarrow{is\_AGENT\_for}$	ETW.TRAGEN_01
	BIER_01	$\xrightarrow{is\_THEME\_for}$	ETW.TRAGEN_01
	MARTYR.M_01	$\xrightarrow{is\_OWNER\_for}$	BIER_01
SO-706	MARKET_01	$\xrightarrow{is\_AGENT\_for}$	ETW.FORDERN_01
	PATIENCE_01	$\xrightarrow{is\_THEME\_for}$	ETW.FORDERN_01
SO-841	HUMAN.M.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERSORGEN_01
	MAN_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERSORGEN_01
SO-1090	NIPPON_01	$\xrightarrow{is\_AGENT\_for}$	JMD.BITTEN_01
	EUROPE_01	$\xrightarrow{is\_THEME\_for}$	JMD.BITTEN_01
SO-4493	DAMAGE_01	$\xrightarrow{is\_AGENT\_for}$	ETW.ANRICHTEN_01
	STORM_01	$\xrightarrow{is\_THEME\_for}$	ETW.ANRICHTEN_01
SO-6179	PEOPLES.PARTY_01	$\xrightarrow{is\_AGENT\_for}$	ETW.ZURUECKWEISEN_01
	DEMAND_01	$\xrightarrow{is\_THEME\_for}$	ETW.ZURUECKWEISEN_01
SO-9681	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SCHICKEN_01
	ADVERTISER.M_01	$\xrightarrow{is\_THEME\_for}$	ETW.SCHICKEN_01

SO-9792	HUMAN.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERTRETEN_01
	COMPANY_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERTRETEN_01
SO-10744	HUMAN.M.F_01	$\xrightarrow{is\_AGENT\_for}$	ETW.DRAENGEN_01
	WAR.PARTY_01	$\xrightarrow{is\_THEME\_for}$	ETW.DRAENGEN_01
SO-40722	COMMISSION_01	$\xrightarrow{is\_AGENT\_for}$	JMD.AUFFORDERN_01
	GOVERNMENT_01	$\xrightarrow{is\_THEME\_for}$	JMD.AUFFORDERN_01

V.2.2 Verb-Specific OBJA-SUBJ Contexts

SO-360	BIER_01	$\xrightarrow{is\_AGENT\_for}$	ETW.TRAGEN_01
	MARTYR.M_01	$\xrightarrow{is\_OWNER\_for}$	BIER_01
	PROTESTER.M_01	$\xrightarrow{is\_THEME\_for}$	ETW.TRAGEN_01
SO-706	PATIENCE_01	$\xrightarrow{is\_AGENT\_for}$	ETW.FORDERN_01
	MARKET_01	$\xrightarrow{is\_THEME\_for}$	ETW.FORDERN_01
SO-841	MAN_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERSORGEN_01
	HUMAN.M.F_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERSORGEN_01
SO-1090	EUROPE_01	$\xrightarrow{is\_AGENT\_for}$	JMD.BITTEN_01
	NIPPON_01	$\xrightarrow{is\_THEME\_for}$	JMD.BITTEN_01
SO-4493	STORM_01	$\xrightarrow{is\_AGENT\_for}$	ETW.ANRICHTEN_01
	DAMAGE_01	$\xrightarrow{is\_THEME\_for}$	ETW.ANRICHTEN_01
SO-6179	DEMAND_01	$\xrightarrow{is\_AGENT\_for}$	ETW.ZURUECKWEISEN_01
	PEOPLES.PARTY_01	$\xrightarrow{is\_THEME\_for}$	ETW.ZURUECKWEISEN_01
SO-9681	ADVERTISER.M_01	$\xrightarrow{is\_AGENT\_for}$	ETW.SCHICKEN_01
	HUMAN.F_01	$\xrightarrow{is\_THEME\_for}$	ETW.SCHICKEN_01
SO-9792	COMPANY_01	$\xrightarrow{is\_AGENT\_for}$	ETW.VERTRETEN_01
	HUMAN.F_01	$\xrightarrow{is\_THEME\_for}$	ETW.VERTRETEN_01

SO-10744	WAR.PARTY_01	$\xrightarrow{is\_AGENT\_for}$	ETW.DRAENGEN_01
	HUMAN.M.F_01	$\xrightarrow{is\_THEME\_for}$	ETW.DRAENGEN_01
SO-40722	GOVERNMENT_01	$\xrightarrow{is\_AGENT\_for}$	JMD.AUFFORDERN_01
	COMMISSION_01	$\xrightarrow{is\_THEME\_for}$	JMD.AUFFORDERN_01

V.3 Sentences with PP-Attachment Ambiguity

V.3.1 Generalised COMITATIVE Contexts

PP-3025	PHYSICAL.OBJECT_01	$\xrightarrow{is\_COMITATIVE\_for}$	FORCE_01
PP-3277	GROUP_01	$\xrightarrow{is\_COMITATIVE\_for}$	ABSTRACT_01
PP-3839	ABSTRACT_01	$\xrightarrow{is\_COMITATIVE\_for}$	ABSTRACT_02
PP-7177	ABSTRACT_01	$\xrightarrow{is\_COMITATIVE\_for}$	PHYSICAL.OBJECT_01
PP-7650	HUMAN.M.F_01	$\xrightarrow{is\_COMITATIVE\_for}$	PHYSICAL.OBJECT_01
PP-17512	HUMAN.M.F_01	$\xrightarrow{is\_COMITATIVE\_for}$	ABSTRACT_01
PP-19569	HUMAN.M.F_01	$\xrightarrow{is\_COMITATIVE\_for}$	GROUP_01
PP-23135	ABSTRACT_01	$\xrightarrow{is\_COMITATIVE\_for}$	PHYSICAL.OBJECT_01
PP-28600	ABSTRACT_02	$\xrightarrow{is\_COMITATIVE\_for}$	ABSTRACT_01
PP-31611	ABSTRACT_01	$\xrightarrow{is\_COMITATIVE\_for}$	LANGUAGE_01

V.3.2 Generalised INSTRUMENT Contexts

PP-3025	PHYSICAL.OBJECT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-3277	GROUP_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-3839	ABSTRACT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01

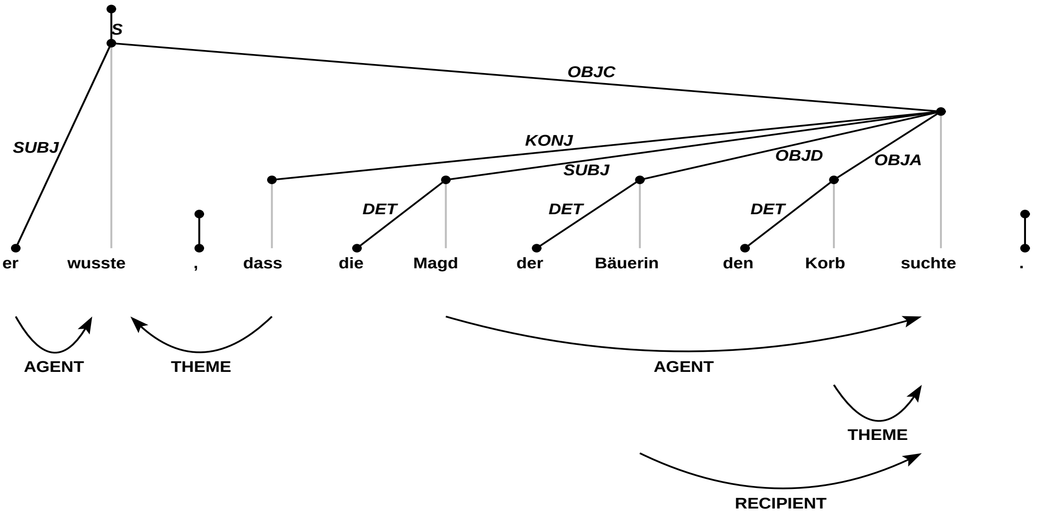
PP-7177	ABSTRACT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-7650	HUMAN.M.F_01	$\xrightarrow{is\_INSTRUMENT\_for}$	UNARY.SITUATION_01
PP-17512	HUMAN.M.F_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-19569	HUMAN.M.F_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-23135	ABSTRACT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-28600	ABSTRACT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01
PP-31611	ABSTRACT_01	$\xrightarrow{is\_INSTRUMENT\_for}$	BINARY.SITUATION_01

VI Parse Trees

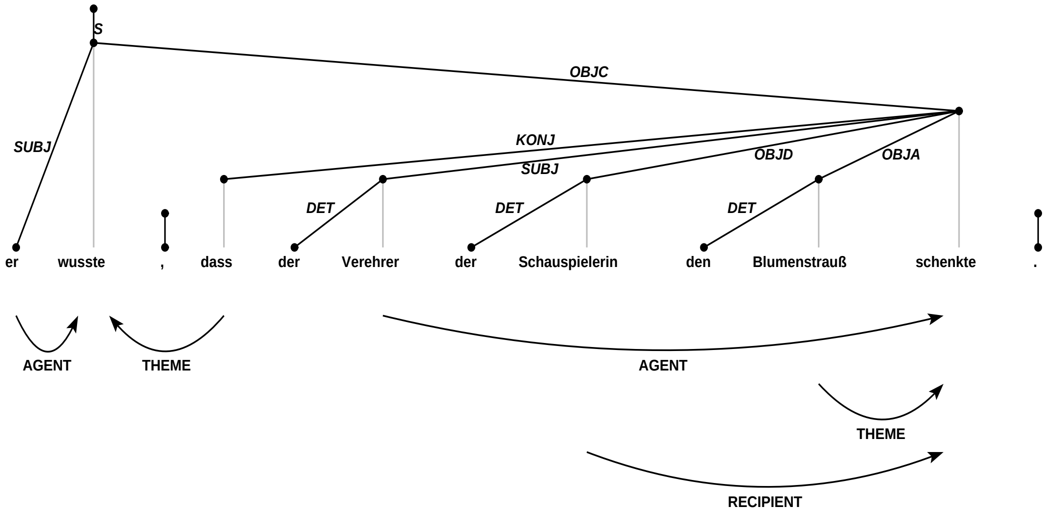
VI.1 Sentences with GMOD-OBJD Ambiguity

VI.1.1 Integration of an Empty Visual Context

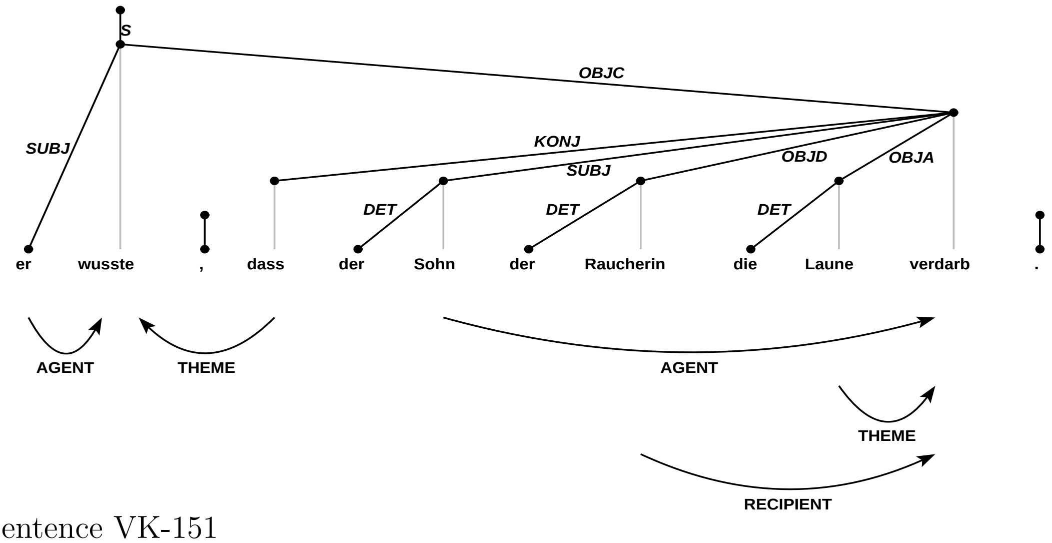
Sentence VK-011



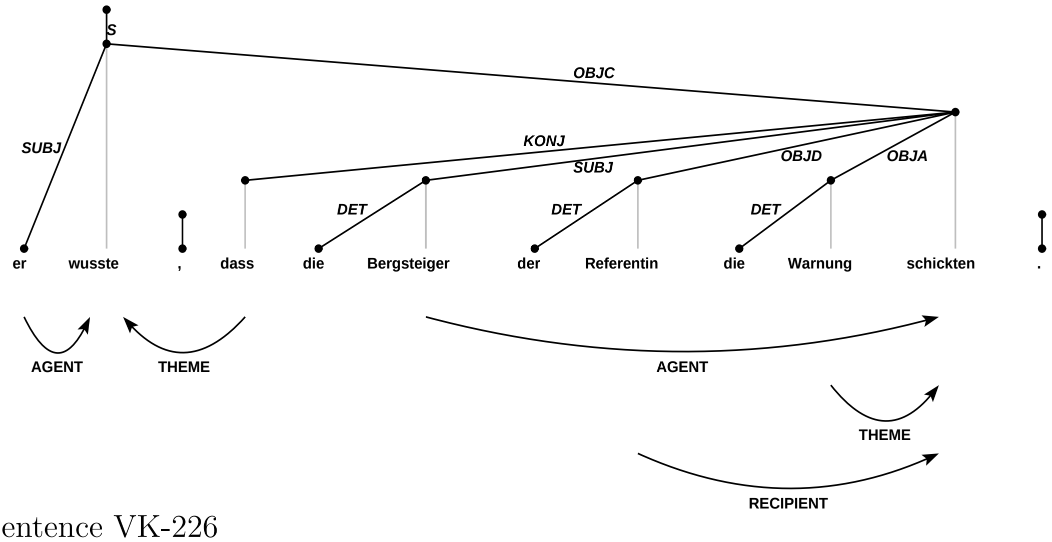
Sentence VK-100



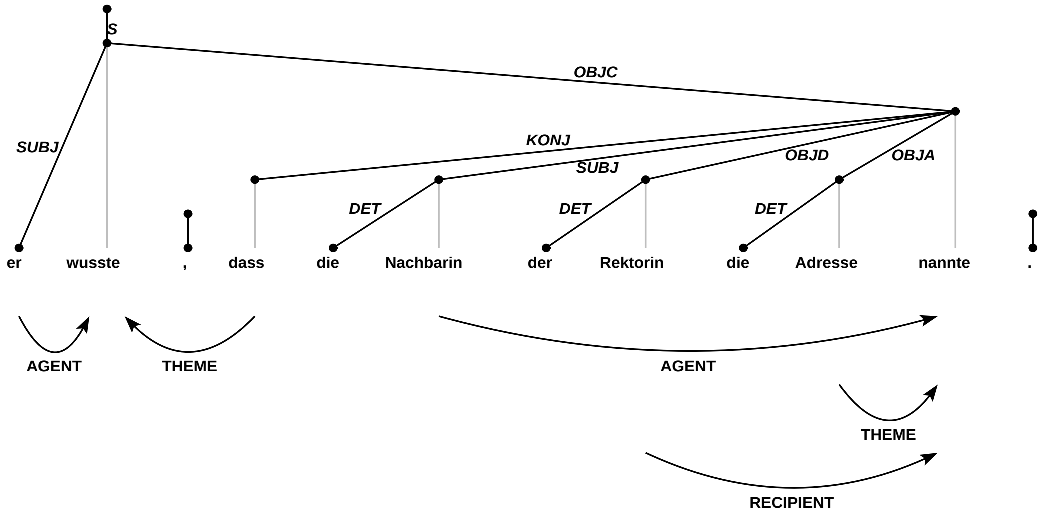
Sentence VK-111



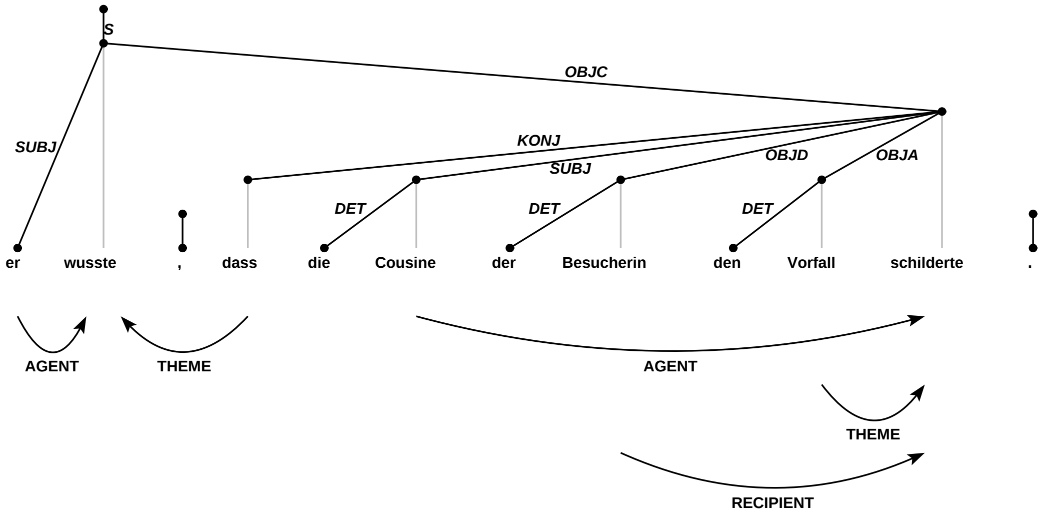
Sentence VK-151



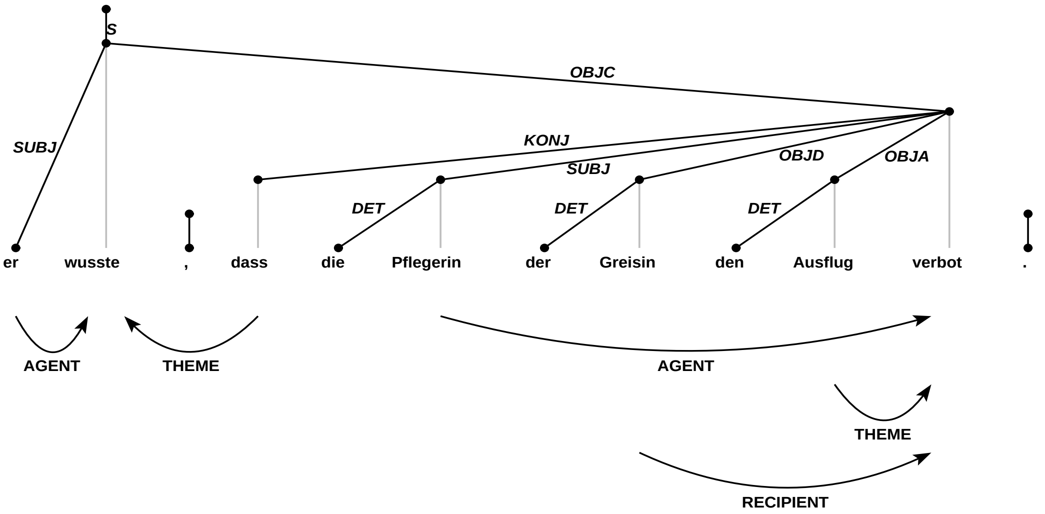
Sentence VK-226



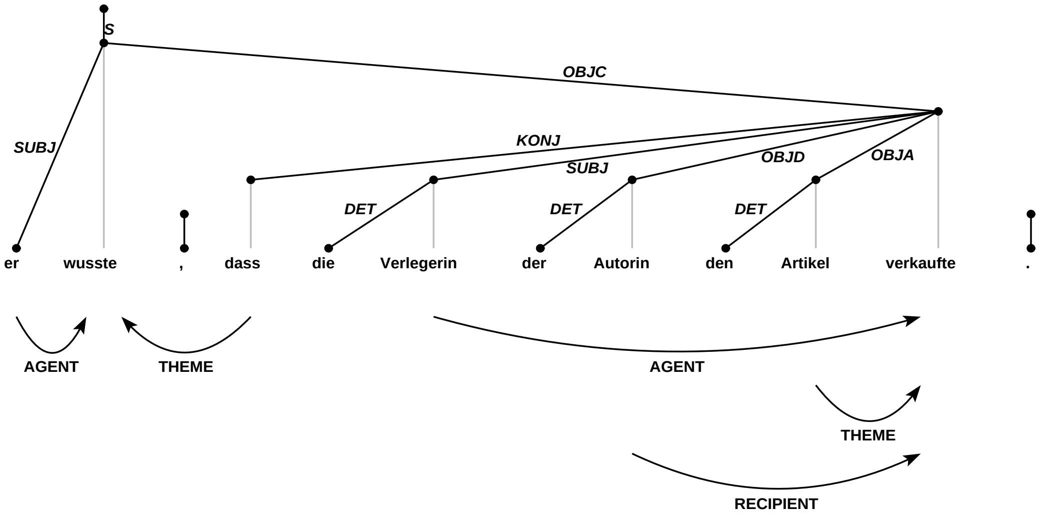
Sentence VK-233



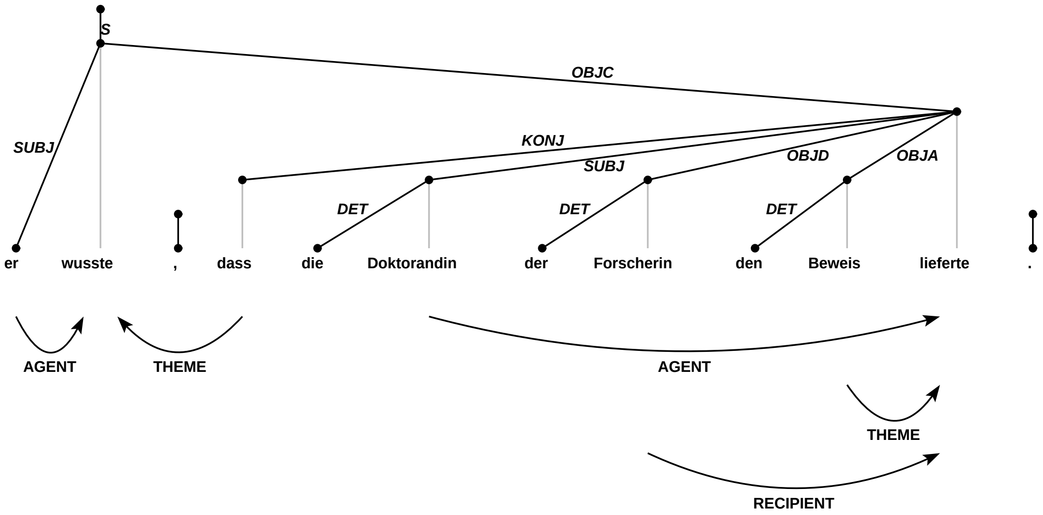
Sentence VK-247



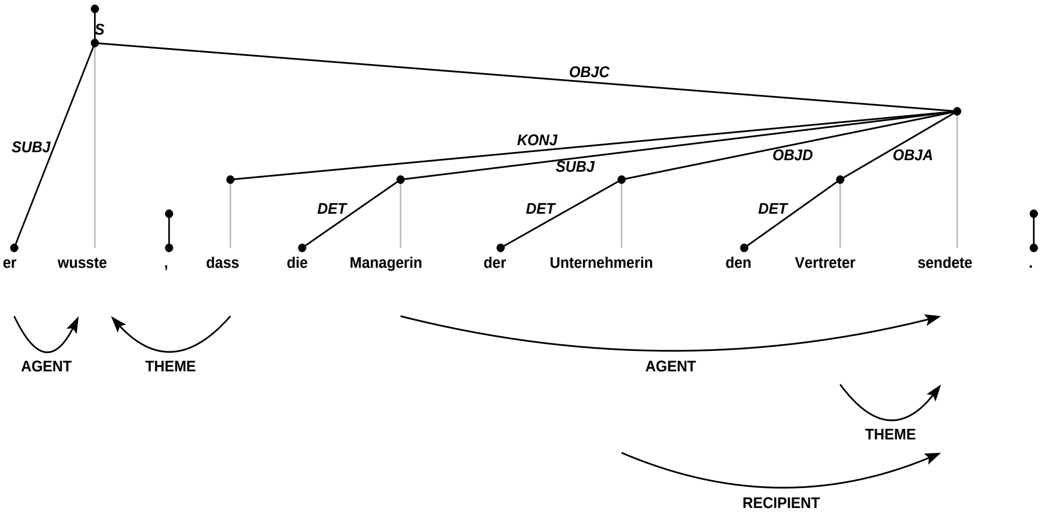
Sentence VK-263



Sentence VK-274

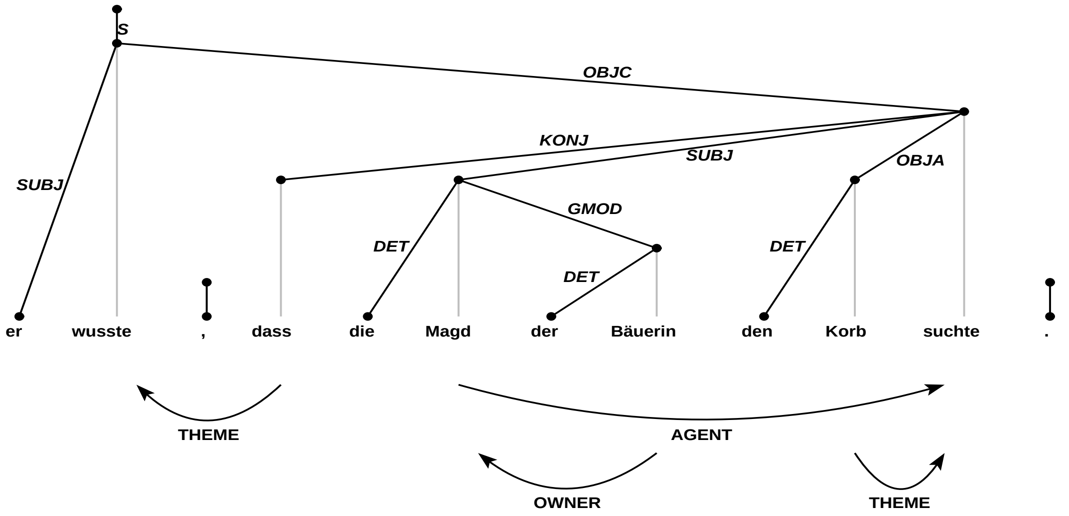


Sentence VK-306

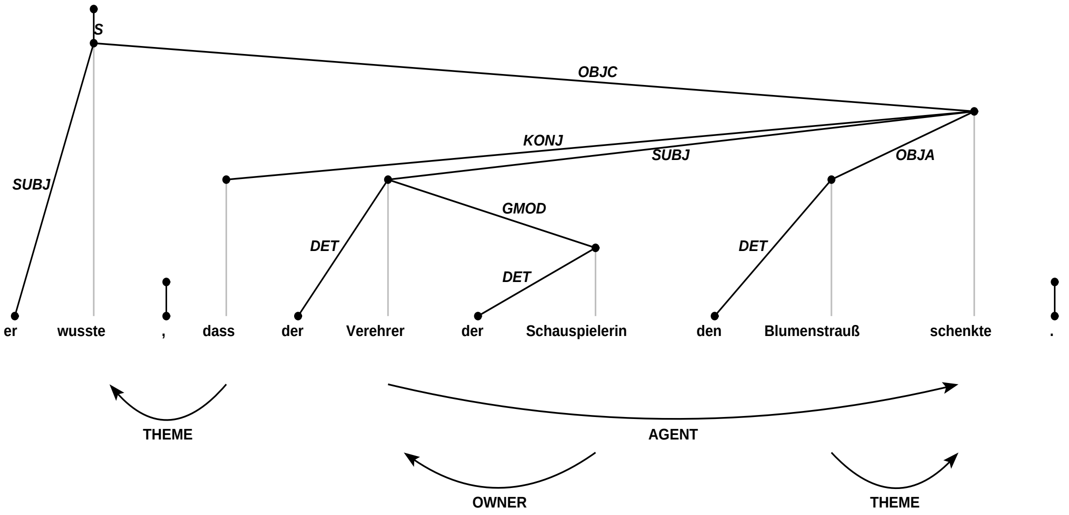


VI.1.2 Hard Integration of Verb-Specific Binary Contexts with Three Entities, Two of which Participants

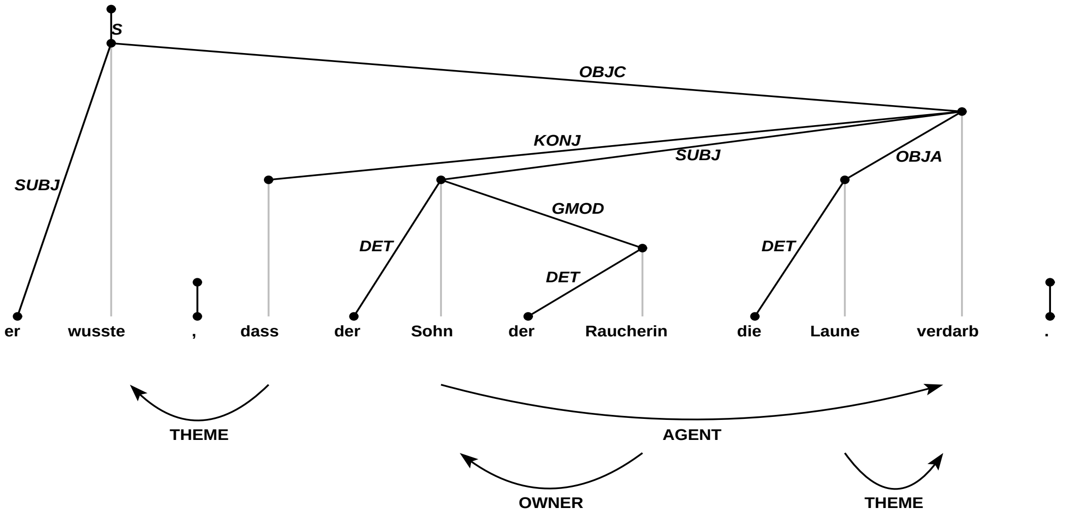
Sentence VK-011



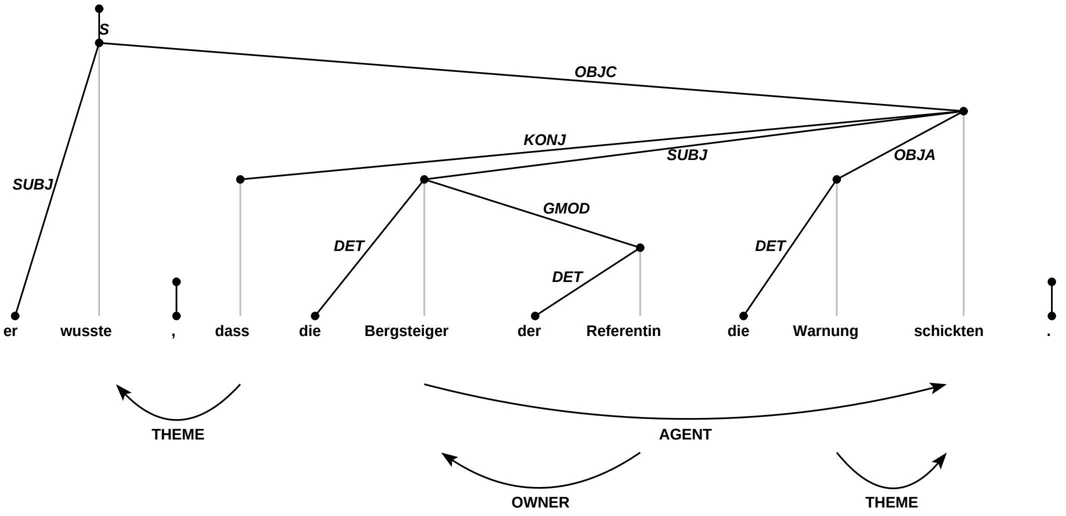
Sentence VK-100



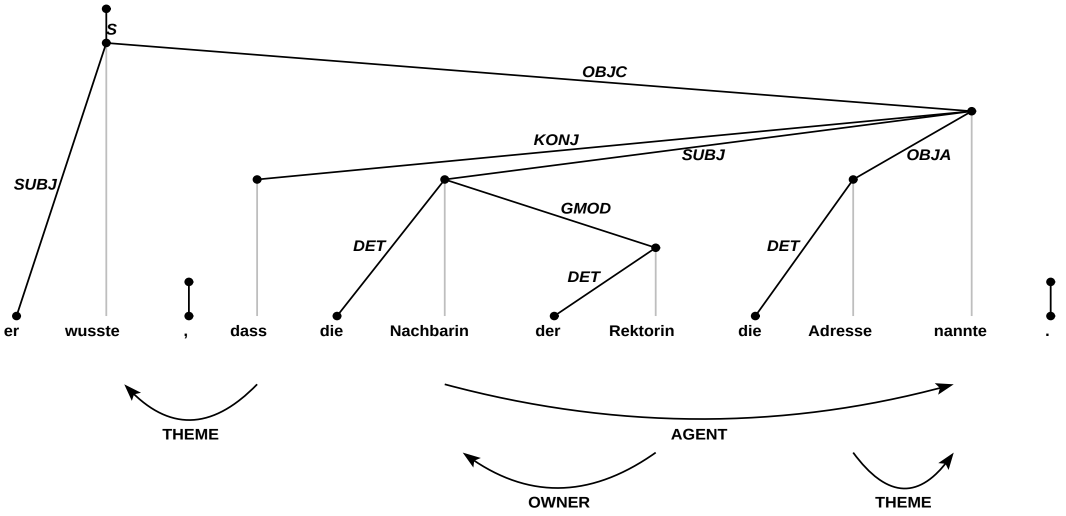
Sentence VK-111



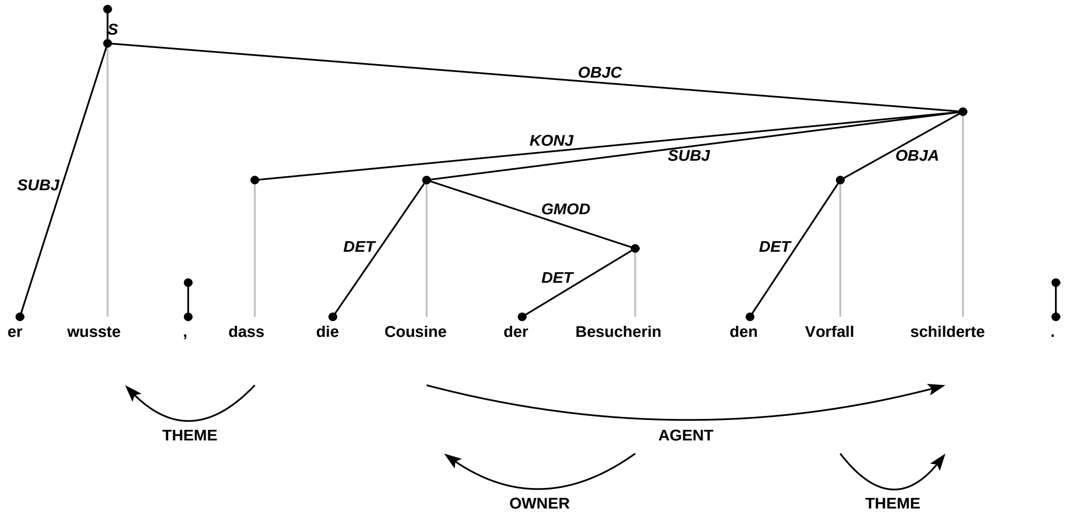
Sentence VK-151


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Bergsteiger der Referentin die Warnung schickten .
THEME
AGENT
OWNER
THEME

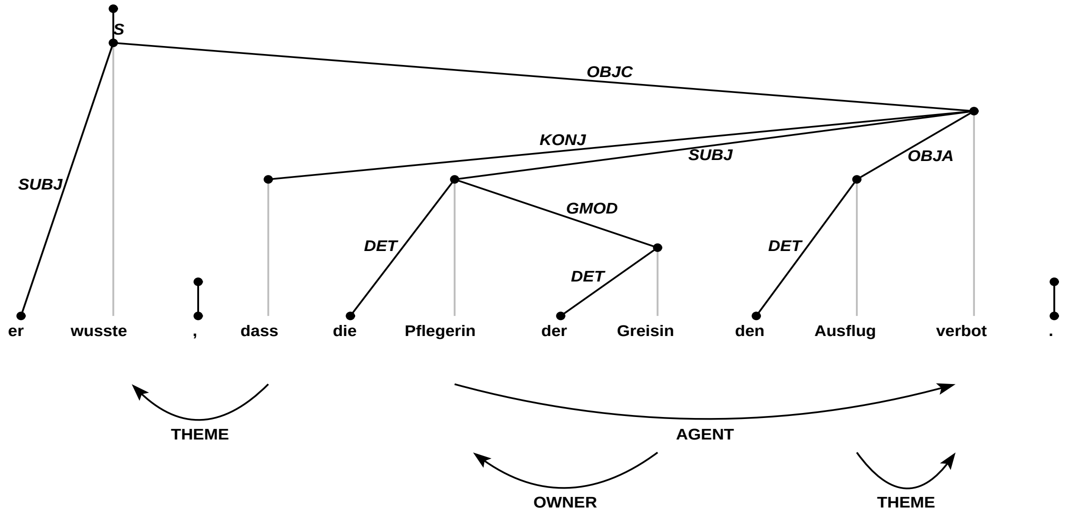
Sentence VK-226


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Nachbarin der Rektorin die Adresse nannte .
THEME
AGENT
OWNER
THEME

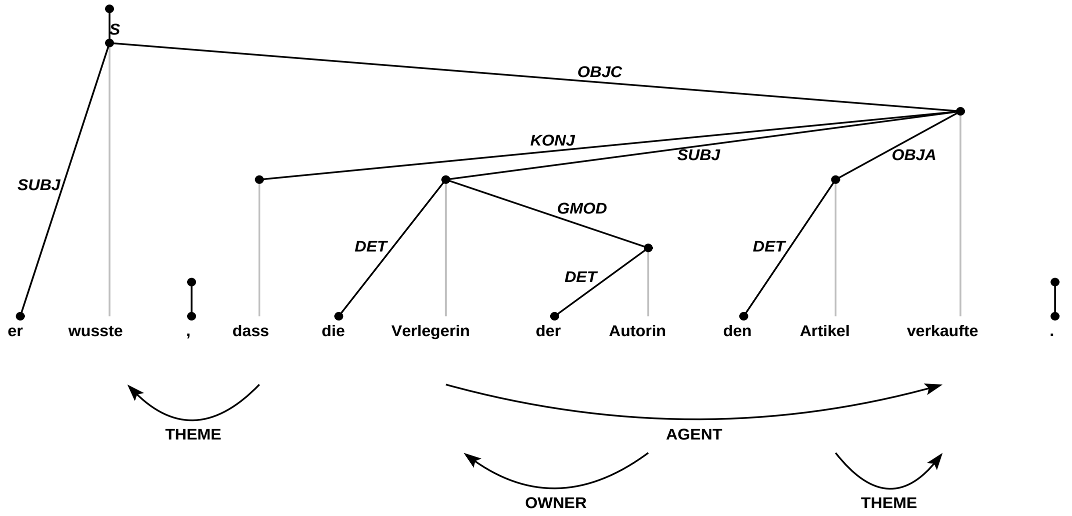
Sentence VK-233



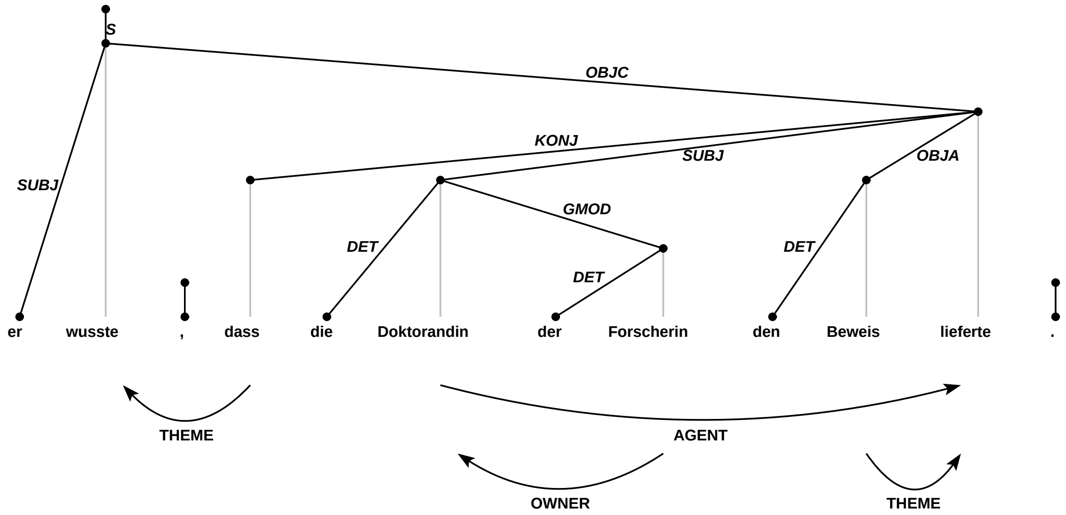
Sentence VK-247



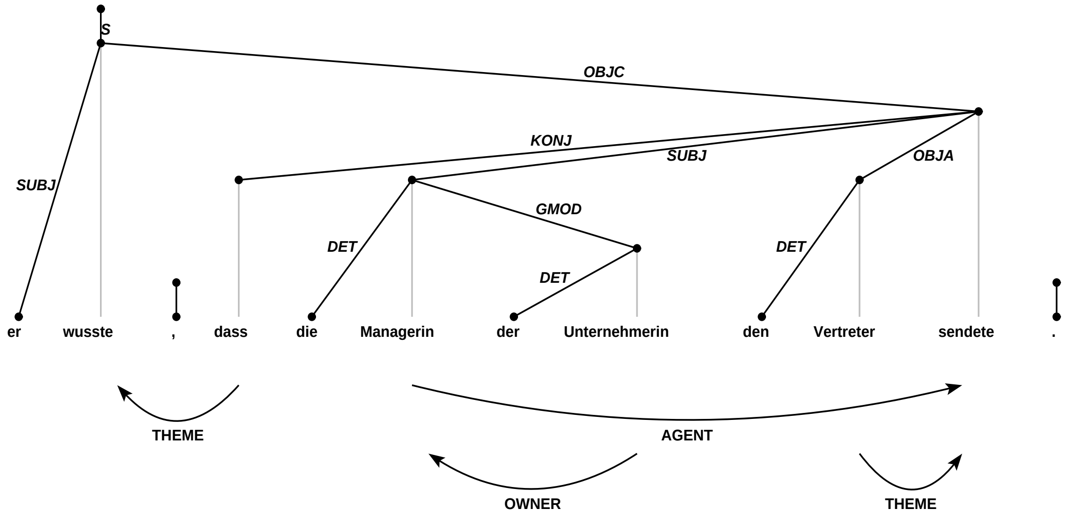
Sentence VK-263



Sentence VK-274

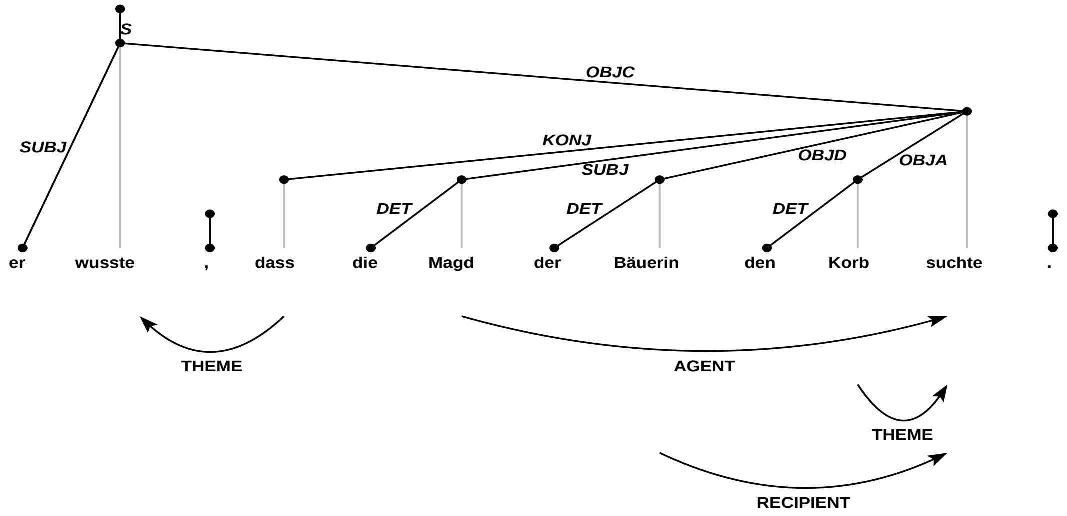


Sentence VK-306

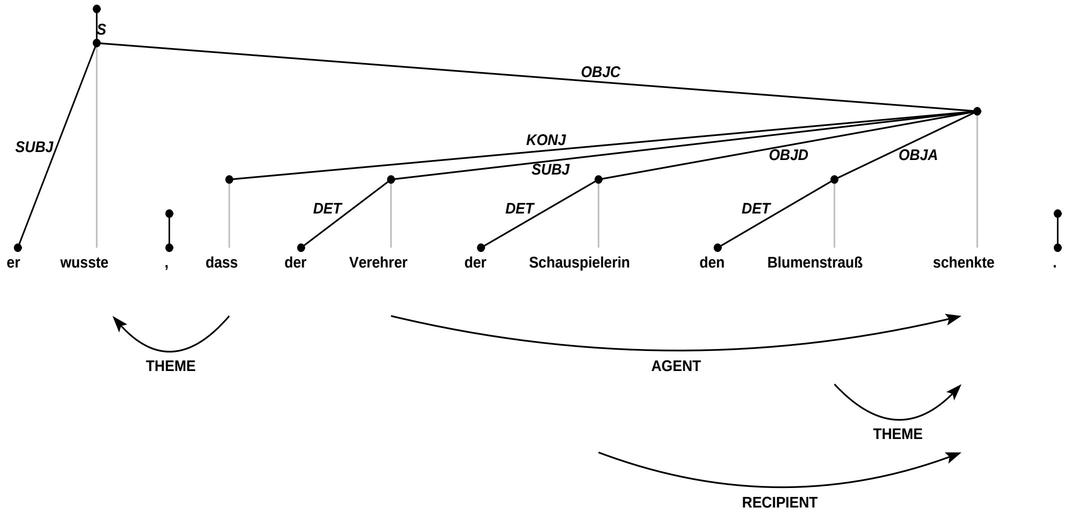


VI.1.3 Hard Integration of Verb-Specific Ternary Contexts with Three Entities, all of which Participants

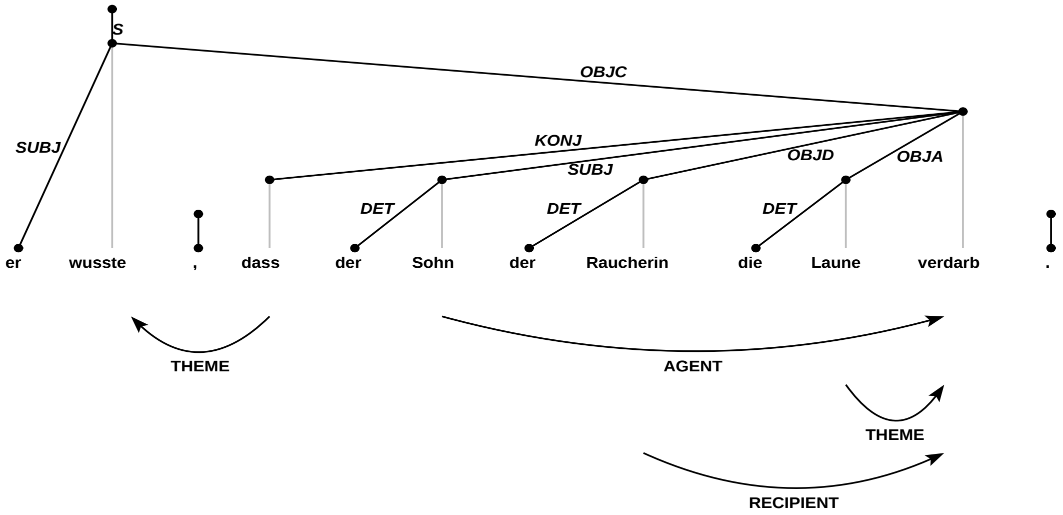
Sentence VK-011



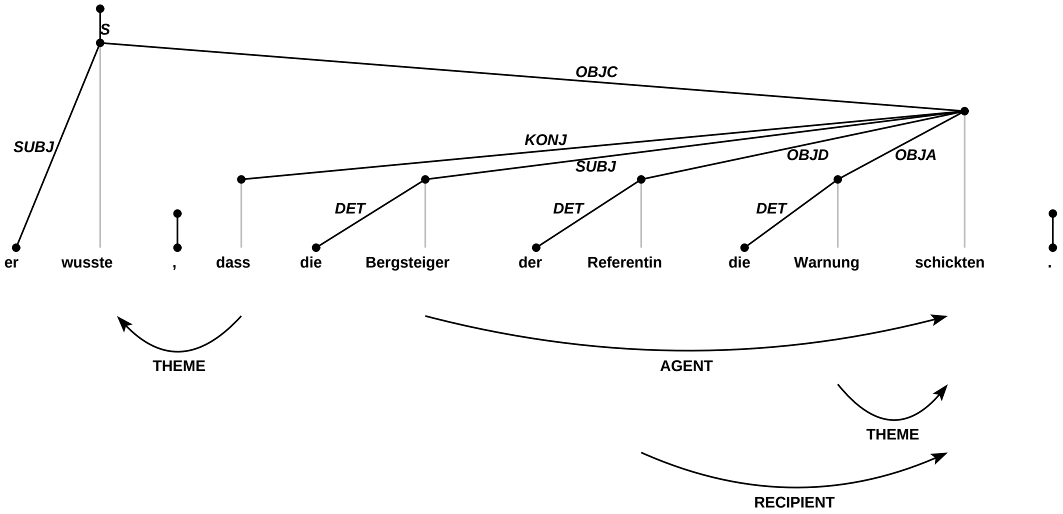
Sentence VK-100



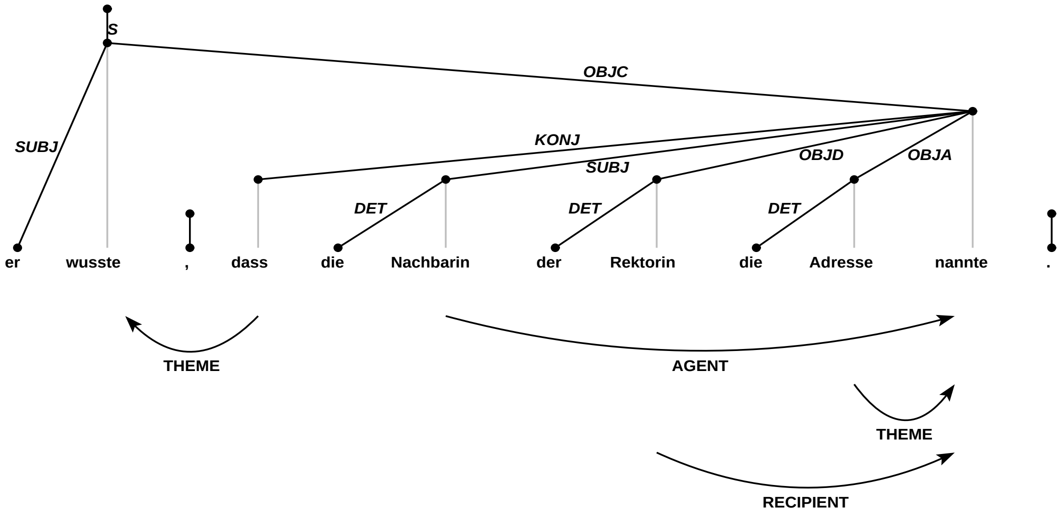
Sentence VK-111



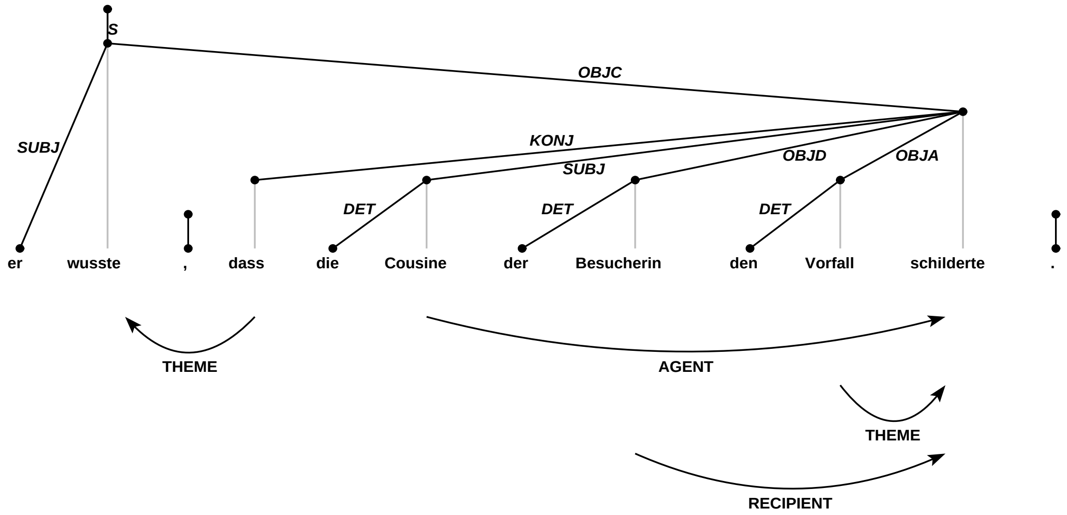
Sentence VK-151



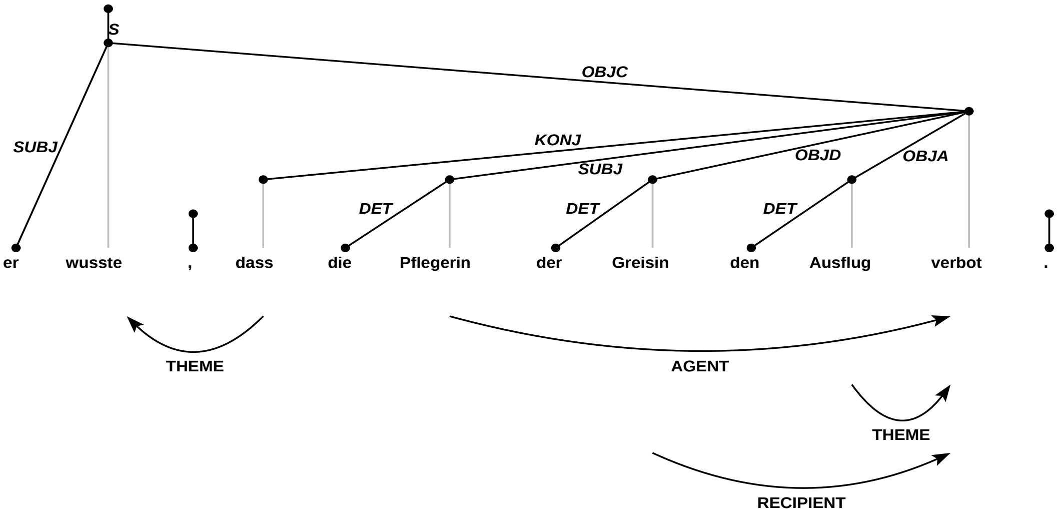
Sentence VK-226



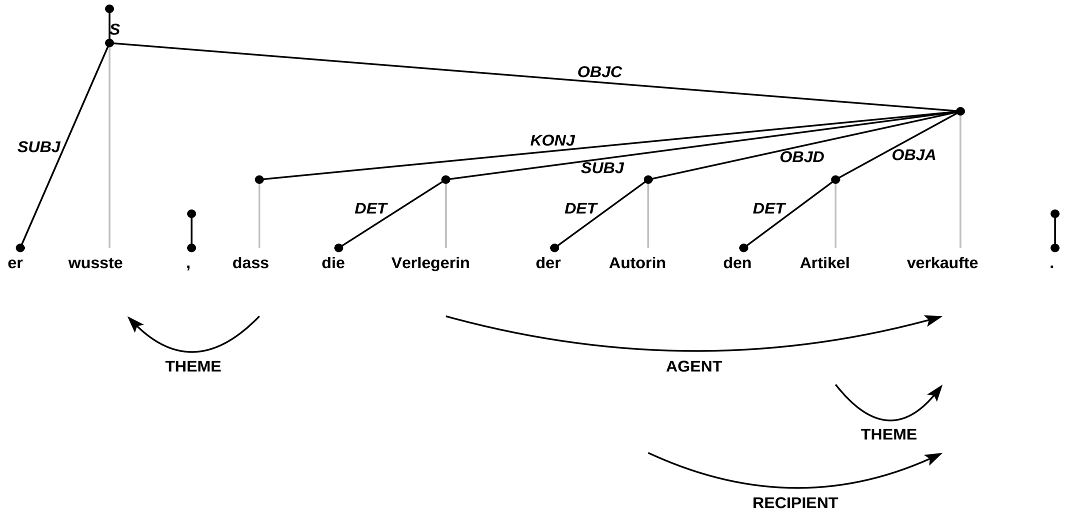
Sentence VK-233



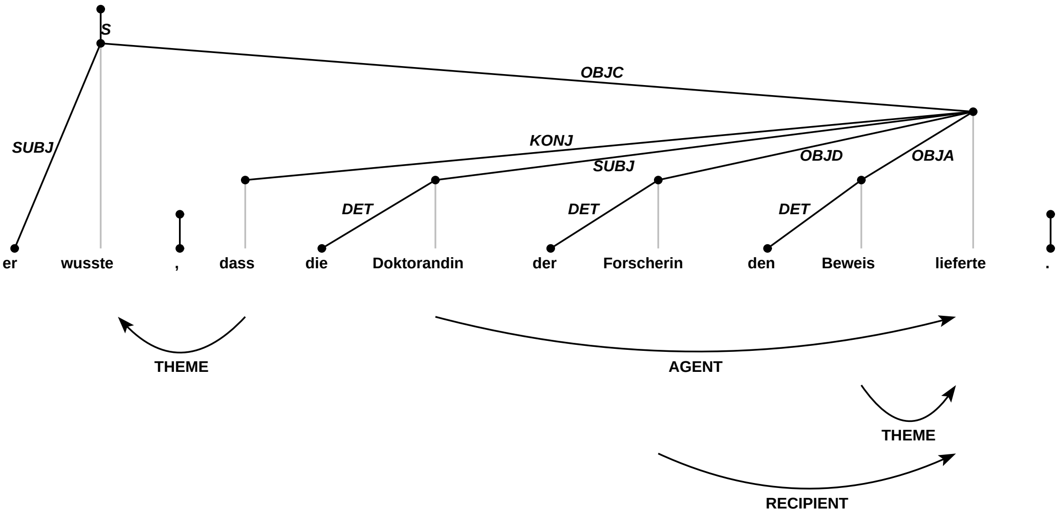
Sentence VK-247



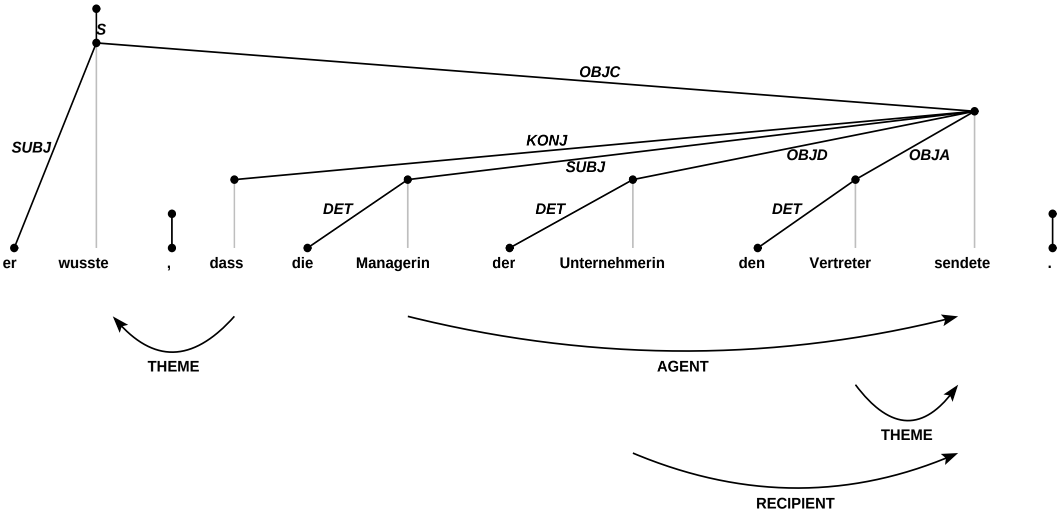
Sentence VK-263



Sentence VK-274

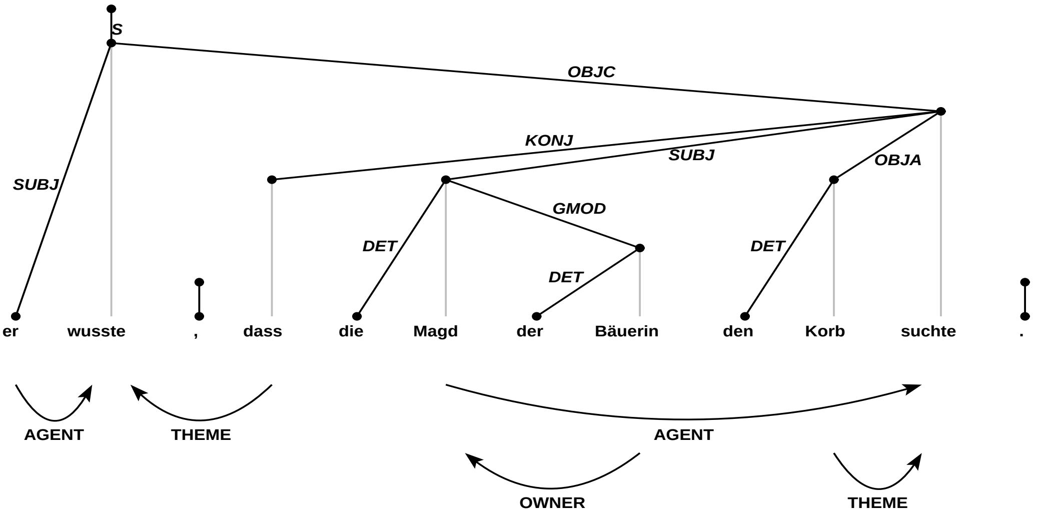


Sentence VK-306

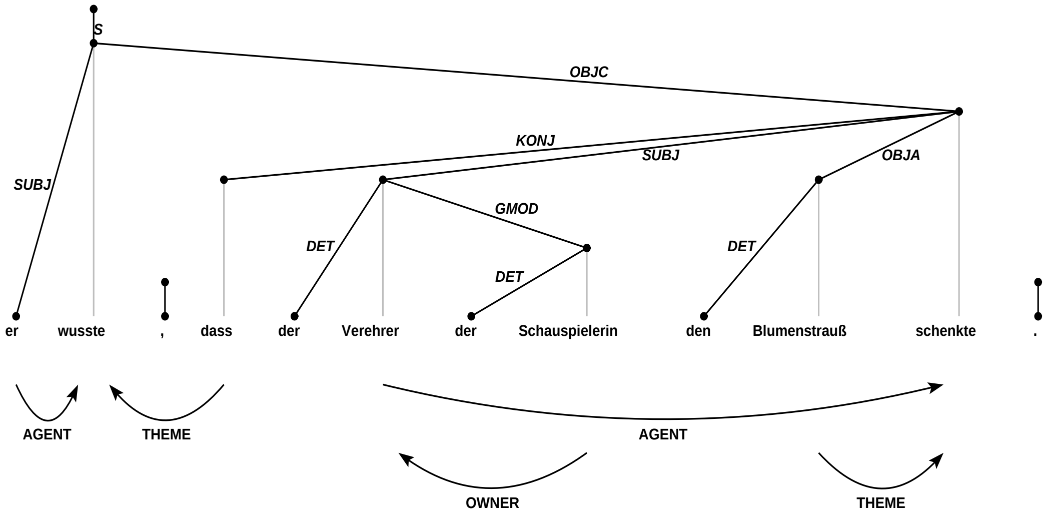


VI.1.4 Soft Integration of Verb-Specific Binary Context with Three Entities, Two of which Participants

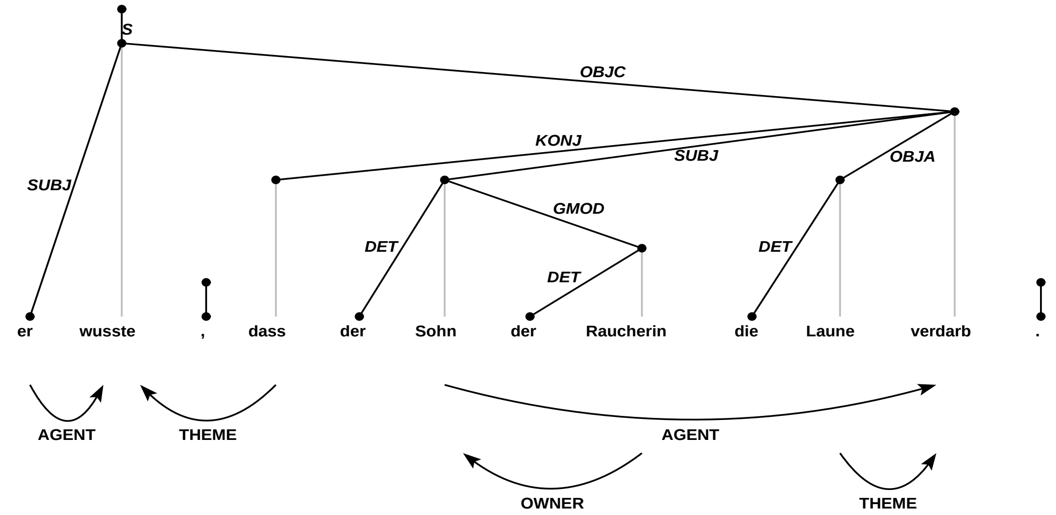
Sentence VK-011



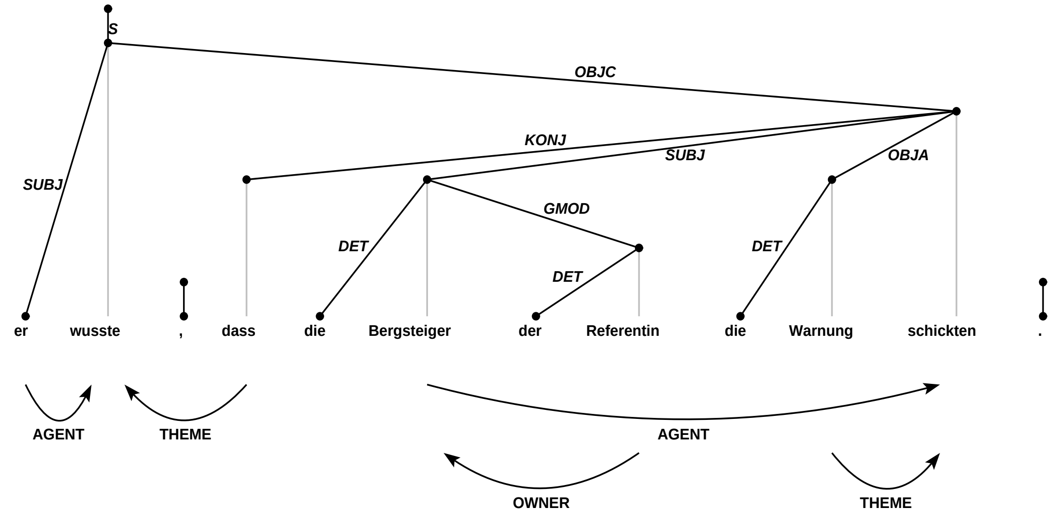
Sentence VK-100



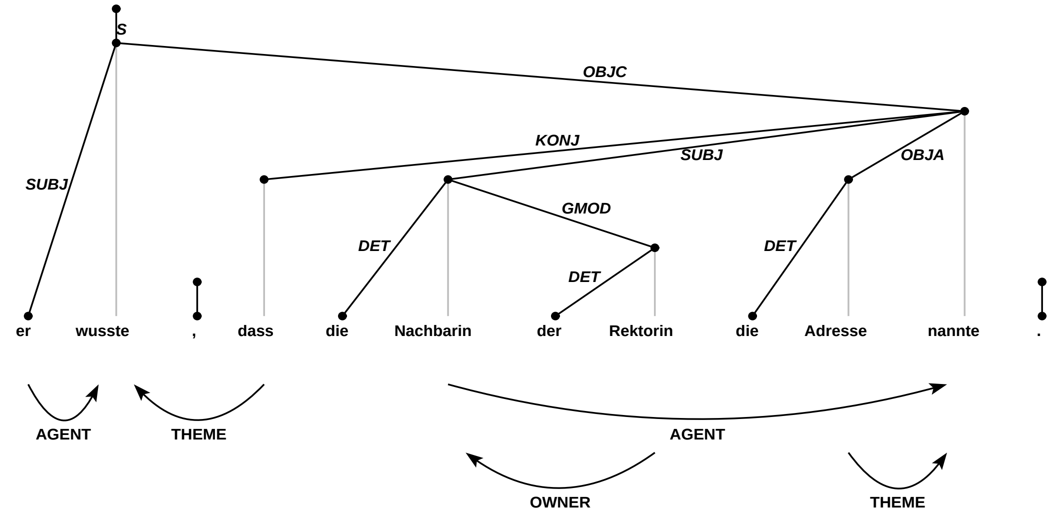
Sentence VK-111



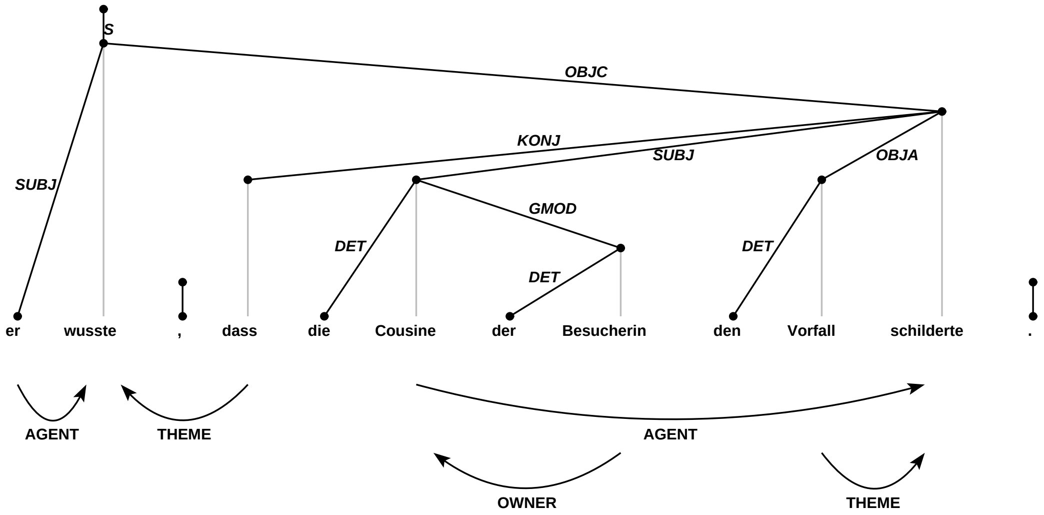
Sentence VK-151



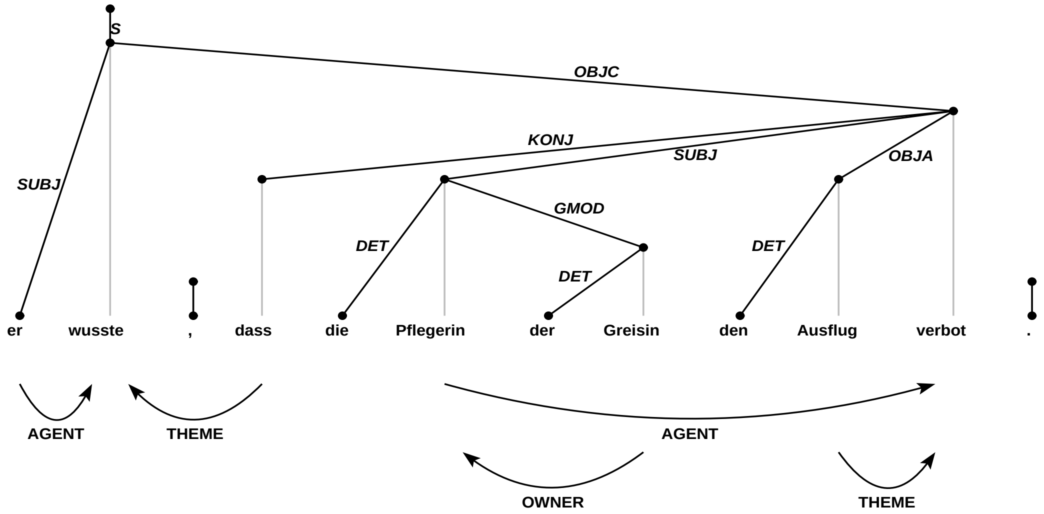
Sentence VK-226



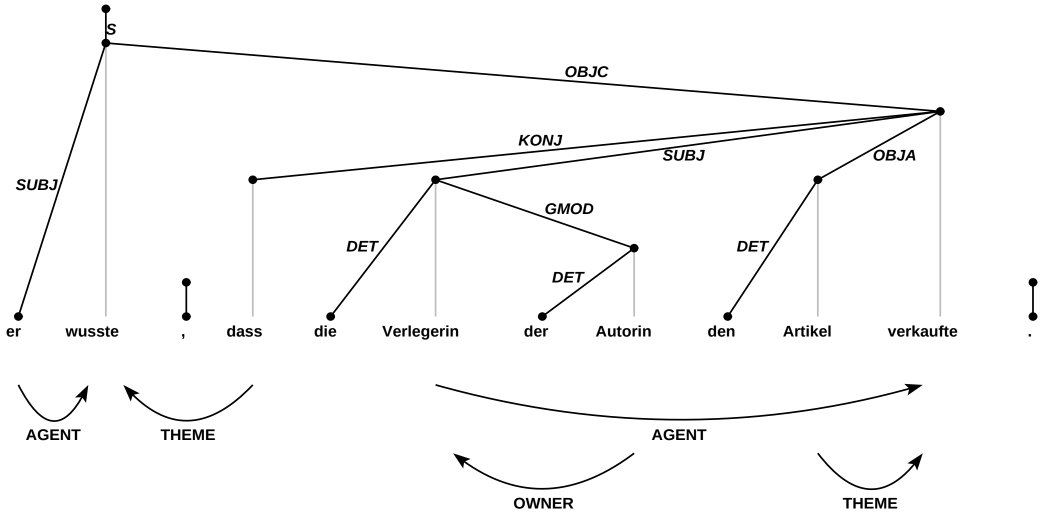
Sentence VK-233


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Cousine der Besucherin den Vorfall schilderte .
AGENT
THEME
AGENT
OWNER
THEME

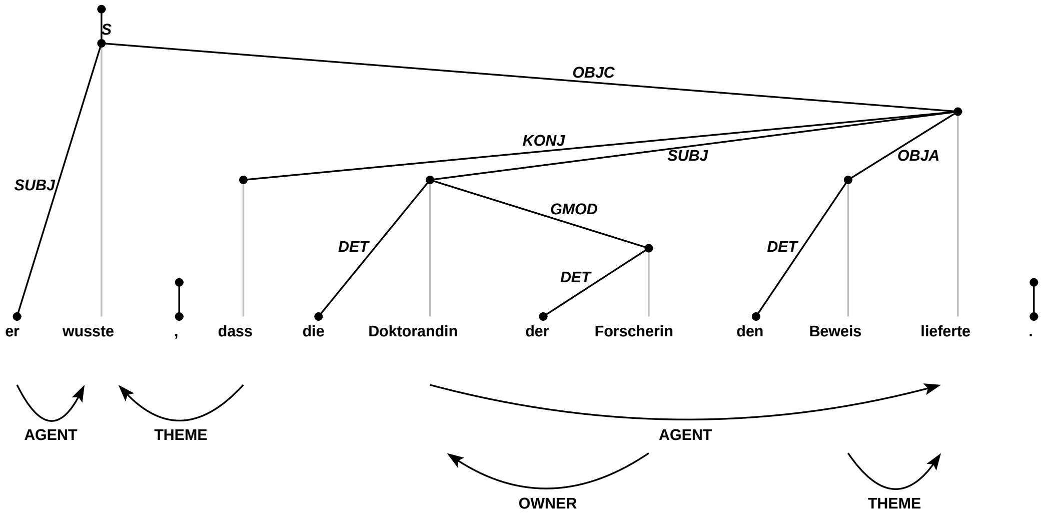
Sentence VK-247


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Pflegerin der Greisin den Ausflug verbot .
AGENT
THEME
AGENT
OWNER
THEME

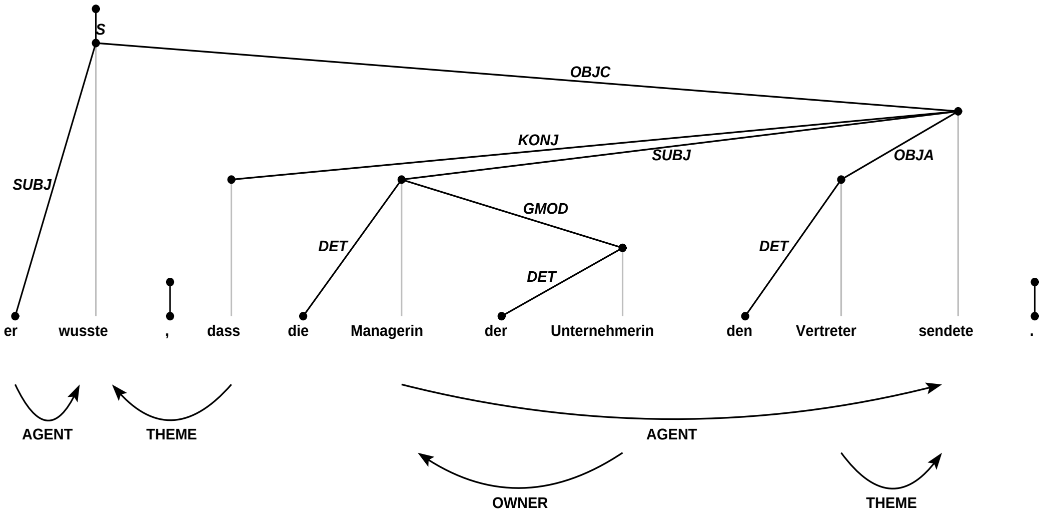
Sentence VK-263


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Verlegerin der Autorin den Artikel verkaufte .
AGENT
THEME
AGENT
OWNER
THEME

Sentence VK-274

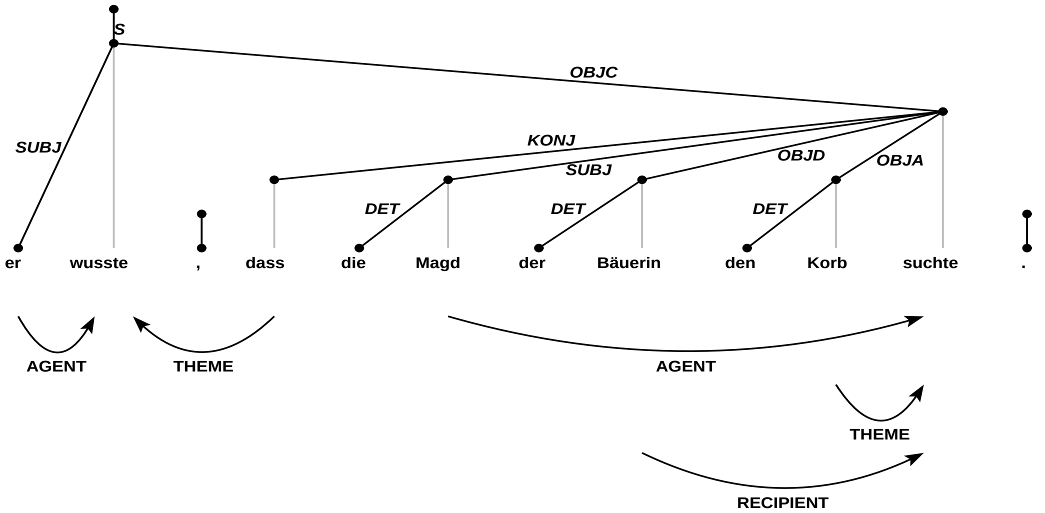


Sentence VK-306

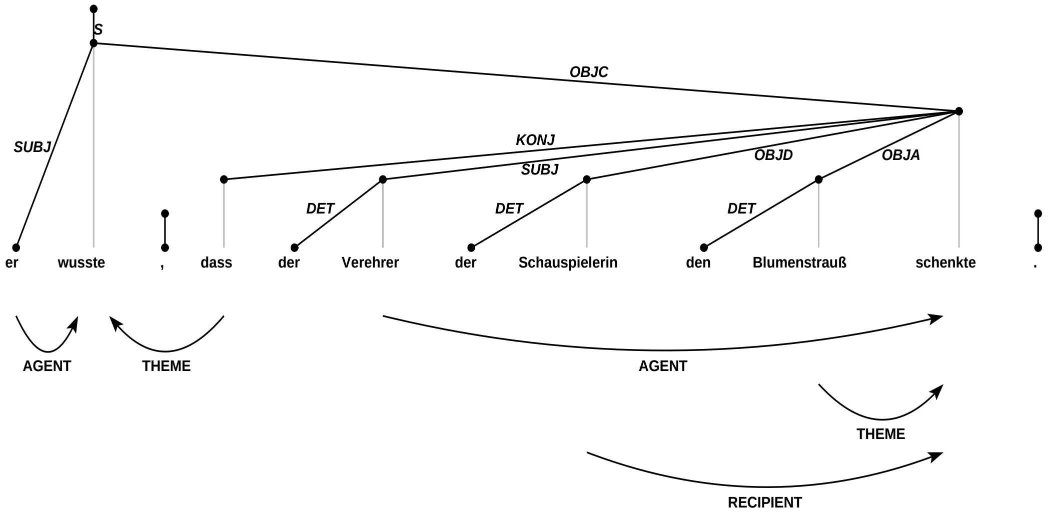


VI.1.5 Soft Integration of Verb-Specific Ternary Contexts with Three Entities, All of which Participants

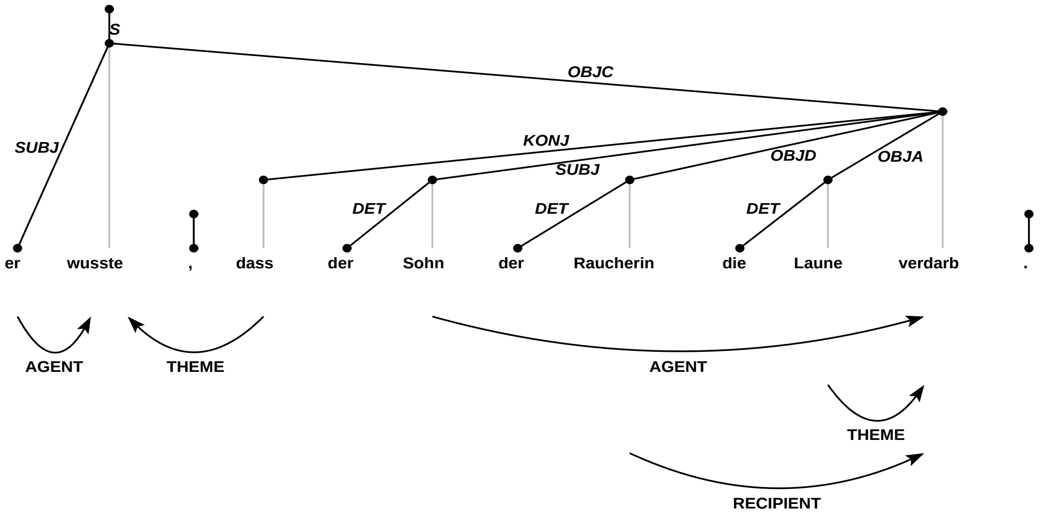
Sentence VK-011



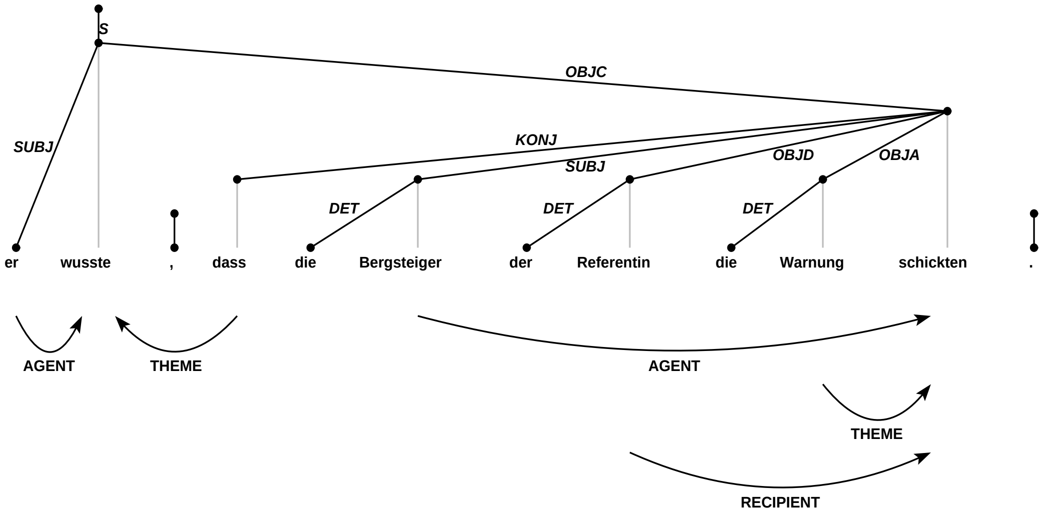
Sentence VK-100



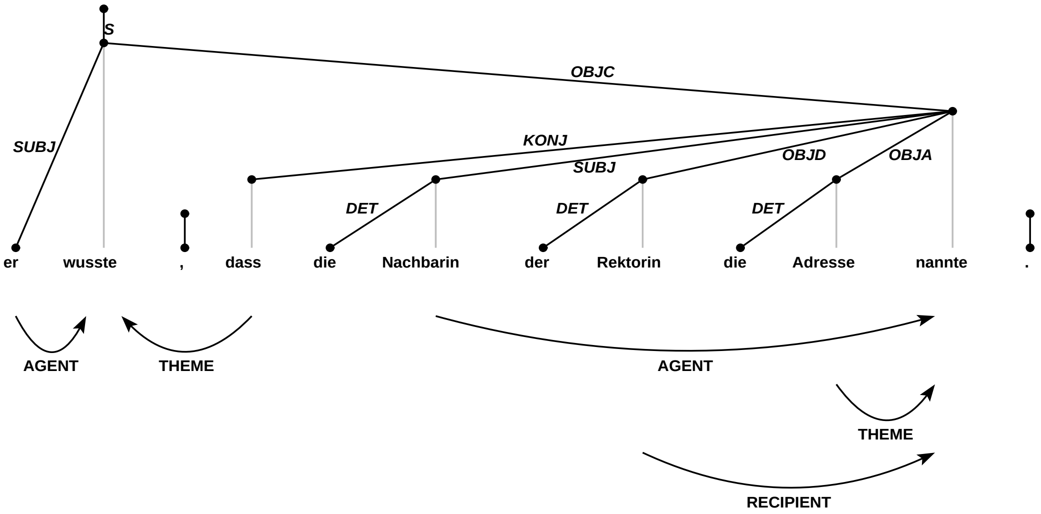
Sentence VK-111



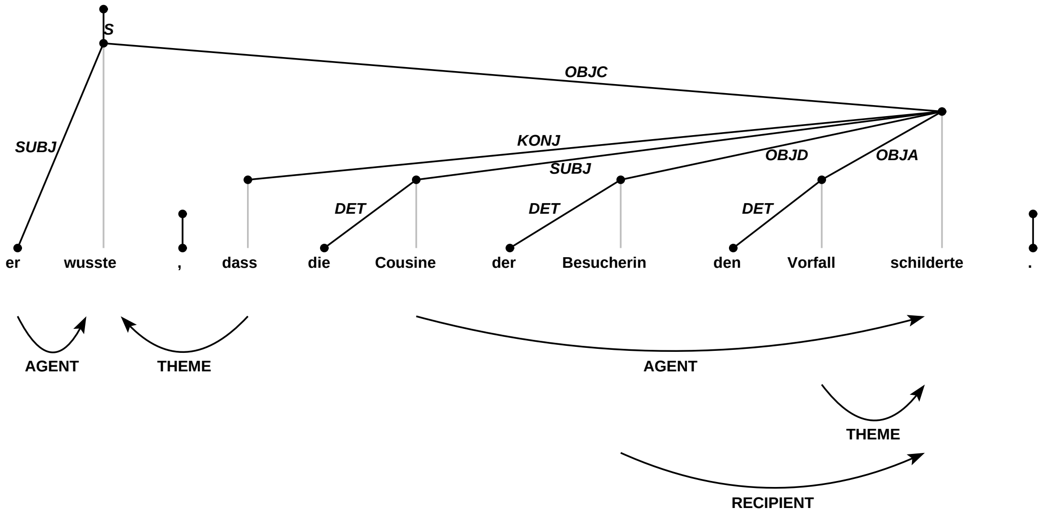
Sentence VK-151



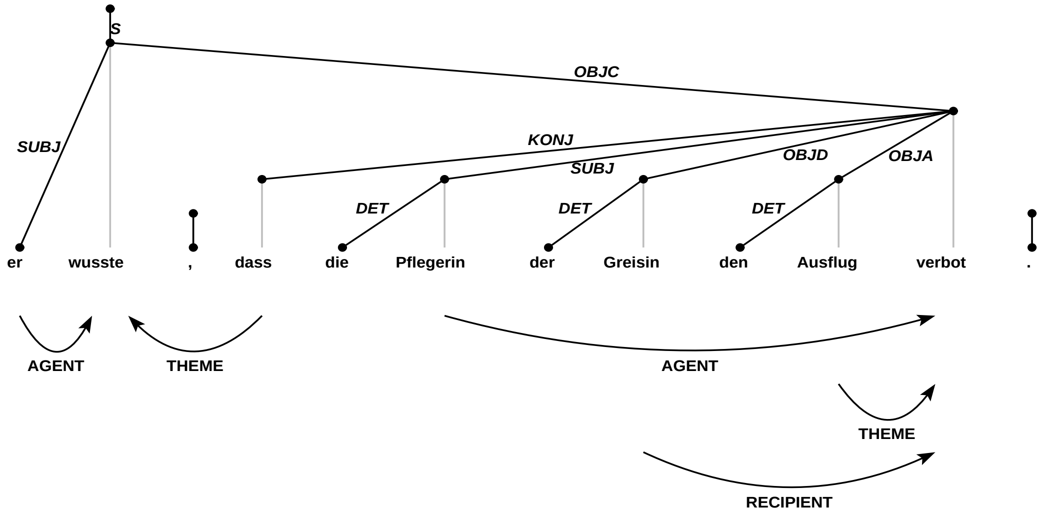
Sentence VK-226



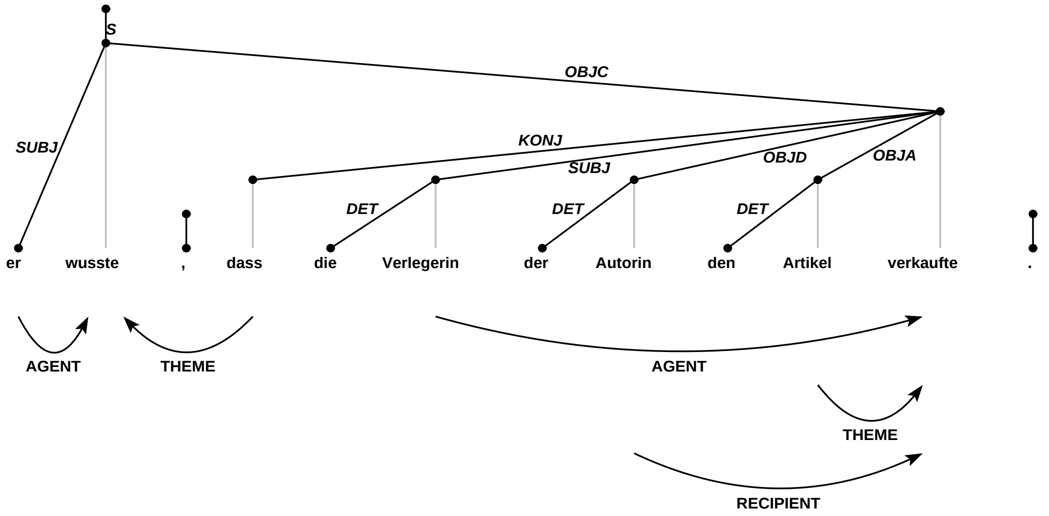
Sentence VK-233



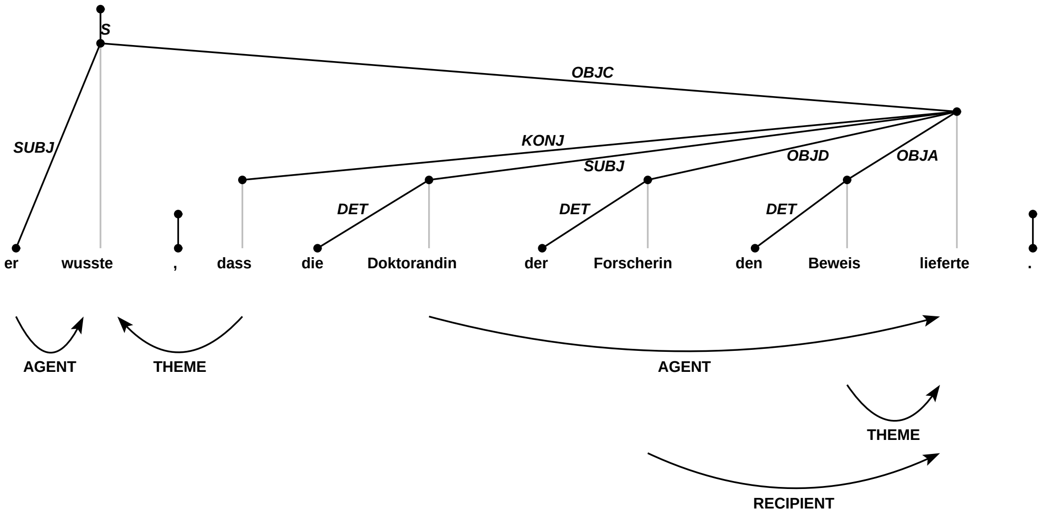
Sentence VK-247



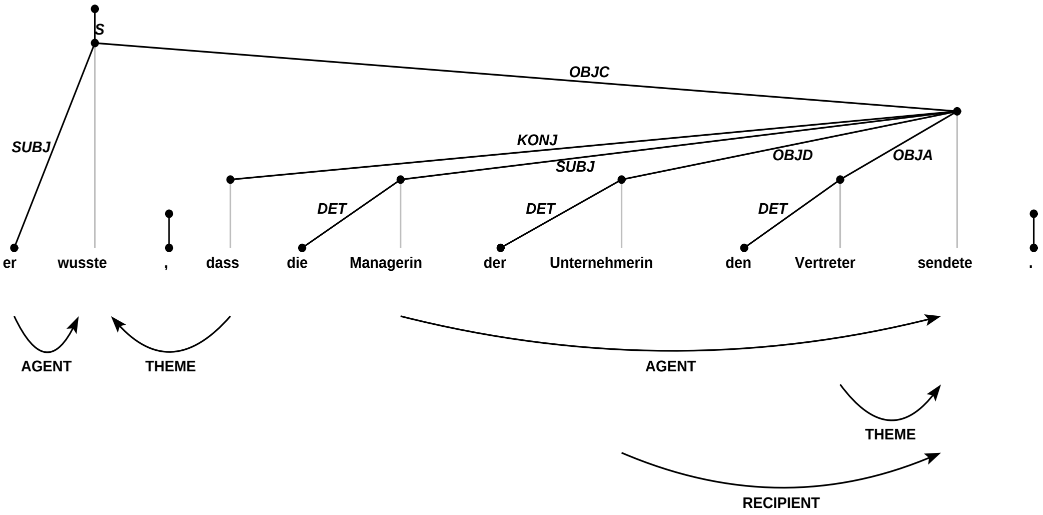
Sentence VK-263



Sentence VK-274

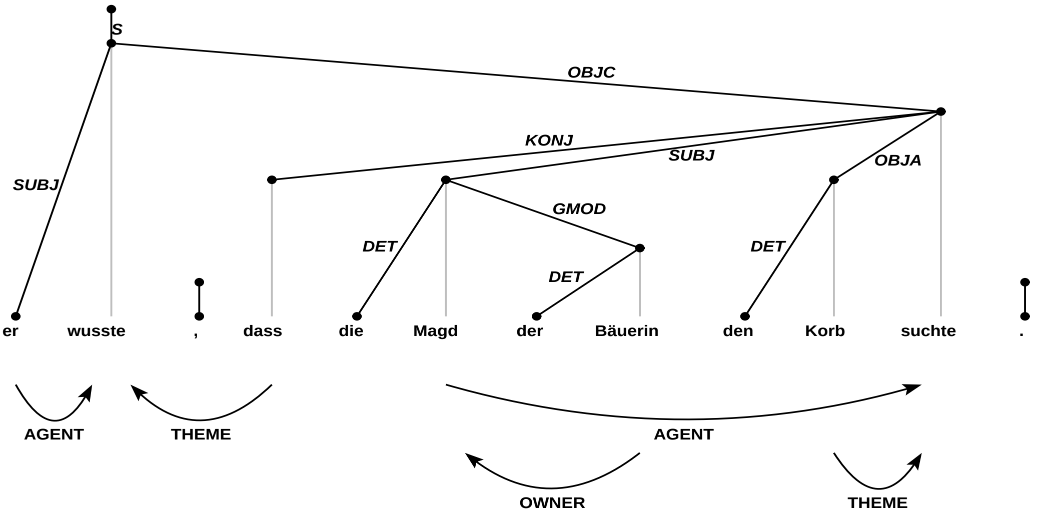


Sentence VK-306

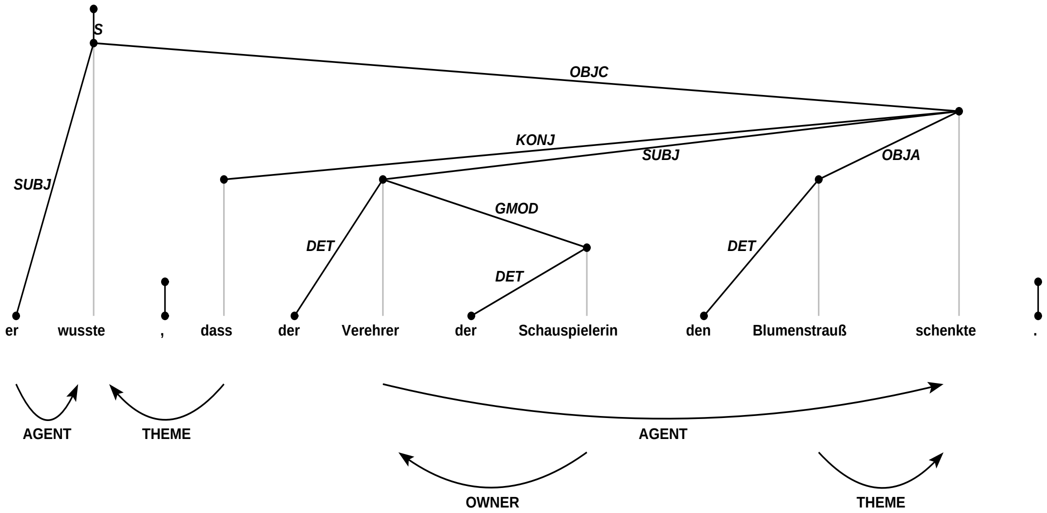


VI.1.6 Integration of Generalised Three-Entity Contexts Centred around an Instance of BINARY.SITUATION

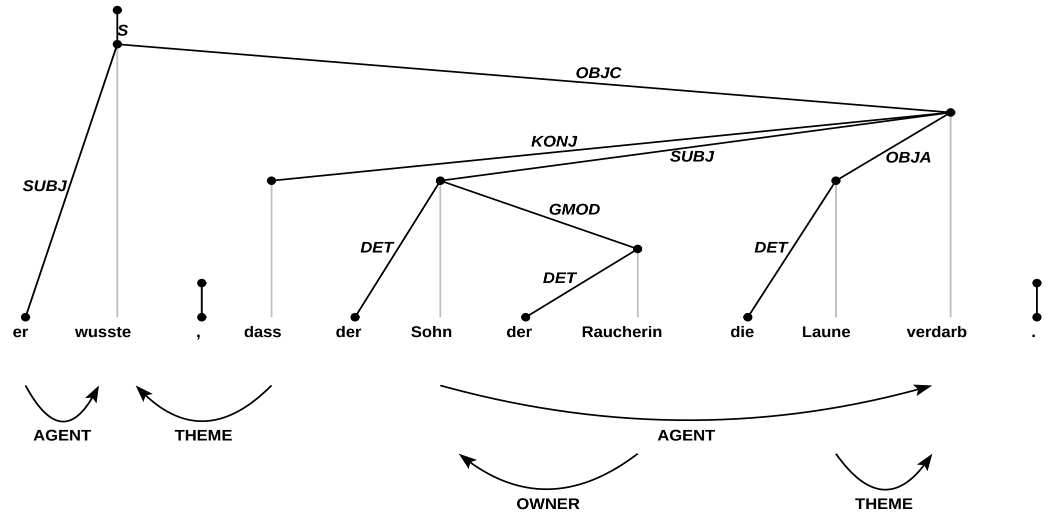
Sentence VK-011



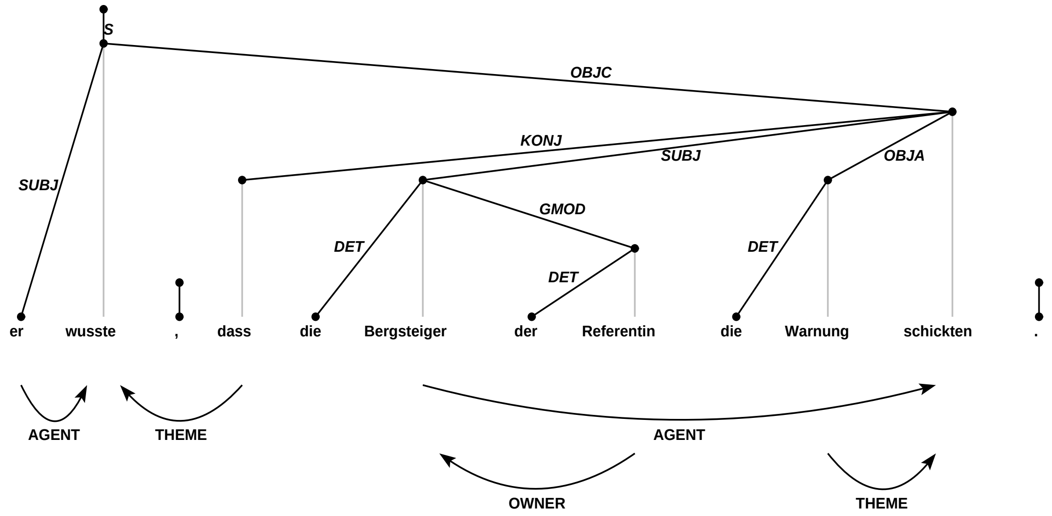
Sentence VK-100



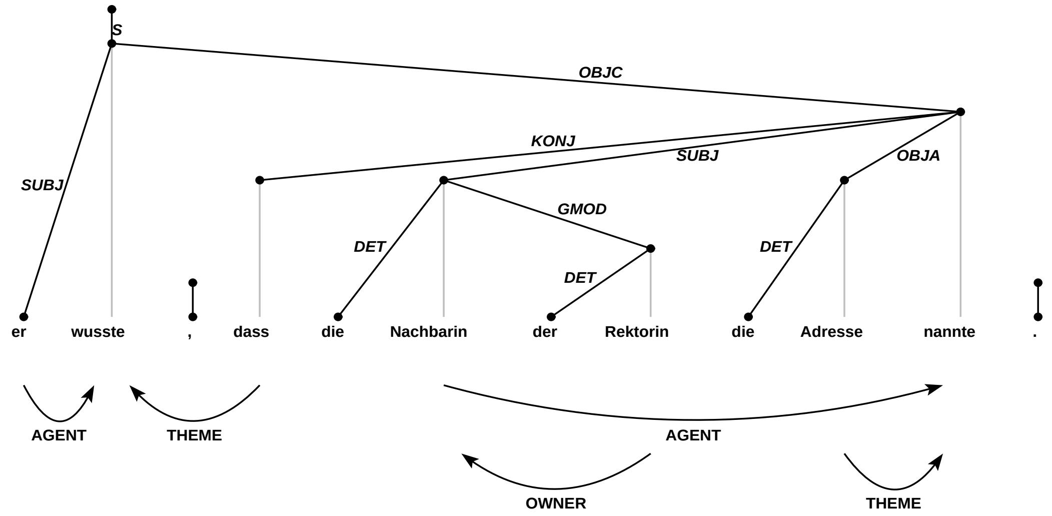
Sentence VK-111



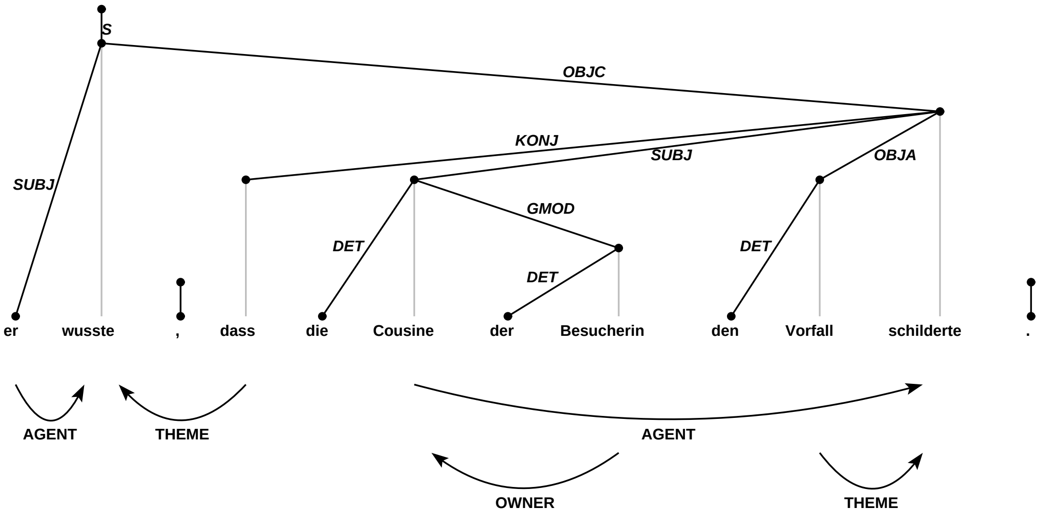
Sentence VK-151



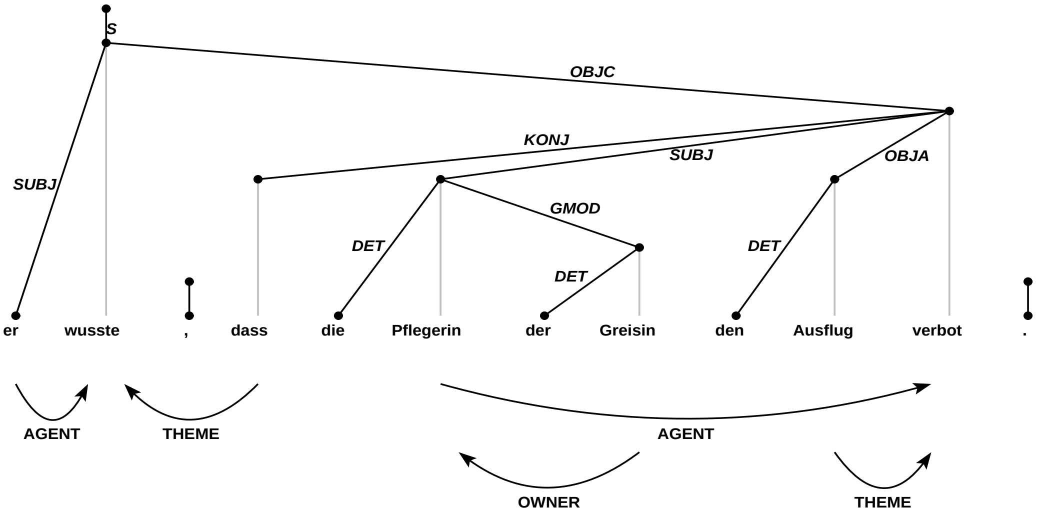
Sentence VK-226



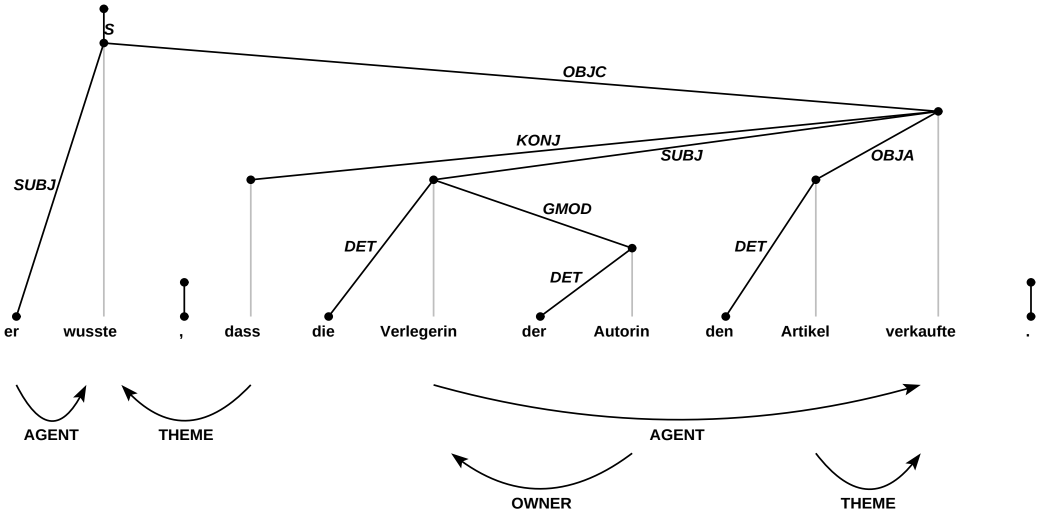
Sentence VK-233



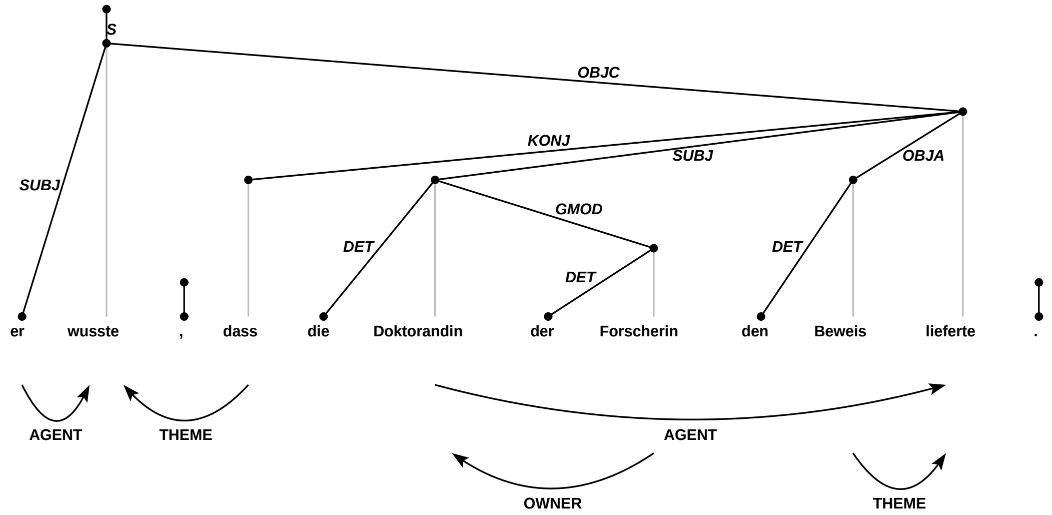
Sentence VK-247



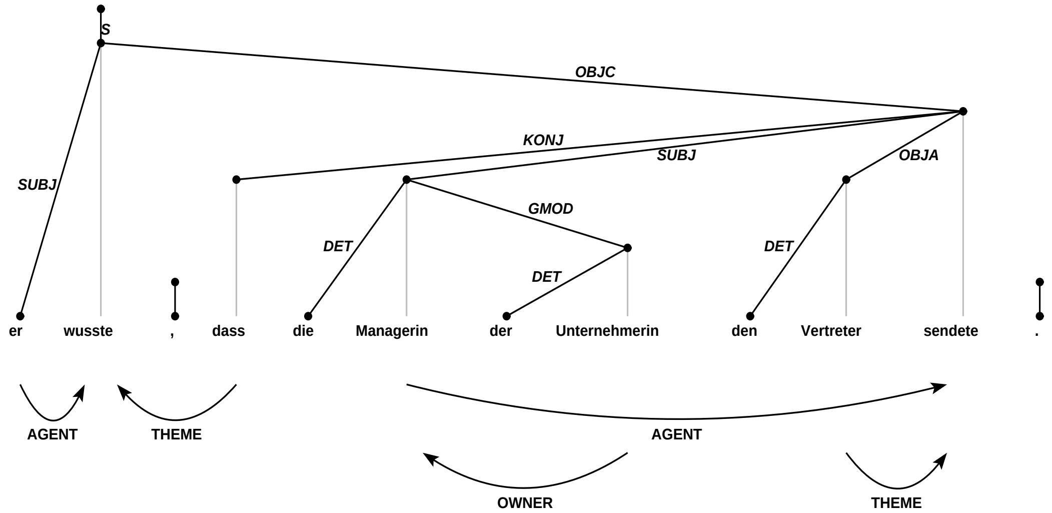
Sentence VK-263



Sentence VK-274

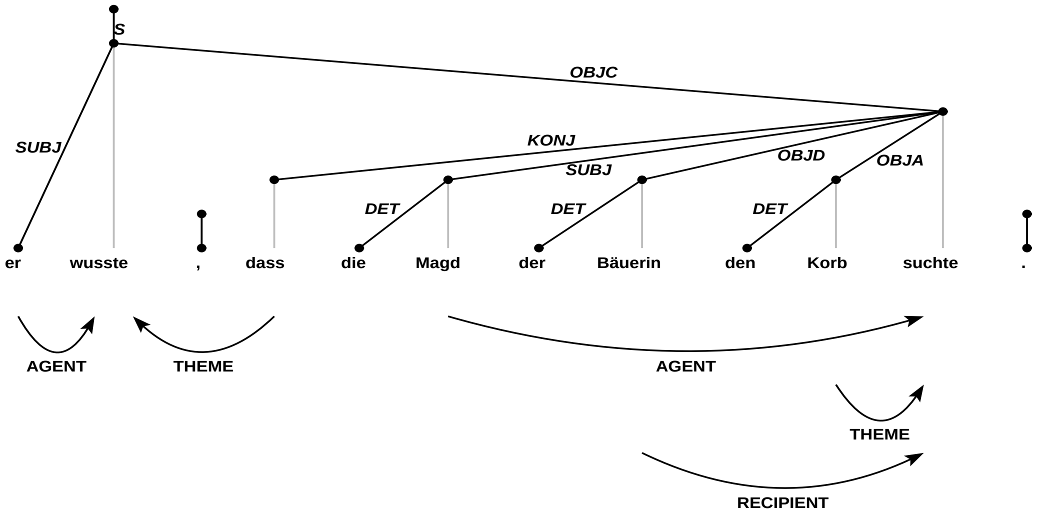


Sentence VK-306

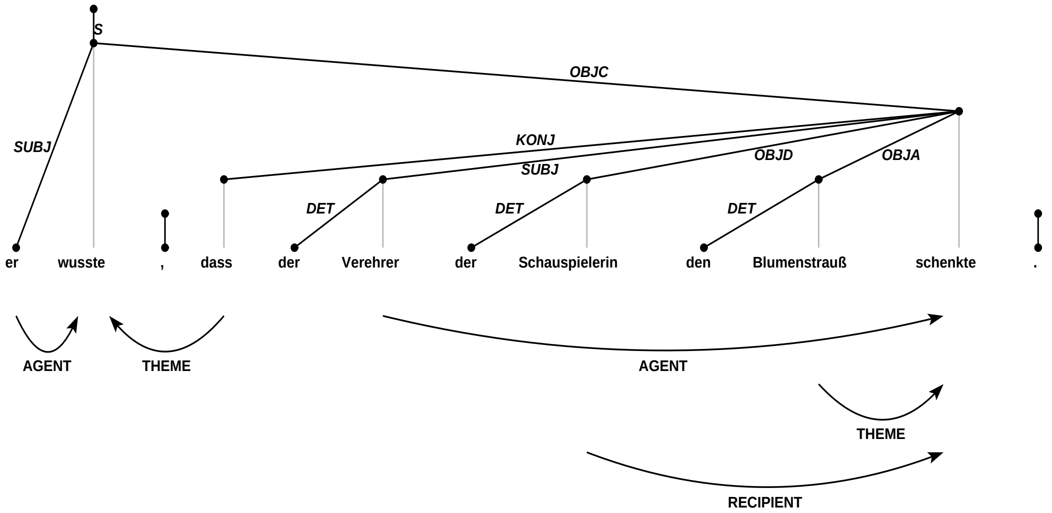


VI.1.7 Integration of Generalised Three-Entity Contexts Centred around an Instance of TERNARY.SITUATION

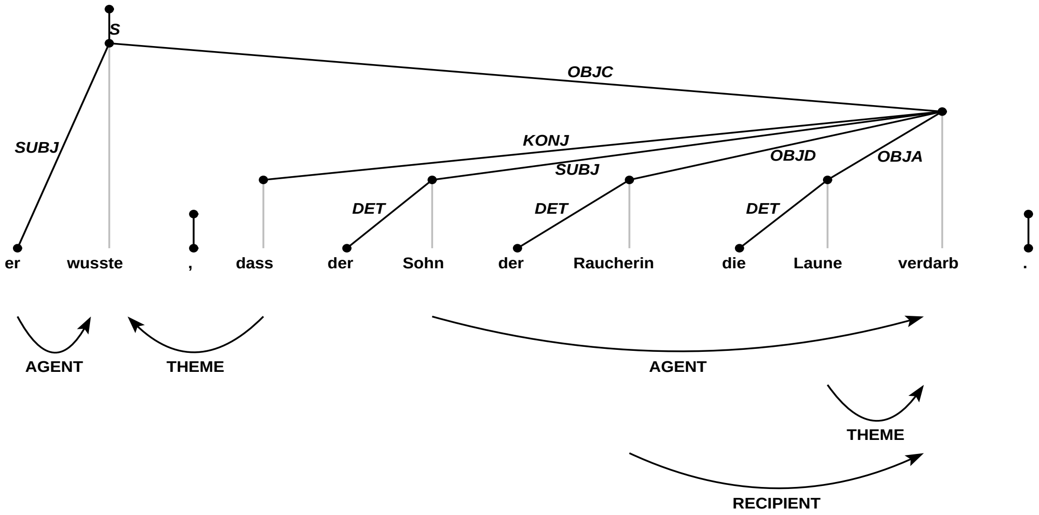
Sentence VK-011



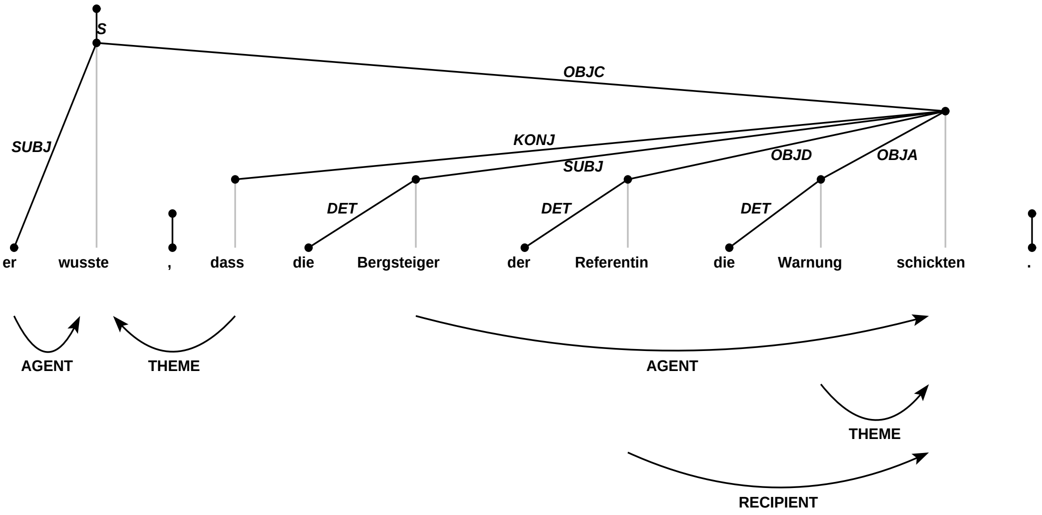
Sentence VK-100



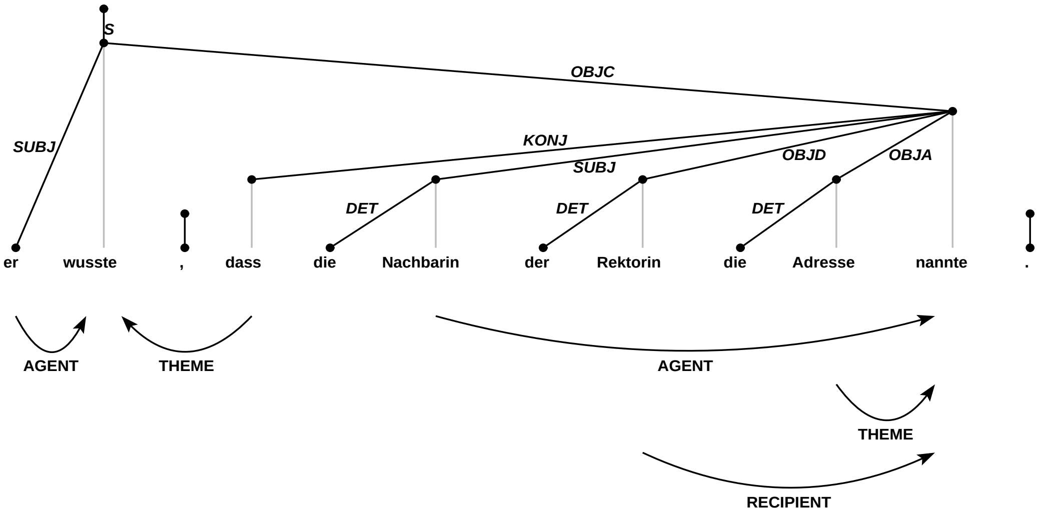
Sentence VK-111



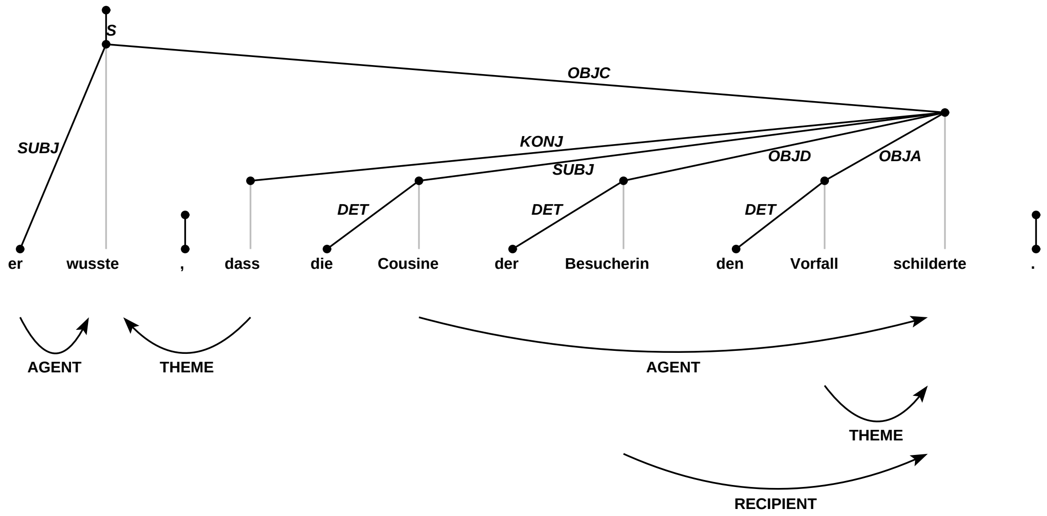
Sentence VK-151


S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Bergsteiger der Referentin die Warnung schickten .
AGENT
THEME
AGENT
THEME
RECIPIENT

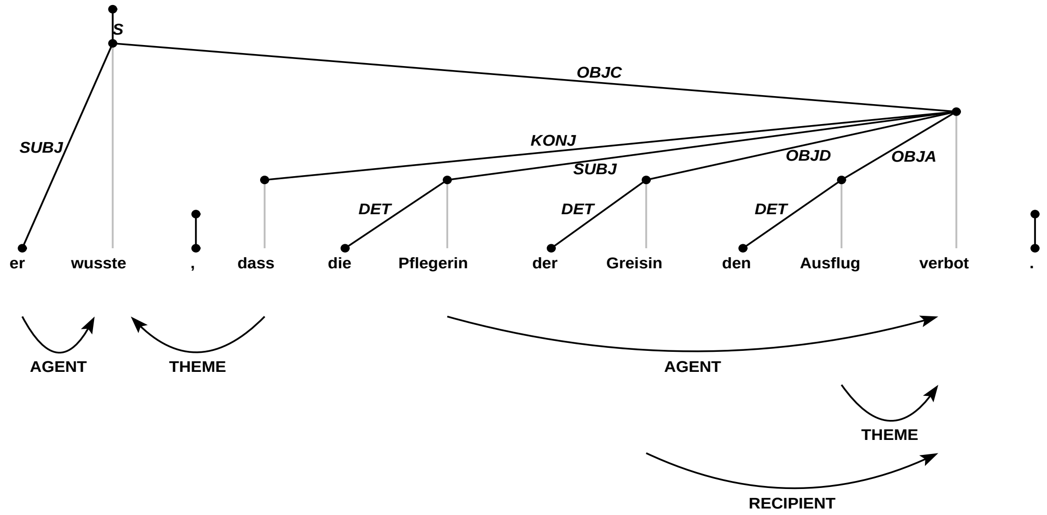
Sentence VK-226


S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Nachbarin der Rektorin die Adresse nannte .
AGENT
THEME
AGENT
THEME
RECIPIENT

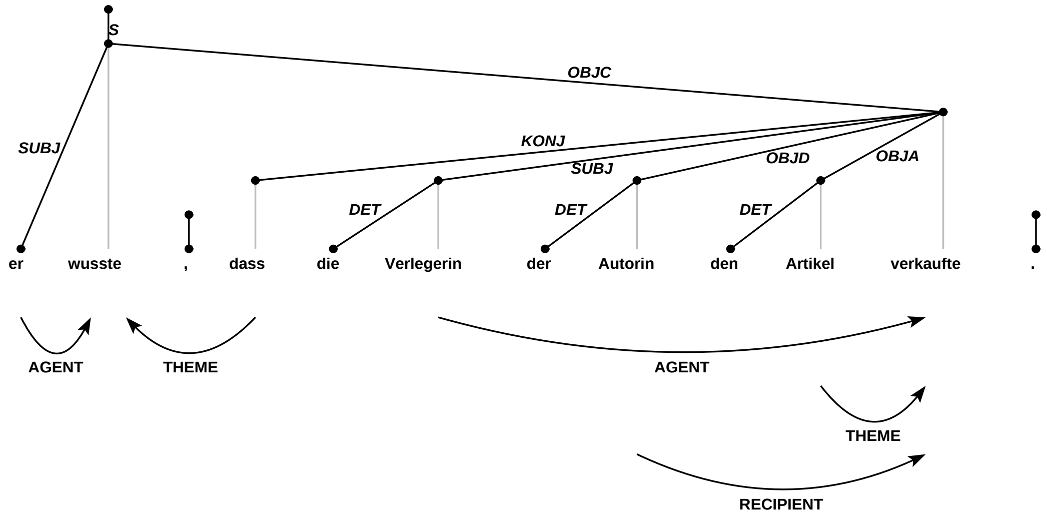
Sentence VK-233



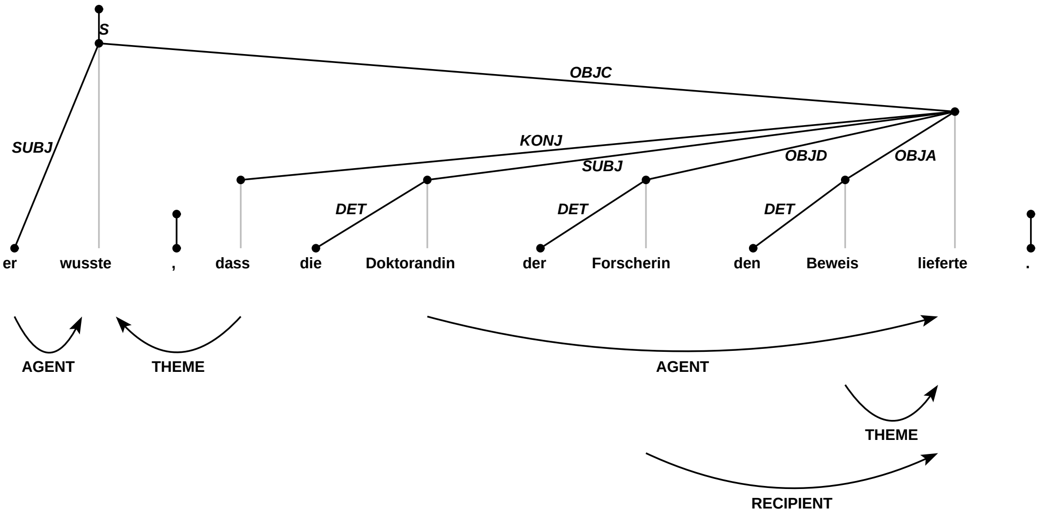
Sentence VK-247



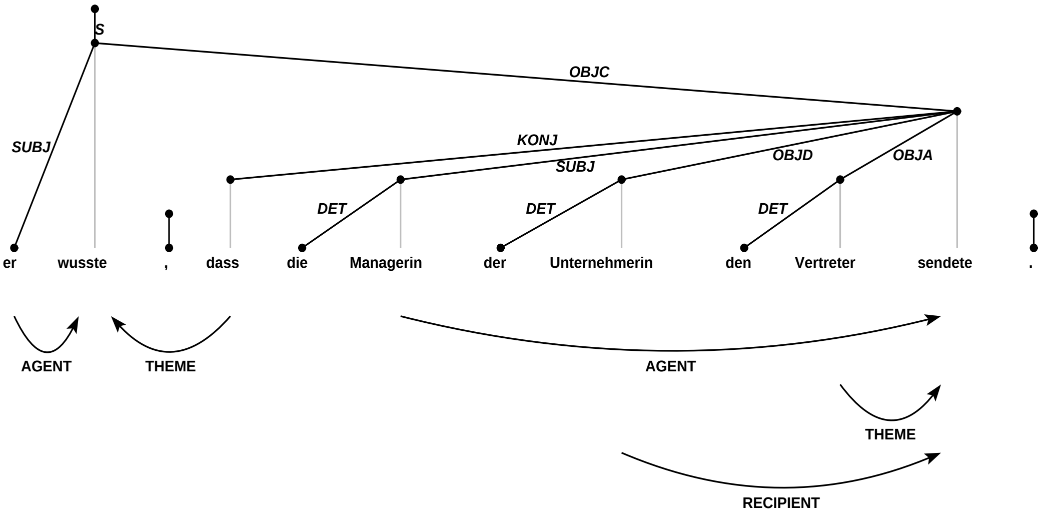
Sentence VK-263



Sentence VK-274

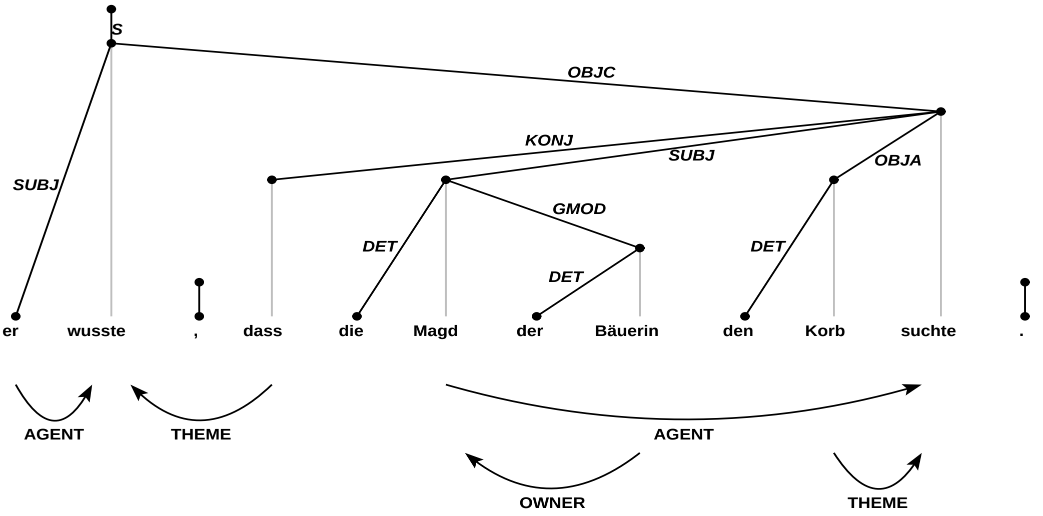

S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Doktorandin der Forscherin den Beweis lieferte .
AGENT
THEME
AGENT
THEME
RECIPIENT

Sentence VK-306

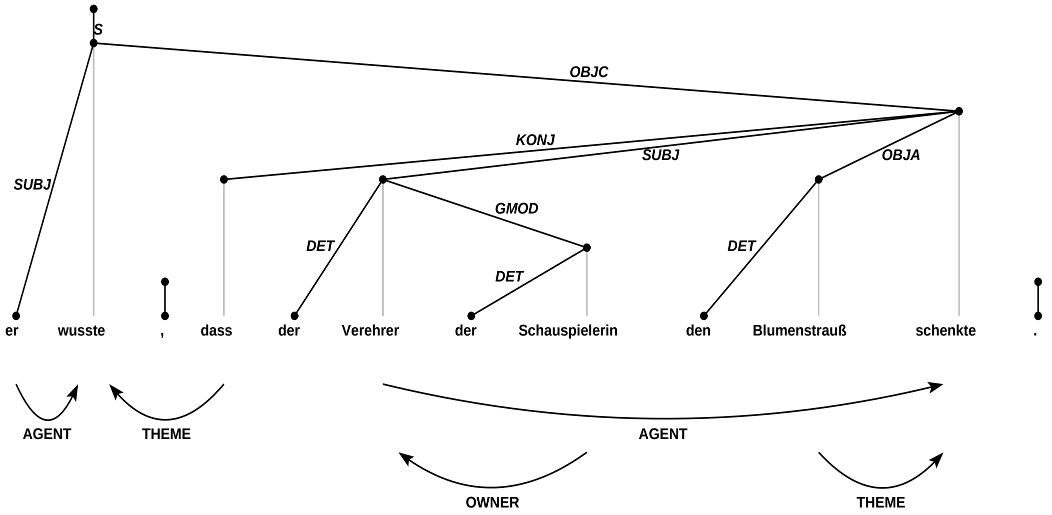

S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Managerin der Unternehmerin den Vertreter sendete .
AGENT
THEME
AGENT
THEME
RECIPIENT

VI.1.8 Integration of Generalised Two-Entity Contexts Centred around an Instance of BINARY.SITUATION

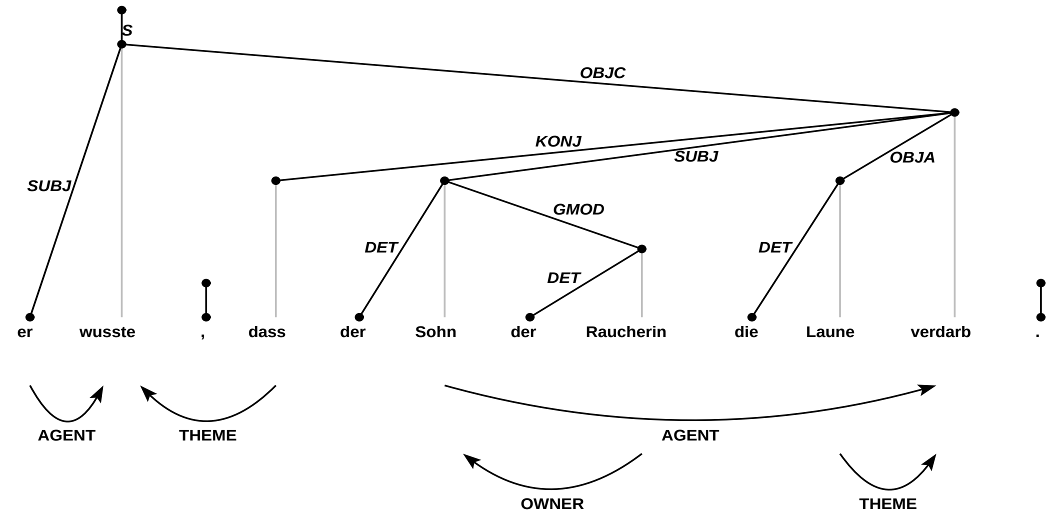
Sentence VK-011



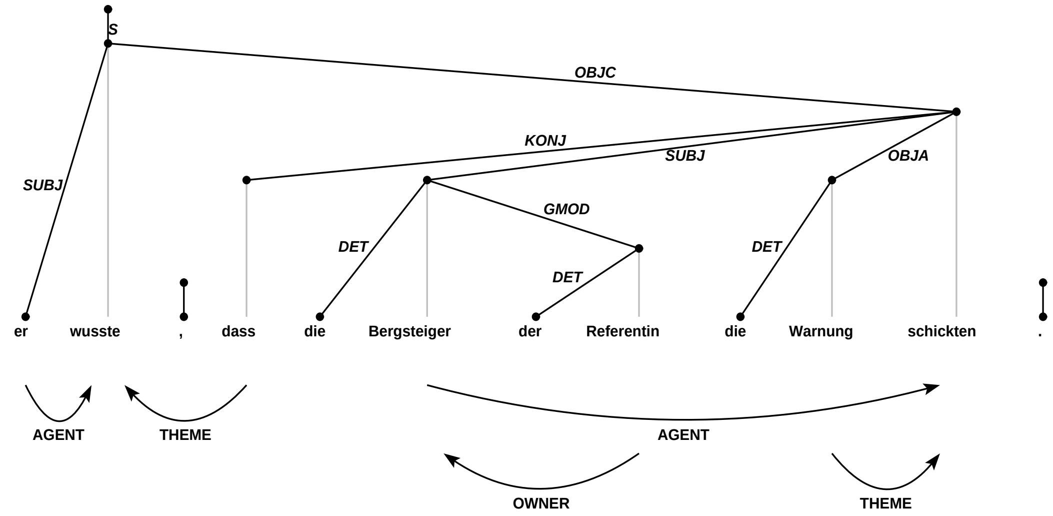
Sentence VK-100



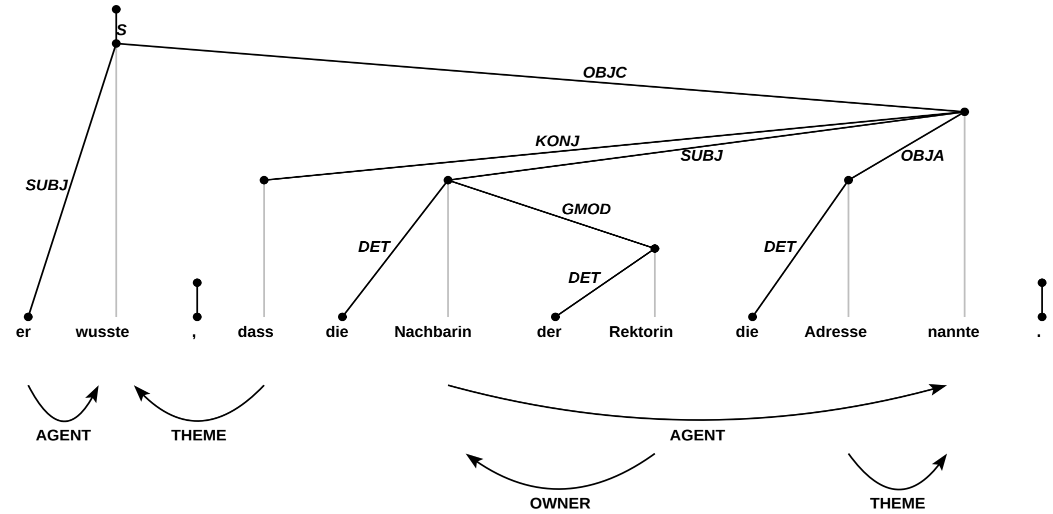
Sentence VK-111



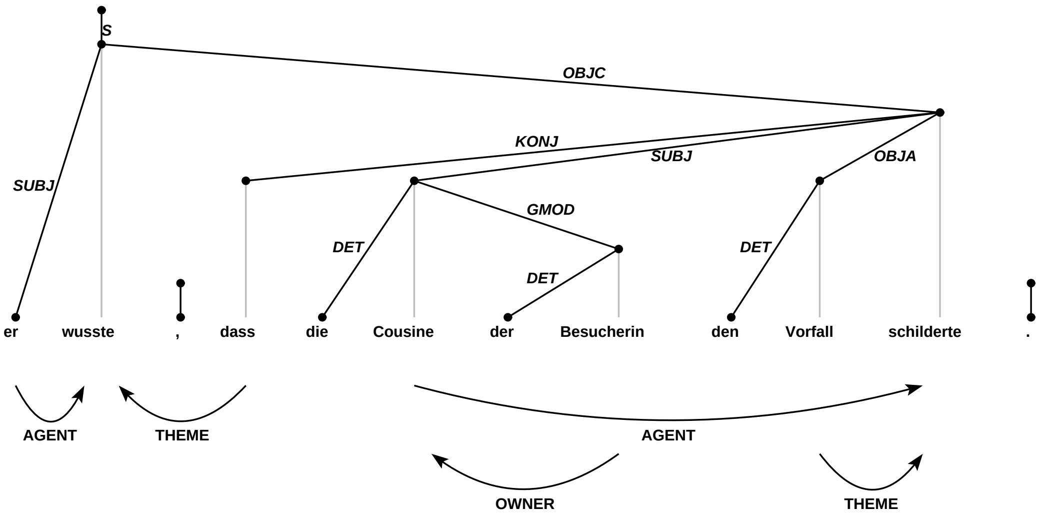
Sentence VK-151


S
OBJC
KONJ
SUBJ
SUBJ
OBJA
GMOD
DET
DET
DET
er wusste , dass die Bergsteiger der Referentin die Warnung schickten .
AGENT
THEME
AGENT
OWNER
THEME

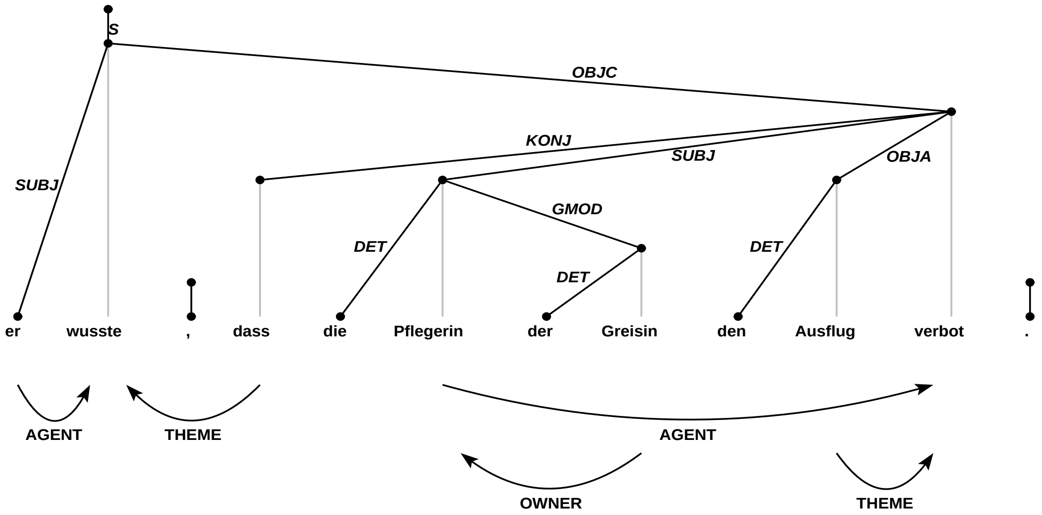
Sentence VK-226


S
OBJC
KONJ
SUBJ
SUBJ
OBJA
GMOD
DET
DET
DET
er wusste , dass die Nachbarin der Rektorin die Adresse nannte .
AGENT
THEME
AGENT
OWNER
THEME

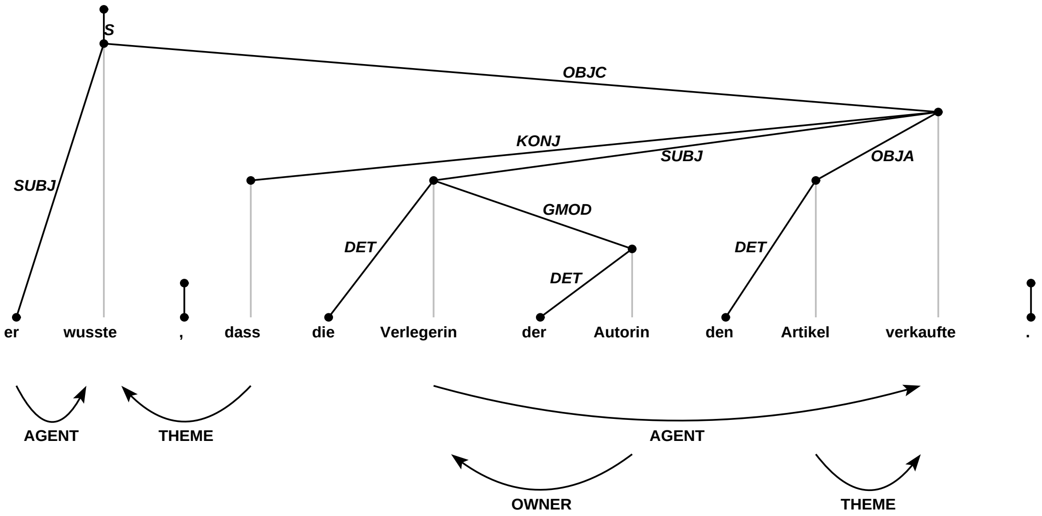
Sentence VK-233


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Cousine der Besucherin den Vorfall schilderte .
AGENT
THEME
AGENT
OWNER
THEME

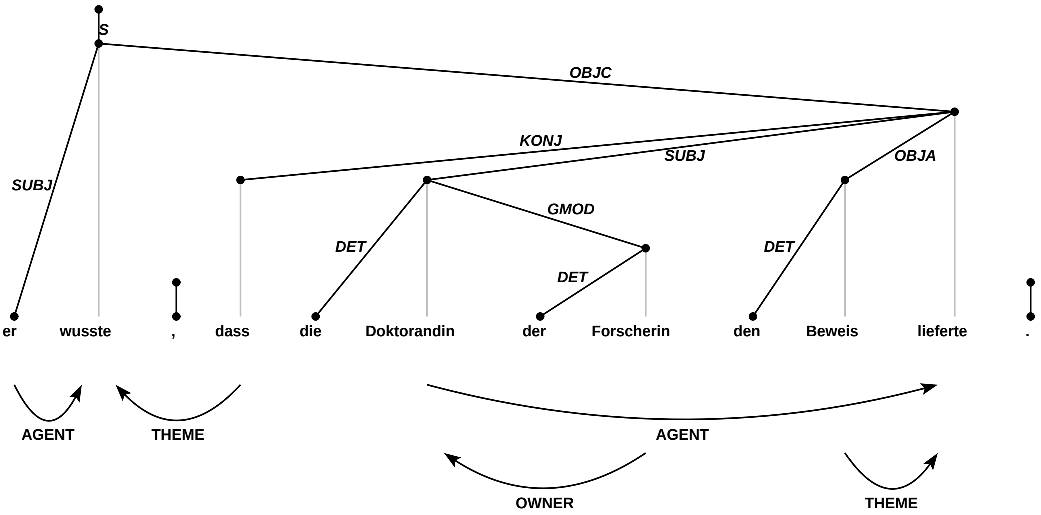
Sentence VK-247


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Pflegerin der Greisin den Ausflug verbot .
AGENT
THEME
AGENT
OWNER
THEME

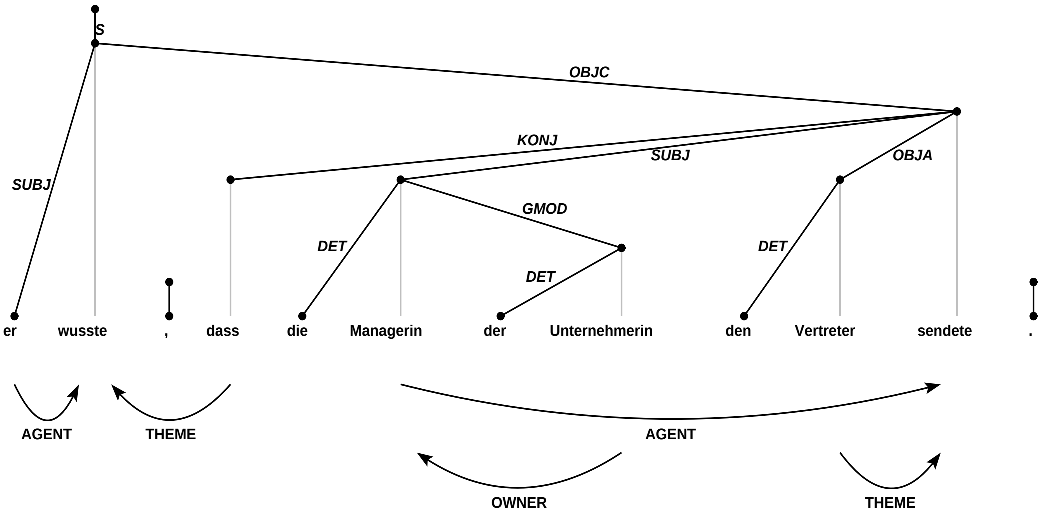
Sentence VK-263


S
OBJC
KONJ
SUBJ
OBJA
SUBJ
GMOD
DET
DET
DET
er wusste , dass die Verlegerin der Autorin den Artikel verkaufte .
AGENT
THEME
AGENT
OWNER
THEME

Sentence VK-274

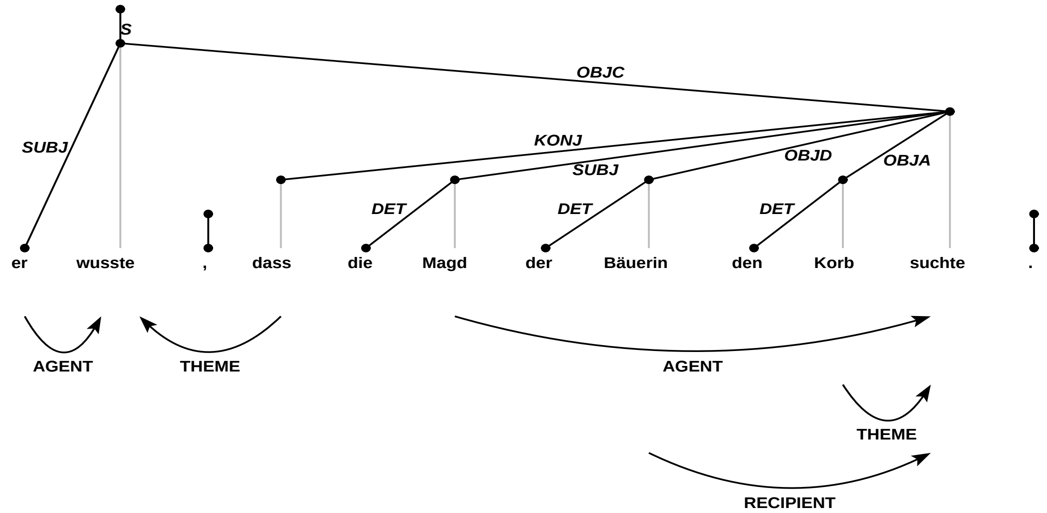


Sentence VK-306

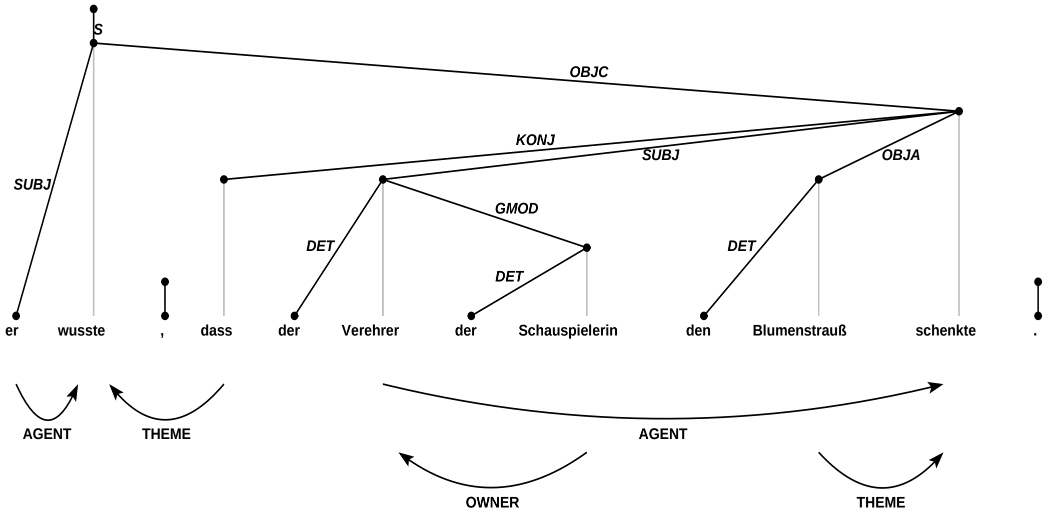


VI.1.9 Integration of Generalised Two-Entity Contexts Centred around an Instance of SITUATION.CONCEPT

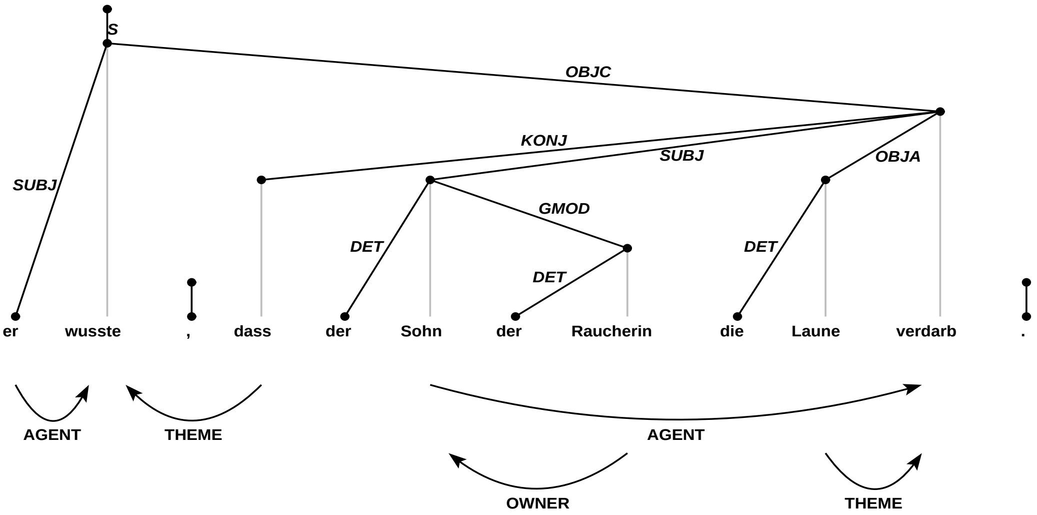
Sentence VK-011



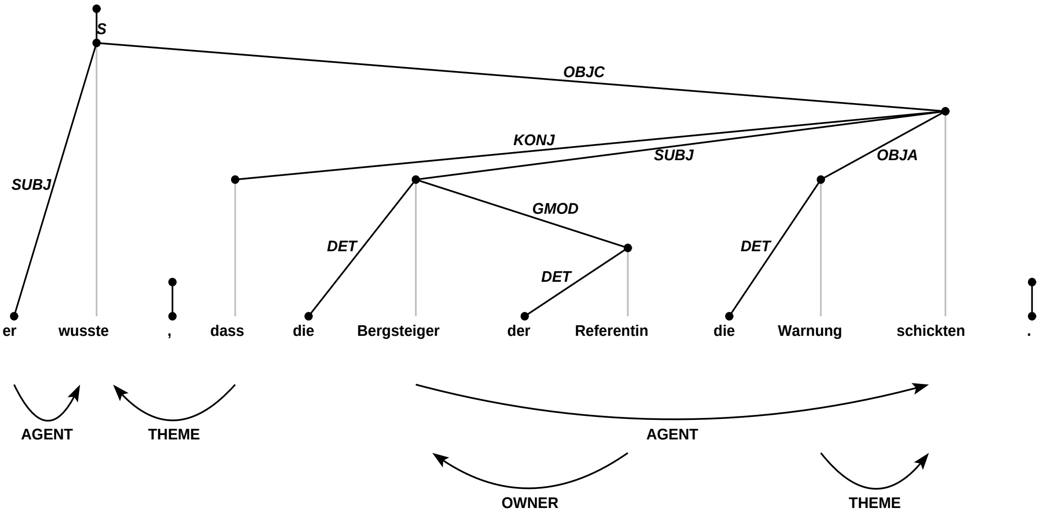
Sentence VK-100



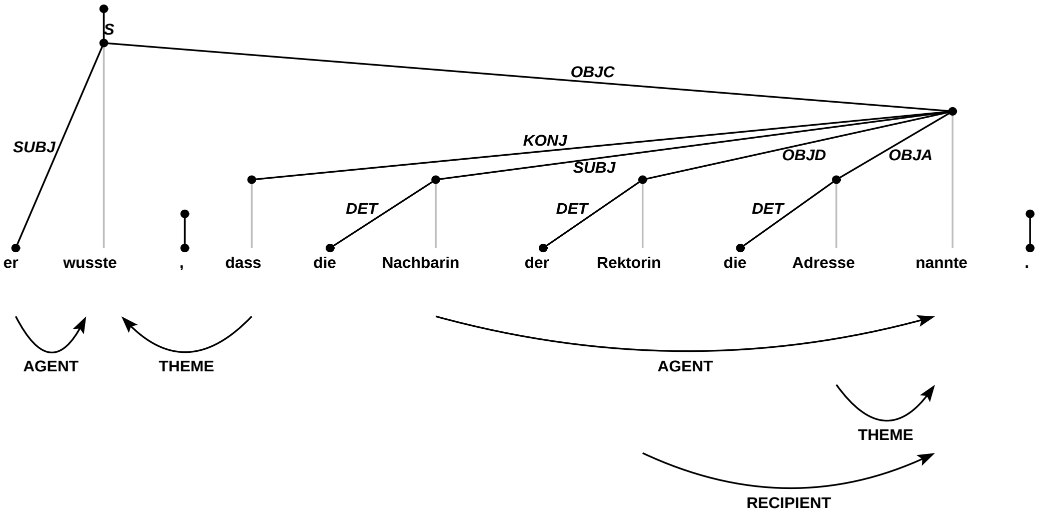
Sentence VK-111



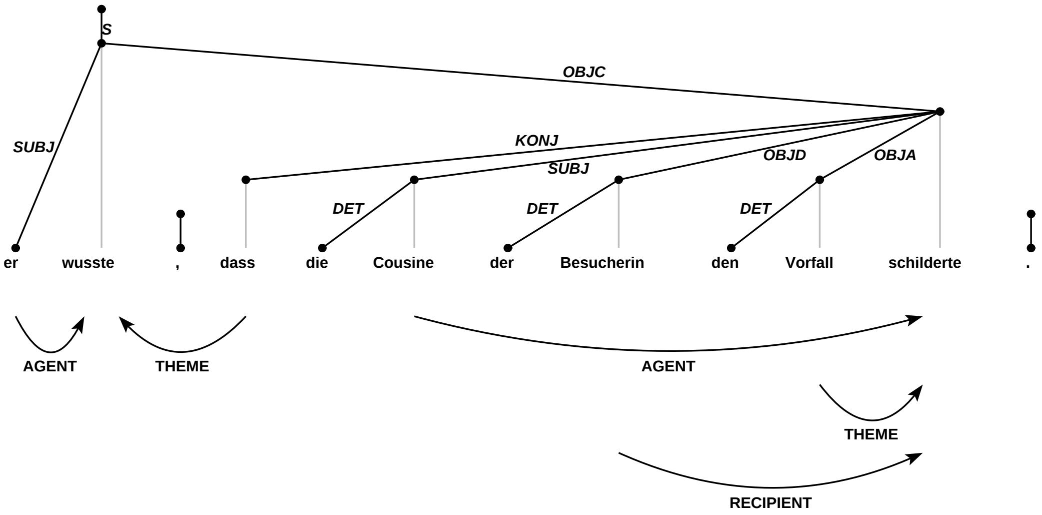
Sentence VK-151



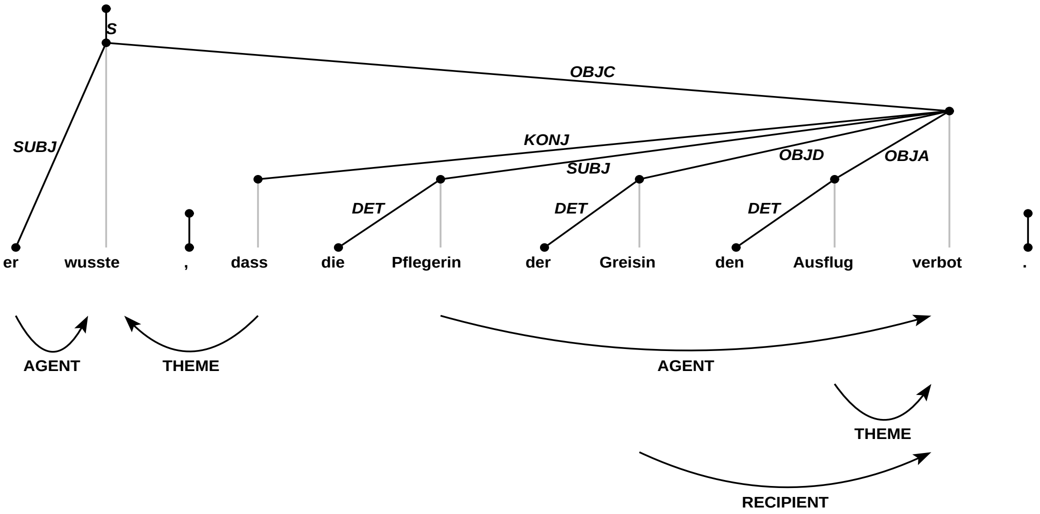
Sentence VK-226



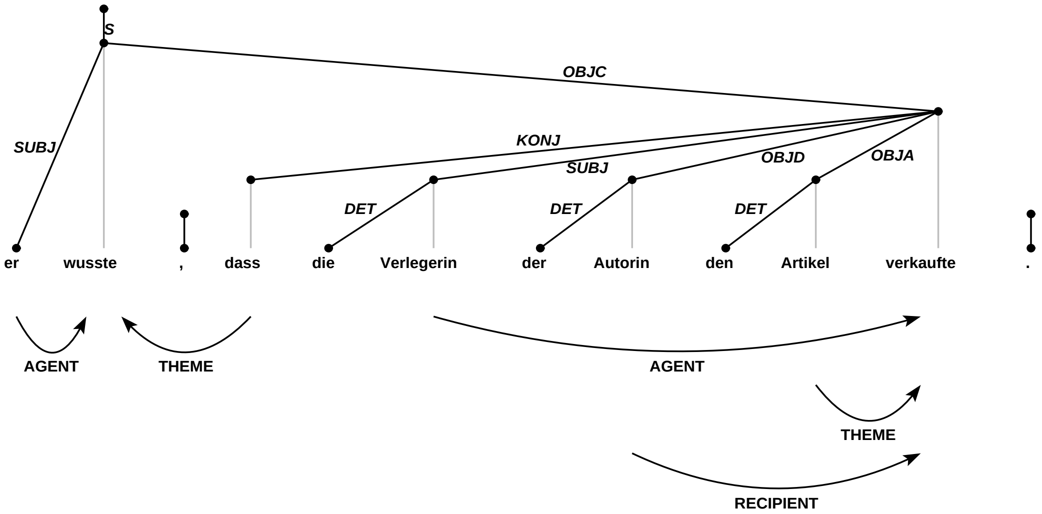
Sentence VK-233


S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Cousine der Besucherin den Vorfall schilderte .
AGENT
THEME
AGENT
THEME
RECIPIENT

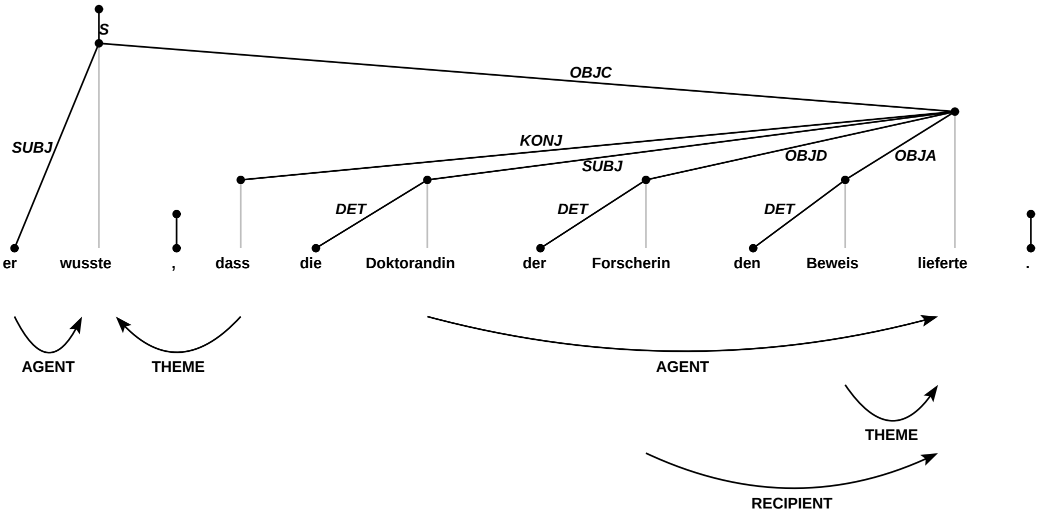
Sentence VK-247


S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Pflegerin der Greisin den Ausflug verbot .
AGENT
THEME
AGENT
THEME
RECIPIENT

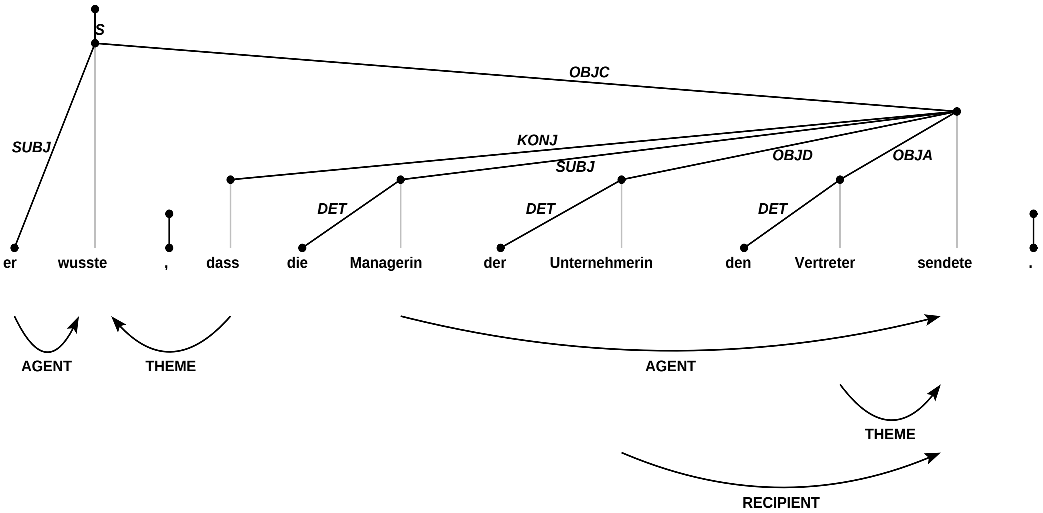
Sentence VK-263


S
OBJC
SUBJ
KONJ
SUBJ
OBJD
OBJA
DET
DET
DET
er wusste , dass die Verlegerin der Autorin den Artikel verkaufte .
AGENT
THEME
AGENT
THEME
RECIPIENT

Sentence VK-274



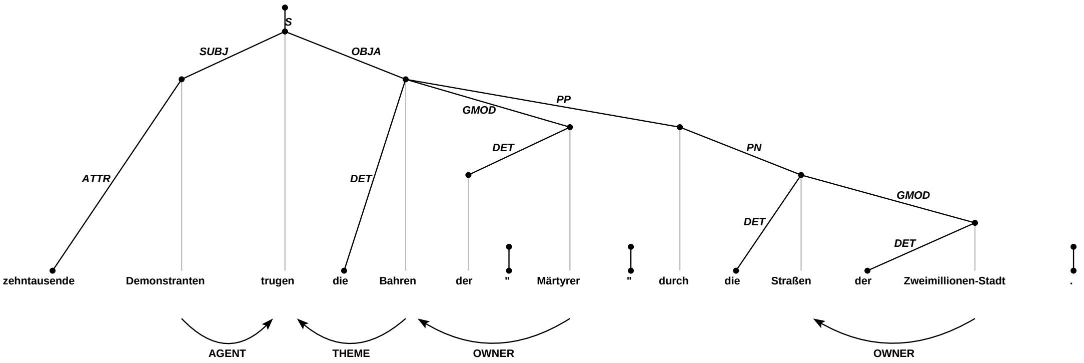
Sentence VK-306



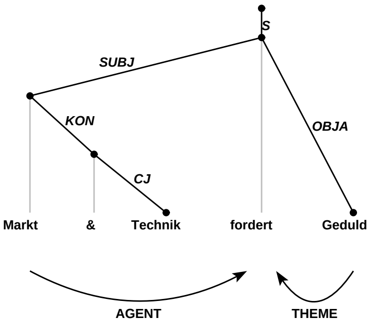
VI.2 Sentences with Subject-Object-Ambiguity

VI.2.1 Integration of an Empty Context Model

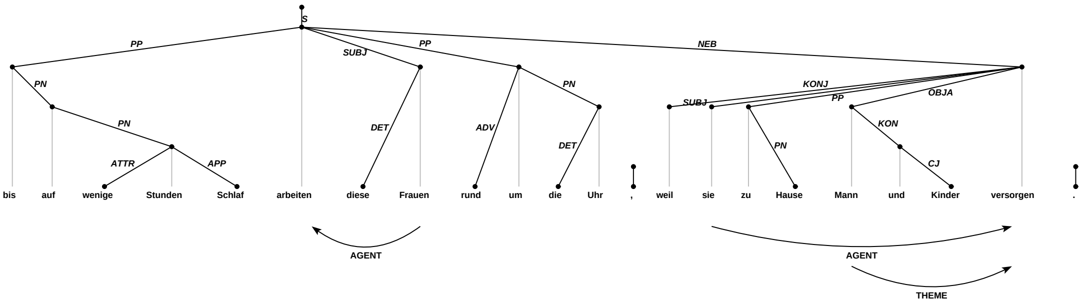
Sentence SO-360



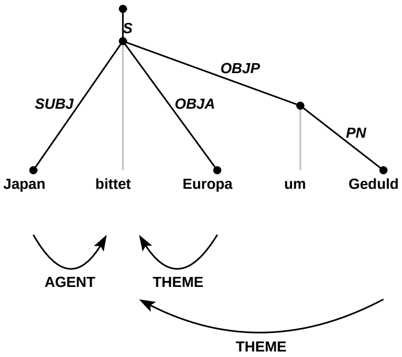
Sentence SO-706



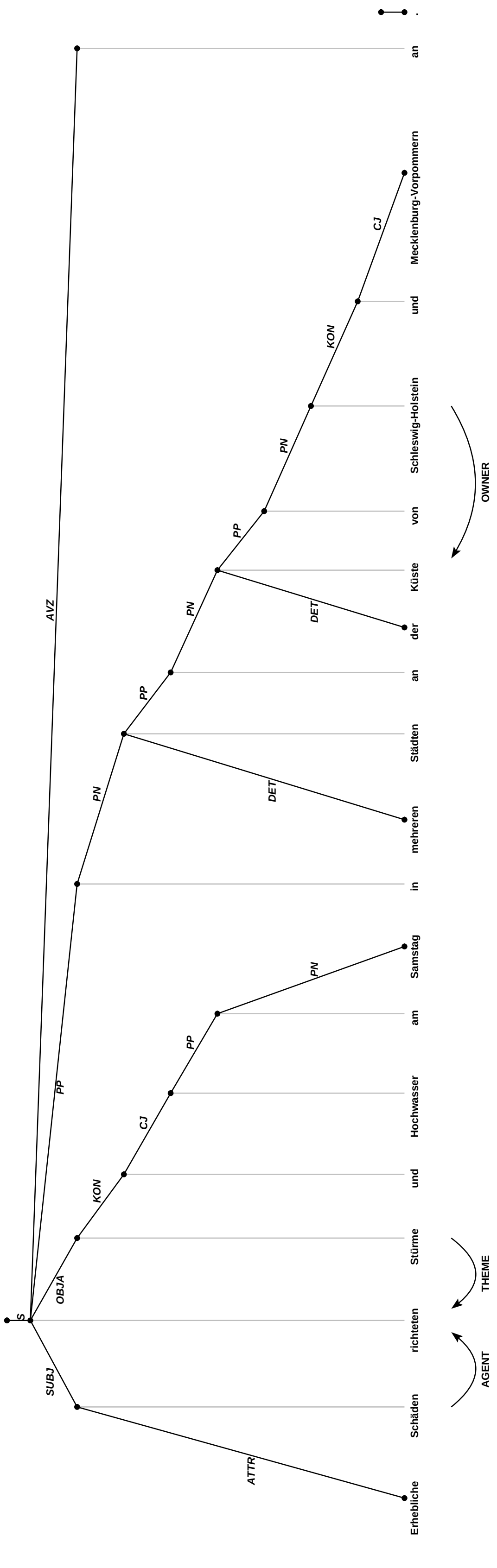
Sentence SO-841



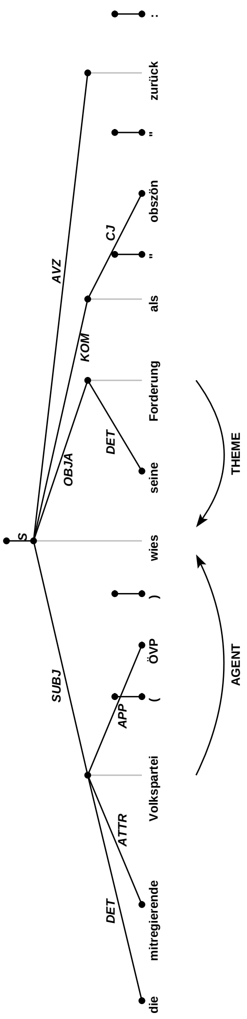
Sentence SO-1090



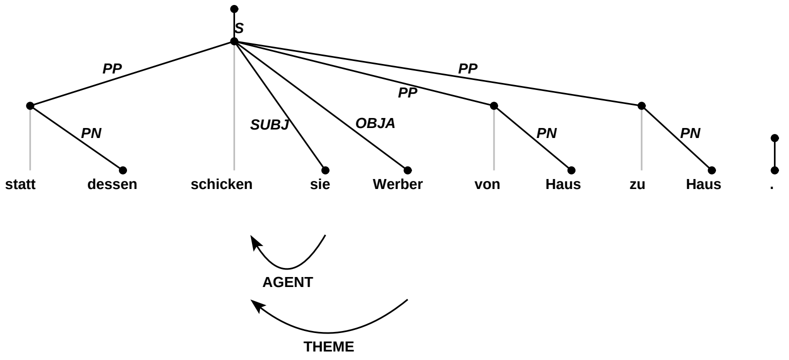
Sentence SO-4493



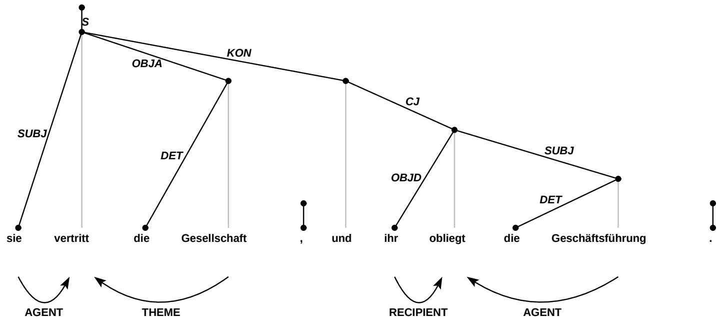
Sentence SO-6179



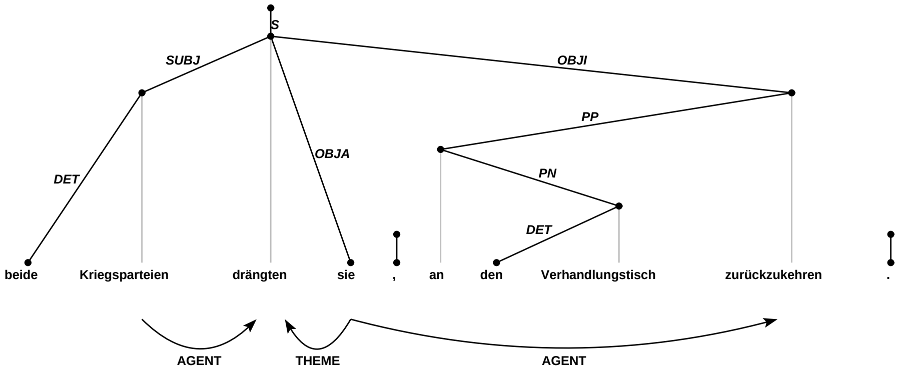
Sentence SO-9681



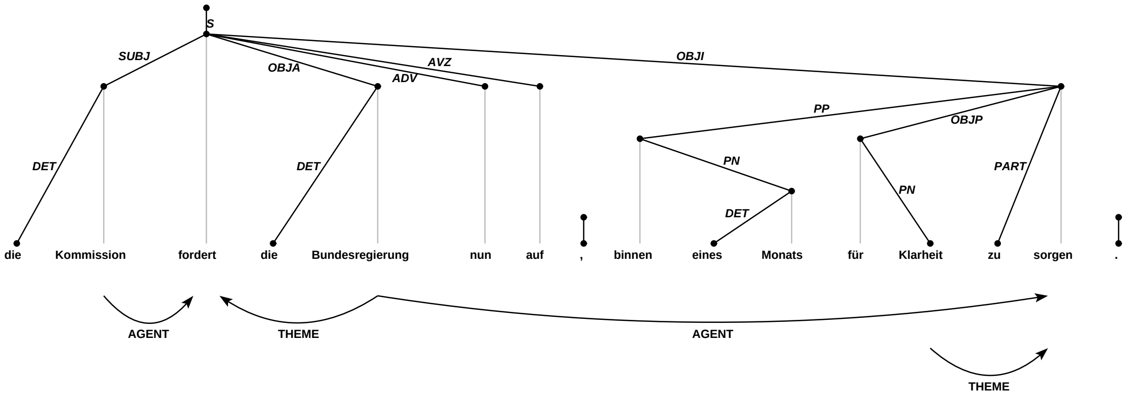
Sentence SO-9792



Sentence SO-10744

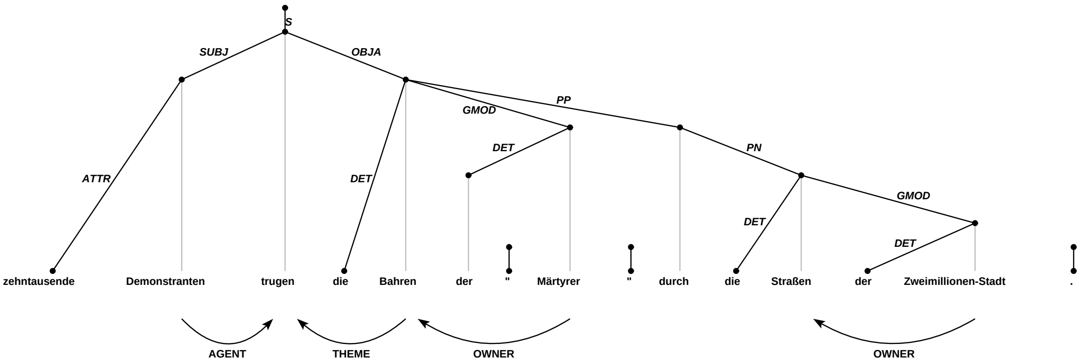


Sentence SO-40722

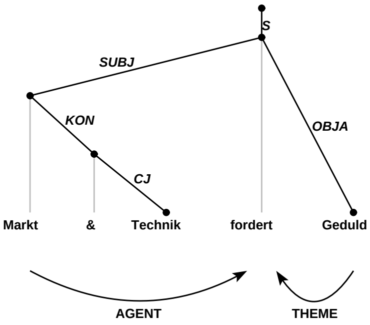


VI.2.2 Soft Integration of SUBJ-OBJA Contexts

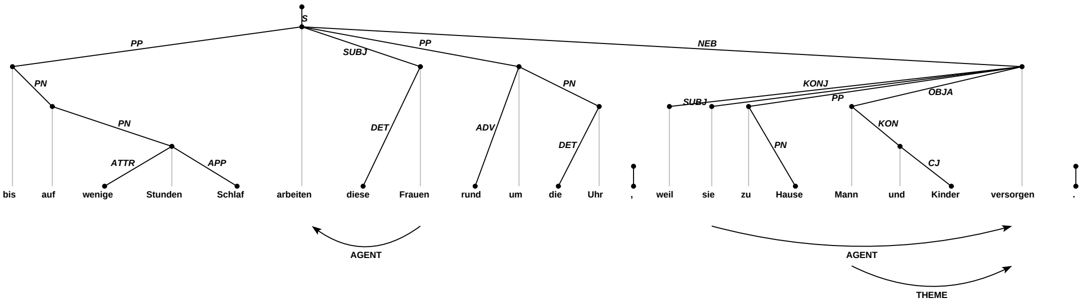
Sentence SO-360



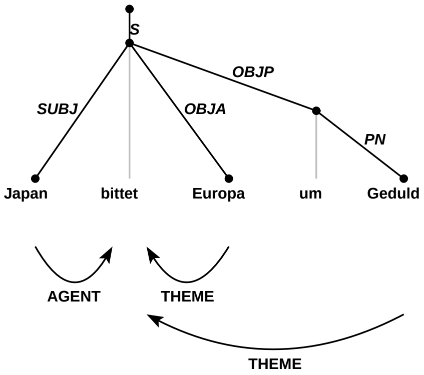
Sentence SO-706



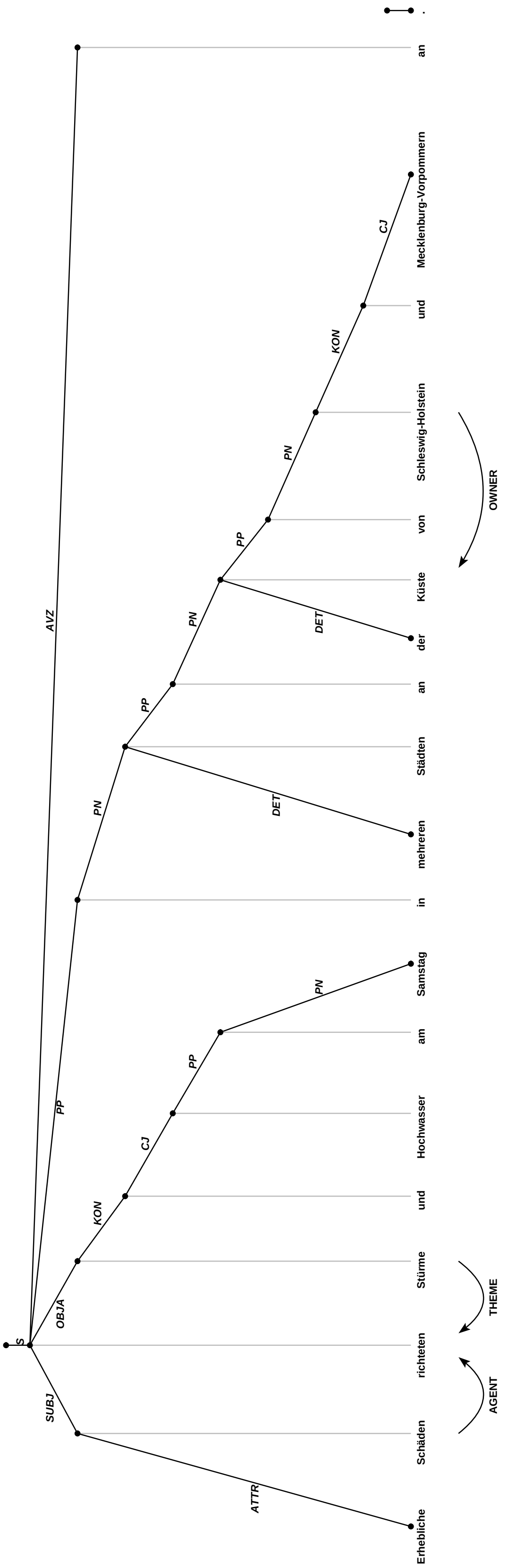
Sentence SO-841



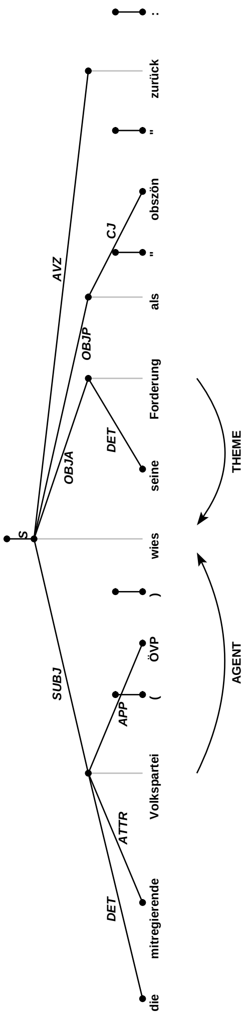
Sentence SO-1090



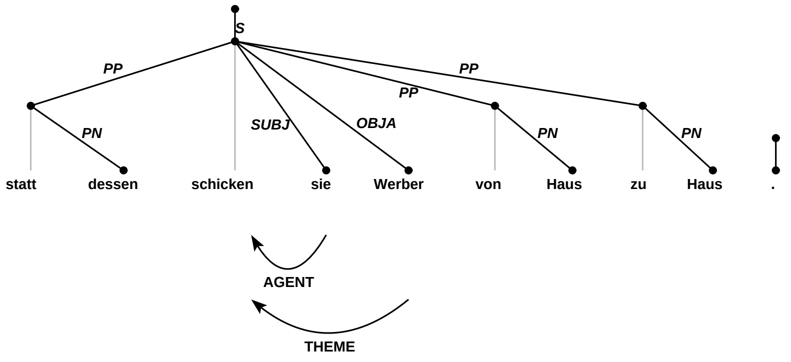
Sentence SO-4493



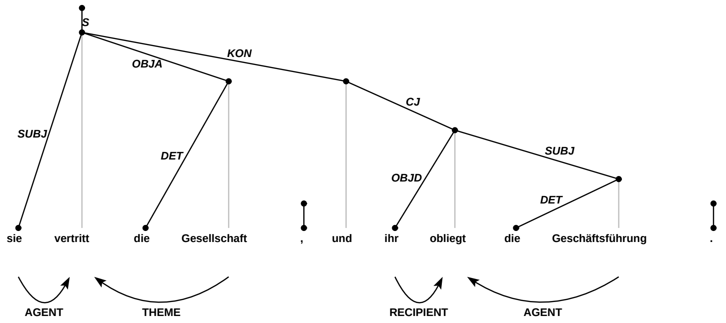
Sentence SO-6179



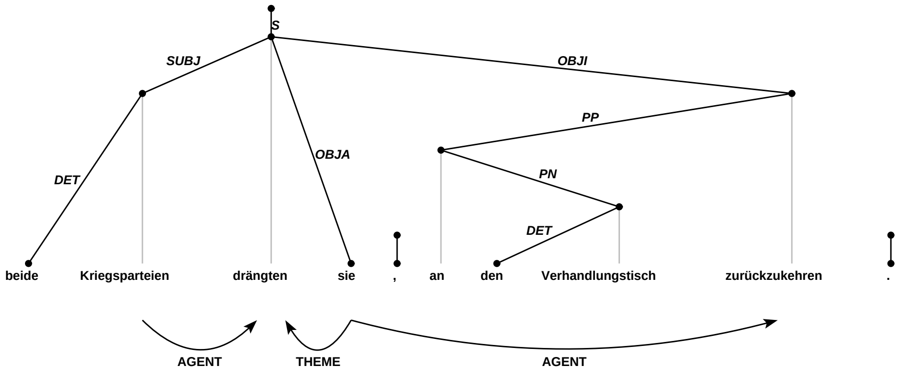
Sentence SO-9681



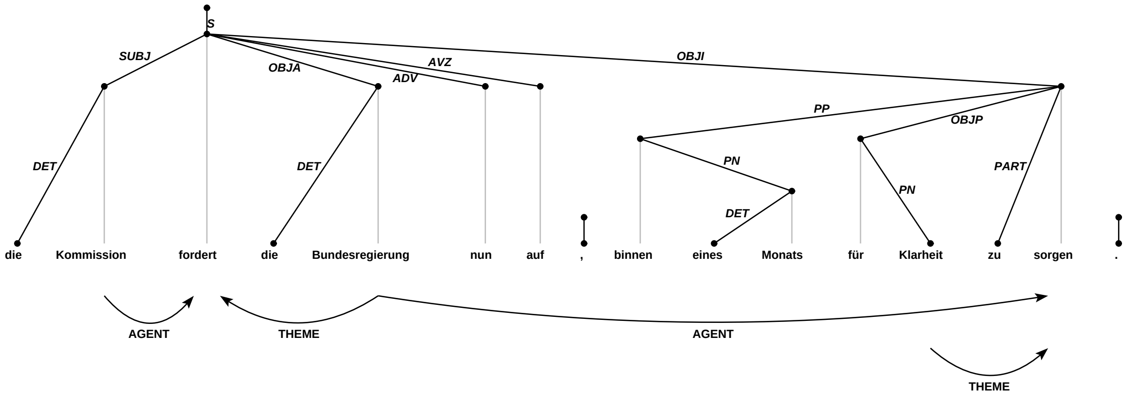
Sentence SO-9792



Sentence SO-10744

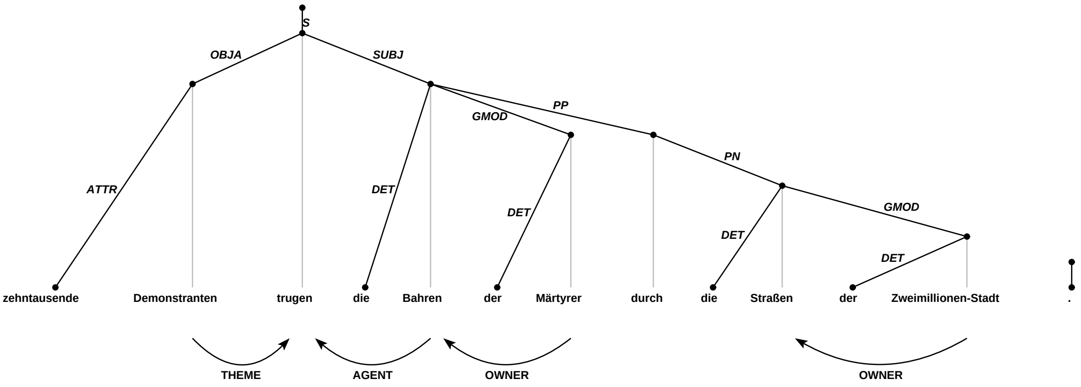


Sentence SO-40722

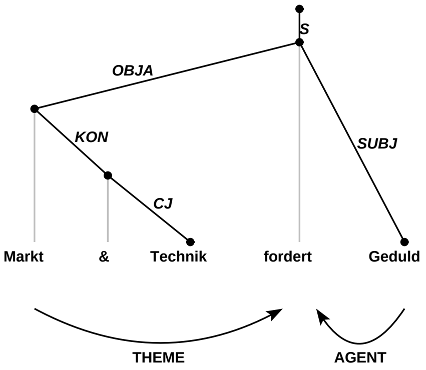


VI.2.3 Soft Integration of `OBJA`-`SUBJ` Contexts

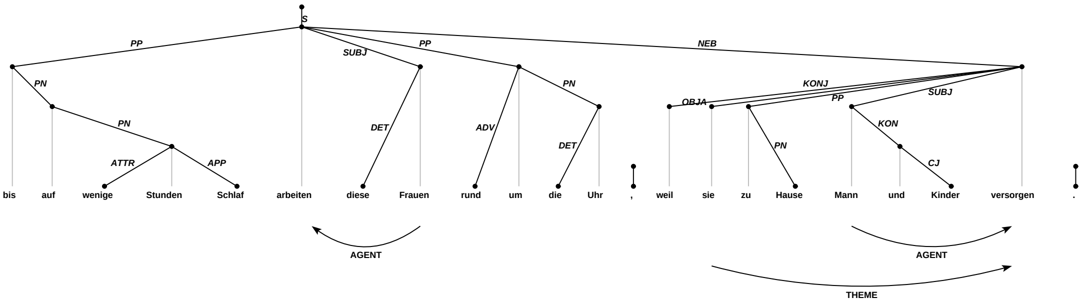
Sentence SO-360



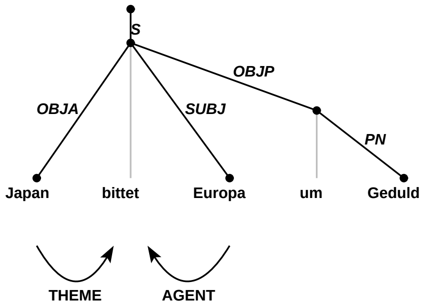
Sentence SO-706



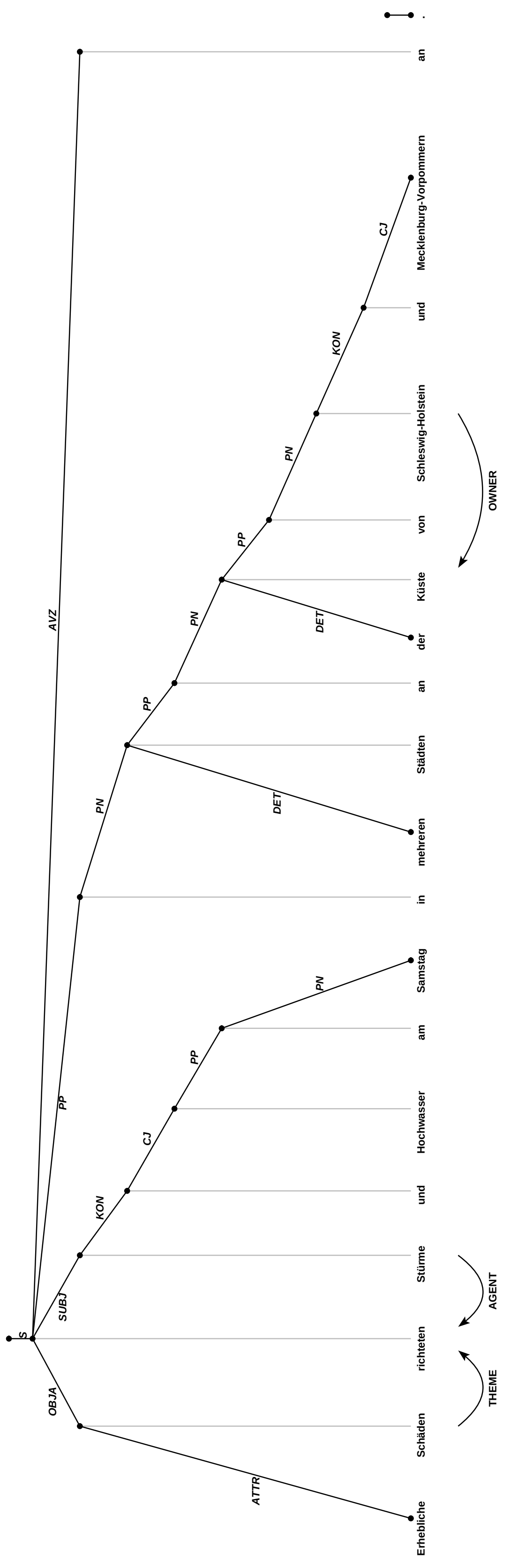
Sentence SO-841



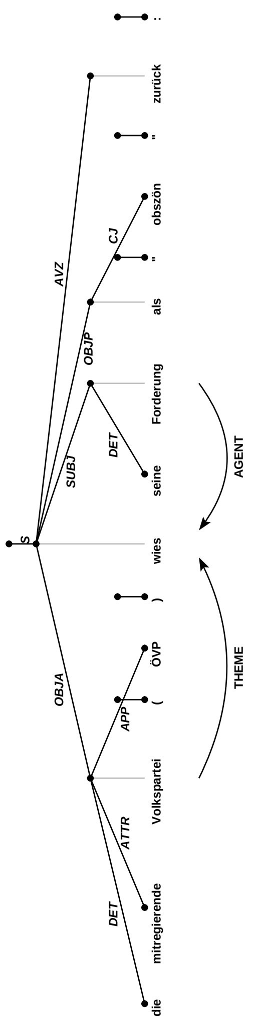
Sentence SO-1090



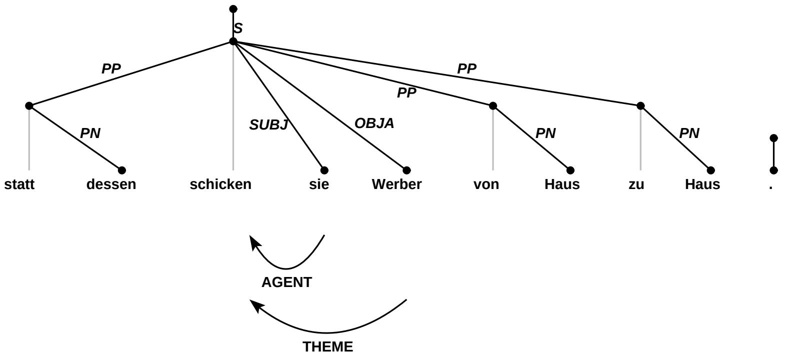
Sentence SO-4493



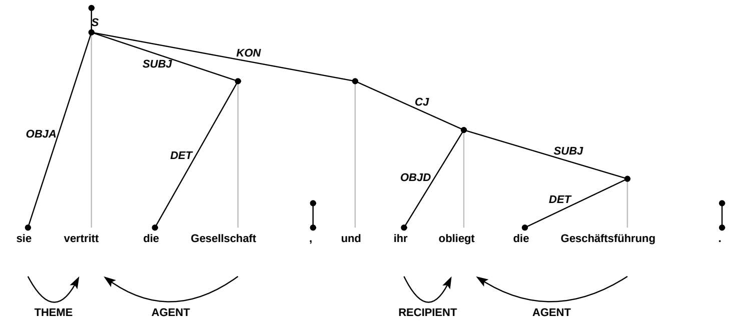
Sentence SO-6179



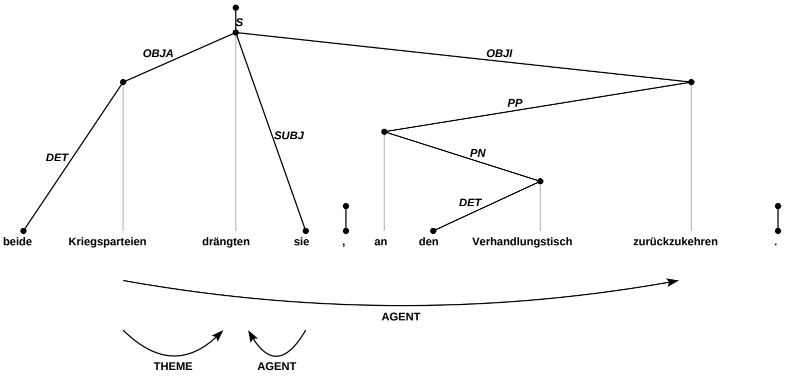
Sentence SO-9681



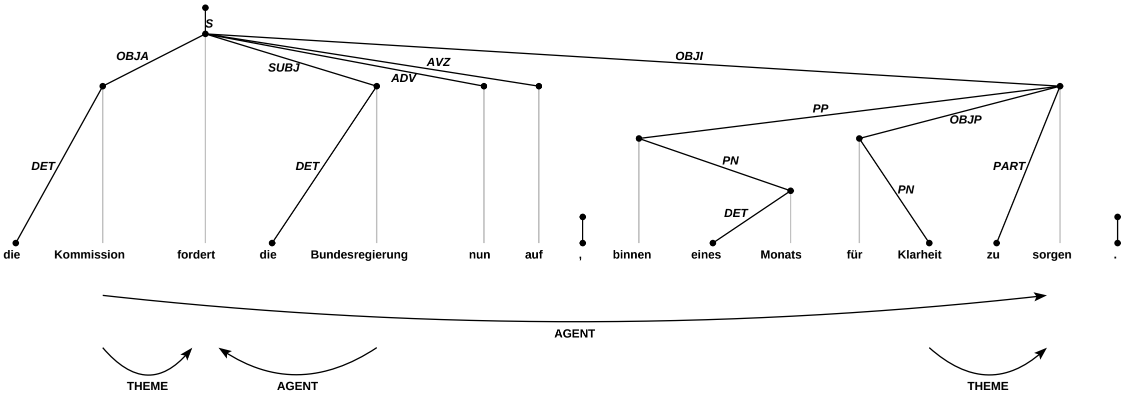
Sentence SO-9792



Sentence SO-10744



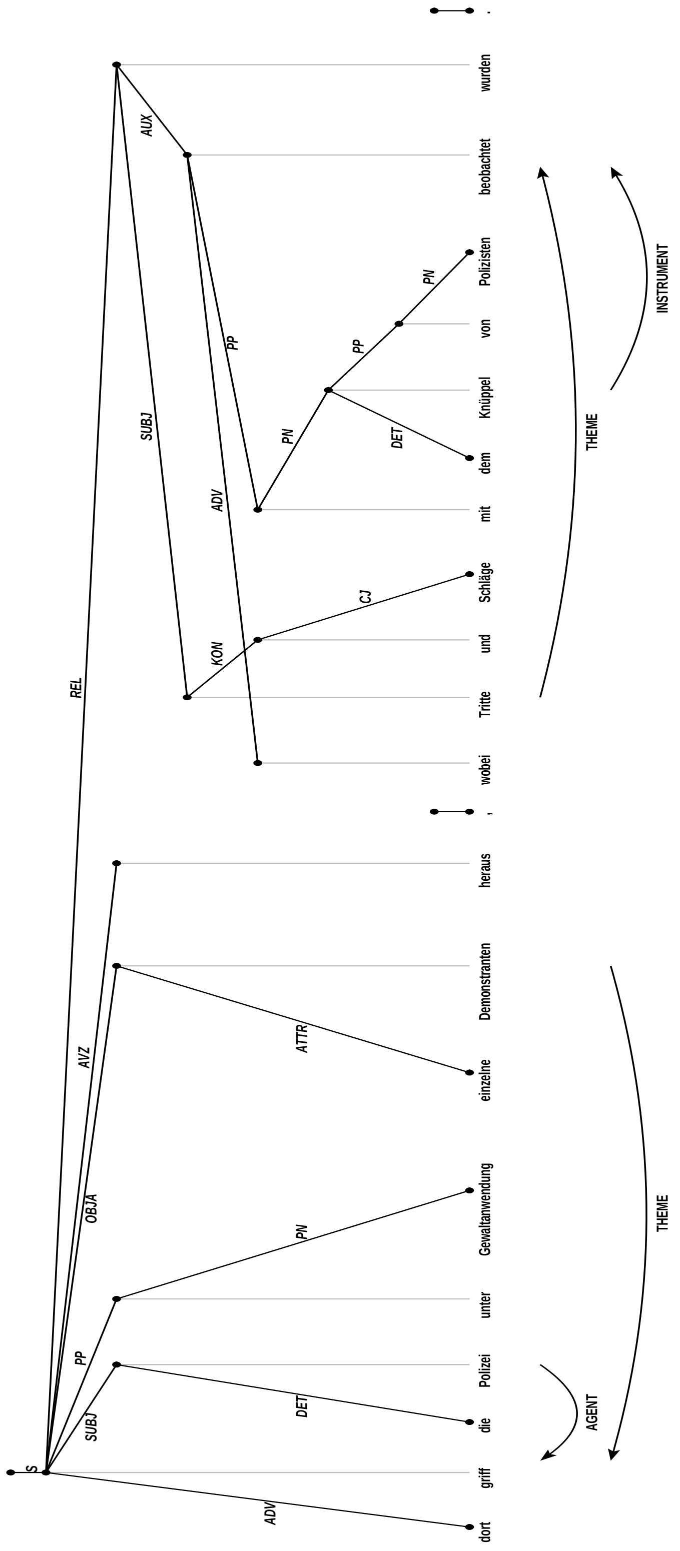
Sentence SO-40722



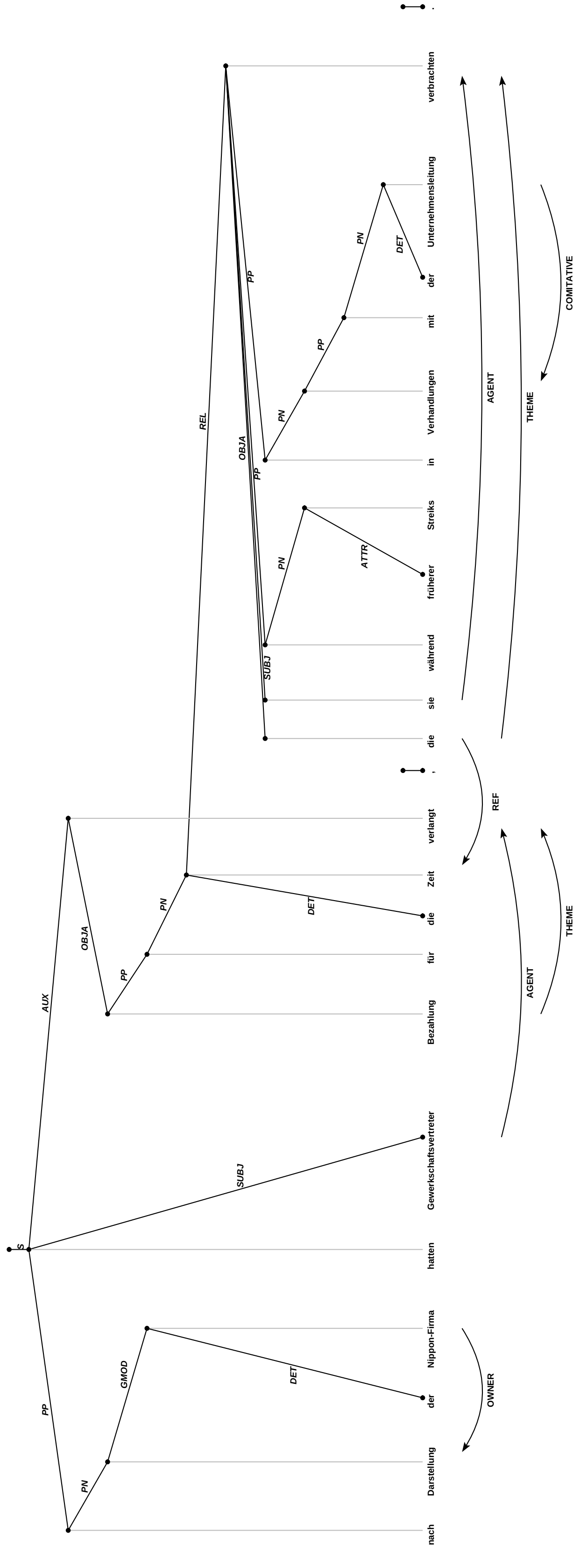
VI.3 Sentences with PP-Attachment Ambiguity

VI.3.1 Integration of an Empty Context

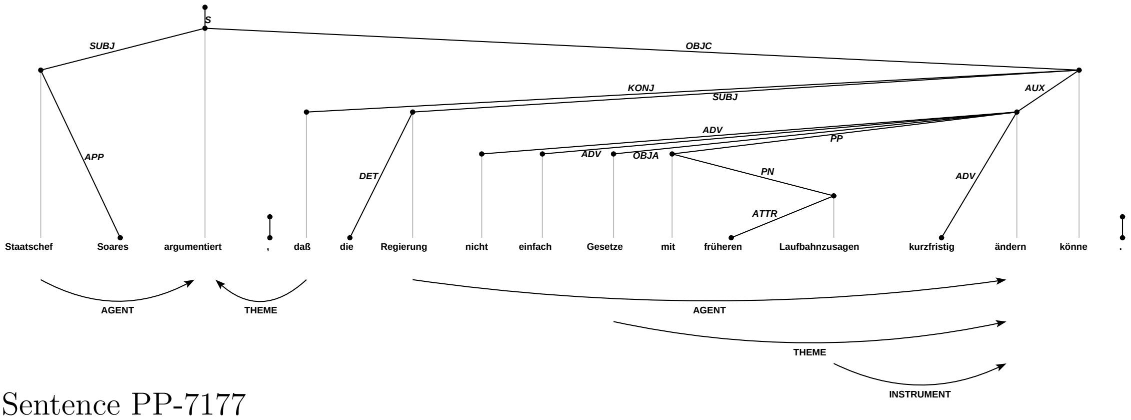
Sentence PP-3025



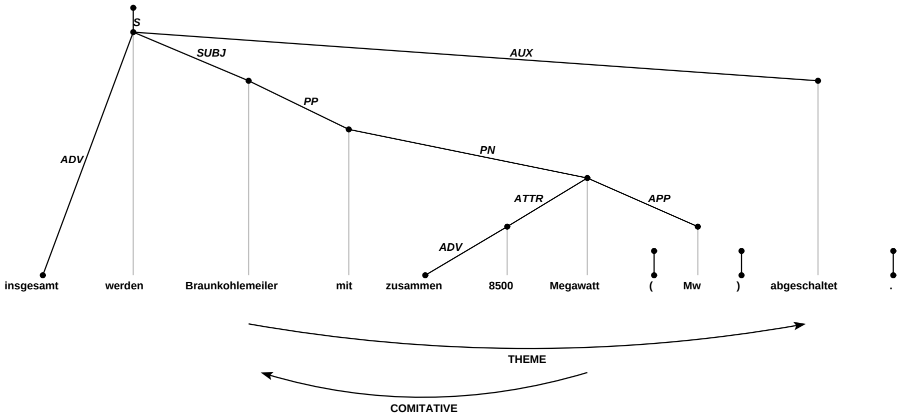
Sentence PP-3277



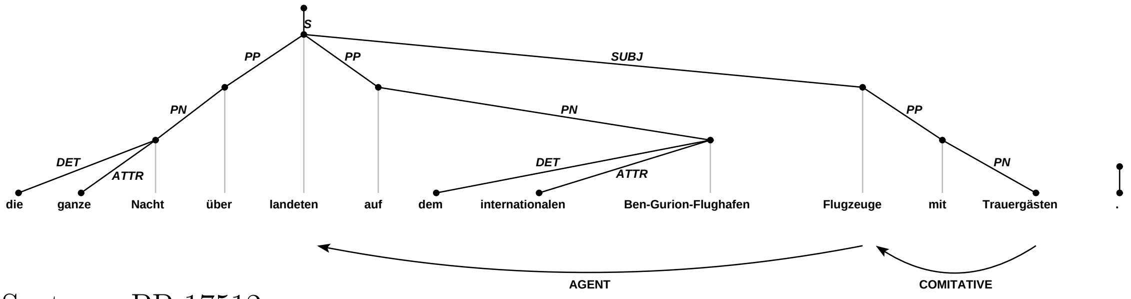
Sentence PP-3839



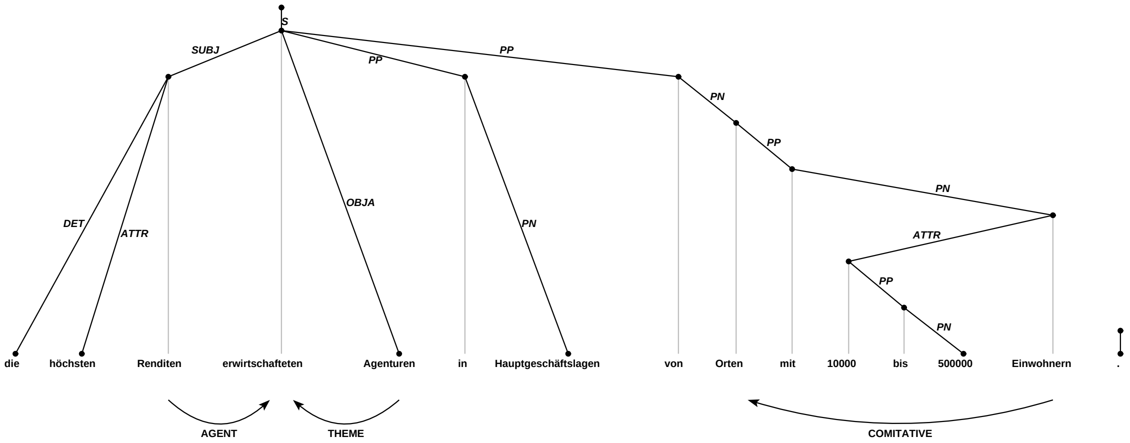
Sentence PP-7177



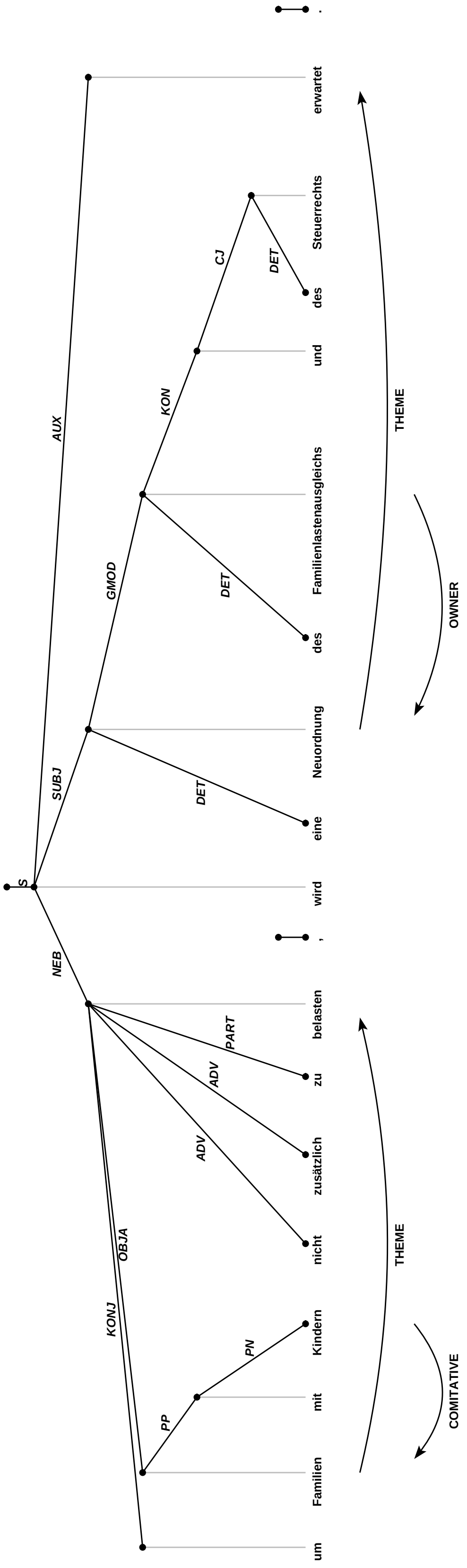
Sentence PP-7650



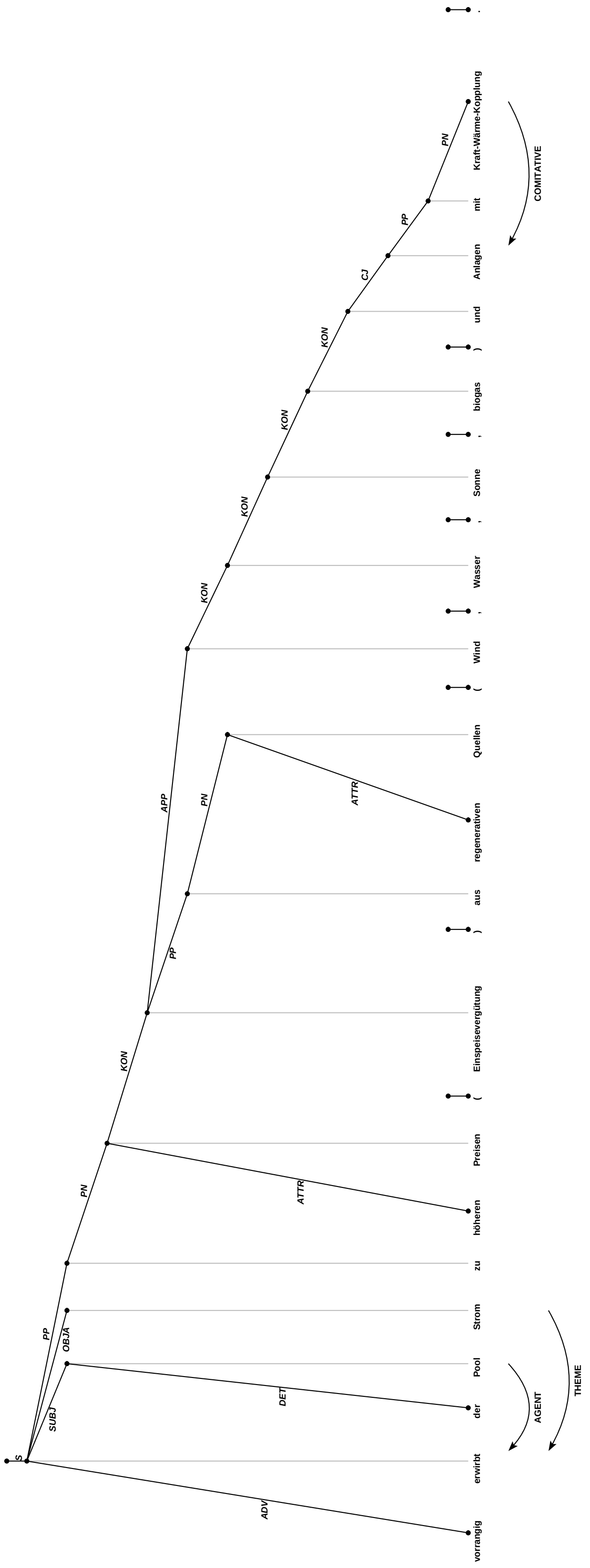
Sentence PP-17512



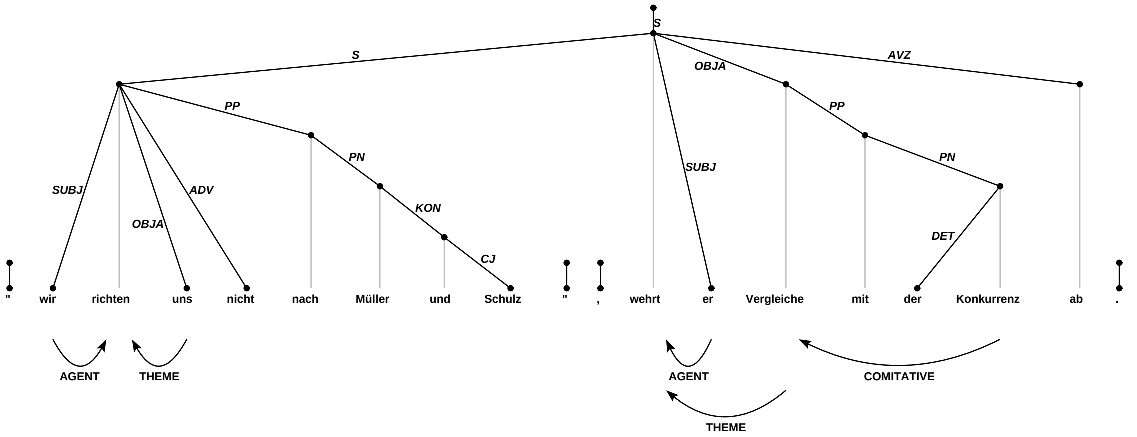
Sentence PP-19569



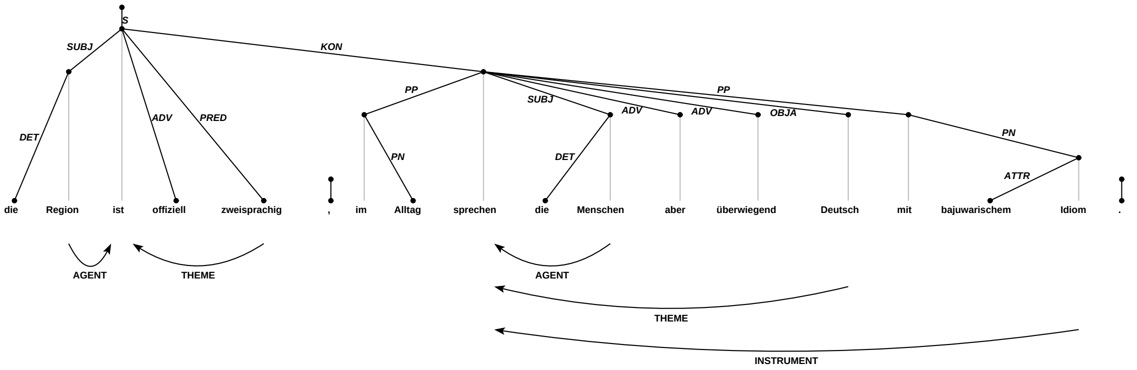
Sentence PP-23135



Sentence PP-28600

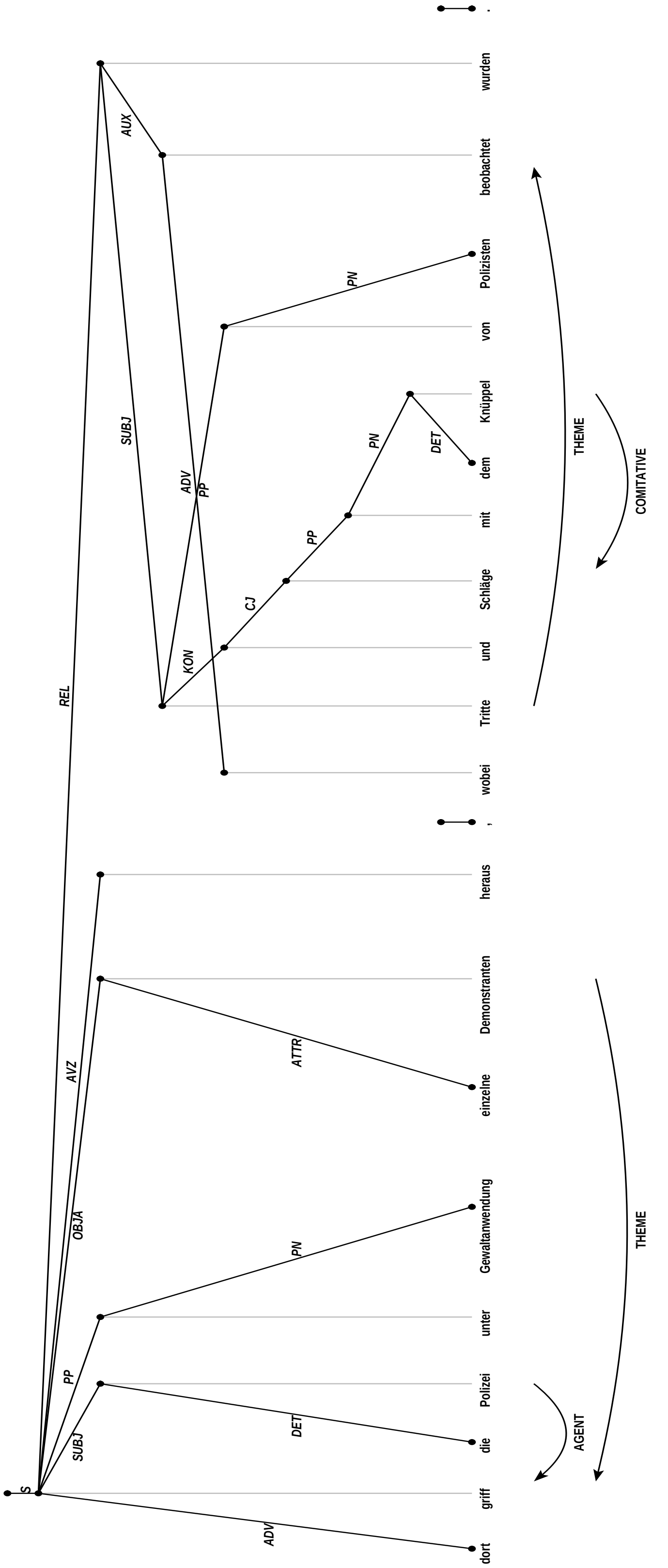

S
S
AVZ
OBJA
PP
PP
PN
PN
SUBJ
ADV
SUBJ
OBJA
KON
DET
CJ
" wir richten uns nicht nach Müller und Schulz " , wehrt er Vergleiche mit der Konkurrenz ab .
AGENT
THEME
AGENT
COMITATIVE
THEME

Sentence PP-31611

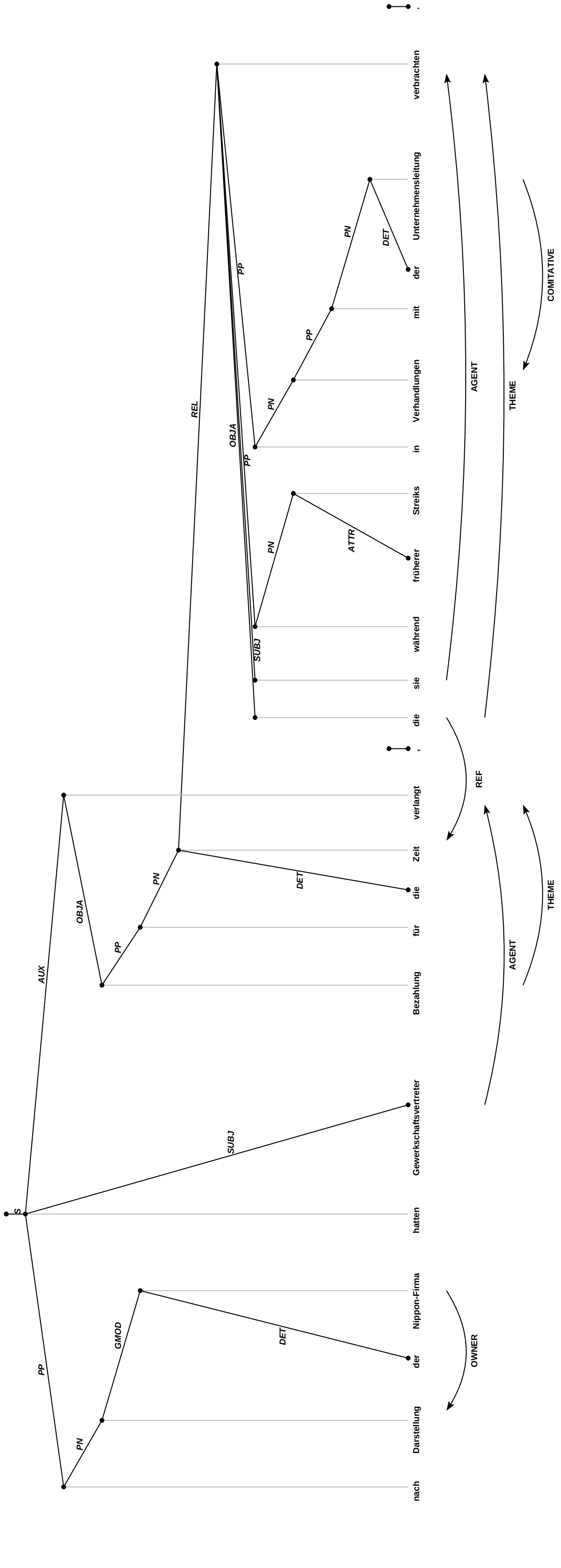

S
SUBJ
KON
PP
PP
SUBJ
ADV
ADV
OBJA
DET
ADV
PRED
PN
DET
PN
ATTR
die Region ist offiziell zweisprachig , im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom .
AGENT
THEME
AGENT
THEME
INSTRUMENT

VI.3.2 Soft Integration of Generalised Comitative Contexts

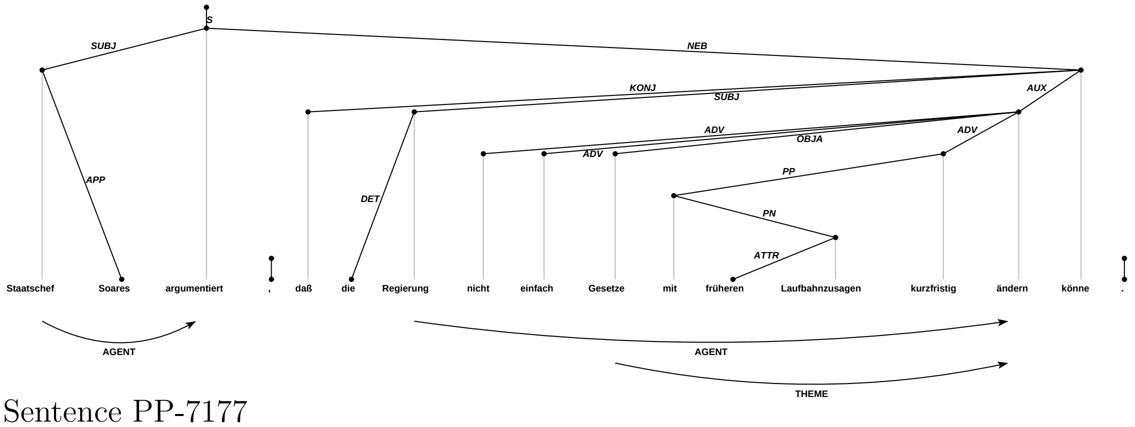
Sentence PP-3025



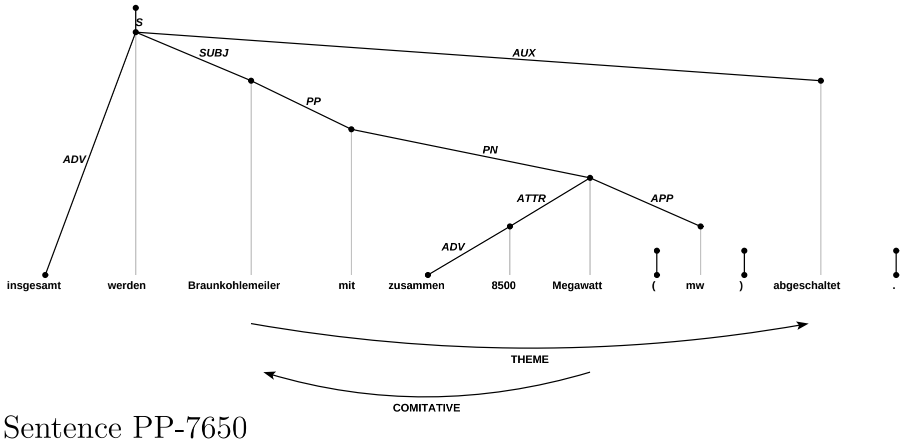
Sentence PP-3277



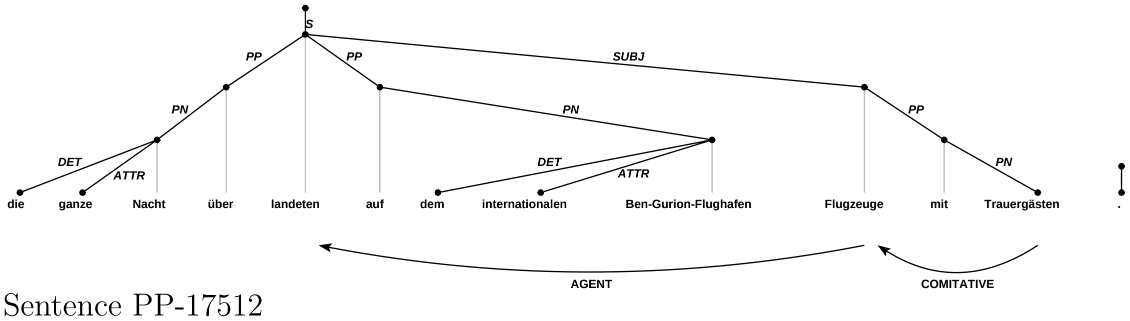
Sentence PP-3839



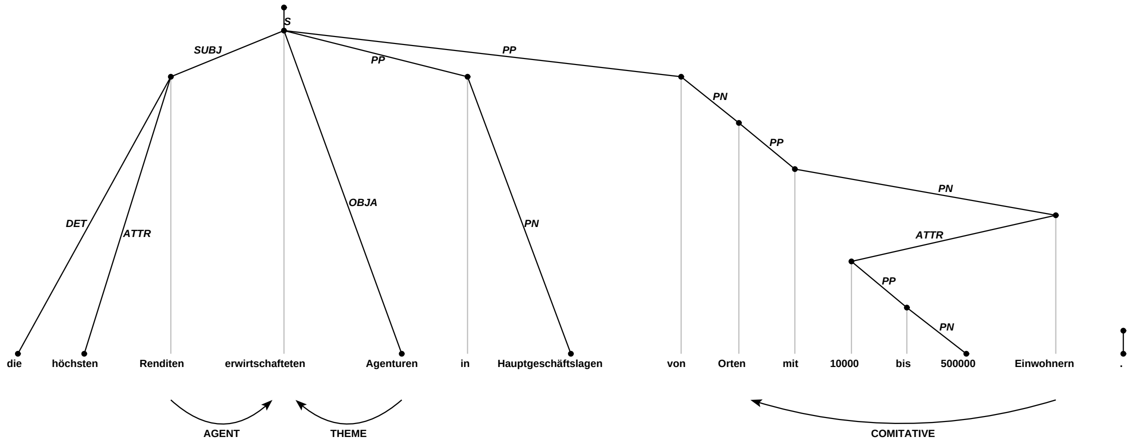
Sentence PP-7177



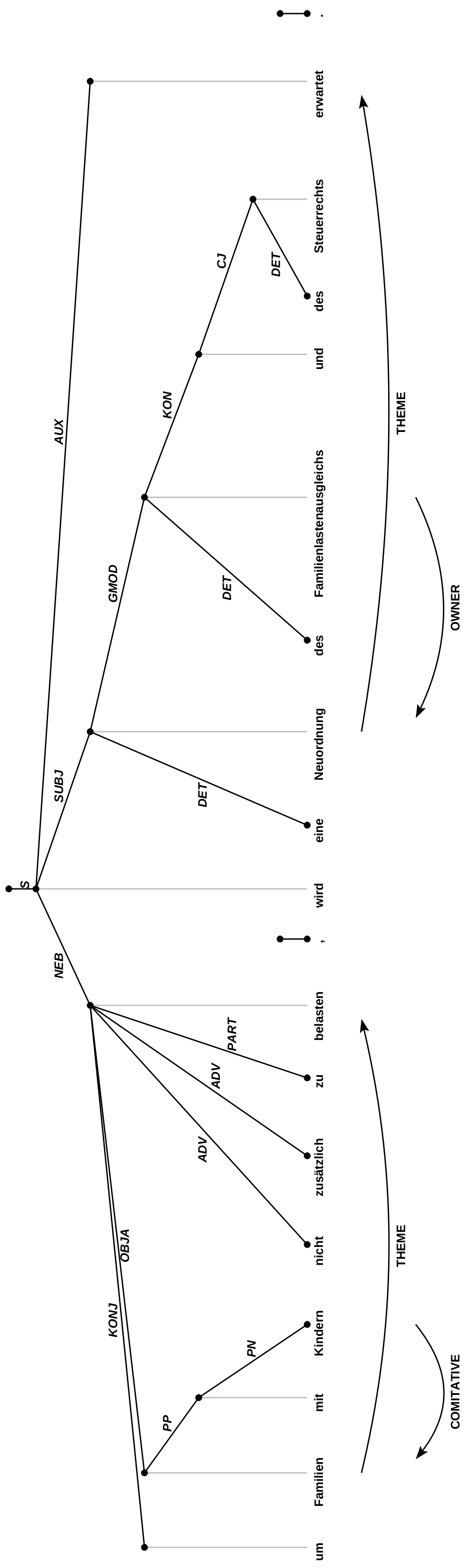
Sentence PP-7650



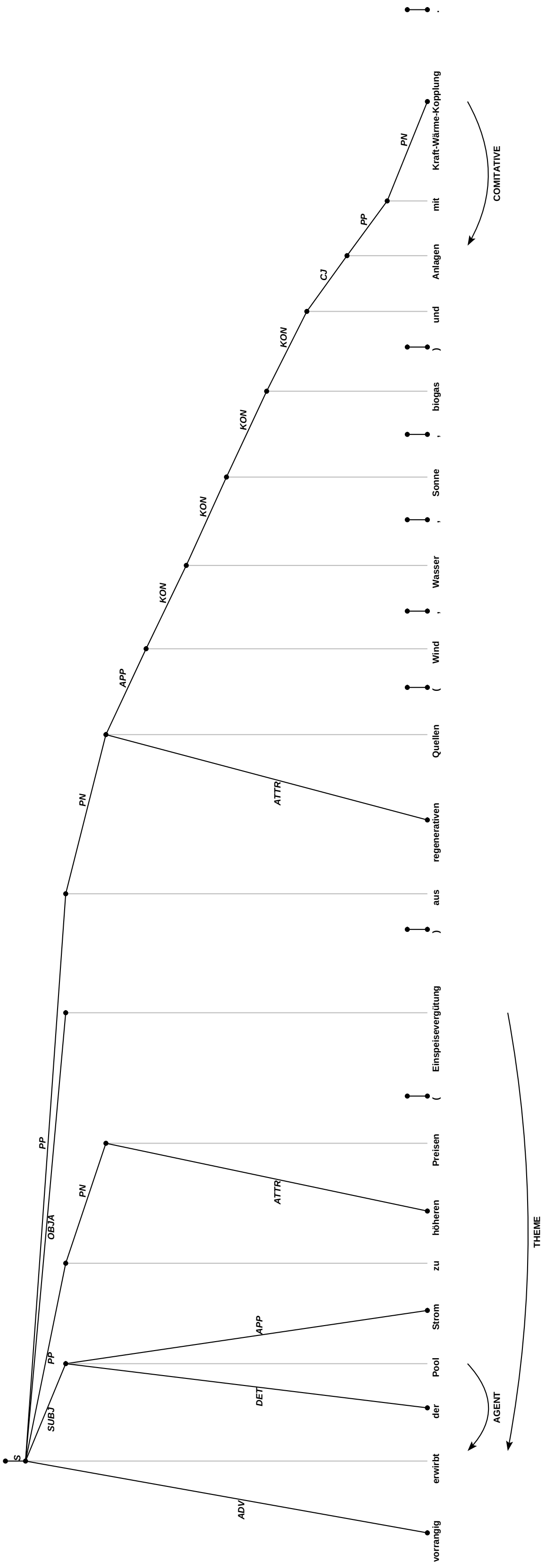
Sentence PP-17512



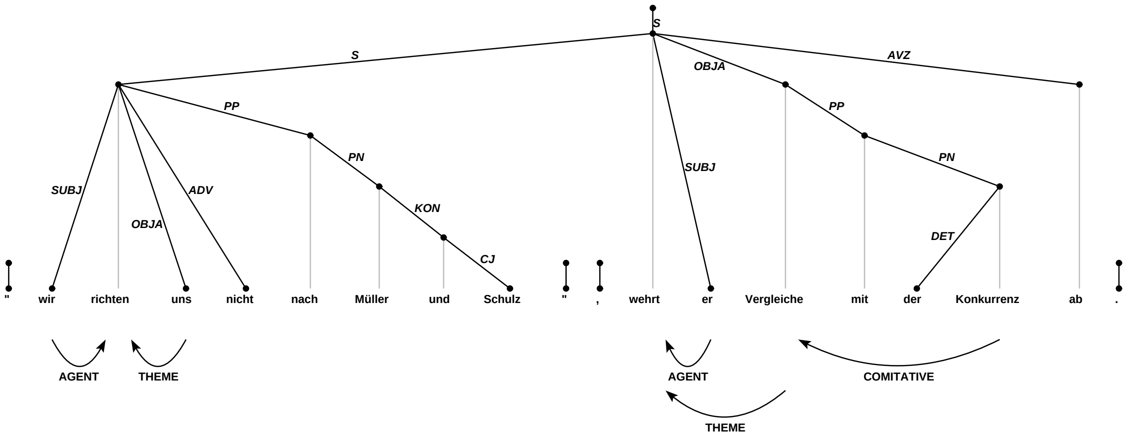
Sentence PP-19569



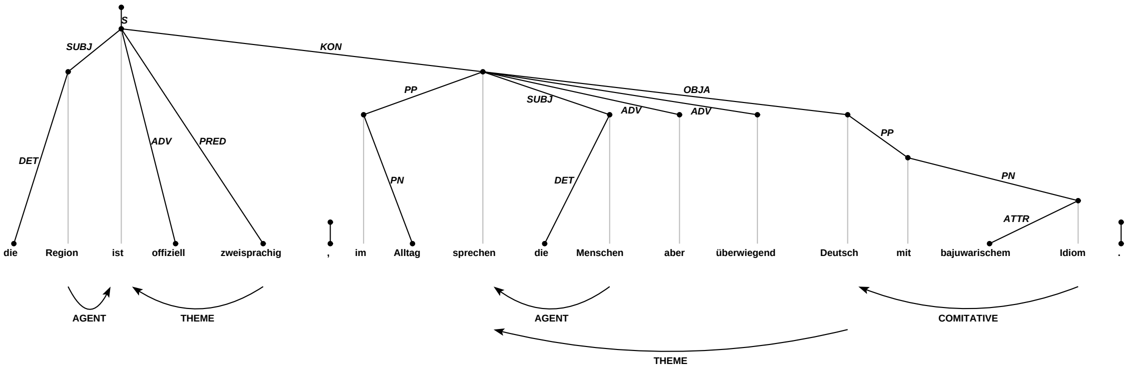
Sentence PP-23135



Sentence PP-28600

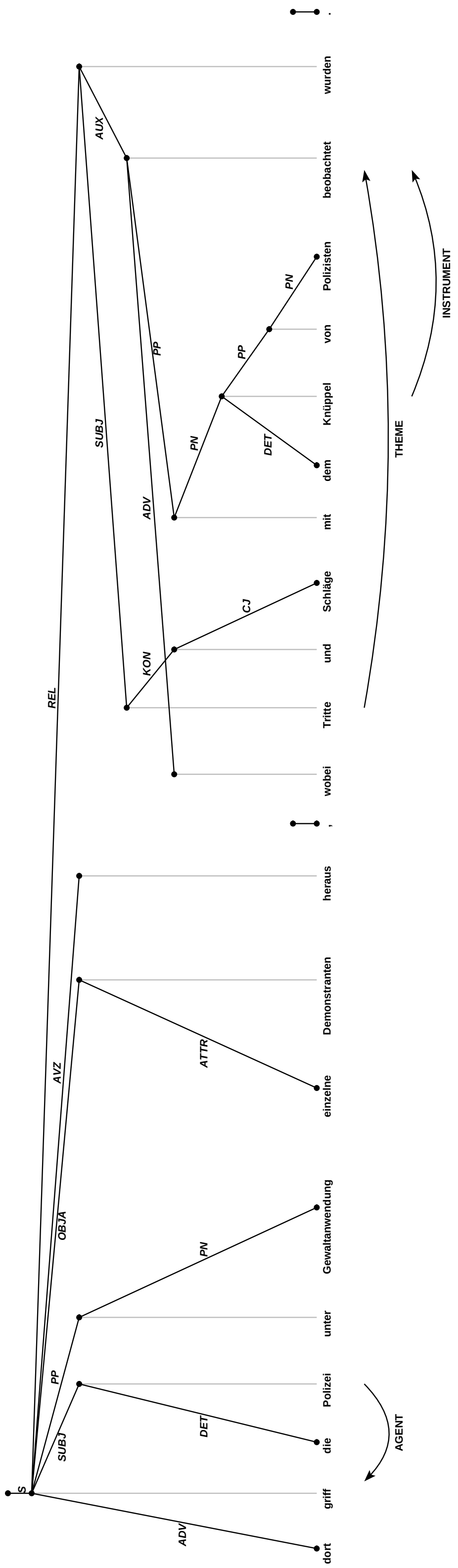

S
S
PP
OBJA
AVZ
SUBJ
OBJA
ADV
PN
KON
CJ
SUBJ
PP
PN
DET
" wir richten uns nicht nach Müller und Schulz " , wehrt er Vergleiche mit der Konkurrenz ab .
AGENT
THEME
AGENT
COMITATIVE
THEME

Sentence PP-31611

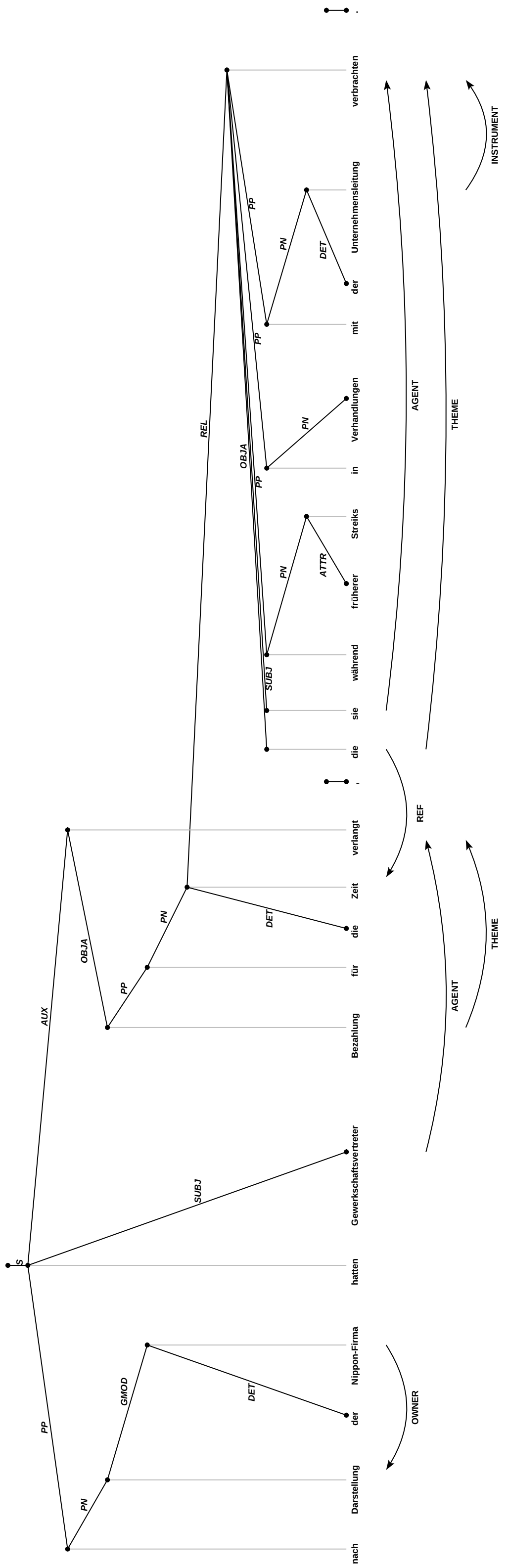

S
SUBJ
KON
DET
ADV
PRED
PP
SUBJ
ADV
ADV
OBJA
PN
DET
PP
PN
ATTR
die Region ist offiziell zweisprachig , im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom .
AGENT
THEME
AGENT
COMITATIVE
THEME

VI.3.3 Soft Integration of Generalised Instrument Contexts

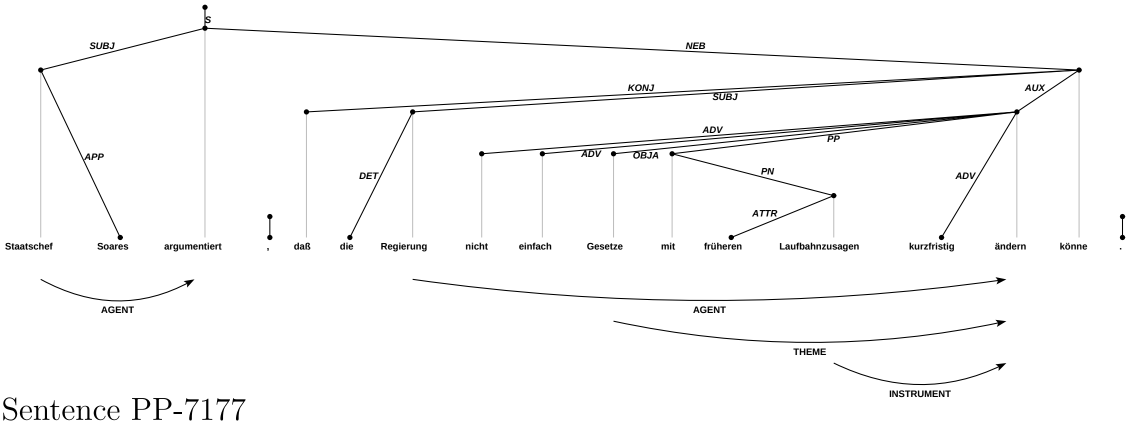
Sentence PP-3025



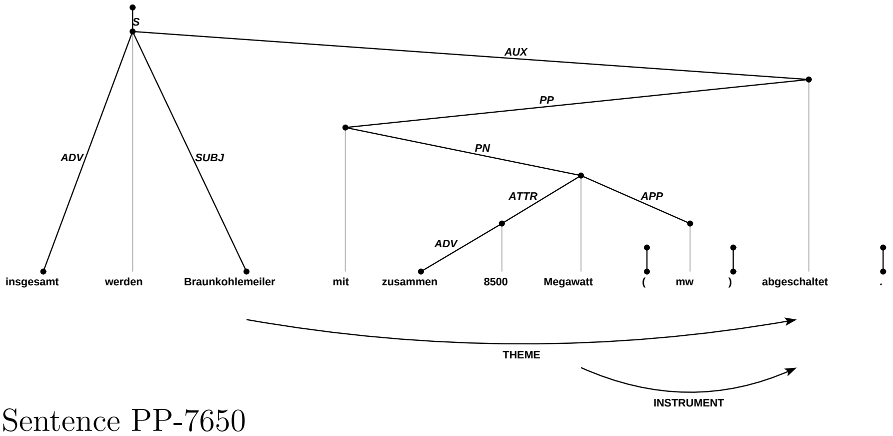
Sentence PP-3277



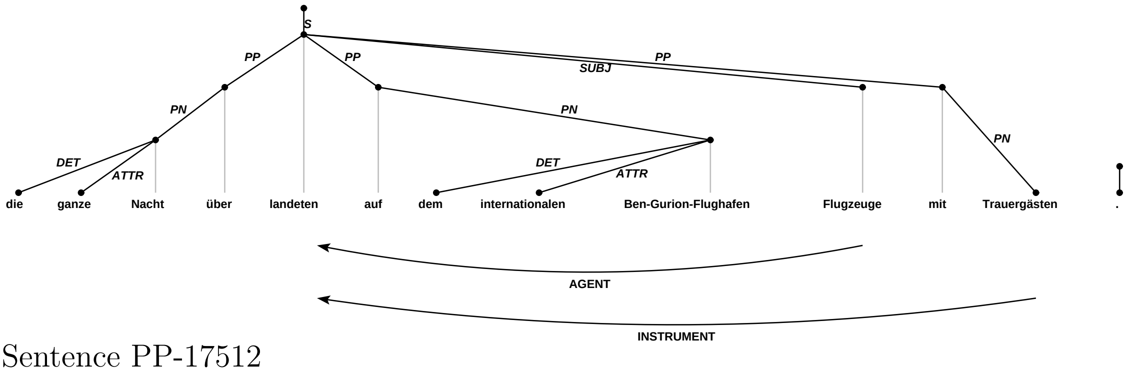
Sentence PP-3839



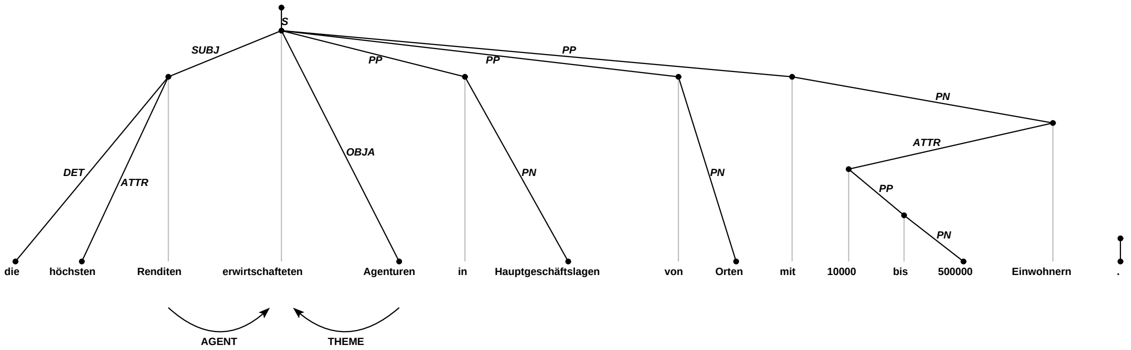
Sentence PP-7177



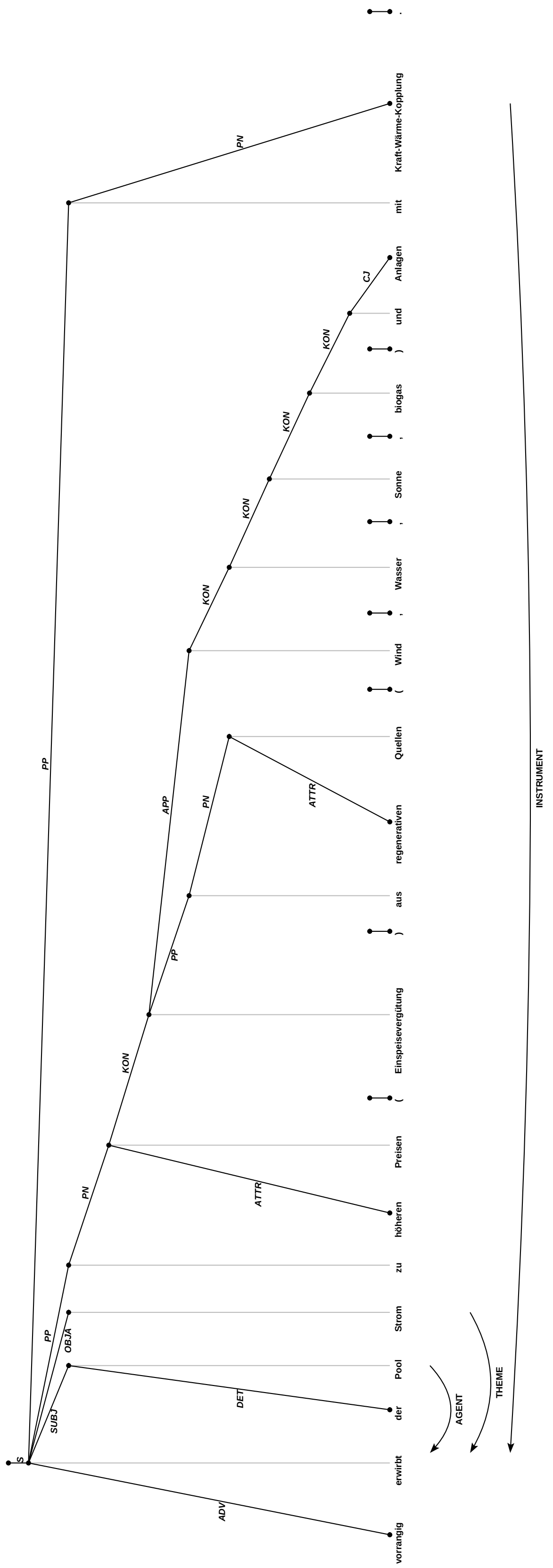
Sentence PP-7650



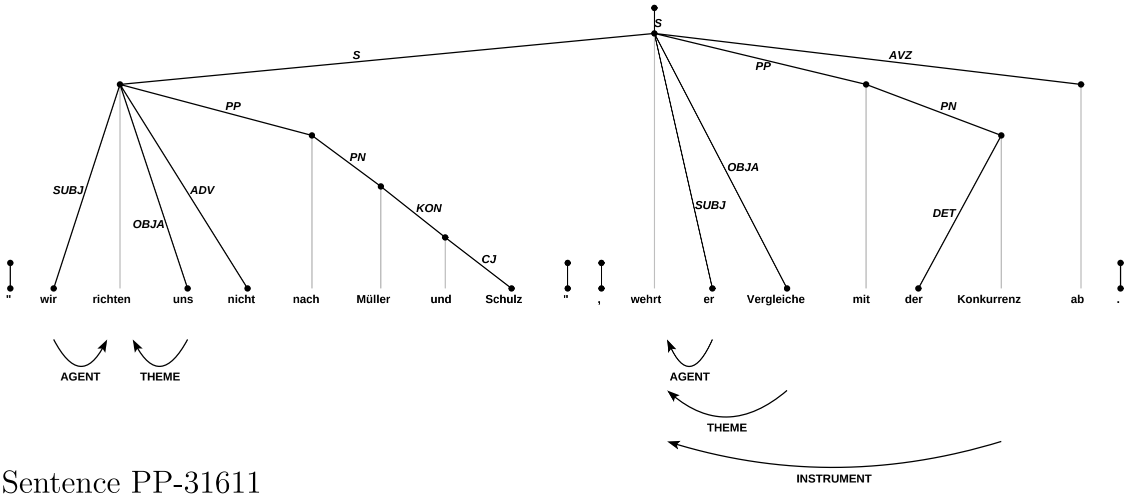
Sentence PP-17512



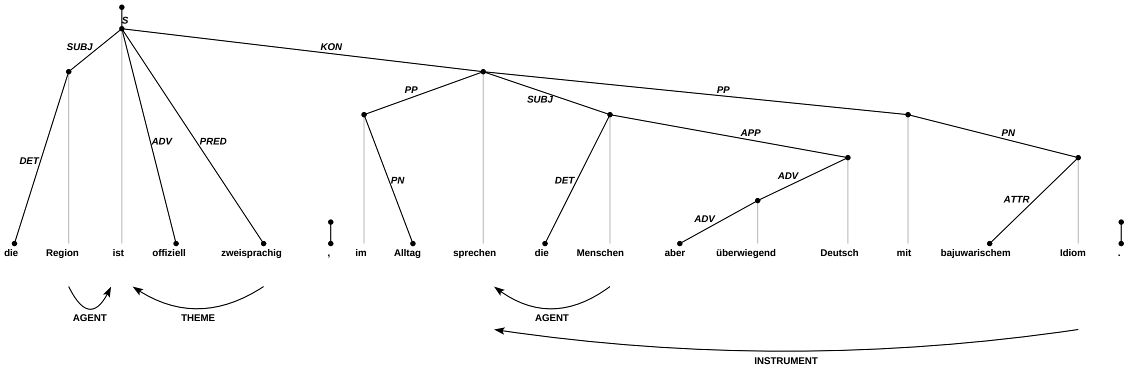
Sentence PP-23135



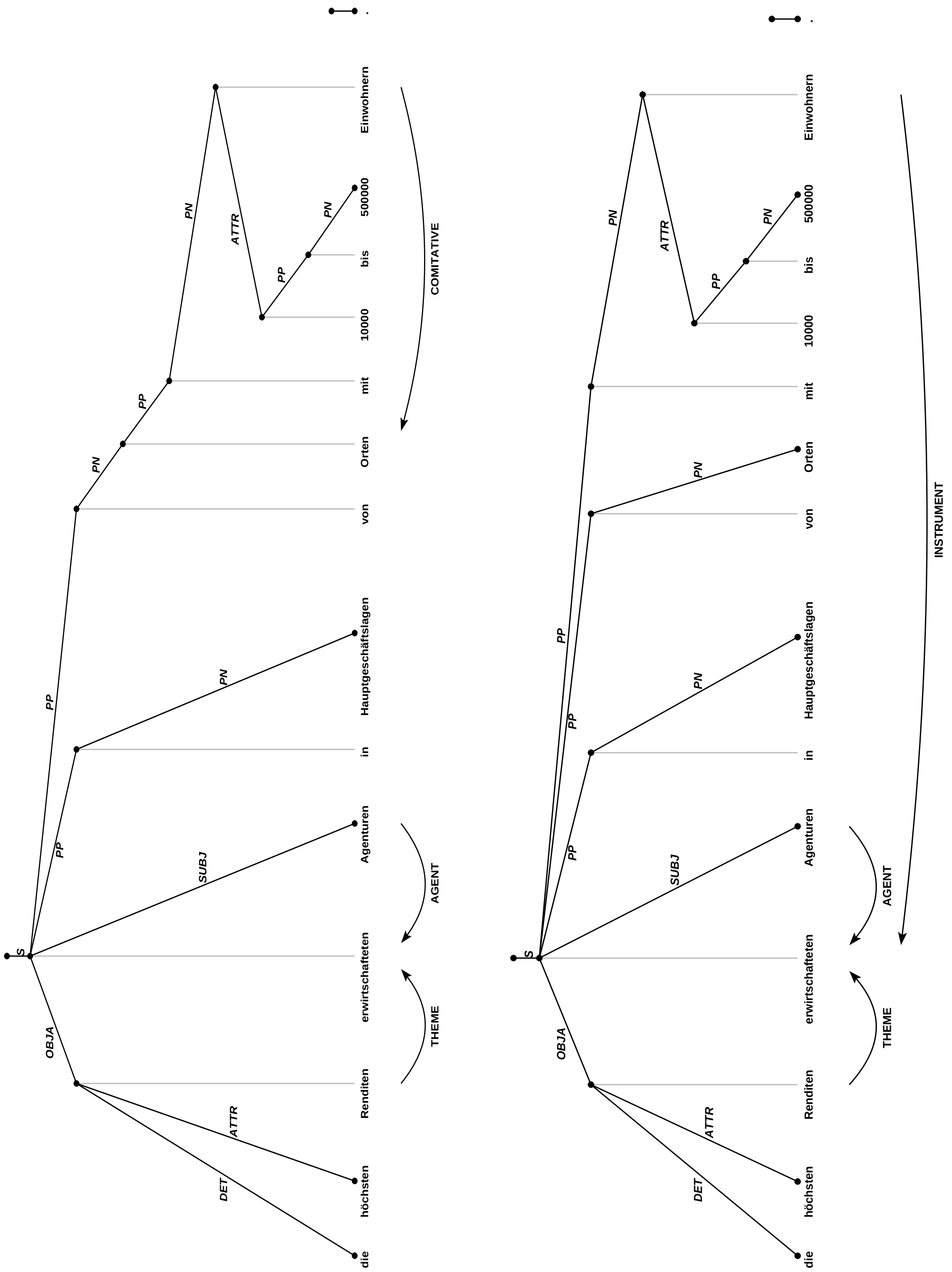
Sentence PP-28600



Sentence PP-31611



VI.3.4 Soft Integration of Augmented Contexts for PP-17512



Augmented `Comitative` Context

Augmented `Instrument` Context

VII Experimental Data

VII.1 Experiment 2

Sentence ID	Average Processig Time [in sec]				
	Empty	Binary	Ternary	$\frac{\text{Binary}}{\text{Empty}}$	$\frac{\text{Ternary}}{\text{Empty}}$
VK-011	12.067	20.586	16.630	1.706	1.378
VK-100	14.107	18.930	13.714	1.342	0.972
VK-111	10.723	17.031	11.963	1.588	1.116
VK-151	14.560	21.669	14.518	1.488	0.997
VK-226	12.561	17.641	14.375	1.404	1.144
VK-233	11.132	16.184	12.669	1.454	1.138
VK-247	10.472	14.053	10.988	1.342	1.049
VK-263	13.150	21.027	16.307	1.599	1.240
VK-274	11.921	19.123	13.595	1.604	1.140
VK-306	11.089	19.297	14.535	1.740	1.311

Table 12: Average processing times in seconds for hard context integration.

Sentence ID	Number of Structural Candidates				
	Empty	Binary	$\frac{\text{Binary}}{\text{Empty}}$	Ternary	$\frac{\text{Ternary}}{\text{Empty}}$
VK-011	$1.001 \cdot 10^{+31}$	$1.019 \cdot 10^{+21}$	$1.018 \cdot 10^{-10}$	$1.019 \cdot 10^{+21}$	$1.018 \cdot 10^{-10}$
VK-100	$3.116 \cdot 10^{+27}$	$4.951 \cdot 10^{+18}$	$1.589 \cdot 10^{-09}$	$3.300 \cdot 10^{+18}$	$1.059 \cdot 10^{-09}$
VK-111	$4.844 \cdot 10^{+29}$	$1.478 \cdot 10^{+20}$	$3.051 \cdot 10^{-10}$	$9.856 \cdot 10^{+19}$	$2.035 \cdot 10^{-10}$
VK-151	$1.168 \cdot 10^{+28}$	$1.855 \cdot 10^{+19}$	$1.588 \cdot 10^{-09}$	$1.237 \cdot 10^{+19}$	$1.059 \cdot 10^{-09}$
VK-226	$2.760 \cdot 10^{+28}$	$8.082 \cdot 10^{+18}$	$2.928 \cdot 10^{-10}$	$3.592 \cdot 10^{+18}$	$1.301 \cdot 10^{-10}$
VK-233	$1.580 \cdot 10^{+27}$	$8.369 \cdot 10^{+17}$	$5.297 \cdot 10^{-10}$	$8.369 \cdot 10^{+17}$	$5.297 \cdot 10^{-10}$
VK-247	$5.560 \cdot 10^{+26}$	$2.945 \cdot 10^{+17}$	$5.297 \cdot 10^{-10}$	$2.945 \cdot 10^{+17}$	$5.297 \cdot 10^{-10}$
VK-263	$4.300 \cdot 10^{+29}$	$4.374 \cdot 10^{+19}$	$1.017 \cdot 10^{-10}$	$4.374 \cdot 10^{+19}$	$1.017 \cdot 10^{-10}$
VK-274	$7.220 \cdot 10^{+30}$	$4.421 \cdot 10^{+20}$	$6.123 \cdot 10^{-11}$	$2.948 \cdot 10^{+20}$	$4.083 \cdot 10^{-11}$
VK-306	$4.300 \cdot 10^{+29}$	$4.374 \cdot 10^{+19}$	$1.017 \cdot 10^{-10}$	$4.374 \cdot 10^{+19}$	$1.017 \cdot 10^{-10}$

Table 13: The number of structural candidates prior to frobbing for hard context integration as quoted by WCDG2.

Sentence ID	Number of Constraint Evaluations [in 10^6]					
	Empty Context		Binary Context		Ternary Context	
	Unary	Binary	Unary	Binary	Unary	Binary
VK-011	17.427	5.919	19.958	12.340	18.595	9.404
VK-100	16.584	8.714	17.721	12.494	15.910	8.484
VK-111	16.849	5.060	18.965	10.374	16.928	6.490
VK-151	21.692	7.193	23.596	12.863	20.907	7.192
VK-226	19.407	6.177	20.934	10.329	19.852	7.691
VK-233	15.980	5.863	17.615	9.971	16.419	7.162
VK-247	15.188	5.688	16.369	8.676	15.309	6.198
VK-263	21.406	5.768	23.520	11.477	21.930	7.871
VK-274	17.503	5.650	19.932	11.362	17.963	7.048
VK-306	17.515	5.245	19.925	11.477	18.336	7.871

Table 14: Number of unary and binary constraint evaluations under hard context integration.

VII.2 Experiment 3

Sentence ID	Average Processig Time [in sec]				
	Empty	Binary	$\frac{\text{Binary}_{soft}}{\text{Binary}_{hard}}$	Ternary	$\frac{\text{Ternary}_{soft}}{\text{Ternary}_{hard}}$
VK-011	12.067	14.829	0.720	10.341	0.622
VK-100	14.107	15.176	0.802	12.244	0.893
VK-111	10.723	13.567	0.797	10.734	0.897
VK-151	14.560	16.559	0.764	13.006	0.896
VK-226	12.561	13.100	0.743	10.436	0.726
VK-233	11.132	12.684	0.784	8.996	0.710
VK-247	10.472	11.614	0.826	8.372	0.762
VK-263	13.150	16.990	0.808	11.932	0.732
VK-274	11.921	14.497	0.758	9.895	0.728
VK-306	11.089	15.136	0.784	10.107	0.695

Table 15: Average processing time in seconds for soft context integration and a context compliance of `0.8`.

Sentence ID	Number Structural Candidates				
	Empty	Binary	$\frac{Binary_{soft}}{Empty}$	Ternary	$\frac{Ternary_{soft}}{Empty}$
VK-011	$1.001 \cdot 10^{+31}$	$1.001 \cdot 10^{+31}$	1.000	$1.001 \cdot 10^{+31}$	1.000
VK-100	$3.116 \cdot 10^{+27}$	$3.116 \cdot 10^{+27}$	1.000	$3.116 \cdot 10^{+27}$	1.000
VK-111	$4.844 \cdot 10^{+29}$	$4.844 \cdot 10^{+29}$	1.000	$4.844 \cdot 10^{+29}$	1.000
VK-151	$1.168 \cdot 10^{+28}$	$1.168 \cdot 10^{+28}$	1.000	$1.168 \cdot 10^{+28}$	1.000
VK-226	$2.760 \cdot 10^{+28}$	$2.760 \cdot 10^{+28}$	1.000	$2.760 \cdot 10^{+28}$	1.000
VK-233	$1.580 \cdot 10^{+27}$	$1.580 \cdot 10^{+27}$	1.000	$1.580 \cdot 10^{+27}$	1.000
VK-247	$5.560 \cdot 10^{+26}$	$5.560 \cdot 10^{+26}$	1.000	$5.560 \cdot 10^{+26}$	1.000
VK-263	$4.300 \cdot 10^{+29}$	$4.300 \cdot 10^{+29}$	1.000	$4.300 \cdot 10^{+29}$	1.000
VK-274	$7.220 \cdot 10^{+30}$	$7.220 \cdot 10^{+30}$	1.000	$7.220 \cdot 10^{+30}$	1.000
VK-306	$4.300 \cdot 10^{+29}$	$4.300 \cdot 10^{+29}$	1.000	$4.300 \cdot 10^{+29}$	1.000

Table 16: The number of structural candidates prior to frobbing as quoted by WCDG2 for soft integration with a context compliance of 0.8.

	Number of Constraint Evaluations [in 10^6]					
	Empty Context		Binary Context			
Sentence ID	Unary	Binary	Unary	$\frac{\text{Unary}_{soft}}{\text{Unary}_{hard}}$	Binary	$\frac{\text{Binary}_{soft}}{\text{Binary}_{hard}}$
VK-011	17.427	5.919	18.150	0.909	7.696	0.624
VK-100	16.584	8.714	16.700	0.942	9.338	0.747
VK-111	16.849	5.060	17.881	0.943	7.361	0.710
VK-151	21.692	7.193	22.129	0.938	8.611	0.669
VK-226	19.407	6.177	19.300	0.922	6.471	0.626
VK-233	15.980	5.863	16.433	0.933	7.043	0.706
VK-247	15.188	5.688	15.519	0.948	6.565	0.757
VK-263	21.406	5.768	22.295	0.948	7.986	0.696
VK-274	17.503	5.650	18.448	0.926	7.491	0.659
VK-306	17.515	5.245	18.700	0.939	7.986	0.696

Table 17: The number of unary and binary constraint evaluations under soft integration of empty and binary contexts (context compliance = 0.8).

	Number of Constraint Evaluations [in 10^6]			
	Ternary Context			
Sentence ID	Unary	$\frac{\text{Unary}_{soft}}{\text{Unary}_{hard}}$	Binary	$\frac{\text{Binary}_{soft}}{\text{Binary}_{hard}}$
VK-011	16.333	0.878	4.080	0.434
VK-100	15.136	0.951	7.186	0.847
VK-111	16.524	0.976	5.147	0.793
VK-151	20.437	0.978	5.607	0.780
VK-226	18.222	0.918	4.179	0.543
VK-233	14.818	0.902	3.894	0.544
VK-247	14.021	0.916	3.708	0.598
VK-263	20.309	0.926	3.911	0.497
VK-274	16.474	0.917	3.693	0.524
VK-306	16.715	0.912	3.911	0.497

Table 18: The number of unary and binary constraint evaluations under soft integration of ternary contexts (context compliance = `0.8`).

Index