On-Line Cross-Modal Context Integration for Natural Language Parsing

Dissertation

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

eingereicht beim Fach-Promotionsausschuss Informatik von

Christopher Baumgärtner
aus Hamburg

Januar 2013

Gutachter:
Prof. Dr. Wolfgang Menzel
Prof. Dr. Ralf Möller
Hamburg, den 02.07.2013 (Tag der Disputation)

To my parents

Zusammenfassung

In dieser Arbeit wird ein System zur Integration visueller Information in den Prozess der Analyse deutscher Sprache vorgestellt. Dabei wird eine Repräsentation des visuellen Kontextes genutzt um die Interpretationen eines Sprache verarbeitenden Systems zu beeinflussen.

Nicht-sprachliche Informationen auf diese Weise zu benutzen ist einfach sofern offensichtlich ist, welche visuellen Informationen genutzt werden sollen um das Verständnis eines deutschen Satzes zu verbessern. Komplexe visuelle Umgebungen enthalten jedoch in der Regel auch solche Informationen, die für das Verständnis sprachlicher Eingabe irrelevant sind. Da ein verarbeitendes künstliches System den Beschränkungen durch begrenzte Ressourcen wie Rechenleistung und Speicherplatz unterliegt ist eine Verarbeitung der gesamten visuellen Eingabe unter Umständen nachteilig. Darüber hinaus kann eine Menge visueller Informationen auch einander widersprechende Teile enthalten. Um unter diesen Umständen die Integration zu realisieren, sollte das System eine für das Sprachverständnis nützliche Untermenge der wahrgenommenen visuellen Kontextinformation auswählen.

Die Wahl dieser geeigneten Untermenge ist zumindest teilweise abhängig von den Ergebnissen der sprachlichen Analyse. Wann immer das sprachverarbeitende System eine Interpretation seiner Eingabe ändert, zum Beispiel durch Eingabe zusätzlicher sprachlicher Information wie neu wahrgenommenen Worten, kann dies die Auswahl der passendsten Untermenge visueller Information beeinflussen. Der Prozess der Integration wird hierdurch bi-direktional: visuelle Referenten beeinflussen die Resultate der Sprachverarbeitung, welche wiederum die Auswahl visueller Einheiten beeinflussen. Um dieses Verhalten abzubilden, orientiert sich das hier beschriebene System an der Integration visuellen Kontextes zur Sprachverarbeitung beim Menschen.

Die vorliegende Arbeit beschreibt ein Modell, das bezüglich seiner Anwendbarkeit nicht auf einen Ausschnitt der deutschen Sprache beschränkt ist, sondern die Integration visueller Informationen für alle Sätze des Deutschen ermöglicht. Um die durch die Fusion der beiden Arten von

Informationen auftretenden Effekte zu visualisieren, integrieren wir ein Modell der menschlichen visuellen Aufmerksamkeit welches den visuellen Fokus innerhalb eines Bildes aufzeigt. Die Auswahl der im Fokus liegenden Region erfolgt hierbei sowohl durch die im Bild vorhandenen Objekte, als auch durch Informationen die über die Sprache übermittelt werden. Um die Vorteile des Systems aufzuzeigen, testen wir die Verarbeitung von Sätzen, deren Interpretation ausschließlich über sprachliche Information schwierig ist. Diese Experimente zeigen die Vorteile des Systems bezüglich dreier Bereiche: Die Verbesserung der Sprachanalyse durch Integration visueller Informationen, die sprachgesteuerte Auswahl von Teilen der visuell wahrnehmbaren Umgebung und die beiderseitige Fusion von visueller und sprachlicher Information zur Steuerung der visuellen Aufmerksamkeit.

Abstract

In our work we present a system for the integration of visual information into the process of parsing natural language sentences of German. A high-level representation of visual context knowledge is used to guide the interpretations of the parser to an analysis of its input that conforms with the provided descriptions of visually perceived surroundings.

Using non-linguistic information in this way is easy whenever it is evident which parts of the information is applicable for a useful integration that improves understanding of a given sentence. In a complex environment however useless information, not relevant for a given sentence, is perceived as well as useful information. As the processing capability of every system is limited, processing every visual entity available in the representation is undesirable as it consumes resources best to be used for other, possibly more important tasks. Furthermore, some information present in context might contradict other visual information, leading to conflicting influence on the interpretation of language input. The challenge thus is to choose from all available context knowledge the subset of entities that is useful for parsing the input sentence.

This choice of an applicable subset of information is in part dependent on the result of language processing. Whenever the language processor changes an interpretation, either by incorporating additional language information such as newly perceived words and phrases or by changing the current analysis of a sentence, the set of visual entities that is applicable might change. Thus, the process of integrating the information becomes bi-directional: Visual referents influence the outcome of parsing which will have an effect on the choice of applicable visual units. An implementation of this behavior can benefit from aspects of context integration found in humans. Therefore, we adopt findings in this area for the design of our system.

The thesis at hand will describe a model that works on a wide range of German utterances using descriptions of visual context. To visualize the effects of language on visual reference resolution, a model of human visual attention is integrated, showing the current focus of attention as well as a saliency landscape, influenced by bottom-up visual features of a given picture as well as top-down information received from the language channel.

To show the benefits of the developed model, an implementation is applied to several problems of natural language processing in different visual contexts. The experiments performed show the benefits of integration of language and vision with regard to three areas: Integration of visual context to improve the processing of sentences of German, integration of semantic information derived from language input that is used to find the correct referent in the description of visual context and on-line integration of visual and language information to improve the decision for a focus of attention in the visual field at each point in time.

The system thus uses information from both language and vision and integrates them in a bi-directional way, thereby solving problems arising in each modality during processing.

Acknowledgements

I would like to express my very great thanks to my supervisor Wolfgang Menzel without whose constant support and encouragement this thesis would not have been possible. His tremendous knowledge and insightful explanations were a great help for me during my time as a graduate student.

Special thanks are also in due to my second supervisor Ralf Möller from the Hamburg University of Technology (TUHH) for accepting the task of reviewing this thesis on such short notice.

Many thanks go out to the members of the CINACS Graduate Research Group for providing me with different views on the topics of cross-modality. I would especially like to thank Jianwei Zhang for making this work possible as well as Christopher Habel and Carola Eschenbach for their advice on my work.

I am greatly indebted to my many colleagues in the CINACS group and at the departments of NatS, WSV and TAMS, who supported me by giving advice, encouraged me whenever I was on the brink of giving up and shared the food of our cafeteria (which sometimes was the hardest part of all). It is because of you that I enjoyed my years as a researcher.

I am particularly grateful for the assistance given by the staff of the University of Hamburg especially Tatjana 'Lu' Tetsis and Hildegard Westermann. A special character is required to put up with people who are great at creating complicated models of the world, but are at a total loss at almost everything else (also known as researchers).

My special thanks are extended to Kirsten Albrecht who did an amazing job reading the last draft of this thesis and who patiently listened to all complaints I had over the years and to my student assistant Rörd Hinrichsen, whose exceptional programming skills were a tremendous help.

Last, but not least, my deepest gratitude goes to my parents who supported me in everything. Words can not express how much you helped me during my whole life.

Contents

List of Figures

List of Tables

1 Introduction

Giving machines at least a semblance of human capabilities to interact with and reason about the world is the goal of the field of artificial intelligence. The traditional approach to attain this goal is to divide the area of AI into a large number of sub-problems. This approach gave rise to the formation of the various fields in the area of AI like reasoning, knowledge representation, learning, perception and natural language processing. Working and researching in one of these fields is not without consequence on the view a person has on the whole field of AI. Ask two researchers from the sub-fields of knowledge representation and computer vision how they would approach the problem of building an agent that interacts with an environment perceived by a camera and you are likely to witness a very long discussion about probable solutions and the best application of a wide array of tools.

From a technical viewpoint, this diversity of opinions might not be considered to be that much of a problem. After all, if a field specific method to tackle a given problem solves it in a way that is deemed sufficient for an application in that field, it might not be of interest for the practical engineer to employ a method invented to approach a similar problem in another area of AI. As all these methods are means to give machines human-like abilities, the question arises, if a person is actually able to show intelligent behavior in one area, while being completely incapable to show it in any other area, as is the tacit assumption when modeling human behavior exclusively with regard to a specific field. To pose some more specific questions in this line: Can a human process language without studying its environment for a time in order to learn that language? Is a human capable of understanding language referring to visual properties of its environment without actually having common sense knowledge about said environment? Is it possible for a person to reason about information from the visual field without being able to update this information by perceptual processes?

A researcher in any of the subfields will probably argue that he kept this dependency in mind and works on the assumption that every resource his model depends on to be functional is facilitated by a model of another subfield. To give an example, a reasoning system for spatial relations is of course only viable if information about spatial relations in the real world is made available to it. This information in turn must be procured by a system with access to sensors, like a camera, and the ability to extract information from this sensory input. But, to continue the argument, as long as the reasoning model clearly states what kind of information it needs in which form, the reasoning model does not have to include any assumption about the functionality of the perceptive system.

This rationale is perfectly sound under one important assumption: that the operational details of one model are absolutely independent of the workings of the other model. As soon as the processes of one model influence the outcome of the other, the situation has completely changed. As long as the reasoning and perceptual models of our example are working independently with the exception of exchanging the final results of the respective processes, the specific workings of one model are of no interest to the other. If, on the other hand, reasoning processes about spatial relations influence how the environment is perceived and vice versa, each model has to take the intermediate processing steps of the other one into account.

This kind of interlocking behavior is what we find when examining the principles of operation of the human cognitive subsystems. With regard to the human processing systems for language and the human visual system one can observe a strong dependency of the interpretation from information received in one modality on what is observed in the other one. This is especially evident for an input with several possible interpretations, where ambiguity is resolved only after information from a different channel is received. Interpretation of the often used example sentence *The man sees the woman with the telescope.* is highly dependent on whether the listener sees a man with a telescope or a woman carrying one. The influence of language on visual processing is observable as a listener changes his visual focus when receiving language information about things he can see. When looking at a car while listening to someone describing it, the observer might change his visual focus as soon as he hears words referring to the engine, the wheel or the seats.

Early theories of the cooperation of the different subsystems [29] suggested a very loose coupling between the different subsystems as an explanation of these effects. Only after input is interpreted in one system, so it was assumed, it would influence other systems. Newer theories [93] suggest a stronger coupling between the processing modules. According to these theories, information is integrated at the earliest moment possible. As an utterance is perceived over time, each single word or even phoneme can influence how the listener interacts with his visual environment, while at the same time the visual information influences language interpretation. In this work we will investigate if and how this strong interaction of input from different modalities can be used as a model for building a hybrid language-vision system for artificial agents.

The problems of approaching the goal of interaction by assuming a human-centered stance arise from the highly difficult mapping between information units of different channels. The connection between two entities usually is highly ambiguous. A word of an input sentence can refer to several contextual objects (i.e. the word *table* while several tables are visible) and two words of a sentence might be possible candidates for the same object (in the sentence *Sheila put the book on the table after she finished reading it* the word *it* is referring to the same object as *book*). The problem becomes even more evident when the connection between entities can only be made by using human-like world-knowledge. For instance the fact that the word *furniture* is referring to an object identified by a visual processing system as a table might be obvious for a human agent, but is a complex issue for an artificial system.

Another issue is the highly dynamic nature of the content of information in both modalities. Humans do not receive language information all at once, but sentences heard are unfolding over time while the listener receives the words and his perception and interpretation is influenced by the words already heard. At the same time the visual context is changing as objects move in and out of the visual field, spatial relationships between objects change and persons engage in different actions at different points in time. Even a relatively static scene might change its semantic content over time as an observing agent re-evaluates what it is seeing, resulting in a different interpretation after some time has passed.

Another problem to be tackled is the misinterpretation of data from one or both channels. Both natural and artificial systems are prone to errors when processing sensory input which could result in words being misunderstood or objects being confused. Generating a mapping under these circumstances is a highly complex endeavor.

The language aspect of our system is modeled by the Weighted Constraint Dependency Grammar(WCDG) formalism [90]. Information about language is stated in the form of rules input into WCDG. These weighted constraints influence the result of the analysis of a sentence of the target language by evaluating logical predicates that a syntactic [90] or semantic [68] interpretation of a sentence may or may not violate. We consider these rules to be a primary candidate for integrating visual information, as they are easily adapted to integrate visual information int their evaluation.

Another important feature is WCDGs ability to process language information incrementally [69]. This feature is crucial for the task of modeling the bi-directional influence between information from both modalities at a very early point in time (i.e. as soon as a word is received and not when the whole input sentence is already obtained).

Visual context is modeled by describing objects and processes in the visual field as individuals and relations between them in a knowledge base. These visual entities are connected to a taxonomy of concepts, relating types of individuals to each other. The visual field described by these entities is extracted from pictures showing actions and participants as well as a 3d-model of objects. In order to demonstrate the effects of the bi-directional impact of one modality on the other we also integrate a model for human visual attention [47]. The attentional focus of this model changes depending on the result of the language analysis.

The complete model is then applied to input sentences in conjunction with visual information. These stimuli have already been used in experiments on human participants. The goal of this application is to show that the behavior of the system with regard to language processing, reference resolution and influence on visual attention is comparable to experimental results about the human cognitive system. Furthermore, we will also investigate how changes in contextual information and language rules influence the results of the system. Finally an application to a virtual environment will illustrate the usefulness of the implemented model with regards to human-machine interaction.

2 The Integration of Language and Vision in the Human Cognitive System

Although we are not able to build machines exhibiting the whole range of a humans capability to integrate different types of information, it is the position of this work that an artificial system for cross-modal interaction may benefit of an underlying model that is inspired by processes of natural cognitive systems. It is not our aim to build a complete model of the human processes of that interaction, but the architecture presented in this thesis is motivated by findings reviewed in this section.

Starting off from a short overview of the attentional effects of language processing in humans and how these effects led to the invention of a new paradigm which is used to investigate the link between language and contextual information from the visual channel, we will review the results of the application of the paradigm on a number of experiments showing the interaction of language and vision. We will finish the section by explaining the conclusions we derive from reviewing human cross-modal interaction with regard to the architecture of our model.

2.1 Attention and the Visual World Paradigm

Although everyone is familiar with the phenomenon of attention, it is quite hard to come up with a precise definition. In the following we use the term as the capacity of the human cognitive system to choose a subset of available input that is processed with a higher share of cognitive resources than anything else perceived. In this we follow Desimone and Duncan [24] who state about visual attention: "The first basic phenomenon is limited capacity for processing information. At any given time only a small amount of the information available on the retina can be processed and used." ([24], 193)

The problem of limited resources explains the functional role of attention in organisms like humans. Complex cognitive systems suffer from two aspects of information overload: On the one hand, the sensory apparatus is capable to absorb a vast amount of input and secondly this input can be matched to the large quantity of information stored in the memory of the organism. The mechanism of attention is required to handle this information overload in cognitive systems with tightly limited resources. The same line of thought can be applied to artificial systems: the amount of data

input into such a system can easily exceed the systems ability to compute solutions for a given problem in real-time. Therefore, attentional mechanisms of the human cognitive system that enable humans to match complex visual input with complex language input can be a model for the same kind of integration in an artificial agent.
Orienting is the process of choosing the subset of available input. Sometimes the decision to attend to specific things might be called "voluntary" (like in watching tv), at other moments our attention is "drawn" to things without us making the decision to do so (like when a person feels pain and cannot concentrate on anything else). These processes are called endogenous (or top-down) and exogenous (or bottom-up) control respectively.
Exogenous control is dependent of orienting stimuli in one or several modalities. Examples of such stimuli are bright light, loud noise or unexpected movements. Attention controlled by this mechanism is usually prone to constant shifts of the attentional focus, moving to a new object after about 100-200 msec.
The endogenous control mechanism is goal-driven based on either some cue in another modality (like hearing the name of a person, which results in looking at the person) or as the result of thinking processes (like someone wants to leave his house, tells himself not to forget the keys and starts shifting his visual attention to the place where he left them).
Although the shift in attention can be obvious to the outside (for example shifting ones head), humans can also direct their attention by covert orienting for instance when a person is listening without indicating to what source by standing still. With regard to visual attention the overt orienting explains findings about human eye-movements and the visual field. Although humans can attend to locations of the visual field currently not in focus of vision, 'looking at the corner of ones eye' so to speak, usually humans focus on visual entities they are attending to.
Filtering is the attentional aspect that results in extraction of more information from the subset attended to than to any other perception. One well know effect stemming from this filtering mechanism is the cocktail party phenomenon [17] which describes how people listen to one conversation while perceiving several others.
The question arises, how attention is distributed in the visual field. Humans can attend to a large object or scene, or they can focus their attention to a small part of it. This zooming in and out effect gave rise to the spotlight metaphor of visual attention [74] which can be adjusted to the task currently at hand. The size of the

spotlight seems to depend on the effort the task generates in the cognitive system.
One influencing factor of the aforementioned exogenously controlled attentional process is the information extracted from the language input of a listener. When people listen to spoken language, they tend to look at objects referred to by the language modality. What is now commonly referred to as the visual world paradigm was used in a large number of psycholinguistic experiments investigating this effect. The experimental setup in these studies consists of participants hearing sentences while looking at pictures displaying objects and scenes. The eye movements of the subjects are recorded in order to investigate the influence of language on human visual attention. This paradigm was successfully applied in a wide range of psycholinguistic experiments. The goal of these studies was to investigate the relationship between utterances and visually perceived scenes. The main advantage of the paradigm is the close time-lock between spoken language and eye-movements, indicating the target area of visual attention. Although there is a delay of saccadic eye-movements after the person shifts its visual attention, it is possible to investigate the time-course of attentional shifts using the paradigm.

One area of research with regard to reference resolution and integration of contextual information into language processing concerns the moment when information from one modality is integrated into processing of knowledge of the other modality. Two different points of view were prevalent. One is the modular theory [29], stating that the brain mechanisms for understanding language are encapsulated from other processing systems, leading to late integration of contextual information. According to this position the human processing systems of language and vision work mostly independent of each other. Only after information from one channel is processed to a certain degree, it is used to influence other information channels. On the other hand, early-integration theories [93] argue for a combination of modality specific information at the earliest time point possible, which therefore backs early integration theories. As soon as knowledge from one modality might be useful for understanding information in another modality, it is integrated.

In an early study [100], the visual world paradigm was applied to experimental setups where people had to find referents for ambiguous sentences using different pictures. Participants listened to sentences like *Put the apple on the towel in the box* while looking at pictures containing one or two possible referents (see Figure 1). Eye-movements of test subjects showed that the interpretation of the phrase *on the towel* as the location of the apple or the destination of the putting action depends on the context presented. In the one referent condition (containing only one apple), participants would look to the empty towel upon hearing the phrase (indicating that they thought this to be the destination of the action). In the two referent condition hearing the phrase would result in looks to the towel with an apple on it. The outcome of these experiments demonstrated that the processing of instructions was done incrementally, using visual context information as early as possible. The integration of context knowledge would lead to intermediate interpretations with regard to the understanding of an instruction and with regard to finding the correct reference for part of an utterance.

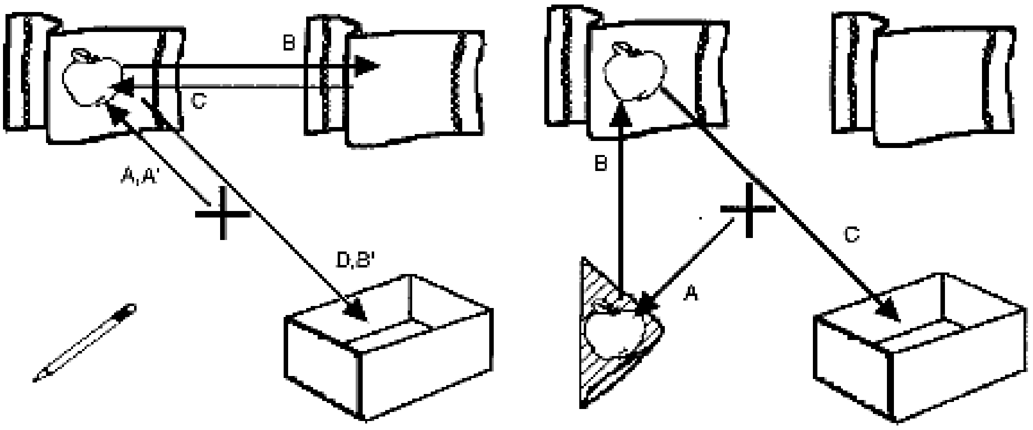


Figure 1: One referent and two referent context for the sentence *Put the apple on the towel in the box* (Taken from [100])

The analysis of the time-course of eye-movements is largely dependent on the aim of the study. It is possible to take a look at the likelihood of participants to shift their gaze to a certain region, the moment the shift is undertaken, the frequency of looks at a certain object or the time spent looking at a part of a picture. To interpret the data received from the studies, the results are averaged over participants and trials, thus producing the likelihood of gaze direction to a certain region in a certain time-window, showing the influence of context-dependent language understanding on visual attention.

2.2 The Bidirectional Dependency of Language and Vision

In this section we will discuss the relationship between the processing of language and vision in the human cognitive system. We are specifically interested in the condition under which information in the language modality will influence whether a possible referent is attended to by an observer and how knowledge derived from what an observer sees influences the understanding of language this person is hearing.
In our treatment of this topic, we discriminate between effects that language has on vision, the effects of visual information on language processing and those cases where language and visual processing are co-dependent, making the connection a bidirectional one.
That humans can establish the connection between a certain word they understand and a visual property is well known [18]. An obvious candidate for language information that is used to find referents in the context seen is word-level information that is matched to the knowledge about lexicalisations of objects (i.e. those words and phrases that are commonly used when humans speak about a specific object). [44] termed this process of matching words heard with knowledge learned as the *phonological hypothesis*: "... phonological representations are retrieved on the basis of acoustic information and visual information [..], and attentional shifts to pictures are made when there is a match between representations retrieved from the two modalities" ([44], 461)
This retrieval due to word-level information is straightforward whenever the received word and the lexicalisation of an object are the same (i.e. hearing the word *fork* while a **fork** is seen). [1] found evidence that lexical access can also happen when word and lexicalisation do not match. They instructed participants to move objects presented on a computer screen (with instructions like *Pick up the beaker*). The objects contained a referent (**beaker**) a cohort competitor with a name beginning with the same vowel (**beetle**) a rhyme competitor (**speaker**) and something unrelated (**carriage**). Observed eye-movements showed that participants were more likely to direct their visual attention to the cohort and rhyme competitor than to the unrelated object. The authors interpreted this behavior as evidence that cohorts and rhymes compete for lexical activation with those words in memory that perfectly match the input heard.
This language-to-vision connection can be made between a word and an object (like car and the vehicle seen), a word and a property of an object (hearing green and looking to a green object) and even entities in the visual field that are not perceptible

as such, but are the result of reasoning about what a person can see. One example of the last case would be the ability to connect a word with a relationship between objects that a person sees at the moment. For instance two persons speaking about a car they are looking at and one person saying *and to the right of it* resulting in the other person guiding its attention to an object (or objects) that are spatially related to the car in a way complying with the term *to the right*.

At the same time it is possible that the visual properties influence the understanding of language. Referring again to the sentence *The man sees the woman with the telescope.* the interpretation can be dependent on a visual context involving a male and a female person, one of them carrying a telescope.

In [101] participants were instructed to touch blocks like the ones depicted in (Figure 2). Sentences provided contained disambiguating information at a different word for each of the presented conditions. Hearing the instruction *Touch the starred yellow square.* the correct referent was non-ambiguous at the marking adjective (i.e. *starred*) in the early condition, the color adjective (i.e. *yellow*) in the mid condition and the noun (i.e. *square*) at the late condition.

Results showed that eye-movements of the test subjects were directed to the correct referent sooner at the early condition than in the mid condition, indicating that the rapid integration of visual context with language information would induce a change in visual attention as early as possible.

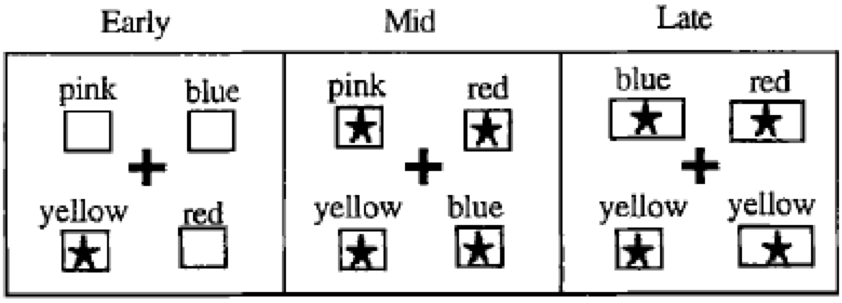


Figure 2: Ambiguous visual information for early, mid and late condition (Taken from [26])

In [91] the influence of contextually-defined contrast on the processing of natural language was investigated. Participants listened to instructions containing scalar adjectives (e.g *tall, big, small*) while looking at a number of objects. Scalar adjectives differ from the color and marking adjectives used in [101] because they have no inherent meaning: while the color of an object is a fixed property, declaring an object as tall is dependent on a comparison with other objects (either those nearby, or those present in memory).

The results showed that during incremental parsing of the sentences, the correct referent for an instruction was already found during adjective onset, due to contrasting effects of the objects presented. For instance if a subject saw two glasses, a tall one and a small one, the participant would shift his gaze already when hearing the word *tall* before onset of the noun. This indicates, that humans use visual contrasting information during reference resolution , giving meaning to adjectives after reasoning about context information from the visual channel.

(a) Choose the cow with the stick. (Modifier Bias),
(b) Feel the frog with the feather (Equi Bias),
(c) Tickle the pig with the fan (Instrument Bias).

Figure 3: Examples for verbs with a bias with regard to PP-attachement (Taken from [94])

In [94] adult participants were given the task to listen to sentences that were ambiguous with regard to the attachment of a prepositional phrase. Sentences were chosen to include verbs that presented a certain bias towards either instrument or modifier use of the PP-phrase (see Figure 4). This bias for a certain kind of PP-attachment was

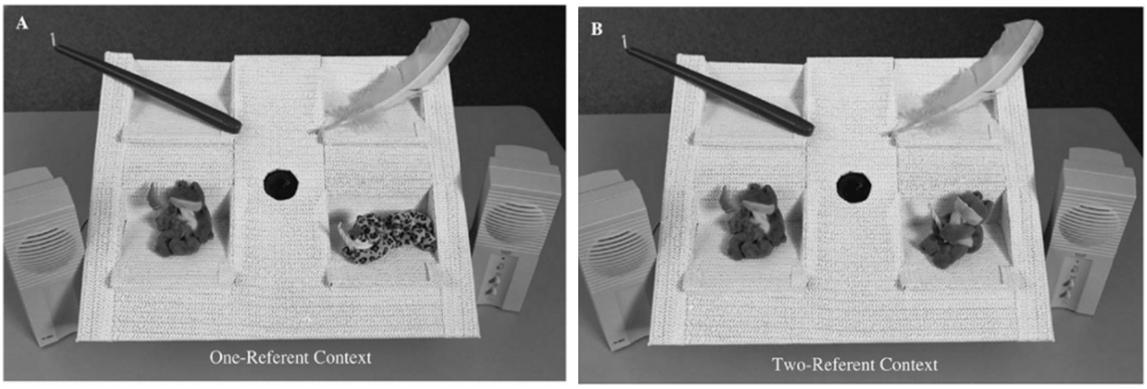


Figure 4: One-referent and two-referent context for the sentence *Feel the frog with the feather* (Taken from [94])

identified in an earlier study where test subjects were asked to complete a fragment of a sentence(e.g. *Touch the teddy bear with...* (taken from [94])). If at least 75% of subject decided for a modifier prepositional phrase (*with the feather*) for the verb, this was judged to have a modifier bias. The opposite applied for instrument bias. If decisions were under 75% for either role, the equi bias was assumed.
At the same time the instructions were heard, participants were presented with pictures (see Figure 4) containing either a one-referent or a two-referent context, as well as a target instrument and a distractor instrument. In the one-referent context, instead of a second possible referent, a distractor was used (like the cat in Figure 4). Results showed that adult humans would use information from language (e.g. verb bias) as well as from referential context, to decide about the attachment of the PP-phrase. The influence from bias and context was found to be additive. The context containing one referent scenes would increase the likelihood of the instrument interpretation of the prepositional phrase, compared to the two-referent context. At the same time verbs with an instrument bias (like in sentence (a)) would also increase the likelihood of the instrument interpretation. This shows that the choice for attachment of prepositional phrases is dependent on the semantics of language as well as on the presented visual context.

[55] investigated the effects of visual context on incremental thematic role descriptions. Sentences used were structurally and role-ambiguous until onset of the second noun phrase during incremental parsing (see Figure 6). Pictures presented showed

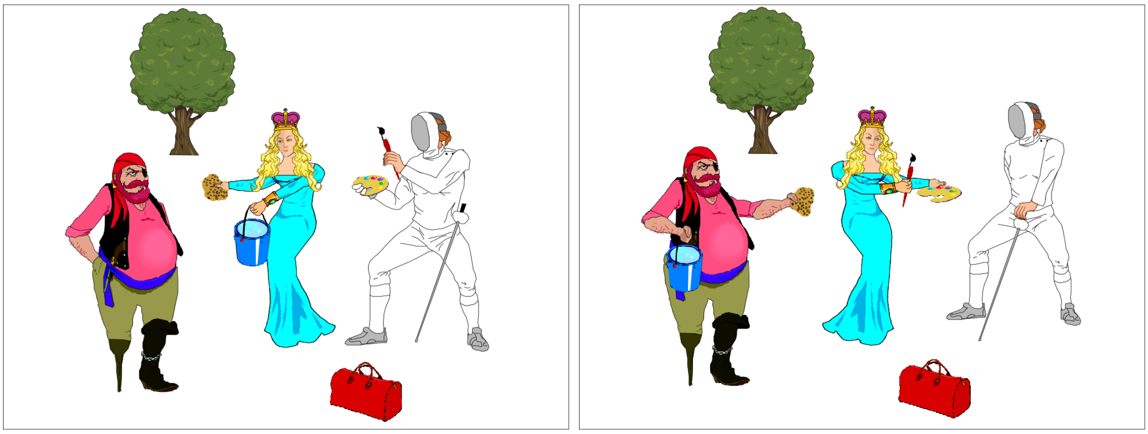


Figure 5: Pictures containing characters with different action-roles: one agent, one patient and one character that is agent and patient (taken from [54])

two actions and three persons that were either carrying out an action (AGENT), receiving the action (PATIENT), or both (AMBIGUOUS) (see Figure 5). Results showed that test subjects identified the character not yet addressed in the sentence already at verb onset, thus showing that they already identified the thematic roles of each character at his moment. The only way participants could assign the correct role to the first noun of the sentence during incremental parsing was to incorporate information from the visual context.

Die Prinzessin wäscht offensichtlich den Pirat
Literally: The princess (subject) washes apparently the pirate (object)
'The princess is apparently washing the pirate'

Die Prinzessin malt offensichtlich der Fechter
Literally: The princess (object) paints apparently the fencer (subject)
'The princess is apparently painted by the fencer'

Figure 6: Sentences describing the scene in Figure 5 (Taken from [54])

2.3 The Predictive Properties of Reference Resolution

The hypothesis that people are able to predict upcoming information depending on prior linguistic input has been investigated with regard to the language modality by studying eye-movements during reading ([78], [79]). Our focus in this work is on those cognitive effects of visual reference prediction that are beneficial with regards to processing of information in one of the modalities. We will distinguish between cases where vision influences language processing, those were language influences visual processing and those where the influence is bi-directional. Regarding the effect of language information on reference resolution, we are interested in the information in the language modality that will result in a prediction for a specific referent.

Furthermore we are looking for effects depending on a higher level of reasoning than merely recognizing words of a sentence. It is imaginable (and as we will see highly probable) that knowledge derived from the sentence listened to (e.g. syntactic knowledge about how words are interconnected or semantic knowledge about the specific meaning of the sentence) might have an effect on the choice of a future referent.

With regards to the visual side of processing we are interested in how the number and

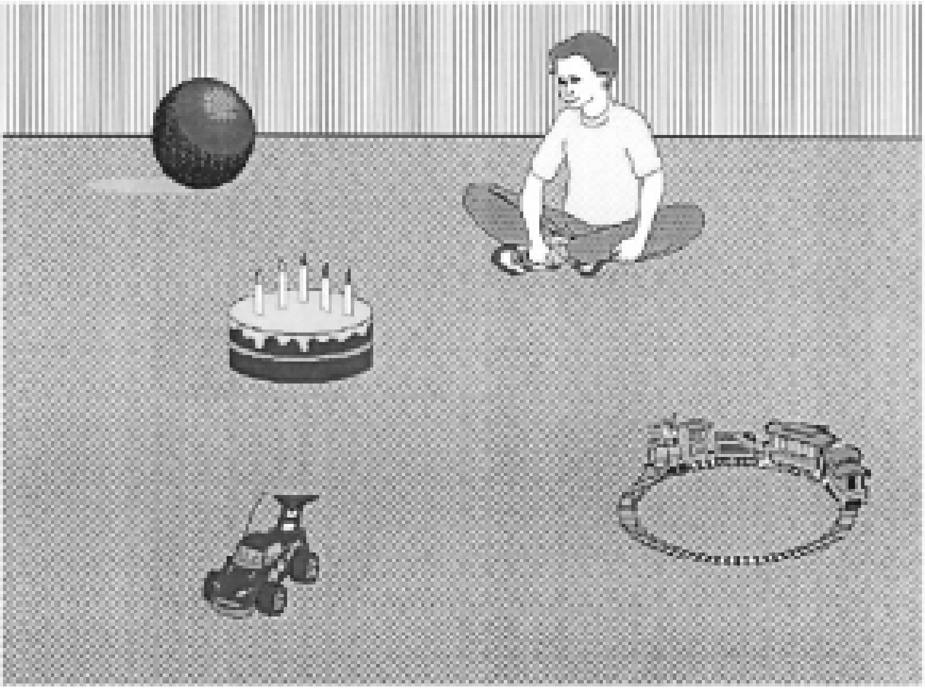


Figure 7: Visual context for the sentence *The boy will eat/move the cake* (taken from [2])

type of possible referents influences the prediction. If only one perceivable referent has any kind of connection with the information presented in the sentence already spoken the choice might be a simple one, but this will probably change when several possible referents are present that are more or less fitting to the language information. Another topic is the bi-directional effect of one modality on the other. Assuming that language processing has an effect on which referent is chosen for parts of a sentence and the referent chosen influences language processing, how does the human cognitive subsystems interact in a situation where a specific understanding of a sentence will influence the choice of predicted referents and a specific referent in turn influences processing?

Humans do not need to hear a word referring to a certain visual object to guide their attention to the correct referent. This effect was first shown by([2]). They presented participants with pictures containing various objects (see Figure 7). While hearing sentences as *The boy will eat the cake.* The findings showed that eye-movements of the subject were already going to the cake at hearing the word *eat*. If the word *move* was substituted at the verb position, eye-movements to the cake were triggered sig-

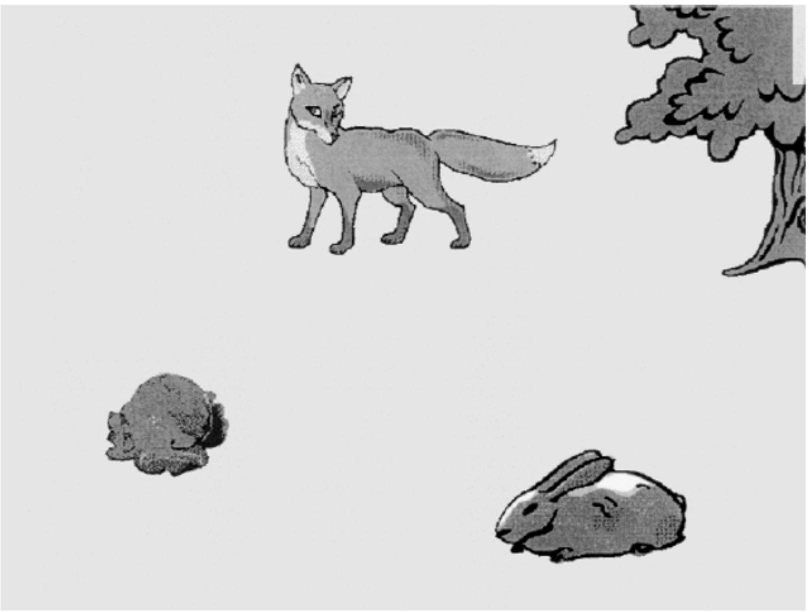


Figure 8: Visual context for the sentence *Der Hase frisst gleich den Kohl*(taken from [52])

nificantly later.

Thus, the verb seems to be constraining the set of possible future referents. Investigating the effects of information on anticipatory reference resolution beyond verbs, [52] conducted experiments using sentences differing in case marking of preverbal nouns. One example used was *Der Hase frisst gleich den Kohl.* Literally: The hare(nom) eats shortly the cabbage(acc). 'The hare will shortly eat the cabbage.' *Den Hasen frisst gleich der Fuchs.* Literally: The hare(acc) eats shortly the fox(nom). 'The fox will shortly eat the hare.' At the same time, participants were presented with pictures like (Figure 8). [52] argued that differences in participants' eye-movements during verb onset would suggest an influence of case marking on the choice of possible upcoming referents. In the given example the participants would direct their gaze to objects that are typically eaten by a hare (for the nom-acc case) or to objects that typically eat hares (in the acc-nom case).

Furthermore, experiments were conducted to show the effects of morphosyntactic marking of the verb on reference resolution. The same scenes as in the preceding experiment were used, in this case giving the participants English sentences like *The hare will eat the cabbage./ The hare will be eaten by the fox.* Both experiments showed

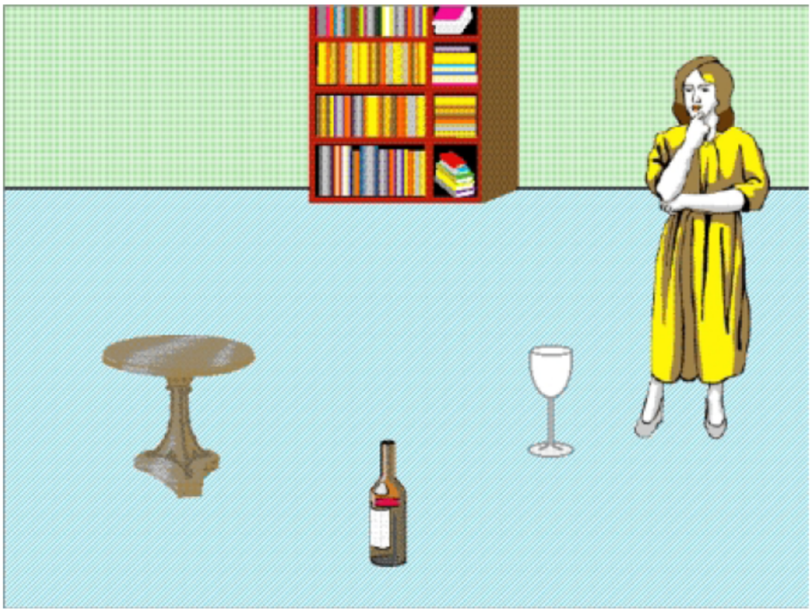


Figure 9: Visual Context for the sentences in Figure 11 (taken from [4])

anticipatory eye-movements to the semantically fitting object during verb onset. Interpreting these results, it seems that humans integrate morphosyntactic marking of pre-verbal noun phrases or verbs with semantic information from the verb in order to find future referents.

2.4 The Conceptual Level

Now that we have compiled evidence for the tight interaction between language and vision, one issue remaining is how these different information types are combined in the cognitive system.

Superficially the answer to this question seems to be simple: humans have knowledge about concepts of the world learned from their experience ([60], [28]). It is easy to assume that the knowledge about a certain concept is related to its visual properties (like the shape of a banana as well as its color) and also language information somehow connected to this concept (i.e. the word *banana*). The issue gets more complicated when we keep in mind that the relationship between words and visual entities is not one-to-one but many to many. As one and the same type of object can have many words applying to it (as is the case for the words *dove* and *pigeon* which refer to the same type of animal) one word can refer to a lot of different types of objects (the word *car* and the wide range of different vehicles existing). Furthermore, different people might have different concepts of the same things. Making the matter even more complex, the classification of things in the world into different concepts is highly dynamic as a person learns things that might rearrange the relationships between concepts. Furthermore, the classification is highly dependent on the situation perceived(see [92]). An example of this would be an everyday object considered to be a piece of art when presented at an exhibition.

In [43] human test subjects were presented with pictures (see Figure 10) while hearing sentences that contained one word that was either referring to one of the objects in the picture (e.g. hearing the word *piano* while looking at a picture containing a **piano**) or was semantically related to one of the objects (the word *piano* while a **trumpet** can be seen) or both (the picture contained a **piano** as well as a **trumpet**). Results showed that participants would increase their fixations to the target object in the first condition. Under the second condition, eye-fixations increased for the object that is semantically related and decreased for the other three objects in the picture, suggesting that 'hearing *piano* activated semantic information which overlapped with the semantic information encoded within the mental representation of the concurrent trumpet.'([43], 30). A similar effect was found when both types of objects were present in the picture: although the number of looks would only increase for the target object (the **piano**) and decrease for the other three objects (including the **trumpet**) the number of eye-fixations for the semantically related object was still significantly higher than for the unrelated objects. Investigating the conceptual overlap between the concept of the reference object and the concept denoted by the word, the authors detected a correlation between the degree of relatedness of the concepts and the probability of looking at the object. A close semantic relationship resulted in a higher probability of visual attention being guided towards the corresponding object showing that conceptual similarity influences reference resolution.

[4] investigated how language and visual context are linked, especially with regard to internal event-representations. The authors argued that internal representations are liable to dynamic change that can, under certain circumstances, compete with visually perceived surroundings. They paired pictures like Figure 9 with two kinds of sentential conditions: Either the sentence described an interaction of the depicted person with a depicted object that was moved before (see Figure 11 (A)), or the

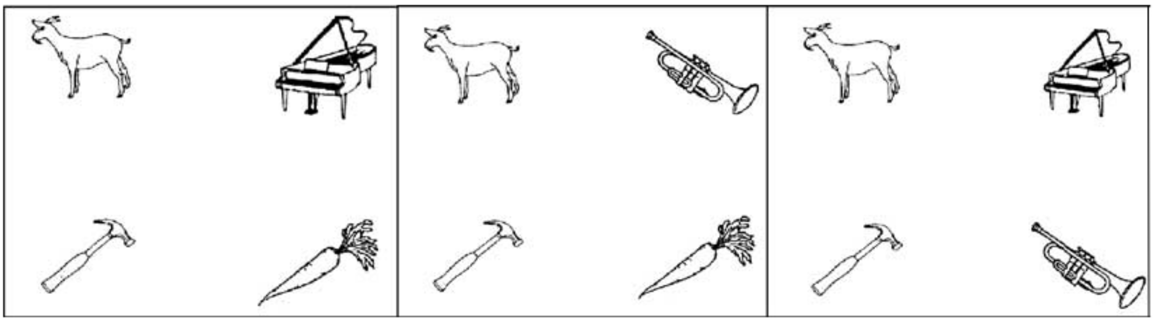


Figure 10: Pictures containing possible referents for the word *piano* (taken from [43])

description would clarify that the object was not moved (see Figure 11 (B)). After onset of the verb (*pour* in our example) participants eye-movements were more likely to be directed to the table in the moved condition than in the unmoved condition. This effect was detected when participants were looking at the picture while listening to the sentence as well as under the condition that the picture was removed before the onset of spoken language. The authors argued that '[...] the eye movements we have observed in these studies reflect a mental world whose contents appear, at least in part, to be dissociable from the concurrent, remembered, or imagined visual world [...]' ([4], 16). Therefore it can be assumed that the indirect mapping of visual and linguistic content onto an abstract conceptual level of representation is subject to dynamic change, not necessarily from the update of information through a modal-specific channel, but also due to reasoning mechanisms undertaken by the human agent.

(A) *The woman will put the glass onto the table. Then, she will pick up the bottle, and pour the wine carefully into the glass.* ['moved' condition].
(B) *The woman is too lazy to put the glass onto the table. Instead, she will pick up the bottle, and pour the wine carefully into the glass.* ['unmoved' condition].

Figure 11: Moved and unmoved sentences for the scene in Figure 9

Following the suggested interaction of language and vision mediated by a conceptual level of semantic knowledge about entities of the wold, the question remains how conflicting information from phonological, semantic and visual-shape information influences the behavior of a subject. Information from any of these areas will influence

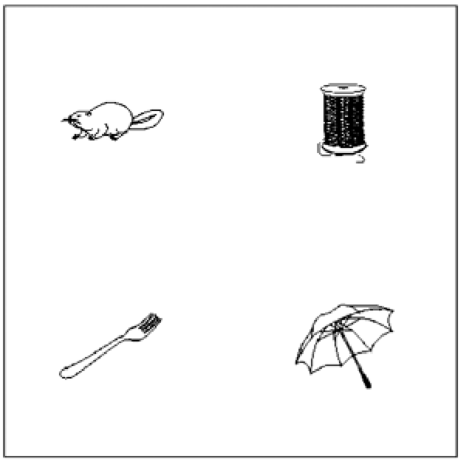


Figure 12: Display containing a beaver (phonological competitor), a bobbin (visual-shape competitor), a fork (semantic competitor) and an umbrella (unrelated competitor) (Taken from [44])

the focus of visual attention of an observer, if presented alone. This holds true for phonological, semantic and shape information. In [44], the time-course of retrieval of each class of information was examined. Participants were presented with pictures of four objects (see Figure 12) and dutch sentences that would contain a word related to some of the objects in the accompanying picture. For a sentence like *Uiteindelijk keek ze naar de beker die voor haar stond.*(Eventually she looked at the beaker that was in front of her.), the picture would contain a phonological competitor (the beaver), a shape competitor (the bobbin), a semantic competitor (the fork) and an object not related to the critical word on any of these knowledge dimensions (the umbrella) (see Figure 12).

The results of the experiments conducted showed two different behaviors depending on the moment the participant was presented the picture. When the subjects were able to see the visual display from the onset of the sentence, attention shifted to the phonological competitors first, and to shape and semantic competitors later. Presenting the picture only 200 ms before the onset of the critical word of the sentence would result in more fixations to the shape and semantic competitors. The authors argued that this could be evidence, that the '[...]storage and/or retrieval of phonological knowledge is independent from the storage/retrieval of conceptual knowledge.' ([44]). They reasoned, that if all knowledge about a concept (i.e. language, visual and

semantic features) would be retrieved at the same time, the time-course of attentional shifts would not differ regardless of when the picture is presented.
The study not only shows that the language driven changes of visual focus are dependent on different knowledge types, but also that the moment of information retrieval (whether information is given early or late during sentence processing) is a factor on the degree of influence of a certain knowledge type.

2.5 Suggested Properties

The review of cognitive findings in the area of cross-modal interaction of language and vision gives rise to implications for modeling this interaction on an artificial system. We propose eight properties that are crucial for creating an artificial framework that is inspired by workings of a natural cognitive system.

1. Humans process language information incrementally, using visual information to guide interpretations

The perceived meaning of a sentence changes as new language information (such as additional words) is received. During every step of incremental processing, available visual information influences the interpretation of a sentence.
Our model should include the means to parse a sentence incrementally while at the same time accessing information from a visual processing module.

2. Humans integrate visual context into the process of language understanding at the earliest moment possible

Contrary to earlier theories, the experiments conducted using the visual world paradigm strongly suggest a closely time-locked interaction between both modalities. As soon as information from the visual field can be used to understand a sentence, it is integrated and will change the interpretation of language information received by the listener.
A model of the interaction between both kinds of information should therefore also make use of any information as soon as it is available.

3. Humans find referents for words and properties

Humans connect language information to objects perceived through the visual channel. The visual world paradigm provides evidence that a listener will shift his visual attention to any visual entity that is a proper referent for a given word. This can either be an object itself, or properties of an object like certain colors, a size or a spatial relationship with other objects perceived. One crucial point to address is, that not every part of the sentence will be a candidate for reference resolution. For instance, a single article (like „the“) at the beginning of the sentence will not cause a shift in attention by a test subject. We interpret this as a humans ability to distinguish between language information that can have a referent and information that, in itself, can not.
Our model should therefore choose possible referents in the visual context for parts of the language input whenever it is viable.

4. Humans find referring actions for verbs

In addition to the properties discussed in point two, humans do not just choose referents that are visually perceptible entities, but also processes defined by what is perceived. As we have described above, this is obvious for the connection of verbs of a sentence and actions observed in the context. The reason why we distinguish between actions and objects as referents is the different effect on the visual attentional focus. It is obvious which part of the visual field is currently attended to when the entity is an object like a chair or a table. But actions are not confined to a specific part of the visual field. A person can look at objects and participants that are part of an action, not at the action itself. Therefore any choice of action as a referent induced by a listened word might result in shifts of attention to parts of a given scene neither not yet attended to nor already mentioned in previous linguistic input.
Our model has to consider these effects of referential resolution for abstract concepts like actions. Therefore our model should include the ability to guide its visual focus to those parts of a given scene that are relevant to an action referred to by a verb of

the input.

5. Humans use context to interpret words

As we have seen, words are interpreted depending on the possible referents available. This holds true for conceptual relatedness between objects named by a word and those seen in the context as well as mismatches between received words and possible descriptions of an object seen.
Our model should therefore be able to connect language and visual information despite slight literal or conceptual mismatches.

6. Humans anticipate upcoming referents

We have evidence that possible referents are attended to even before words naming them are received during incremental parsing. The choice of applicable referents is dependent on the semantic content of the sentence fragment already received as well as on the referents chosen for words of the fragment.
The model should show the same behavior by assigning referents to anticipated parts of the sentence.

7. Humans include information from the whole sentence in order to establish reference

All of the previous considerations do not only concern one specific part of the language input, like a single word or a phrase, but each decision for attentional shifts and choice of referents is always dependent on all information available by interpreting those parts of the input sentence already received. Possible referents for an action, for instance are not only dependent on the verb of the sentence, but also on the participants of action either already addressed or anticipated by the system.
The model should include every information deduced from language input received when choosing a possible referent for a word or phrase.

8. Humans include world knowledge and conceptual relations into the process of connecting both modalities

Humans not only apply their knowledge about conceptual relations when connecting words to their referents, but also whenever phonological and conceptual information are competing with regard to reference resolution. We therefore propose a model that does not link words to objects directly, but is mediated by a conceptual model, where different types of information and their relations can be represented.

It is our assumption that modelling these properties when creating an artificial system of cross-modal interaction will result in a behavior that is highly similar to the one outlined above. This similarity should affect each and every aspect, be it incremental processing, context integration or reference resolution.

3 Related Work

In the following section we discuss several approaches to combine natural language processing and visual context processing. The scope of the subsections is largely defined by the similarity of the discussed approaches with the model we are aiming at. Although a large number of models are somewhat related to the interaction of language and vision, ranging from commenting visual scenes, speech guided robots working in a visual environment or using natural language for video search, we are mainly interested in a comparison of our model with those architectures that connect natural language parsing and visual representations in a bi-directional way. We therefore focus this section on models we consider to have similar aims with regard to the integration of these different modalities. At the end of the section, we briefly introduce further approaches connecting these modalities.

3.1 Early approaches

3.1.1 Winograd

In [104] a system name SHRDLU is introduced that uses language input for reference resolution in a simulated blocks world environment (see Figure 15). This blocks world is described on a representational level as a database of facts in the form seen in Figure 13

(IS B1 BOX)
(IS B2 PYRAMID)
(IS B3 BLOCK)
(IS B4 BLOCK)
(COLOUR-OF B2 BLUE)
(COLOUR-OF B3 RED)
(COLOUR-OF B4 BLUE)
(CONTAINS B1 B2)

Figure 13: Example of the SHRDLU database of facts (Taken from [38])

The database does not distinguish between general knowledge (i.e. rain is wet) and specific knowledge (i.e. Peter owns a goldfish). The system receives questions and instructions, parses these and attempts to find referents for identified phrases of the sentence. This is done by translating the parsed sentence into a theorem which is then proved depending on the information found in the knowledge base.

For instance a description like *a block which is in the box and red* would be trans-

```
(THGOAL (#IS $?X #BLOCK))
(THGOAL (#IN $?X :BOX))
(THGOAL (#COLOR $?X #RED))
```

Figure 14: Example of a formula with variables to be instantiated with visual entities (Taken from [104])

lated into the theorem in Figure 14 where the ?X is a variable to be instantiated with a possible referent. In our example case a possible referent should satisfy the three conditions specified by the goals of the theorem (e.g. is a block, is red, is in a box). When reference resolution is successful, the system acts accordingly by either answering the question or executing the instruction.

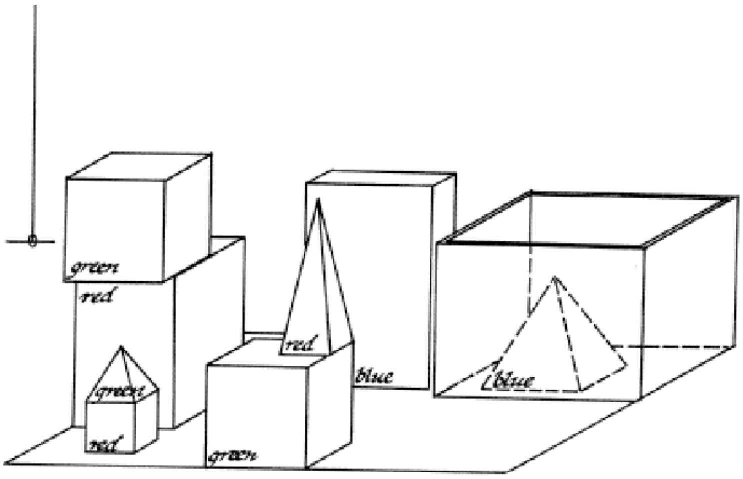


Figure 15: Visual context for the system SHRDLU (Taken from [104])

It is always possible for SHRDLU to be unable to resolve reference. Although the easiest explanation for a failure of reference resolution is that the instruction does not fit to the visual scene (for instance telling the model to find a blue block in a box when the only block in a box is red) another reason for this failure is the wrong translation of an input instruction into a theorem such as in Figure 14. This incorrect translation will always happen whenever the parsing component of the system chooses a syntactic interpretation of the instruction not intended by the user. The problem arises for instance with the two possible interpretations of the sentence *Put the blue pyramid on the block in the box* (see Figure 16) which mirrors human understanding difficulties with regard to PP-attachment while observing visual context as researched in [100] (as we have already discussed in Section 2). Depending on whether the parser decides to choose the syntactic reading a. or b., the translation would either require that the pyramid is on the block or that the block is in the box. Whenever such a theorem can not be proved with regard to the visual context (which happens when a referent is not found) the system will feed back this information to the parser which will result in a re-evaluation of the analysis. If there actually is a blue pyramid standing on a block, thus presenting an applicable referent for the noun phrase in interpretation a., this reading of the sentence is chosen. If the only blue pyramid is not standing on a block, interpretation b. is the preferred one. After switching syntactic interpretations, a new theorem is generated which might be more successful with regard to reference resolution.

a. Put [the blue pyramid on the block]$_{\text{NP}}$ [in the box]$_{\text{PP}}$
b. Put [the blue pyramid]$_{\text{NP}}$ [on the block in the box]$_{\text{PP}}$

Figure 16: Example of two different parsing decisions for the sentence *Put the blue pyramid on the block in the box* (Taken from [38])

3.1.2 Haddock

Haddocks [37] work uses Combinatory Categorical Grammar (CCG) as described by [96]. It consists of a lexicon associating each word with a syntactic category and a grammar which specifies rules for combining categories to form new categories. The lexicon distinguishes between words and functions: Words can be arguments which are linked to atomic categories. For example the word *car* is associated with the atomic category N (for noun). Functions, on the other hand, describe syntactic mappings between categories. For instance an article like *a* is modeled as a function in the form NP/N, indicating that the article combined with a noun on its right will result in the syntactic category of a noun phrase. These functions can be arbitrarily complex like the word *drives* which is categorized as (S\NP)/NP, indicating that a noun phrase on its right (/NP) will result in a function which, if given a NP on its left (\NP) will result in a sentence (S).

Furthermore, CCGs include rules for combinatory operations such as function composition, which is used to generate syntactic categories for word combinations where each word is considered to be a function in itself. An example of this is the fragment *the fast*. For this particular case the CCG contains a rule that combines the article (NP/N) with the word *fast* (N/N) into the function NP/N. Simply put this means that the fragment *the fast* given a noun on its right results in a noun-phrase.

Haddock extends this formalism by associating semantic content with each word in the grammar. The noun car would be defined as

car := $N_{e2} : [car(e2)]$

where e2 is a semantic variable. Function categories work in this regard as mappings between semantic variables. The preposition

in := $(N_{e1}\backslash N_{e1})/NP_{e2} : [in(e1, e2)]$

links the variable e1 of a preceding noun with e2 which is the prepositional object. The fragment *rabbit in a box* applied to this rule would result in the semantic content of the words *rabbit* and *box* being linked to the semantic variables e1 and e2 respectively. The semantic analysis of the phrase would thus be **in(rabbit, box)**.

Furthermore, the semantic rules that accompany each word not only describe mappings between possible referents (as a rabbit and a box in our last example) but also specify rules that constrain referential resolution. One example is the constraint **unique(e1)** given with the definite article *the*, which holds when the set of possible referents for e1 consists of a single element.

Using this model can account for problematic structures such as the phrase *the rabbit in the hat*. Using the context in Figure 17, [37] showed that the incremental semantic interpretation of the sentence with the presented context using this model would indeed find the correct referent for the hat.

rabbit(r1), rabbit(r2), rabbit(r3), hat(h1), hat(h2), box(b1), in(r2,h1), in(r3,b1)

Figure 17: Context for the phrase *take the rabbit in the hat* (Adapted from [37])

The analysis of the phrase (the reader is referred to [37] for a full exposition of all processing steps) results in the facts [**rabbit(e1), in(e1,e3), hat(e3)**] or, simply put, of all the possible referents the parser chooses the rabbit and the hat that are related by the **in** predicate.

3.2 Recent Models

3.2.1 Gorniak and Roy

The system Bishop [33] operates in a domain of different colored cones. The basis for the system was an experiment in which subjects were required to describe randomly generated visual scenes in a virtual environment. Strategies used by participants to identify a single object or a group of objects were analyzed and led to a number of requirements for a system implementing reference resolution for natural language.
Bishop consists of a visual display of a number of purple and green cones placed on a board (see Figure 18). It receives descriptions of a specific cone or a number of cones in the visual scene, parses these and tries to find the correct referents. The parser builds a parse incrementally which is checked for consistency with regards to the scene presented.
To resolve reference, the system incorporates several cues:
Colors are expressed through adjectives. Using sentences like *the purple cone* the system is able to find the correct candidate in a picture contained cones of different colors.
The system can resolve references to specific regions of the depicted scene. Input like *the cone on the right side*, combined with others cues is used to find the referent.
The system can also comprehend grouping descriptions like *the green cones* or *the five in front*. This information can be used to identify cones that are part of the group (*the front-most of the three green cones*) or are related to it (*the purple cone to the left of the five green ones*).

Using descriptions of spatial relationships in combination with other descriptive

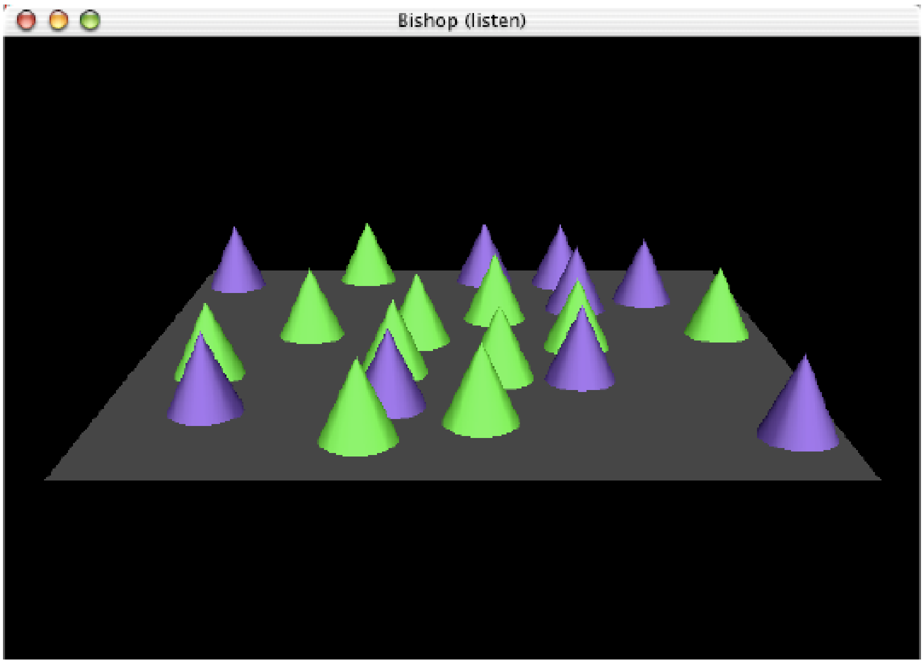


Figure 18: Visual context of the Bishop system (Taken from [33])

strategies is another way to resolve referential ambiguity. A simple example would be the utterance *the green cone to the left of the green cone* while showing a picture of two green cones.

3.2.2 Scheutz

In [86] a robotic model was proposed that applies cognitive findings of experiments on human reference resolution to build a system with traits of human attentional processes. One core assumption incorporated in this model is the conclusion that '...parse trees may contribute minimally or not at all to comprehension in communicative situations in which the referential context is visually co-present with both the listener and the speaker, and, therefore, highly accessible; i.e. it does not need to be maintained in working memory'([86], 145)

The system chooses referents for natural language instructions in a simple Blocks World Domain (see Figure 19). In this domain, blocks of different color are placed so that they are in spatial relationships such as **on**, **under** or **next-to** each other. During evaluation the system is presented with conditions where instructions like *Put*

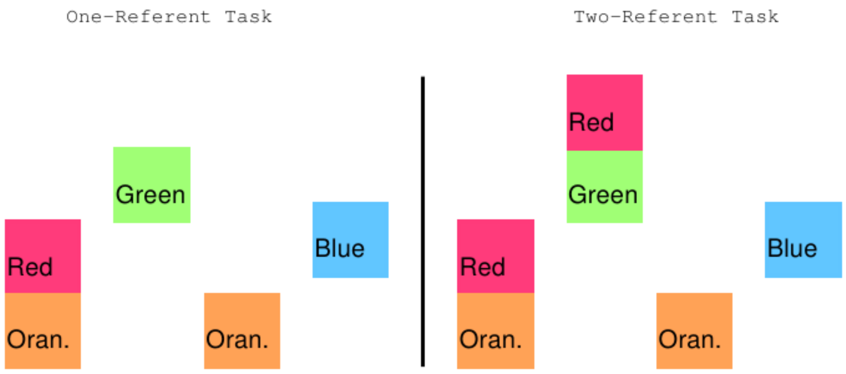


Figure 19: Example of a one-referent and a two-referent context for the sentence *Put the red block on the green block on the blue block*(Taken from [86])

the red block on the orange block on the blue block are applied to block arrangements containing one referent or two referents (see Figure 18). Evaluation showed that the model exhibited the same kind of shifts in attention as found in humans, but only in the above-mentioned, simplified domain of blocks in a picture with a restricted set of spatial relations (like **on**) and no complicated semantic content (i.e. actions and objects of several different classes)

3.2.3 McCrae

[68] proposes a model for the influence of cross-modal information on syntactic parsing of natural language. Semantic information, received for instance from the visual modality, is used to guide the processing of language by a parser of German. The language parser used is based on the Weighted Constraint Dependency Grammar (WCDG) and outputs its results as a forest of dependency trees in the form seen in Figure 20 where words of the input sentence are modeled as nodes and syntactic relations between words as edges. [68] extended the syntactic parsing process by introducing a level of shallow semantic interpretations showing thematic relations between words. This semantic interpretation can be modified by external knowledge, received from the visual channel. The external information is integrated by a predictor-based approach: information from a knowledge representation of visual context is received at the beginning of the parsing process and subsequently changes the outcome of a

parse by modifying the probability of connecting two words by a specific edge of the structure.

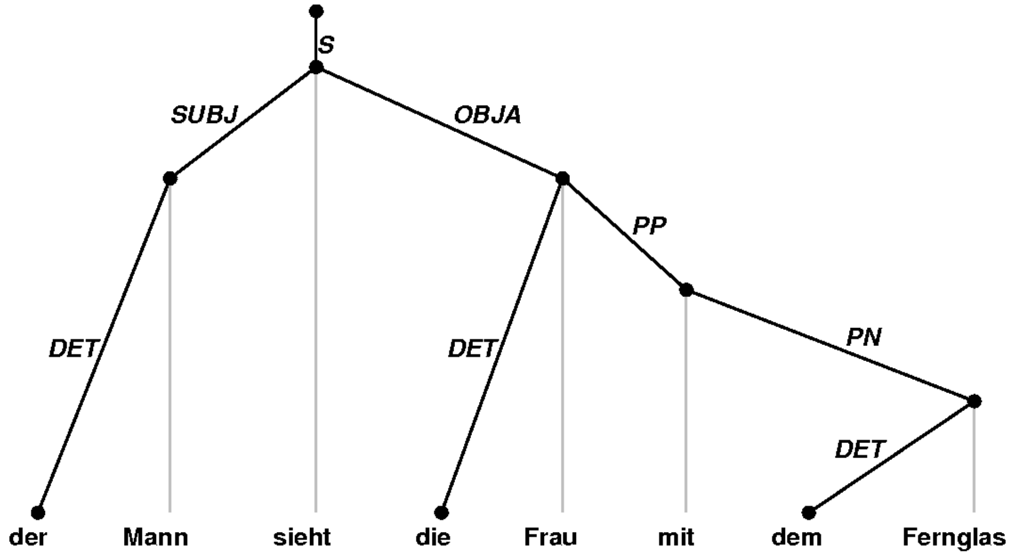


Figure 20: Dependency tree for the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope)

Although the integration results in a higher performance of the parser the system has a number of shortcomings that require further investigation:
The system receives contextual information only at the beginning of the parsing process and leaves it unchanged during sentence processing. It is thus unable to cope with a dynamic environment where information regarding external events is constantly changing. This is a highly unrealistic scenario not in line with a cognitively motivated model that receives information from its environment dynamically, which is immediately integrated into the processing of natural language.
The information received is unambiguous in nature. It is always evident which part of the context influences which part of a sentence analysis. In natural systems, agents will have to decide which part of contextual information is referred to by a given sentence. In a real-life environment a large number of visual entities are present. It is not evident from the visual context itself, which of these entities are relevant for the presented linguistic input and which are irrelevant or even misleading. Since it considers the connection from language to contextual information as static, the system is not able to reevaluate its choice of information to be used in language parsing, by

selecting different parts of the context as influential knowledge. This is not in agreement with the cognitive findings of human reference resolution presented in Section 2, where we discussed the human ability to change referents for linguistic information depending on language processing results.

3.3 Context-dependent Speech Recognition

With regard to the the synthesis of spoken language and referential information, several models exist.
[95] describes a method that identifies referents in a visual scene of puzzle objects for sentences like

a. Take the piece in the middle on the left.
b. Take the piece in the middle.

Each object in the scene is represented by a number of object features such as size, length, height and also topological features like distance to certain landmarks and to other objects. The system grounds language in vision by learning visual semantics: depending on how words of a training corpus were used in conjunction with a visual scene, a feature vector is computed from the object features that represents the visual semantics for each word. E.g. for the word left, the feature vector would contain a value representing the relevance of the relation between horizontal position of the object and the center of the board on which the objects are arranged.

The framework proposed in [89] which is based on a world model consisting of entities and their properties and relations incorporated into a probabilistic language model that can be used for the task of speech recognition. The model works incrementally on a static representation of a context domain while the integration of speech and context information is realized by means of Hidden Markov Models. Language is used to find possible referents in a contextual model by traversing edges that model relations between sets of possible entities (see Figure 21). For instance in Figure 21, which describes an domain of three files(f1, f2, f3) on a computer, the set of possible referents is found by navigating through the graph along edges corresponding to properties of the input sentence. Thus, a sentence containing the noun phrase *a*

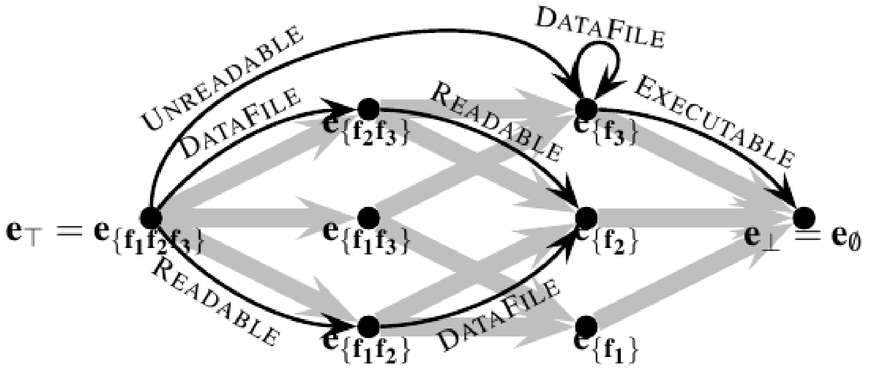


Figure 21: Subsumption lattice including relations (adapted from [89])

readable file should narrow down the set of possible referents for the word *file* to (f1, f2). The referential context found is then used to influence the transition probabilities in a hidden Markov model used for speech recognition.

[25] presented the interface FIGLET, which interprets English instructions in a drawing domain. The system performs incremental interpretation of instructions like *Put a circle below the eyes* with regard to a visual context that shows caricatures of drawn faces. The system not only finds referents in the picture (reasoning, for instance, that two circles are referred to by the word *eyes*), but also finding appropriate actions for spoken instructions. The last task includes the match of ambiguous phrases with specific spatial locations, as is the case for the PP *below the nose* and the exact place where something is to be drawn (i.e. how far below the nose should the system place the new object). In order to do this the model disambiguates natural language expressions with information about contextual entities. For example, prepositional phrases of instructions like *put the circle to the left of the nose* are interpreted as modifying a noun or a verb, depending on positions of objects referred to by the sentence.

[59] integrated contextual knowledge into a model for situated dialogue between a human and a robot. The presented architecture consists of a number of subsystems where each one is responsible for a specific task (such as vision, dialogue processing, manipulation, spatial reasoning, planning, coordination and binding). The binding

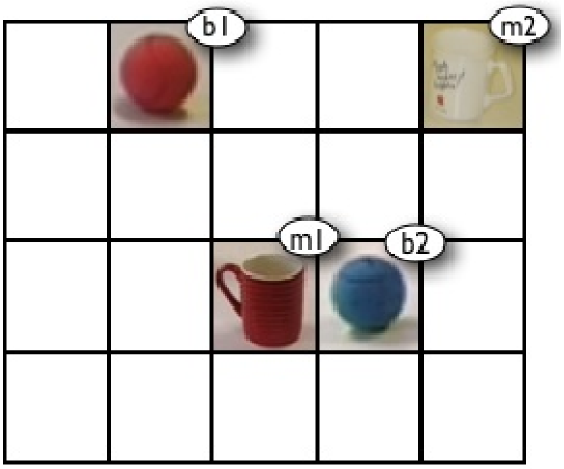


Figure 22: Visual context from [59]

subsystem of the model combines representations of different types of information contained in the subsystems of the modalities.
The system was used to identify objects in visual setups such as Figure 22. Utterances consisted of commands (*put the mug to the left of the ball*) or assertions (*the mug is red*) that can be used to identify possible referents in the visual context.

[53] describe a computational model that can interpret expressions containing spatial prepositions. The model is used to find referents for targets (the object to be located) and landmarks (objects relative to the targets location) of phrases such *The man*(target) *near the table*(landmark).
The system receives input from a vision subsystem and a speech interpretation pipeline. The speech module transforms spoken input into a string representation which is processed by a language parser to generate a representation of the syntactic analysis.
The vision subsystem detects and categorizes objects in the visual field and generates a representation, including geometric positioning information for each visual entity.
The connection (and thereby reference resolution) between these two types of information is realized by a component for spatial reasoning that translates one kind of representation into the other one.

3.4 Conclusions

As we have seen, a wide range of different models exist that connect language and vision. We suggest that all these systems perform poorly when compared to cross-modal interaction in humans as any of the proposed approaches lacks one or more of the following traits:

- The ability to process language incrementally
- Using visual context to influence language processing
- Using visual information from a wide range of possible domains (and not only from a blocks world environment)
- Re-interpretation of already made parsing decisions due to changes
- Reacting to dynamic changes in both modalities
- Using semantic information
- Including world knowledge into processing

The approach in this thesis is to improve upon the state of the art with a model that includes all these traits, thereby approaching human-like capabilities of integrating information from two types of information.

4 Fundamentals

In the following section we introduce the subsystems that are the basis for our architecture of cross-modal interaction. We model the language modality using a parser for natural language, which covers a large set of possible human utterances. The description of visual context is implemented by a knowledge representation formalism which is used to model world knowledge as well as an abstract description of the visual scene perceived. A model driven by visual cues is then used to capture the effects of bottom-up visual attention on the choice of the dynamic focus in the visual field.
The basic systems are described in this section as stand-alone models showing the as-is state. We will discuss the extensions made to achieve our goal of establishing an interactive interplay between modalities in section 5.

4.1 Knowledge Representation

As discussed in Section 2, evidence from human interaction of language and vision strongly suggest an indirect link of both information sources. Therefore we mediate the connection by using a level of knowledge representation. Our basis is the abstract description of world knowledge and visual entities introduced in [68] where the Web Ontology Language (OWL) (see [73]) was used for a high level description of visual context. This semantic representation consists of two distinct but related sources of information: firstly, the world knowledge is described as a hierarchy of concepts (the t-box) and secondly the situation dependent knowledge of information derived from an assumed visual input source is described as a set of visual entities and relations between them (the a-box). The aforementioned link of the two information sources is realized as an **INSTANCE- OF** relation between visual entities and concepts of the taxonomy: every visual object and every visual relationship that is derived from perception is always connected to a concept of the hierarchy. In this section we will give a short overview of both knowledge sources as already discussed in [68].
The taxonomy of concepts consists of a set of classes which are related by **IS-A** relations thereby creating a hierarchy like the one in Figure 23. Concepts always describe entities perceivable by the visual channel, but not necessarily concrete objects. Thus, a concept might denote a class of objects (such as **Chair**), an action that is undertaken by persons in the visual field (e.g. **Painting**) or be a description for abstract things related to visual objects (for instance the concept **Danger**).

```
Thing
   →  Animal
   |       →  Snail
   →  Human
   |       →  Cousin.m.f
   |       →  Guest
   |       →  Inhabitant.m.f
   |       →  Member
   |       →  Named.Johnson
   |       →  Named.Klitschko
   |       →  Patient.m.f
   |       →  Ruffian.m.f
   |       →  Tourist.m.f
   |       →  Visitor.m.f
   →  Physical.Object
   |       →  Aeroplane
   |       →  Bag
   |       |       →  Medical.Bag
   |       →  Ball.Concept
   |       →  Barn
   |       →  Barrel
   |       →  Basket
   |       →  Bench
   |       →  Bier
   |       →  Bouquet
   |       →  Bucket
   |       →  Building
   |       |       →  Hospital
   |       |       →  Store
   |       |       →  University
   |       →  House
   |       →  Hydrant
   |       →  Leg
   |       →  Newspaper
   |       →  Plant
   |       |       →  Tree
   |       →  Shelter
   |       →  Stick
   |       →  Table
   |       →  Tent
   |       |       →  Awning
   |       →  Tool
   |       |       →  Spatula
   |       |       →  Trowel
   |       →  Vehicle
   |       |       →  Car
   |       →  Wall
   |       →  Wand
```

Figure 23: Example Taxonomy

$$\begin{array}{lcl} \text{Woman\_01} & \xrightarrow{\text{is\_AGENT\_for}} & \text{tragen\_01} \\ \text{Book\_01} & \xrightarrow{\text{is\_THEME\_for}} & \text{tragen\_01} \end{array}$$

Figure 24: Description for the visual context representing a woman carrying a book. A woman (**Woman__01**) is the agent of the carry action (**tragen__01**) while a book (**Book__01**) is the theme

[68] defined fourteen relationships that can hold between visual entities. These are used to model semantic relations between visual entities or concepts of them. The relations describe thematic roles (such as AGENT, THEME/PATIENT, INSTRUMENT, OWNER and COMITATIVE) of participants engaging in visually perceivable actions. In addition to the thematic relations, a concept can be assigned a lexicalisation by asserting the **has__Lexicalisation** relation between an individual representing a word or a phrase and the concept.

With regard to the situation-dependent visual context information, the introduced model takes the stance of being a description of visual percepts after cognitive processes have already extracted high level knowledge from the visual channel and preselected visual entities that are relevant for disambiguating linguistic input. The context is represented as a set of visual individuals and their semantic relations.

Every individual is the instantiation of exactly one concept described in the conceptual hierarchy. These individuals are connected according to the thematic roles they represents by relations such as **is__AGENT__for**.

For an example of a visual situation represented as a set of individuals and thematic roles, Figure 24 describes the perceived action of a woman carrying a book. The woman is modeled as the AGENT of the carrying action, while the book is the THEME.

4.2 WCDG

To demonstrate the effects of incorporating visual context information into the processing of natural language, a system capable of processing unrestricted human utterances is needed. The system should be able to receive external information in order to select one of several different possible interpretations of a given natural language

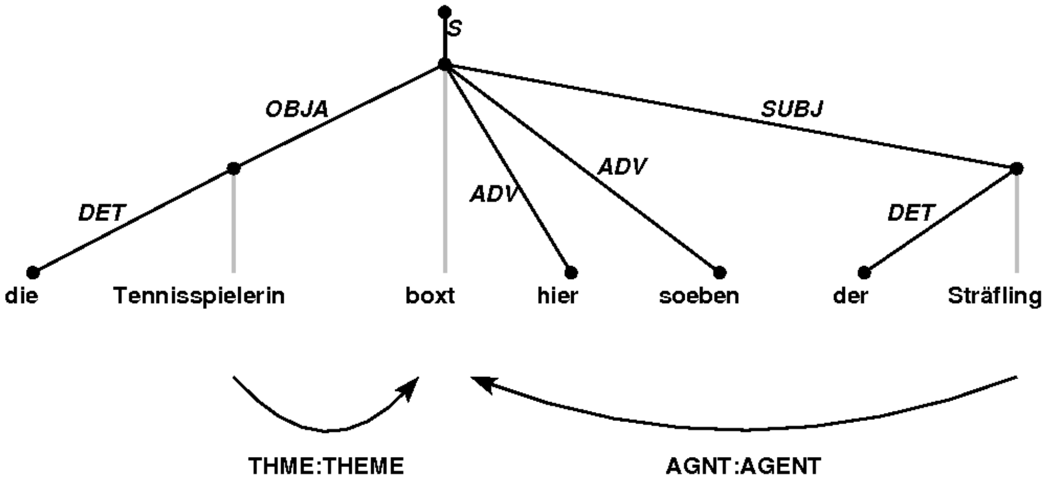


Figure 25: Dependency tree for the sentence *Die Tennisspielerin boxt hier soeben der Sträfling* (Literally: The tennis player boxes here just now the convict)

sentence. Contextual knowledge can be used as a factor to change an analysis of a sentence to an interpretation that is more fitting with regards to visual knowledge. The Weighted Constraint Dependency Grammar (WCDG)-Parser (see [90]) is our choice for this task. The input of the system consists of a grammar and a lexicon of the language that the parser should work on, as well as a sentence to be parsed. The output is a forest of dependency trees as seen in Figure 25 showing lexemes of words as well as their dependencies, indicating the most plausible interpretation of a sentence. The nodes of the tree represent lexical entries found in the WCDG lexicon while the edges show syntactic and semantic dependencies between them (e.g. that the word *Tennisspielerin* (tennis player) is the object of the verb *boxt* (boxes) is indicated by an OBJA edge connecting both words).

4.2.1 Levels of Analysis

As each word in an dependency structure can only have one modifier it is necessary to generate more than one tree for a single sentence, whenever different properties apply to the same word. This is particularly commonplace whenever a word of a sentence takes on one of the thematic roles mentioned above. In Figure 25 the dependency tree shows the word *Sträfling* (convict) as being the subject of the sentence (indicated by the label SUBJ at the connecting edge) while at the same time being the AGENT of the boxing action. Although the graphical representation shows both of the modeling edges in the same representation, the processing of these different types of information is confined to two separate dependency structures, in which each word of the sentence has exactly one word it modifies.

The implementation of WCDG used in our work uses five different levels of analysis:

- The level SYN is the level for representation of syntactical properties of the sentence like denoting the subject and object of the sentence and the attachment of prepositional phrases
- The level REF shows the attachment of relative pronouns to those words they refer to.
- The level AGNT describes part of the semantics of the processed sentence by showing the acting participants of actions denoted by a verb.
- The level THME shows those parts of the sentence affected by an action denoted by a verb.
- The INST level differs from the other two semantic levels by being responsible for the representation of three distinct classes of edges: the INSTRUMENT edge denotes those resources that are part of carrying out an action, the COMITATIVE edge denotes those persons that accompany an action, the RECIPIENT edge denotes the person receiving the action and the OWNER edge shows the possessor of some object.

4.2.2 Constraints

WCDG formalizes the rules of a language by specifying constraints in the grammar. Each constraint models a property of the modeled language. This is accomplished by defining a condition in form of a predicate logic formula. This formula is applied to one (unary constraints) or two (binary constraints) edges of the dependency tree. An analysis is said to satisfy a specified constraint whenever the formula evaluates true for all edges of the structure. A constraint as shown in Figure 26 consists of:

- A list of variables that are instantiated with edges during parsing, including restrictions on which level of analysis a constraint should be applied as well as

```
{X : SYN/\Y : SYN} : syn_det_zahl : 0.0 :
~(X.label = DET&
Y.label = DET)
```

Figure 26: Constraint expressing that a word can only have one determiner

the direction of the edge in the tree and (for binary constraints) how the two edges should be connected to each other

- A name which is used internally to uniquely identify the constraint as well as to display constraint violations during processing of a sentence
- A weight stating which penalty should be applied to an analysis which violates this constraint.
- The formula implementing the grammatical rule of the target language

The formula may contain predicates and functions testing specific properties of an edge. Predicates evaluate to true or false while functions evaluate to strings or numbers. Predicates as well as formulas can be dependent on information beyond that of a single edge (i.e. found somewhere else in the dependency tree). Therefore, constraints can be classified as either context-sensitive (those that are dependent on the whole structure of the tree) and context-insensitive (only dependent on the edges the constraint is applied to).

4.2.3 Frobbing

The standard solution procedure of WCDG is the frobbing algorithm. Starting from an initial analysis of the input sentence, the algorithm improves this analysis by repairing violations of the constraints described above. Constraints are evaluated by checking the formula for each edge (or combination of two edges in case of binary constraints) of an analysis. A constraint violation occurs whenever the formula of a constraint of the grammar is evaluated false for the current analysis.

The algorithm choses as the violation to be repaired the one which violates the lowest weighted constraint. A violation is repaired by exchanging those edges in the dependency tree that are deemed responsible for the violation. After the exchange, the old and the repaired analysis are rated in order to determine which one is better. Rating

$$\mathbf{R(A) = h|\prod_{i=1}^{n} g_i}$$

R(A)=Rating of analysis A
h=Number of violated hard constraints
n=Number of violated soft constraints
g_i=Weight of constraint i

Figure 27: Formula for rating an analysis

an analysis is accomplished by evaluating all constraints in the grammar with regard to the edges of the dependency structure, computing a score using the formula in Figure 27. Several factors are considered for evaluation:

1. The number of hard constraints violated. When comparing two analyses, the one which violates less hard constraints is always assumed to be the better one, regardless of the number of soft constraints violated.

2. The product of all weights of violated soft constraints. Whenever two analyses violate the same number of hard constraints, the one with the lower product of these weights is considered to be the better one.

3. Whenever two analyses violate the same number of hard constraints and the product of the violated soft constraints is also the same, the one with the smaller total of violations is considered to be the better one.

If the new analysis is rated better than the old one, the procedure uses the new analysis as a starting point for further repair steps. This continues until either all violated constraints are repaired, an analysis is deemed to be near enough the highest possible value that can be reached, the user has interrupted the process or the number of repair steps exceeds a previously specified upper limit.

4.2.4 Incremental Processing

The findings discussed in Section 2 show that incorporating visual knowledge into language understanding happens at the earliest time possible (i.e. immediately when a word or phrase is received, without the listener waiting for a sentence to finish).

The ability of a listening person to process an utterance before the speaker stops his statement (in the following called incremental processing) has a number of advantages which can also be used in the field of natural language processing in artificial systems. In this section we will discuss the advantages and drawbacks of incremental processing as well as the implementation of an incremental processing mode in the language parser WCDG.

Incremental processing is most useful when applied to problems encountered in the field of human-computer interaction. The real-time input of speech offered by a human participant requires the system to continually update its current interpretation of what it received. Although it is possible for a machine to sub-divide the incoming input waiting for the speaker to stop what he is saying, processing the input already at the time it is received has a number of advantages:

Processing input immediately should result in a better performance of the processing system. As each interpretation of a sentence fragment is based on the analysis of the preceding fragment, processing time should be lessened. Of course this is highly dependent on the time it takes to re-evaluate an interpretation that is disproved by a newly received word.

A second advantage is the time it takes for a system to react on language input. Responding to the input before it is complete might be beneficial for a machine. For instance a robot receiving an order might already interact with its environment, dependent on the incomplete input by manipulating objects or adjusting sensors such that information required for executing the command can be extracted from the environment.

WCDG includes its own processing mode for incremental parsing of language input, introduced in [12]. Processing starts with the first fragment of the sentence (the first word) and will add words one by one. A crucial property of this mode is the successive generation of its output: any analysis of a sentence fragment is the basis for the analysis for the next fragment.

Another important feature of the implementation of incremental processing in WCDG is the ability to predict upcoming elements of any incomplete fragment. As seen in Figure 28 the fragment *Der Mann kauft*(The man buys) will result in an analysis showing syntactic properties of the words as well as a prediction about the missing object of the sentence. This prediction is realized by introducing virtual nodes into the process. These nodes are placeholders for words that might be received at a later time point. The advantage of these nodes is the possibility to generate edges of the dependency tree to words not yet existing in the current fragment.

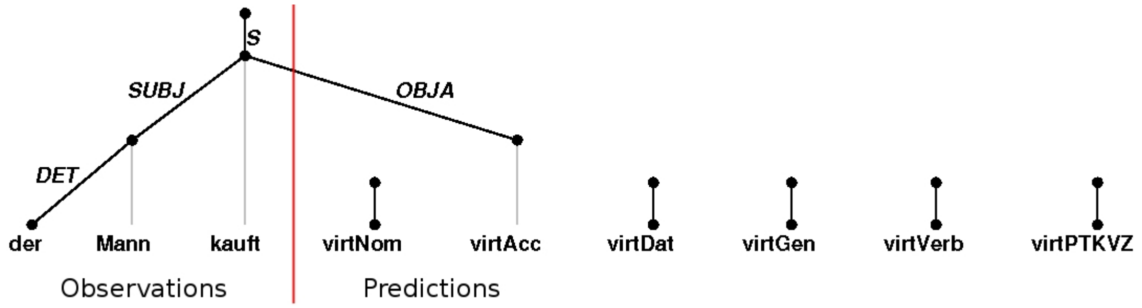


Figure 28: Prediction of an object (OBJA) for the fragment *Der Mann kauft* (The man buys)

4.2.5 Context Integration

In this section we give an overview of the predictor based approach to integration of visual context presented in [68]. At the beginning of processing a sentence the visual context predictor is invoked by WCDG. The task of this component is to link information of the language parser with the semantic information found in the knowledge representation component, deriving probability scores for possible dependency edges before the above-mentioned frobbing algorithm is started. At the beginning of the process, words and phrases found in the input sentence are mapped to individuals found in the context. This process of grounding is dependent on the conceptualization of the visual entities described in the situation model.

The words found in the sentence to be processed are matched to visual entities via concept compatibility. As outlined above, each concept of our knowledge representation is related to at least one lexicalisation by the **has_Lexicalisation** relation, which specifies words by which individuals of this concept are identified. Depending on these lexicalisations, the words of the sentence are matched to individuals connected to the corresponding concepts. One example of this matching process would be the sentence *Er hört die Männer singen.* (He is hearing the men sing) (adapted from [68]). With the visual context in Figure 29 that connects individuals to concepts (e.g. **Man_01** is an instance of **Man.sg**) as well as describing a visual scene (a

Man_01 $\xrightarrow{\text{instance\_of}}$ Man.sg

Man_02 $\xrightarrow{\text{instance\_of}}$ Man.pl

Etw.Hoeren_01 $\xrightarrow{\text{instance\_of}}$ Etw.Hoeren

Singen_01 $\xrightarrow{\text{instance\_of}}$ Sing

Man_01 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Hoeren_01

Man_02 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Hoeren_01

Man_02 $\xrightarrow{\text{is\_AGENT\_for}}$ Singen_01

Figure 29: Context for the sentence *Er hört die Männer singen.* (He is hearing the men sing) (Adapted from [68])

man (**Man_01**) hearing (**Etw.Hoeren_01**) other men (**Man_02**) that are singing (**Singen_01**)).

The result of the matching process would be Figure 30. In this case the word *Er* (he) is matched with the individual of a single man (**Man_01**) while the word *Männer* (men) indicating a plural is matched with **Man_02**, as this individual is an instantiation of the concept representing several man (**Man.pl**).

Whenever the conceptualization of an a-box-individual is compatible with a concept that is a probable referent for a given word or phrase, this individual is included in the set of possible referents for the sentence fragment. Although conceptual relatedness is included in this model, co-reference of the same concept always takes precedence. In a visual scene that contains an actor as well as another man, the word *Mann*(man) might refer to the concept **Man** as well as **Actor**. In this case, **Man** is the preferred choice for a referent as both visual individual and sentence-word are grounded in the same concept.

After grounding is completed, scores for semantic relations between grounded visual entities are computed and by this, between words of the input sentence. The basic idea behind the scores computed by the predictor is that thematic relations between visual entities influence the choice of dependency assignments in WCDG-parsing. The score for a dependency between two words is computed depending on five rules:

- All relations not backed by visual evidence are vetoed by assigning a pre-defined

Er $\xrightarrow{\text{matches}}$ {Man_01}
hört $\xrightarrow{\text{matches}}$ {Etw.Hoeren_01}
die $\xrightarrow{\text{matches}}$ {}
Männer $\xrightarrow{\text{matches}}$ {Man_02}
singen $\xrightarrow{\text{matches}}$ {Singen_01}

Figure 30: Matching the sentence *Er hört die Männer singen.* (He is hearing the men sing) to the context in Figure 29 (Adapted from [68])

penalty score to this relation. For instance an AGENT relation between the words *sing* and *song* would be penalized if the visual context states that the corresponding entities are related via the THEME relation

- If there is contextual evidence that the relation is a feasible one, the value one is assigned to the dependency
- The reverse direction of an relation for which there is evidence (e.g if *man* is AGENT for *sing*, *sing* should not be AGENT for *man*) is penalized.
- All other relations for the same dependent (e.g. if man is AGENT for *sing*, it should not be AGENT for another word or THEME for *sing*) are penalized.
- All other relations for the same regent (e.g. if *man* is AGENT for *sing*, no other word should be AGENT for *sing*) are penalized.

Once computed, these scores are returned to WCDG and are used by constraints to influence results of the parsing process. Using these constraints and the context in Figure 31 describing an INSTRUMENT relation between the visual entities **Fernglas_01** (Telescope) and **Sehen_01** (to see), the sentence *Der Mann sieht die Frau mit dem Fernglas.* (The man sees the woman with the telescope) is interpreted as the man using the telescope to conduct the seeing action on the woman by using the telescope as evident by the instrument edge in Figure 32 and the attachment of the prepositional phrase *mit dem Fernglas* (with the telescope) to the verb of the sentence indicated by the PP-edge of the tree.

$$\begin{array}{lcl}
\text{Man\_01} & \xrightarrow{\text{is\_AGENT\_for}} & \text{Sehen\_01} \\
\text{Woman\_01} & \xrightarrow{\text{is\_THEME\_for}} & \text{Sehen\_01} \\
\text{Fernglas\_01} & \xrightarrow{\text{is\_INSTRUMENT\_for}} & \text{Sehen\_01}
\end{array}$$

Figure 31: Context for the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope)

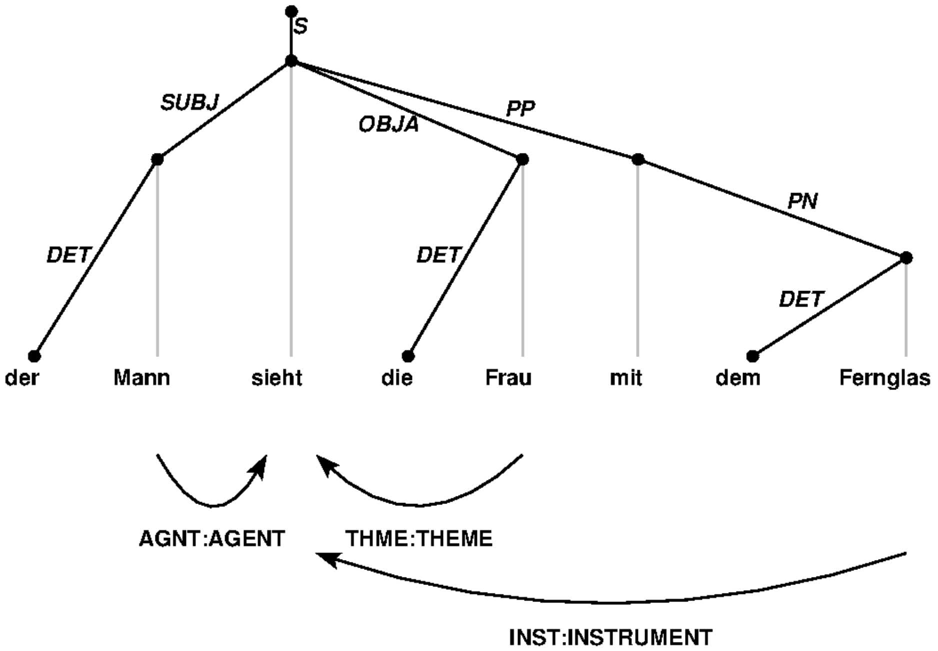


Figure 32: Dependency tree for the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope)

4.3 The Model of Visual Attention

Connecting knowledge derived from language input with a description of visual context mainly consists of finding an applicable set of visual referents for an utterance. But although humans do indeed tend to guide their attention to scenes described in what they are hearing (see Section 2), they do not process every part of the scene at the same time. Instead, the effects of language on human visual attention evolves over time, resulting in constantly changing the focus of attention to referents connected to parts of the utterance perceived. Thus, finding the set of possible referents is only one part of modeling the influence of language on human visual attention. To demonstrate the above mentioned effects we integrate a model of human visual attention. The model to be integrated should provide us with a means of displaying the current focus of attention at every point in time. As human language unfolds over time, the influence it has on human visual attention will also change. Therefore it is essential that the chosen model will be able to change the current focus whenever new information is received from the language channel. Ideally the model receives information that influences the current focus and changes the focus whenever new information comes up.

Our implementation of choice is the model of bottom-up visual attention proposed by [47]. The model predicts human eye-movements with regards to a picture, depending on visual properties found in that picture. The basic idea is that features such as brightness, color and shape influence which parts of a scene are attended to. The input picture is preprocessed, resulting in pictures of different sizes. These are the origin of information extracted from different channels. These channels include intensity, color information for red, green blue and yellow and orientation information of regions in the picture.

In the following we give an overview of the model by discussing every processing step from preprocessing of the initial input picture to the dynamic changes of the predicted focus of attention.

Step 1: Computing the Gaussian pyramid

In the description of a technique for image encoding ([14]) preprocessed the images used in order to compute what they called the Gaussian pyramid. The same method is used in this model. A Gaussian pyramid is computed using the original input picture (g_0) to derive a sequence of images ($g_1, g_2, ...$) from it. Each picture is computed by

Intensity:
$I = \frac{r+g+b}{3}$

Red:
$R = r - \frac{g+b}{2}/2$

Green:
$G = g - \frac{r+b}{2}$

Blue:
$B = b - \frac{r+g}{2}$

Yellow:
$Y = r + g - 2(|r - g| + b)$

r := red channel of the input image
g := green channel of the input image
b := blue channel of the input image

Figure 33: Formulas for the pyramids of intensity and color

applying a low-pass filter to the previous image of the picture. Thus, g_1 is computed by filtering input obtained from g_0, g_2 is computed from g_1, and so on. To stay put with the metaphor of the pyramid, the input picture would be the base of the pyramid, while each derived picture is another level up to the last one which would sit on top of the pyramid. Each pixel of an image g_{i+1} consists of the weighted average of a 5-by-5 window of image g_i. The computed pyramid in the model has 9 levels with scales from 1:1 to 1:256 compared to the size of the original.

Step 2: Computing the pyramids for the channels of intensity, color and orientation
From the 9 pictures of the Gaussian pyramid, three feature channels are derived.
The intensity and color pyramids are obtained using the formulas in Figure 33 resulting in 9 intensity images, one for each of the above mentioned scales and 9 color images for each color.

The information about orientation of features in the picture is generated using Gabor pyramids with four different orientations (0°, 45°, 90°, 135°). Again, the results are 9 different images for each of the four orientations.

Step 3: Computing the feature maps
[47] proposes a set of seven features to be computed from the different channels by comparing pixels on specific levels of the pyramids to their surrounding pixels on other levels. The choice of these features, it is argued, derives from evidence of existence of them in the visual system of mammals. One feature for contrast of intensity, one for red/green contrast, one for blue/yellow contrast and four for contrast regarding local orientation.
For each of these seven features, several maps are computed.

Intensity:
$I(c, s) = |I(c) \ominus I(s)|$

red/green:
$RG(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$

blue/yellow:
$BY(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$

orientation:
$O(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|$

c := level of the center with c $\in$ {2,3,4}
s := level of the surround with s = c + δ and δ $\in$ {3,4}
θ := orientation with $\theta \in$ {0°, 45°, 90°, 145°}
$\ominus$:= subtraction of two maps of different scales by interpolation to the finer scale

Figure 34: Formulas for computing intensity, contrast and orientation difference between pixels on level c and their surrounding pixels on level s

The number of maps for each feature depends on the pyramid levels of the pixels to be compared times the number of levels of the pixels compared to. The model compares the pixels on levels 2, 3 and 4 to their adjacent pixels 3 and 4 levels higher resulting in comparisons between levels 2-5, 2-6, 3-6, 3-7, 4-7, 4-8. Therefore 6 feature maps are computed for each of the seven features using the Formulas in Figure 34. In total, 42 feature maps are computed : six for intensity, 12 for color and 24 for orientation.

Step 4: Combining information across multiple maps
In order to predict attentional shifts, the feature maps are combined into three con-

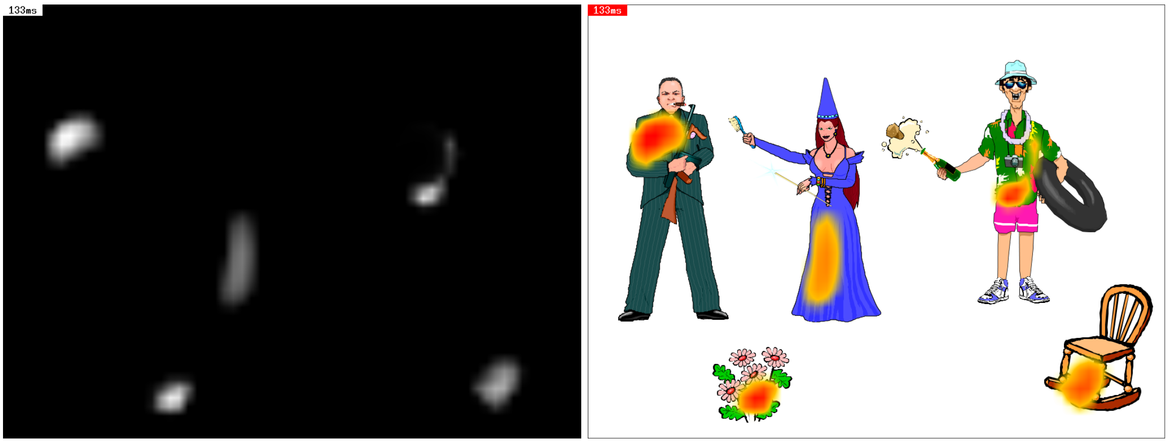


Figure 35: Saliency map (Picture taken from [54])

spicuity maps for intensity, color and orientation (see Figure 36). The normalization operator (denoted by the symbol N(.) in the formulas) indicates pre-processing of each map by a combination scheme chosen from four strategies (Naive Summation, Learned Linear Combination, Contents-based Global Non-linear Amplification, Iterative Localized Interactions). The reason for adjusting the maps by one of these strategies is the difficulty that is inherent when combining information from different information types: 'how should a 10° orientation discontinuity compare to a 5% intensity contrast?' ([47], p. 21). Therefore the data modeled by the maps is adjusted before combining them. These maps are finally combined into the saliency map by the formula in Figure 37. An example is given in Figure 35: on the left the saliency map shows the regions with the highest probability of being in focus as white patches. On the right, the picture that is the basis for computing the map is shown with the salient regions depicted as colored areas.

step 5: The generation of dynamic attentional focus
Deciding which region is the first focus of attention is straightforward: the region containing the highest saliency should be chosen. In the model, this is done by applying an implementation of the winner-takes-all(WTA) neural networks presented in [58]. After the WTA-algorithm selects a region as the maximum of the saliency map, the focus of attention(FOA) shifts to the corresponding region.
Using this procedure on a static picture will always result in a static saliency map with unchanging saliency for every region. The FOA in the saliency map would forever stay on the same region, no matter how much time passes. As this is not the desired

Intensity:
$\overline{I} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(I(c,s))$

Color:
$\overline{C} = \oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} [N(RG(c,s)) + N(BY(c,s))]$

Orientation:
$\overline{O} = \sum_{\theta \in \{0,45,90,135\}} N(\oplus_{c=2}^{4} \oplus_{s=c+3}^{c+4} N(O(c,s,\theta)))$

c := level of the center with c $\in$ {2,3,4}
s := level of the surround with s = c + δ and $\delta \in$ {3,4}
θ := orientation with $\theta \in$ {0°, 45°, 90°, 145°}
$\oplus$:= addition across scales by reducing each map to scale 4 and point-by-point summation
N(.) := normalization operator

Figure 36: Formulas for computing the conspicuity maps for intensity, contrast and orientation

behavior, inhibitory feedback is introduced. Whenever a region is the current FOA its saliency is diminished which will result in another region becoming the new maximum of saliency of the map. Consequently the WTA-algorithm will decide for this region as the new FOA in the picture. The inhibition is time dependent and the saliency will increase to its original value after approximately 500 to 900 ms. As a consequence, attentional shifts might return to one and the same region of the picture at different moments of processing the input.

S := $\frac{1}{3}(N(\overline{I}) + N(\overline{C}) + N(\overline{O}))$

$\overline{I}$:= Intensity conspicuity map
$\overline{C}$:= Color conspicuity map
$\overline{O}$:= Orientation conspicuity map

Figure 37: Formula for computing the saliency map

5 Architecture

We hereby introduce the means to model the behavior we derived from the cognitive findings discussed in Section 2. The implementation of the aspired model requires adjustments to each of the basic systems introduced in Section 4, as well as the creation of a component that realizes the exchange of information from different representations between the subsystems. The final model uses processing results of each of its components by constantly exchanging information between them (see Figure 38). The system finds a scene of reference that is applicable to given language input and uses the information contained therein to influence parsing results as well as showing the focus of attention in the visual field.

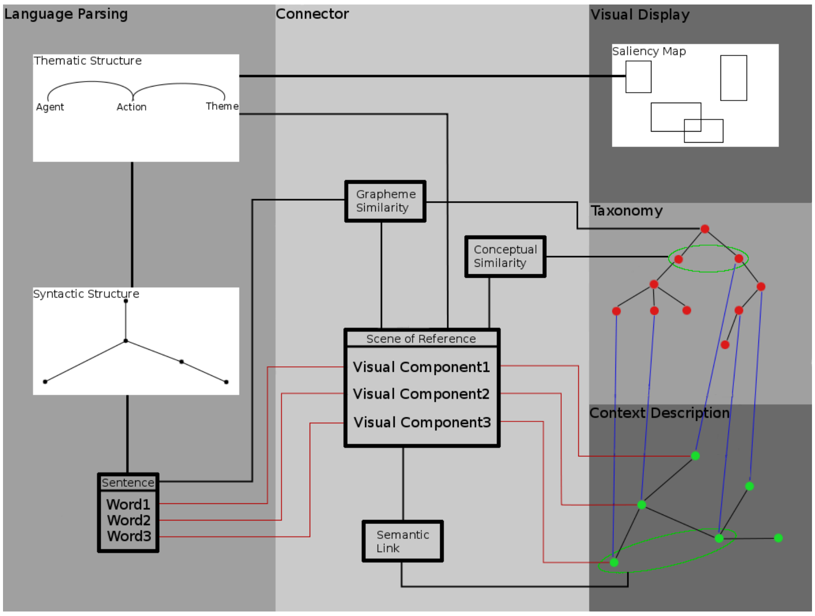


Figure 38: Overview of the architecture

Following the requirements defined in Section 2 the following extensions have to be made to the systems discussed in Section 4:

- Integration of visual information into the language parsing mechanism
  Following requirement 1, the architecture should include the means to integrate extra-linguistic information into the language parsing module.

Although the means to use visual context to enhance processing of WCDG already exists as a predictor, the parser has to be extended to access information from the context model at each online access of incremental processing, using different visual entities to influence parsing outcome whenever a new word is added to the fragment of the input sentence.

- Integration of changed information
  According to requirement 2, humans integrate context information at the earliest moment possible. Whenever contextual representation changes, the updated knowledge about visual percepts should be immediately made available to the language parser. This means that whenever the context model is changed due to new entities being introduced, old ones deleted or relationships between the visual entities changed, the language parser has to have access to these changes in order to incorporate them into processing.

- Finding referents for words of the sentence in an ambiguous context.
  To implement requirements 3 and 4, a link between words of the input sentence and entities in the context model has to be established. This link should be created for objects and their properties as well as processes defined by representations of actions in our context model. This link should even be established if several referents for a word or a phrase are present in the context.

- Interpreting words
  Due to requirement 5, incorrectly spelled words of the input sentence should not prevent the establishment of the above-mentioned link. Whether or not an erroneous written word still contains enough information to be linked to an entity should be dependent on the difference between the word and the entities lexicalisation.

- Anticipating upcoming referents
  According to requirement 6, the system should emulate the human ability to predict those visual entities that will be named by not yet received words of the input sentence.
  As we have seen in Section 4 the parser can predict missing elements of the sentence during incremental parsing by using virtual nodes as placeholder for words. Our model should connect virtual nodes to parts of the context model, by inferring possible referents from the analyses of a sentence fragment.

- Reference resolution by using different aspects of language processing
  Requirement 7 states that humans use information of the whole sentence to establish reference. Therefore, whenever several referents for a word have been found, the system should use disambiguating information derived from other parts of the sentence as well as from the processing results as given by the dependency structure that is the analysis.

- Integration of world knowledge
  In order to meet requirement 8, the system should use the information about conceptual relationships contained in our context model as a basis to integrate knowledge about the world into reference resolution.
  Possible referents are not only dependent on words of the sentence but also on the conceptual distance between a concept denoted by a word and the concept that is the type of an individual in the context model. The similarity between these two concepts influences the plausibility of the individual being a referent for the respective word.

5.1 WCDG

Changing WCDG to adhere to the desired effects is a two-fold process: firstly, visual context is used to influence the results of parsing. Secondly linguistic information is transferred to external models (i.e. the high level representation and the model of visual attention) to choose the correct subset of visual entities that are relevant for the input sentence or its fragment.

As we have seen in Section 4, the CDG formalism relies on a language model provided

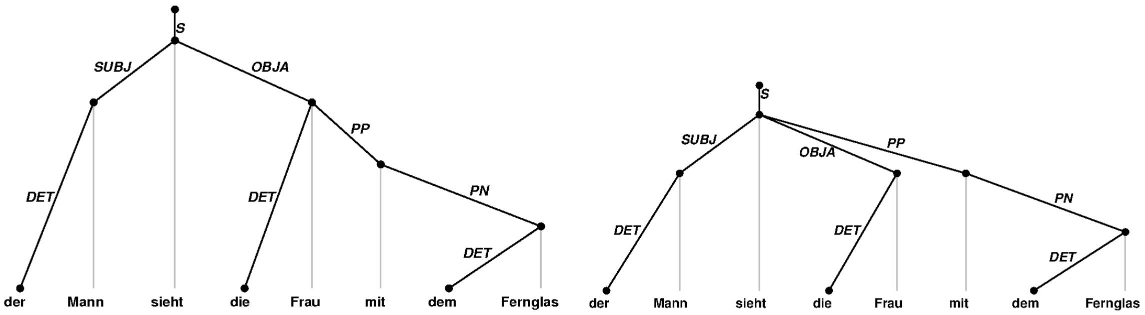


Figure 39: Different attachment of the PP-edge for the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope)

by the grammar specific to any language used. As our external context information is an additional factor that influences language, it is necessary to introduce structures into the grammar for the purpose of connecting context with language.
At first glance, an update of the existing predictor interface to the visual representation seems to be a suitable idea for this purpose. But this approach would ignore the fact that WCDG-predictors provide their results once at the beginning of processing a sentence. They are therefore incapable of reacting to the changes in a dynamic visual environment. Instead, we choose an approach that is capable to link any kind of represented external visual context to the language model contained in the grammar by evaluating constraints depending on specific contextual information.
In order to benefit from the context information accessible, the grammar is adjusted to influence the language processing for a certain domain of language expressions and fitting visual context. For that purpose the syntactic and semantic information of the dependency structure is connected to the visual data received from the context model.
Several new predicates and functions are introduced to accomplish this goal (see Table 1). These include methods for accessing information attached to visual entities referred to by the given sentence, methods to influence the choice of the scene of reference and methods to modulate the saliency in the attentional model.

Name	**Description**	**Example**
eye	Influences the behavior of the attentional model.	eye(X@id, 20) : Increases saliency of the referent for the node (X@id) using the weight 20
visual	Checks visual information represented in the context model.	visual(X@id, X^id, AGENT) : Checks the visual model for a relation between the referents of two nodes (X@id and X^id)
visualChange	Influences which referents are chosen depending on properties of the sentence.	visualChange(X@id, X^id, 5.2, 1, AGENT) : Influences the score of all scenes that have an AGENT relation between the referents of two nodes (X@id and X^id) by the factor 5.2
hasReferent	Tests if a word has a visual referent.	hasReferent(X@id) : Checks if a node (X@id) has a referent in the visual context
moveObject	Moves an object in the 3d-model.	moveObject(table_2, 1.7, 0, 1.6) : Moves the object (table_2) to the position (1.7, 0, 1.6) in the coordinate system of the 3d-model
enforceProperty	Specifies a property that a referent must have.	enforceProperty(num, X@number, X@id); : Enforces a value (X@number) for a property (num) of the referent of a node (X@id)

Table 1: New predicates and functions of WCDG

These methods are used to specify a new class of constraints (in the following called visual constraints) that model the task dependent link between a specific property of the target language (analyzed by WCDG in form of the dependency structure) and specific visual evidence that influences this property. Whenever these constraints are evaluated with regard to the intermediate parsing results (as described in Section 4), they access the context model and integrate the information contained therein to influence the evaluation of the constraint formula. By doing this, the visual scene described in our model directly influences the attachment of edges of the dependency structure. The constraints that realize this link of modalities are modeled according to the findings of Section 2. Therefore it becomes possible to create rules for the integration of spatial knowledge into the interpretation of prepositional phrases (like humans do according to [100],[94]), to assign thematic roles to parts of the sentence (see [54]) and to predict referents for upcoming words (see [52]).

A simple example for this integration into constraint evaluation would be the sentence *Der Mann sieht die Frau mit dem Fernglas* and its two possible interpretations with regard to PP-attachment (see Figure 39) that is parsed with a grammar including the visual constraint in Figure 40. Spelled out in natural language the meaning of this constraint would be : if two edges of the tree are connected on the syntax level of WCDG ($\{X!SYN\backslash Y!SYN\}$), one of which has the label PP, the other PN ($Y.label = PP\&X.label = PN$) and one node of each edge has a referent in the context model ($hasReferent(Y\hat{\ }id)\&hasReferent(X@id)$) then ($->$) the referents should be connected by a relation ($visual(Y\hat{\ }id, X@id, Y@word)$) denoted by a word($Y@word$). The effect of this constraint during evaluation depends on the situation present in visual context: if the referents are indeed connected by this specific relation, the constraint evaluates to true and no further penalty is imposed on this analysis. If, on the other hand, the context does not contain such a relation, the constraint weight (in this case 0.9) is included in the score of the analysis. The violation

```
{X!SYN\Y!SYN} : PP_attachment : 0.9 :
Y.label = PP&X.label = PN
hasReferent(Y^id)&hasReferent(X@id)
- >
visual(Y^id, X@id, Y@word) = 1;
```

Figure 40: Constraint expressing that a prepositional relation between words should be verified by evidence of a relation between the referents of those words in the context model

of this specific constraint influences the ongoing parsing process when the transformation based algorithm is used to repair violated constraints with the possible effect of reattaching the PP-edge, switching between the interpretations shown in Figure 39. Of course this alternative outcome might again violate the same constraint, depending on the visual model.

For the opposite direction where language information is transferred to an external module to influence contextual processing, further changes have to be made, as the predictor-interface discussed in 4 relies only on information about lexemes of words of the sentence, thereby discarding any intermediate parsing results modeled as the WCDG dependency trees. As the link between language and vision does not only depend on word level information, but also on interpretations of phrases (one example is the referential resolution depending on several adjectives referring to the same noun as investigated with regard to humans in [101]), we create an interface that is capable of exchanging much more information (i.e. analyses of sentences including choices of lexical entries for words and edges connecting them). The transfered information is utilized by the connector subsystem (see below) to carry out reference resolution.

Parsing results are one factor that influence the choice of referents for parts of the sentence. This is especially relevant if several referents for a word exist in the visual representation, but these visual entities differ with regard to how they are related to each other or with regard to their properties. For instance, in a sentence containing the words *woman* and *telescope* while two women are present in the context, one of which is the owner of a telescope, the presence or absence of a COMITATIVE edge in the dependency tree should influence which woman is chosen as the more plausible referent. To realize this influence of language processing upon the choice of referents the predicate ***visualChange*** and ***enforceProperty*** (see Table 1) to influence the scoring mechanism of a scene or to demand that a referent has a specific property (such as a color) respectively.

The two mechanisms (influencing language processing with visual knowledge and using sentence analyses to find referents) model the human inspired cross-modal interaction between language and vision only when not used as separate operations of our system, but tightly interwoven processes that constantly influence each other: if the language parser guides the system to referents that are not present in the context (for instance by suggesting relations between them for which no evidence exists in the model) this absence of the expected information should feed back to the parser,

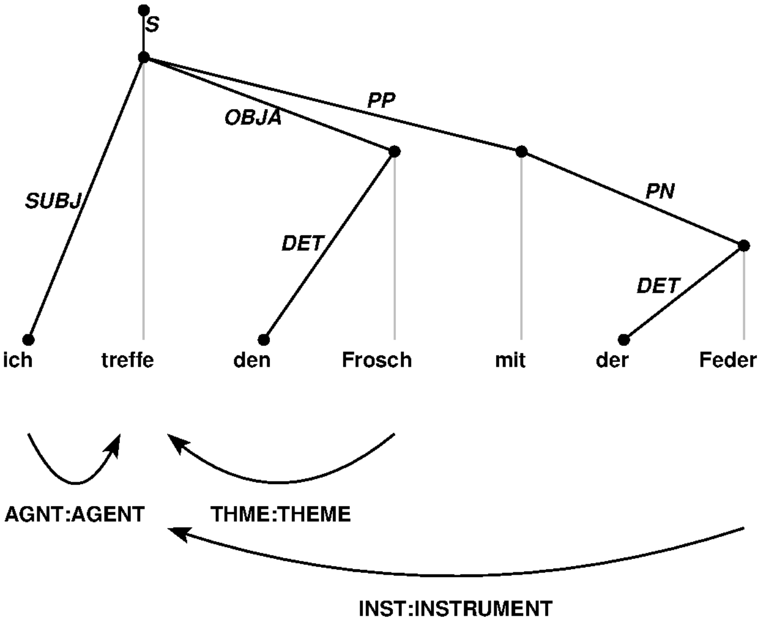


Figure 41: Dependency tree for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather) interpreting the prepositional phrase *mit der Feder*(with the feather) as the instrument of the striking action

resulting in a search for alternative sentence interpretations. If, on the other hand, visual knowledge guides the parser to interpretations that are highly unlikely given the language model, parsing results should lead to a change of the scene of reference. We will now give an example of the impact of language processing on the choice of a specific subset of contextual information using the attachment of prepositional phrases for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather) (Adapted from [94]). Humans choose an attachment due to several factors (see our discussion of [94] in Section 2). The parser decides if the phrase *mit der Feder*(with the feather) is either interpreted as the modifier to the noun of the sentence (interpreting the meaning as a frog that has a feather) (see Figure 42) or as the instrument of the verb (meaning that the action of striking the frog should be carried out with a feather that is present in the vicinity)(see Figure 41).

As we have seen above, constraints in the grammar can be used to choose one of these analyses according to contextual representations. Therefore, given appropriate visual constraints, the context in Figure 43 (containing a feather that is the COMITATIVE of a frog) guides the parser to the analysis in Figure 42, while the context in Figure 44

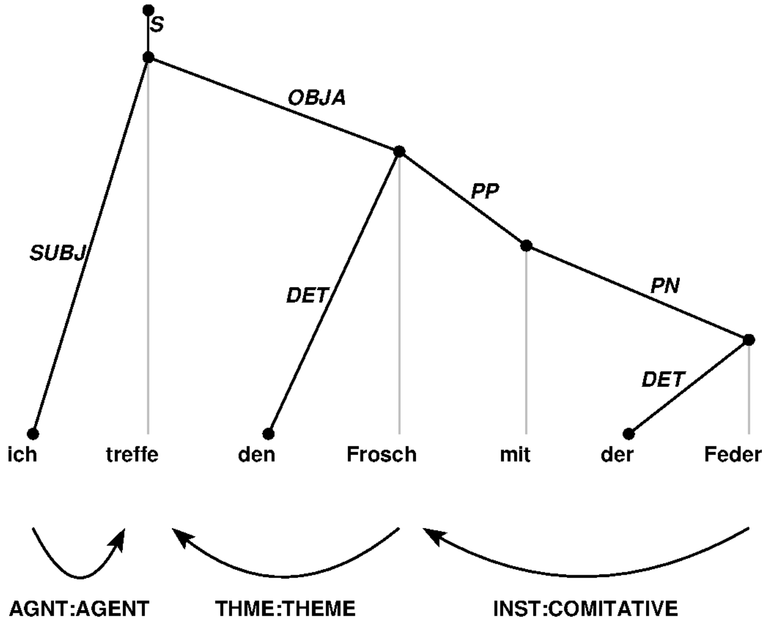


Figure 42: Dependency tree for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather) interpreting the prepositional phrase *mit der Feder*(with the feather) as the modifier of the frog

(containing a frog and a feather that are not related to each other) has no influence on the decision as it contains no visual information that suggests the feather being either instrument of an action or comitative of a character. If a context is introduced that includes visual evidence for one of the two interpretations, while at the same time containing entities that have no influence (see Figure 45 for a context containing two frogs, one of which owns a feather), the parsing results are dependent on the choice of visual entities as referents (i.e. if the frog that is not linked to the feather is chosen, there is no influence, if the frog that is the COMITATIVE of the feather is chosen, parsing is influenced).

If the parser suggests a COMITATIVE relationship between the words *Frosch*(frog) and *Feder*(feather), the referential component should choose the frog with the feather as the correct referent as this choice is favored by the linguistic inference made by our language processor. Two constraints are necessary to induce this integration: the first one (see Figure 46) is used to integrate a COMITATIVE relationship present in the context into parsing, while the one in Figure 47 is used to influence referential resolution according to an existing COMITATIVE edge in the dependency tree. The

$$\text{Frosch\_1} \xrightarrow{\text{is\_COMITATIVE\_for}} \text{Feder\_1}$$

$$\text{Frosch\_1} \xrightarrow{\text{has\_LOCATION\_auf}} \text{Tisch\_1}$$

Figure 43: Context for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather). A frog (**Frosch_1**) is the COMITATIVE of a feather (**Feder_1**) and is located on a table (**Tisch_1**)

effect of integrating these constraints into the grammar while using the ambiguous context in Figure 45 depends on parsing decisions made during incremental processing.

$$\text{Frosch\_1} \xrightarrow{\text{has\_LOCATION\_auf}} \text{Tisch\_1}$$

$$\text{Feder\_1} \xrightarrow{\text{has\_LOCATION\_auf}} \text{Tisch\_2}$$

Figure 44: Unambiguous context for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather). A frog (**Frosch_1**) is located on a table (**Tisch_1**) while a feather (**Feder_1**) is located on another table (**Tisch_2**)

If the parser suggests a COMITATIVE relationship between *Frosch*(frog) and *Feder* (feather), the system will favor the feather having a COMITATIVE relation in the context model (**Feder_1**) as a referent due to the influence of the constraint in Figure 47. Choosing this referent will reinforce the parsing decision for this particular semantic edge as any changes made in later parsing steps (like deleting the COMITATIVE edge in the tree) will result in violating the constraint in Figure 46. The overall effect is, that a parsing decision once made is less likely to change in later processing steps because evidence in the visual context strengthens the assumption already made.

If the parser suggests no COMITATIVE edge, both feathers represented are equally likely to be chosen as referents. In this case choosing **Feder_1**, having a COMITATIVE relation, will influence the parser to create a corresponding edge, while **Feder_2** will not influence language processing.

$$\text{Frosch\_1} \xrightarrow{\text{is\_COMITATIVE\_for}} \text{Feder\_1}$$
$$\text{Frosch\_1} \xrightarrow{\text{has\_LOCATION\_auf}} \text{Tisch\_1}$$
$$\text{Feder\_2} \xrightarrow{\text{has\_LOCATION\_auf}} \text{Tisch\_2}$$

Figure 45: Ambiguous context for the sentence *Ich treffe den Frosch mit der Feder* (I strike the frog with the feather). A frog (**Frosch_1**) is the COMITATIVE of a feather (**Feder_1**) and is located on a table (**Tisch_1**). A second feather (**Feder_2**) is located on another table (**Tisch_2**)

Although the examples presented use a simplified matching strategy by connecting information about semantic relationships found by the parser with semantic relations presented in the context model and then using semantic relations in the model as an influence on the attachment of semantic edges of the dependency tree, the methods in Table 1 can be used to link any type of information found in both representations. Thereby we can use this mechanism to match referents depending on attributes that are mentioned in the sentence (e.g. the analysis of the sentence *the red apple* can influence the matching process by choosing an **apple** as a referent that has the color property **red** even if there are a large number of green apples observable) or choosing attachments of syntactic edges depending of visual knowledge.

$$\boldsymbol{\{X!INST\}} : Connect\_COMITATIVE\_correct : \mathbf{0.5} :$$
$$\boldsymbol{X.label = COMITATIVE\&}$$
$$\boldsymbol{hasReferent(X@id)\&}$$
$$\boldsymbol{hasReferent(X\hat{}id)}$$
$$\boldsymbol{->}$$
$$\boldsymbol{visual(X@id, X\hat{}id, COMITATIVE) > 0.9;}$$

Figure 46: Constraint expressing that if the tree contains a COMITATIVE edge between two words that have referents, there should be evidence about a COMITATIVE relation between the referents in the context model

Furthermore the mechanism is not restricted to the constraint-formulas shown in the above examples. Information fusion can be done on a high level of complexity, making the influence of extra-linguistic information on edge-attachment either very

$\boldsymbol{\{X!INST\}} : Search\_referent\_for\_COMITATIVE : \mathbf{0.5} :$
$\boldsymbol{X.label = COMITATIVE}$
$\boldsymbol{->}$
$\boldsymbol{visualChange(X@id, X\hat{\ }id, 5.2, 1, COMITATIVE);}$

Figure 47: Constraint expressing that a COMITATIVE edge in the dependency tree should be linguistic influence that the referents for the words connected by the edge should also be in a COMITATIVE relation in the context model

general (e.g. any edge of a certain property is always influenced by a certain kind of visual evidence) or very specific (e.g. a certain attachment is only influenced under very narrow set of conditions, both with regard to edge configurations as well as representations of objects and relationships in the visual field).

5.2 Context Model

The component for the representation of visual information that we introduced in Section 4 was adjusted with regard to classes, their properties and relations between them. Changes were made to enable the representation of spatial relations, a possibility not available in the original model introduced in [68]. Furthermore, the choice of a subset of possibly ambiguous information fitting to the input sentence relies on disambiguating characteristics of the visual entities presented. Therefore a number of properties were defined (e.g. color, height) that can be used to select a specific individual out of several ones that belong to the same class.

The system also has to link the information present on this level to other external components such as the attentional model or a 3d-model of the world. Although the model already includes the link to the words of the input sentence (via the **has_Lexicalisation** property), we also link high level visual descriptions to visual models such as the visual attention component and the 3d-virtual universe.

There are two ways to model a spatial relationship between two individuals of the context model: either directly by connecting the entities by one of the relations (see Figure 48), which allows the description of binary spatial relations such as **left**, **right** or **above**, or by introducing the spatial relationship as an individual of a concept (see Figure 49), thereby describing higher arity relations.

The link to external components is realized by a description of the state of the ob-

$$\text{Mann\_1} \xrightarrow{\text{has\_LOCATION\_links}} \text{Auto\_2}$$

$$\text{Mann\_1} \xrightarrow{\text{has\_LOCATION\_rechts}} \text{Auto\_1}$$

Figure 48: Representation of a man (**Mann_1**) standing to the left of a car (**Auto_2**) and to the right of another car (**Auto_1**)

jects in their respective model. For pictures that are used as input of the attentional model the description consists of a representation of the bounding box that surrounds regions of interests (ROI) in the picture. The representation describes the shape of the area, its size and its position with respect to a two-dimensional coordinate system (see Table 2). These ROIs enable us to define regions for objects as well as actions. The example in Figure 51 shows the representation of the region in Figure 50.

The connection to the 3d-model is realized by linking each visual individual in the context model to a 3d-object by a unique name. The difference between these 3d-descriptions and the above-mentioned ROI is the dynamic adjustment to changing object positions. A ROI is static with respect to a specific picture, while the objects of the 3d-model can move. Therefore the latter representation is subject to changes of the coordinate representation as well as the descriptions of spatial relationships between objects.

Name	**Description**
has_LOCATION	Relates two visual entities by a spatial relation. Includes sub-relations for specific relations(such as is_LOCATION_an_for, is_LOCATION_auf_for). Inverse to is_LOCATION_for.
is_LOCATION_for	Relates two visual entities by a spatial relation. Includes sub-relations for specific relations(such as has_LOCATION_an, has_LOCATION_auf). Inverse to has_LOCATION
picture_height	Describes the height of a region in a picture
picture_rotation	Describes the rotation of a region in a picture
picture_type	Describes the shape (rectangular, circular, oval) of a region in a picture
picture_width	Describes the width of a region in a picture
picture_x	Describes the x-position of a region in a picture
picture_y	Describes the y-position of a region in a picture

Table 2: New elements of the context model

$$\text{Mann\_1} \xrightarrow{\text{has\_LOCATION}} \text{zwischen\_1}$$

$$\text{Auto\_2} \xrightarrow{\text{is\_LOCATION\_for}} \text{zwischen\_1}$$

$$\text{Auto\_1} \xrightarrow{\text{is\_LOCATION\_for}} \text{zwischen\_1}$$

Figure 49: Representation of a man (**Mann_1**) standing between (**zwischen_1**) two cars (**Auto_1** and **Auto_2**)

5.3 Visual Attention

In the following paragraph we will describe how our system connects language information with the bottom-up model of visual attention. The pictures used are described in our context model by means of higher level representation. This description contains the individual entities as well as spatial and semantic relations between them.

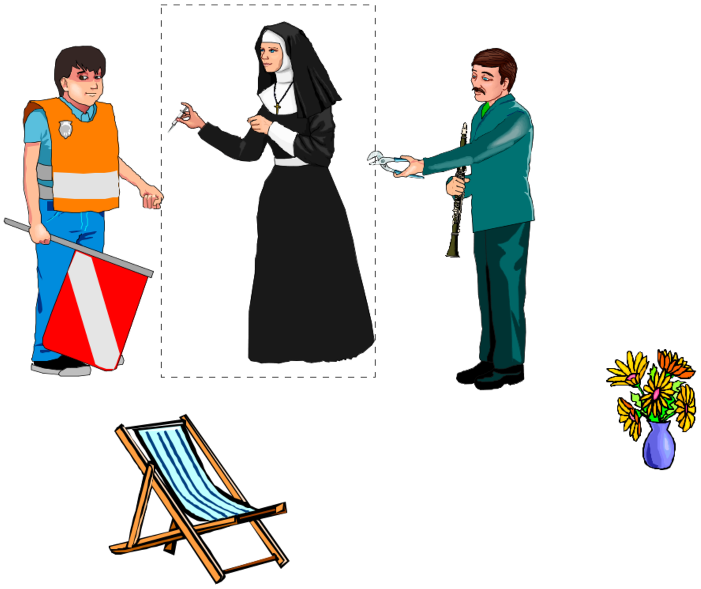


Figure 50: Bounding box surrounding part of a picture where a possible referent is located

The system uses constraints of WCDG to change the saliency of the described region. At crucial points of processing a sentence, the saliency of a region can be influenced

X_coordinate	→	17
Y_coordinate	→	2
height	→	26
width	→	16
type	→	rectangle

Figure 51: Example of the representation of the bounding box seen in Figure 50

by the predicate ***eye*** (see table 1) which increases or decreases the saliency of the referent connected to a specific node of the dependency tree of the processed sentence. For instance to increase the saliency of the region containing the THEME of an action denoted by a verb of a sentence the constraint in Figure 52 would be used. This constraint states that if an edge with the THEME label has a node as a dependent that has a referent in the context, saliency for the corresponding region of that referent should be increased. This has a twofold effect on the saliency map. First of all, the saliency in the corresponding region is increased and secondly saliency for every other position is decreased (see the formula in Figure 54). Three factors affect saliency. The first one is the saliency (S(p)) previously assigned by the bottom-up model of visual attention due to low level cues contained in the picture. Incorporating bottom-up cues has the effect that visual effects not addressed by the language input still have an impact on the saliency landscape.

The second factor that influences the saliency landscape is its global maximum

$\boldsymbol{\{X!AGNT\}}$: $Increase\_saliency\_for\_AGENT\_of\_an\_action$: $\mathbf{0.95}$ **:**
$\boldsymbol{X.label = THEME}$ **&**
$\boldsymbol{hasReferent(X@id)\ -\ >}$
$\boldsymbol{eye(X@id, 20));}$

Figure 52: Constraint expressing that the saliency for the THEME of an action should be increased

(GM). We introduced this factor to adjust the influence depending on the current saliency distribution of the input picture. The effect of using the global maximum is that the increase or decrease of saliency is more pronounced, the higher the global maximum of the saliency map is.

The third factor is the value specified in the WCDG **eye**-predicate (W). As the intended change in saliency is highly dependent on decisions made by the grammar modeler (such as domain specific reasons or influence of specific linguistic phenomena), this value can be freely assigned.

An example of using a grammar containing the constraint in Figure 52 when parsing the sentence *Die Fee bürstet hier den Gangster*(Literally: The fairy(ambiguous) brushes here the gangster(object)) can be seen in Figure 53: The initial distribution of salient pixels (marked as red and yellow areas) shows several regions as likely to be in focus (see the left picture). After the system finds a referent for the word *Gangster* and identifies the thematic role as the THEME of the described action, saliency is increased for the corresponding region and decreased everywhere else.

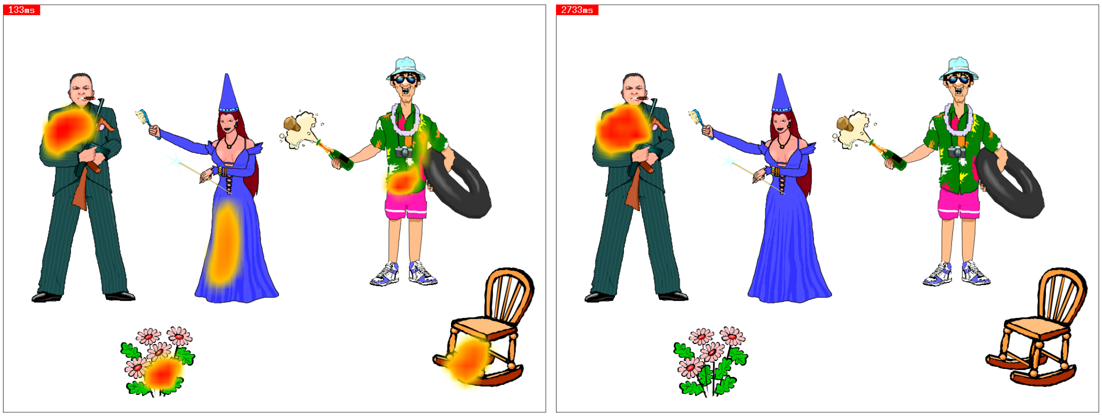


Figure 53: Saliency before and after processing the sentence *Die Fee bürstet hier den Gangster*(Literally: The fairy(ambiguous) brushes here the gangster(object)) (Underlying picture adapted from [54])

5.4 The Visual Universe

Our system applies the results of the modality interaction not only to the bottom-up model of visual attention but also to a three dimensional model that represents the visual context (see Figure 55). With this model it is possible to show a wide range of situations containing a large variety of visual objects belonging to different classes. The 3d-model is connected to a representation of what is in the current field of view.

Formula for adjusting saliency for a pixel in a region specified by the language parser:

$$S(p) = S(p) * W * GM$$

Formula for adjusting saliency for a pixel not in a region specified by the language parser:

$$S(p) = \frac{S(p)}{(W*GM)}$$

S(p)	Saliency of pixel p
W	The weight of the eye predicate specified by the constructor of the constraint
GM	The global maximum of the saliency map

Figure 54: Formulas for computing the saliency of a picture depending on linguistic influence

For instance the scene in Figure 55 is described in the form presented in Figure 56, showing the objects and their spatial relationship.

The 3d-model is not a static representation but can dynamically change. First of all, the angle of view can change according to the actions of a user moving in the three-dimensional universe. This change invalidates the description in the context model as spatial relationships are dependent on the position of visual entities to the observer (i.e. The truck (**Truck**) in Figure 55 is behind the policeman (**Cop**). Looking at the scene from the other side it would be in front). Also, a change in position of the observer might result in objects entering the field of view, while others are not in sight any more.

The second dynamic effect is the movement of objects observed. Scene-objects can move in the universe, changing their positions over time. Again, this might change spatial relationships between them as well as making them observable or unobservable. Both kinds of changes require an update of the context model at specific moments of processing visual input.

The process of updating the context model can be triggered by a user or during processing of an input sentence (like moving an object at a specific point of incremental parsing). The field of view is considered to contain each object situated in a cone with an angle of 45 degrees. The spatial relationships between objects are computed by transforming the global coordinates of an object into a coordinate system centered on the observer position. In this system the distance with regard to each of the three axes is correlated with a specific set of spatial relationships (i.e. the distance of the z-axis corresponds to the above/under-relations, the y-axis corresponds to in front of/behind and the x-axis to left/right). As objects should not be identified as spatially related if they are very close (for instance two objects in contact) or very far apart (which an observer would consider to be unrelated) objects are only considered to be related if the distance falls in a pre-defined margin.

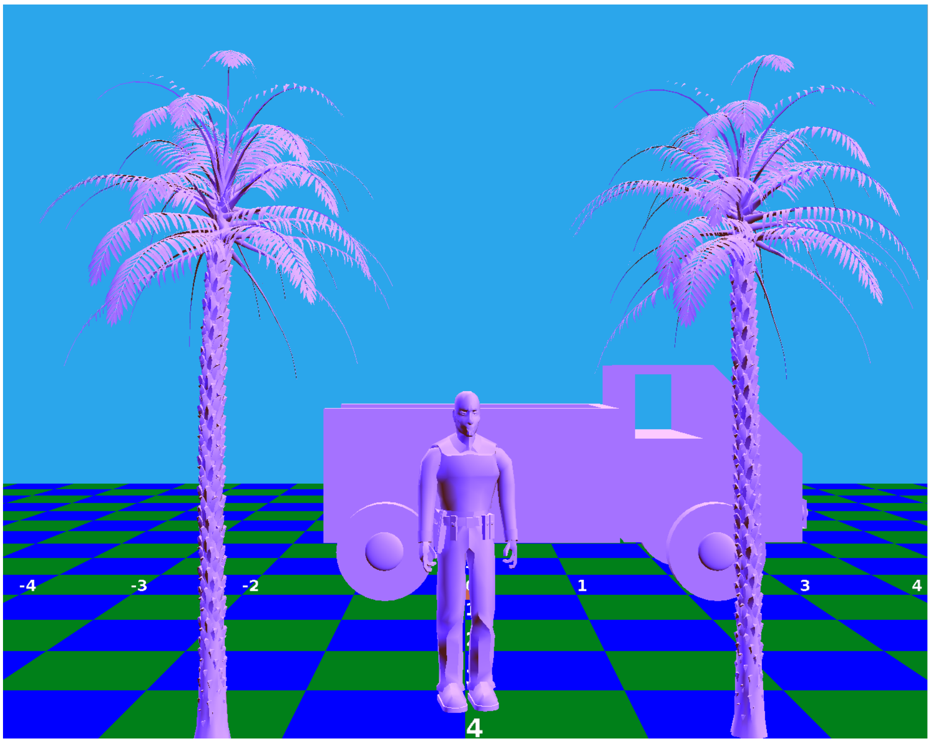


Figure 55: Example of a three-dimensional scene

5.5 Connector

The task of the connecting component (in the following called Connector), is to link linguistic and context information. This mapping can be defined as establishing two highly dynamic processes where information changes over time in both domains. Linking the specific processes of our model is not trivial since the highly different modes

Cop	$\xrightarrow{\text{has\_LOCATION\_vor}}$	Truck
Cop	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	Palm_002
Cop	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	Palm_001
Truck	$\xrightarrow{\text{has\_LOCATION\_hinter}}$	Cop
Truck	$\xrightarrow{\text{has\_LOCATION\_hinter}}$	Palm_002
Palm_001	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	Cop
Palm_001	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	Palm_002
Palm_002	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	Cop
Palm_002	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	Palm_001
Palm_002	$\xrightarrow{\text{has\_LOCATION\_vor}}$	Truck

Figure 56: Representation of the 3d-scene in Figure 55 containing two palm trees (**Palm_001** and **Palm_002**) a policeman (**Cop**) and a truck (**Truck**)

of change have to be considered. Information from both modalities changes over time but in a very different way. A sentence unfolds word by word. Words already received will not change any more (although syntactic and semantic analyses of them might). Changes of visual information on the other hand are not restricted to a certain 'place' of change (unlike in the language domain where new information is attached to the end of the sentence fragment) and the context might even return to an earlier state, where relationships between entities are the same as moments before. These different dynamics make establishing the cross-modal link difficult. Neither can we assume that newly received information from the two channels is always mapped to each other (for instance connecting new words with visual information that is newly received) nor can it be taken for granted that any connections made due to conclusions derived from one domain remains unchanged as long as these conclusions still hold (e.g. if the parser assigns a specific role to a word of a sentence, the referent of this word might change even when its role does not).

We use four different strategies to find the correct referent for a sentence-word:

5.5.1 Grapheme Similarity

In accordance with findings discussed in Section 2 we introduce a means to measure how much a given word of the language input differs from the lexicalisation of a possible referent. Although such differences might occur on the phonetic level, we simulate such effects on the grapheme representation as we do not model pronunciation and phonetic similarity in our system. Several different measures for comparing strings exist. The basic idea of each of these measures is to compute the number of operations necessary to transform the first string to be compared with the second one. These measures differ in the possible operations that can be applied to transform a string. According to [71] four different operations are possible

- Insertion: Inserting a letter, for instance *Tage* changes to *Trage*
- Deletion: Deleting a letter, for instance *Brett* changes to *Bett*
- Substitution: Replacing a letter, for instance *Baum* changes to *Raum*
- Transposition: Swapping adjacent letters, for instance *Keil* and *Kiel*

Examples include Levenshtein distance [62] which allows insertions, deletions and substitutions, Hamming distance [85] which allows only substitutions, Episode distance [23] uses only insertions and Longest common subsequence distance [72] that uses insertions and deletions. For our model we use the normalized Hamming distance to compute the similarity between a word of the sentence and the lexicalisation (see the formula in Figure 57).

We favored the Hamming distance over the presented alternatives because we believe that it is the most viable measurement for modeling the effects described in Section 2 due to its restriction to substitution as the only operation allowed to transform a string. Measures that use the other three operations (insertion, deletion and transposition) will allow matches that are not in agreement with the effects found in any of the studies of the matching effect.

The Hamming distance allows words to be matched to their respective cohort competitors (like *Band* and *Bank*) as well as rhyme competitors like (*Band* and *Rand*). Introducing the other three operations would, admittedly, result in even more matches, especially with regards to rhymes (for instance *Raum* and *Traum* are only matching when the similarity measure allows the insertion of a letter at the beginning of the

$$GS = 1 - (\frac{diff+|wordlength-lexlength|}{max(wordlength,lexlength)})$$

GS	grapheme similarity
diff	number of characters that are different
wordlength	number of characters in the word of the sentence
lexlength	number of characters in the lexicalisation

Examples:
The distance between *Buch* and *Tuch* is 0.75
The distance between *Tag* and *Nacht* is 0.2
The distance between *Tag* and *Tagung* is 0.5
The distance between *Tag* and *betagt* is 0

Figure 57: Formula and examples of the normalized Hamming distance

word). The disadvantage of including these operations would be a high similarity between words that are probably neither rhyme nor cohort competitors for the target word. An example of this would be the high similarity of the words *Keil* and *Kiel* or *Braten* and *beraten*. Thus, allowing additional operations would increase the recall (matching more often in cases where a match is desirable), but at the same time, precision would be reduced (as we would also match more words that should not be matched). As we use several strategies to find correct referents (increasing the likelihood of finding a fitting one) while at the same time applying the system to domains that are potentially containing objects belonging to a large quantity of different types, we are more concerned with finding fitting referents than with finding each and every one of those that are possible.

5.5.2 Conceptual Similarity

To meet the requirement 8 of including possible referents depending on conceptual similarity (Section 2) we need a measure to compare the similarity of two different concepts depending on a concept taxonomy. The finding of [44] that humans are more likely to choose the **fork** than the **umbrella** as a referent for the word *beaker* is only evidence that the concept **beaker** is more closely related to the concept **fork** without giving any clue about the degree of relatedness. Therefore, in order to model this behavior, the chosen measure should feature the means to compute values for the similarity between concepts that can be compared to each other, but these values are not related to a human impression of the similarity between two things. In other words, whether the similarity between **beaker** and **fork** is 0.1 or 0.2 does not matter because neither of these values is representing some kind of real-world similarity. Although promising approaches to model human-like concept similarity exist (see [20] for one example of a measure based on feature norms and [43] for its use), it is very complicated to create such a metric for many concepts. As the authors stated '[...] it is difficult to derive a measure of semantic similarity for a large number of concepts. Semantic similarity has typically been measured subjectively by having participants rate pairwise similarity of a set of concepts when presented with the concepts' names.' ([20], 185). As our model should be adaptable to different domains easily, we use the similarity measure based on taxonomic distance instead.

[13] reviewed several measures: Rada et al. distance [77], Resnik similarity [82], Leacock and Chodorow similarity [61], Wu Palmer similarity [105], Jiang conrath distance [48], Lin similarity [64], Sussna distance [99], Hirst St Onge relatedness [40]. According to [13] each possible measure of distance or similarity of concepts is based on one or more parameters

- The length of the shortest path
  As every taxonomy of concepts can be seen as a tree with concepts as nodes and relations between them as edges, it is possible to compute the shortest path between two arbitrary concepts. The length of this path influences some of the existing measures

- Depth of the most specific common subsumer
  If two concepts have a common ancestor-concept (i.e. the **Animal** as ancestor of both **Cat** and **Dog**), the length of the path between this subsumer and the root of the concept-tree influences some of the measures.

- Density of concepts of the shortest path
  The number of subconcepts of each concept on a path that connects two concepts

- Density of concepts from the root to the most specific common sub-sumer
  Number of subconcepts on the path from root to the most-common ancestor of two concepts

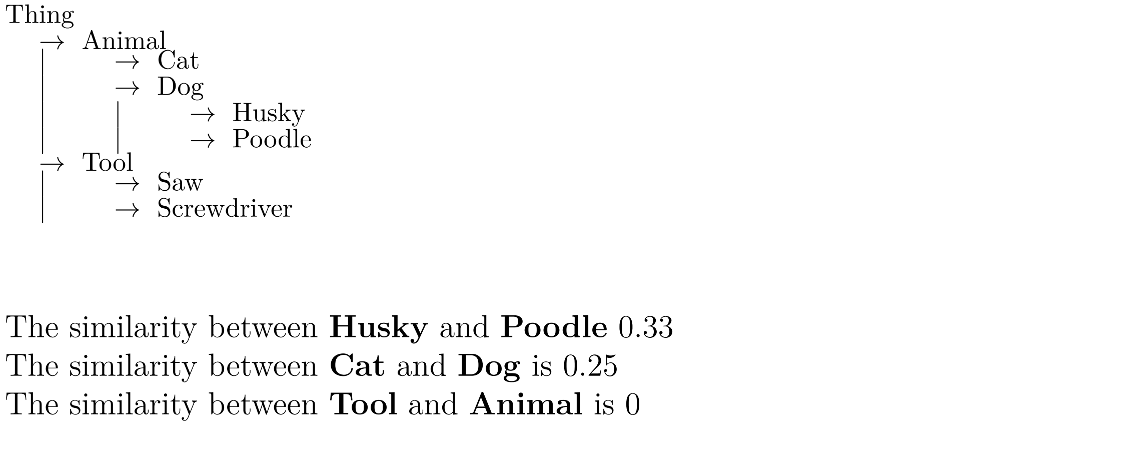
Observing the parameters outlined above, we can partition them into two classes: either a parameter is affected by path length (i.e. number of concepts between two classes), or by density of concepts (i.e. the number of subconcepts a concept on a path has). Our task-dependent ontology is likely to change whenever it is adjusted to a specific kind of visual context. We expect that these adjustments will affect first of all the density of concepts. This will be the case whenever we introduce a new class of visual entities that can appear in the context (e.g. if we adjust the ontology in Figure 59 by including the concept **Horse**, the density of **Animal** will change). The path-length, on the other hand, is less likely to change. Restructuring the ontology (like introducing a new concept that is located between **Animal** and **Dog**), is usually not the consequence of adjustment to a specific kind of visual environment. If we model our view of the world by stating that a **Dog** is a direct descendant of **Animal**, without an intermediate concepts like **Mammal**, this modeling decision, even if it is biologically inadequate, is likely to pertain to all tasks.

These considerations about the dynamics of ontology structure are important with regard to a decision for a specific measure, because the conceptual similarity in our system is an important factor for reference resolution. Thus, a similarity that changes with each new task is likely to change the outcome of reference resolution, even if used to map the same sentence/visual-scene pair. Due to these considerations, we choose the Wu-Palmer similarity for our model as it relies only on path length, discarding density as a parameter. The similarity is computed by the formula in Figure 58 depending on the distance of the two concepts to a common superconcept shared, as well as on the distance of this common ancestor to the root of the hierarchy of concepts (see Figure 59 for an example taxonomy and the results of comparing two of its concepts).

$$CS(C1, C2) = \frac{2*N3}{N1+N2+2*N3}$$

CS	conceptual similarity
C3 (not in formula)	least common superconcept
C1, C2	concepts to be compared with regards to similarity
N1	number of concepts from C1 to C3
N2	number of concepts from C2 to C3
N3	number of concepts from C3 to the root of the hierarchy

Figure 58: Formula for conceptual similarity



The similarity between **Husky** and **Poodle** 0.33
The similarity between **Cat** and **Dog** is 0.25
The similarity between **Tool** and **Animal** is 0

Figure 59: Example for computing the similarity between concepts

5.5.3 Semantic Link

One factor for choosing a referent is its semantic relatedness to other referents. Visual entities are more likely to be chosen as referents for words or phrases, if they are spatially or semantically related to entities already chosen as referents.
To model this, referents are penalized if they are not connected in any way in the representation of the visual context and reinforced if a connection exists. iven the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope) and the context in Figure 60, the choice for the word *Frau*(woman) can be (**Frau_1** and **Frau_2**) as they both have the same lexicalisation, they are probable as referents for the word taking only the conceptual and grapheme similarity described above into account. The most likely candidates for the words *Mann*(man) and *Fernglas*(telescope) are the individuals **Mann_1** and **Telescope_1**. As **Frau_1** relates to **Fernglas_1** via the COMITATIVE-relation(and is thus connected to another referent), it is preferred over **Frau_2** as a referent.

5.5.4 Linguistic Influence

We meet the requirement of linguistic influence by introducing means to guide the reference resolution depending on cues extracted from the sentence.

Frau_1 $\xrightarrow{\text{instance\_of}}$ Frau

Frau_2 $\xrightarrow{\text{instance\_of}}$ Frau

Mann_1 $\xrightarrow{\text{instance\_of}}$ Man

Fernglas_1 $\xrightarrow{\text{instance\_of}}$ Telescope

Frau_1 $\xrightarrow{\text{is\_COMITATIVE\_for}}$ Fernglas_1

Figure 60: Context for the sentence *Der Mann sieht die Frau mit dem Fernglas* (The man sees the woman with the telescope) containing two women one that carries a telescope (**Frau_1**) and one that does not (**Frau_2**)

```
X:SYN : 'ATTR can be dataproperty' : visual : 0.9 : (visual) :
X.label = ATTR
->
visualChange(X^id, X@id, 5.2, 1);
```

Figure 61: Constraint expressing that an ATTR dependency between words is a cue for a relationship between the referents found in the context model

Modeling these cues is done by constraints of the WCDG-grammar containing the predicate **visualChange**. Whenever these constraints are evaluated by an analysis of a sentence the referential resolution is influenced by either inhibiting or amplifying the possible choice of a specific visual entity for a given word. An example of such a constraint can be seen in Figure 61 that expresses the influence of an ATTR edge (indicating a prenominal attribute) on the choice of the set of referents for the processed sentence. In a case of a dependency structure that includes such an edge between two words, the system adjusts the score of scenes that include related referents for these words according to the number specified in the **visualChange**-predicate (in this case 5.2). The result for the example sentence in Figure 62 would be that a scene of reference is more likely to be chosen if it includes a red car and not one that includes a car and something else that is red.

The value specified in the predicate is incorporated into the overall score of the scene according to the formula in Figure 63. Another way to influence referential resolution with the results of language parsing is the predicate **enforceProperty** which deletes

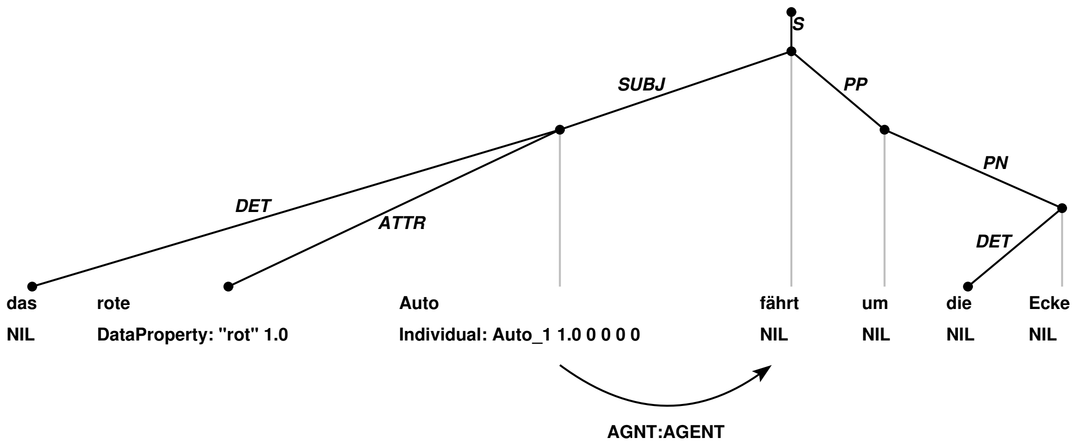


Figure 62: Dependency tree for the sentence fragment *Das rote Auto fährt um die Ecke*(The red car is driving around the corner)

all referents that do not have the specified property. Using this predicate in the constraint in Figure 61 instead of **visualChange** would force the system to discard all possible sets of referents that do not include a red car.
Using these predicates, constraints that implement human inspired top-down influence of language on reference resolution (for instance the effects of the interpretation of prepositional phrases on the choice of referents as in [94]) can be created.

$constr = (1 + \frac{predicateScore}{100})$

constr	The score assigned to a visual entity describing its quality as a referent for a given word
predicateScore	the score specified in a WCDG-constraint for this specific linguistic effect

Figure 63: Formula for linguistic influence

5.5.5 Finding a Scene of Reference

Applying the measures discussed above, the system finds a set of referents for a given sentence. This scene of reference displays those parts of the visual surroundings, that are addressed by the utterance under consideration.

$$V(S) = \prod_{i=0}^{n}(\Theta(GS(x_i), CS(x_i)) * conn(x_i) * constr(x_i))$$

V(S)	value of scene S
n	number of words in the sentence
x_i	referent for word at position i
GS(x)	the similarity of the referent compared to the concept denoted by the word
CS(x)	the distance between the word at position i and the lexicalisation of its referent
conn(x)	a value expressing the connectedness of this referent to the other referents
constr(x)	a value expressing the influence of constraint evaluation induced by the parser upon this referent
$\Theta(a, b)$	Selects either the grapheme similarity (GS(x))) or the conceptual similarity (CS(x)), depending on which is applicable. If the word directly denotes the concept of a visual entity (the word *dog* while a **dog** is in the model), grapheme distance is chosen. If the word denotes a concept that is related to visual entity (the word *dog* while an unspecified **animal** is in the model) conceptual distance is chosen.

Figure 64: Formula for grading a scene of reference

Construction cycle of referential resolution

1. For each word, possible referents are found comparing the word with the lexicalisation of the concept. If this concept has individuals of its type in the context, these visual entities are added to the set of possible referents for this word. A weight is attached to each visual object, based on the string matching distance described above.

2. If concepts with a high similarity with regard to the conceptual measure (see above) have individuals, these individuals are added to the set of referents. Their score depends on the distance between the concepts.

3. From these sets of referents for each word, every possible scene is built. These scenes are graded by means of the formula in Figure 64.

4. The scene graded with the highest score is deemed to be the best fitting one and is used as the scene of reference for the sentence at this point in time. The visual entities of this scene are then used to influence the evaluation of visual constraints as described above.

5. Any change invoked during parsing (when a new word that is added to the unfinished sentence or an improved analysis is found) restarts the search for a scene of reference.

6 Evaluation

In this section we evaluate our model using a range of different experimental setups. We will first discuss how to measure the quality of results of these experiments by discussing the metrics employed in our evaluation. After that we employ these metrics to analyze the results of several experiments.

6.1 Evaluation Metrics

The quality of the integration of both types of information depends on the effects reference has on the representation of modality-specific content. In the following section we will discuss how we measure these effects by introducing different metrics.
The system uses information provided by the language parser to link words of the sentence with entities in the context description. This choice of referents has to be evaluated with regard to its quality. We therefore introduce metrics that show the qualities of a given reference choice by adapting the work of [87].
In order to compute these measures we have to create a standard to which we compare our results. This should consist of the correct referents for each word of a sentence or a label that denotes that no referent exists for this word. We have two options to create such a gold standard.

(a) Create our own set of reference objects for a given input sentence
Using this option we create a set of referents for each used sentence, linking a word of the sentence with a certain visual entity (if a referent for this word exists) or stating that a word has no correct visual referent (by assigning the label NIL to it)

(b) Produce a standard using the module for reference resolution
This approach creates sets of referents for a sentence by processing the input in conjunction with a very restricted context. The used context contains only referents that are easily linked to a given word. To give a simple example, processing the sentence *Die Prinzessin malt offensichtlich der Pirat*(Literally: The princess (subject/object) paints apparently the pirate (subject).) 'The princess is apparently painted by the pirate.' while using a context containing only one pirate(see Figure 65) would result in a non-ambiguous link between the word *Pirat*(pirate) and the corresponding entity.

Name	**Formula**	**Description**
average first correct	**fc:=** $\frac{\sum_{i=1}^{m}\frac{\sum_{j=1}^{n_i}\frac{first_{ij}}{maxPos_i}}{n_i}}{m}$ $first_i$:= position of first correct guess for word j in sentence i m := number of sentences n_i:= number of words in sentence i that are referring to visual entities $maxPos_i$:= number of words in sentence i	This shows at which parsing step the system chooses the correct referent for a given word. The value is the relation of the step number of the first correct guess for a given word and the absolute number of parsing steps computed.
first correct applicable	**fca:=** $\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_j$ m := number of sentences n_i:= number of words in sentence i that are referring to visual entities $x_j := \begin{cases} 1 & \text{if correct referent for word j is found} \\ 0 & \text{if correct referent for word j is not found} \end{cases}$	This numbers specifies how often the average first correct (s.a.) is applicable to a specific word/referent combination. The average first correct is only computed when the correct referent is found at least once during parsing. The first correct applicable is the number of times a correct referent was assigned at least once for a specific word.
average first final	**ff:=** $\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{final_{ij}}{maxPos_i}}{m}$ $final_i$:= position of final correct guess for word j in sentence i m := number of sentences n_i:= number of words in sentence i that are referring to visual entities $maxPos_i$:= number of words in sentence i	This measure shows at which parsing step the correct referent for a word is found, without subsequently changing this decision. As with average first correct, the value is the relation of the parse step of the first correct guess for a given word and the absolute number of parsing steps.
first final applicable	**ffa:=** $\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_j$ m := number of sentences n_i := number of words in sentence i $x_j := \begin{cases} 1 & \text{if correct referent for word j is found} \\ & \text{and not changed later} \\ 0 & \text{if correct referent for word j is not found} \\ & \text{or is found but changed later} \end{cases}$	As for first correct applicable, this measure specifies how often the average first final is applicable to a word/referent combination. It shows the number of times, the correct referent is assigned, without being changed later during parsing.
mean edits per utterance	**eu:=** $\frac{\sum_{i=1}^{m}\frac{\sum_{j=1}^{n_i}\frac{last_i - first_i}{maxPos_i}}{n_i}}{m}$ $first_{ij}$:= position of first correct guess $last_{ij}$:= position where correct referent was found last m := number of sentences n_i:= number of words in sentence i that are referring to visual entities $maxPos_i$:= number of words in sentence i	The system might change the decision for a referent for a given word even after it already found the correct referent. The mean edits per utterance measures how many parsing steps the system needs to decide again for the correct referent.
edit overhead	**eo:=** $\frac{\sum_{i=1}^{m}\frac{\sum_{k=1}^{n_i}\frac{\sum_{j=1}^{l_i} nec}{\sum_{j=1}^{l_i} unnec}}{n_i}}{m}$ $nec := \begin{cases} 1 & \text{if referent was changed} \\ & \text{to correct one} \\ 0 & \text{else} \end{cases}$ $unnec := \begin{cases} 1 & \text{if referent was changed} \\ & \text{to incorrect one} \\ 0 & \text{else} \end{cases}$ m := number of sentences l_i:= number of words in sentence i n_i:= number of words in sentence i that are referring to visual entities	This is the ratio of decisions that are unnecessary (i.e. leading to the choice for an incorrect referent for a given word) and necessary decisions (choosing the correct referent).
correctness	**co:=** $\frac{\sum_{i=1}^{m}\frac{\sum_{k=1}^{n_i}\frac{\sum_{j=1}^{maxPos_i} x_j}{n}}{n_i}}{m}$ m := number of sentences $maxPos_i$:= number of words in sentence i n_i:= number of words in sentence i that are referring to visual entities $x_j := \begin{cases} 1 & \text{if correct referent for word j is found} \\ 0 & \text{if correct referent for word j is not found} \end{cases}$	This is the number of times, the correct referent for a given word was found.

Table 3: Metrics for evaluating reference resolution (Adapted from[87])

$$\text{Princess\_001} \xrightarrow{\text{is\_THEME\_for}} \text{Etw.malen\_001}$$

$$\text{Pirate\_001} \xrightarrow{\text{is\_AGENT\_for}} \text{Etw.malen\_001}$$

Figure 65: Context for the sentence *Die Prinzessin malt offensichtlich der Pirat*(Literally: The princess (subject/object) paints apparently the pirate (subject).) 'The princess is apparently painted by the pirate.' containing a pirate(**Pirate_001**) that is the AGENT of the painting action(**Etw.malen_1**) and a princess(**Pirate_001**) that is the THEME

In the experiments described in this section, we chose option (b) to create a standard. This gold standard is then compared to the experimental results by the measures in Table 3.

One feature of the system is its ability to change visual saliency depending on parsing results. The changes of saliency over time are evaluated with regards to a baseline model which does not use any linguistic input. We already introduced the model by [47] which predicts the influence of visual features on bottom-up visual attention of human observers. The saliency maps produced by this system serve as a baseline for our experiments about visual attention. We compare experimental results of influence of context on this model with this baseline in order to show the effects of natural language on visual attention. Since the model depicts saliency of a given region of the input picture by intensity in that region (regions likely to catch attention are bright white, regions with low saliency are black) comparing maps to each other requires further processing.

We use the output maps to compute three different values for regions of interest (ROI) in the picture. A ROI is defined as any region that exists as a corresponding entity in the higher level description (the a-box of our knowledge representation model) of the picture. Each region has a certain dimension which is established as a rectangle drawn around an object in the picture. For each region three different values are computed.

(a) The percentage of pixels of the region that are salient

(b) The percentage of all salient pixels that are located in the region

(c) The mean saliency of the region

For this purpose a pixel is defined as salient if the brightness value exceeds a predefined threshold. The mean of salience (in (c)) is the mean brightness of the region in percent of the maximal possible brightness.

6.2 Experiments

In this section we will discuss the experiments conducted to show the capabilities of the system. The main method for each of the presented experiments will always be the same. We test the system with natural language sentences, a visual context described by means of our knowledge representation component and, optionally, a set of pictures corresponding to the visual context. The experiments differ in variability regarding the input of the system:

Different sentences
Depending on the goal of a specific experiment, we use different sentences. Each set of sentences used will consist of utterances that contain a certain problem that is difficult to resolve using only language information, thus requiring access to extra-linguistic knowledge sources.

Different visual contexts
Each contextual description contains information that is helpful for interpreting a sentence. We will vary the complexity of these descriptions by adding or removing visual entities. We will present the system with input data that requires increasingly difficult challenges with regard to reference resolution. We predict that a context containing many descriptions of objects and events in the visual context will lead to a behavior that is prone to errors with regard to the search for the correct referent for a word or phrase. Our goal is to show that even errors caused by complex contexts can be corrected using linguistic information provided by the language parser.

Different pictures
Testing the influence of context on visual saliency, we will change the pictures used. It is our goal to show that even if different low-level visual features affect the saliency

landscape, the development of saliency still demonstrates the influence of language on attention in line with human behavior found in cognitive studies.

Different constraints

In order to integrate visual context, language information and saliency in a given picture, constraints are needed that establish this connection. We will change the type and number of constraints in different experimental set-ups. Our goal is to show that it is not only important to use information from one channel for the benefit of the other, but also how this information is integrated. We will investigate how the number of constraints and the level of refinement contributes to the correct reference resolution even in highly complicated experimental tasks.

6.2.1 Reference Resolution

6.2.1.1 Experiment 1 : Finding the correct referent with incremental parsing

In this experiment we investigate if and when the system chooses the intended set of referents for a given sentence. The visual representation for this experiment contains a small set of possible referents (describing the content of a single picture) making the choice of the correct one straightforward. This experiment uses only words as a cue to find the correct referents. Sentences are processed in incremental parsing mode. Referents are found only by matching words of the sentence to the visual entities contained in the representation, without using any linguistic cues provided by constraints. The goal is to investigate how the system will accomplish reference resolution given a simple contextual description when the only information provided are words of the sentence. The expectation is, that the system indeed finds correct referents for each part of the sentence, but only at word onset, without predicting any upcoming referent.

Materials and design

Input for the visual attention model were 48 pictures taken from [54] (see Section 10.3). These pictures contain three characters. One character is carrying out an action (the agent), one is the receiver of an action (the patient) and one is the receiver of an action as well as carrying out an action (the ambiguous character). The language input for the parser consisted of sentences describing the depicted events (see Section 10.2.1). Each action depicted is described by one sentence, which is created according to the pattern shown in Figure 66 (left side). A noun phrase followed by a verb, an adverb and another noun phrase. Sentences are either in subject-verb-object-order or in object-verb-subject-order.

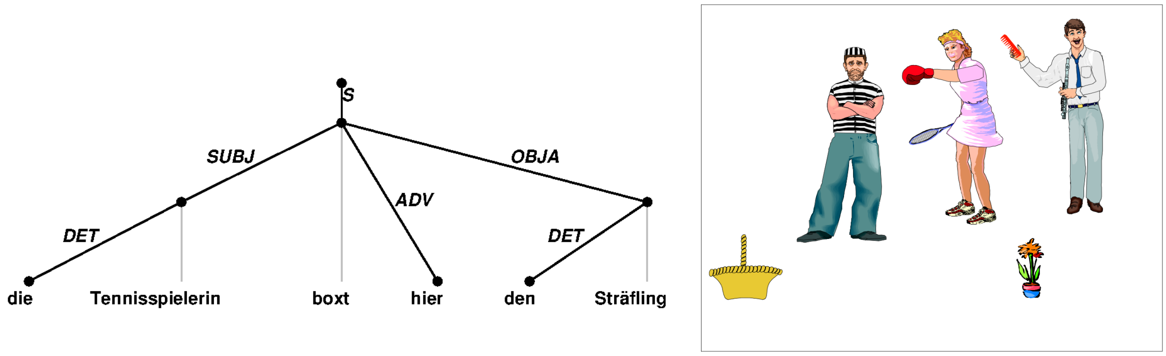


Figure 66: Dependency tree and picture for the sentence *Die Tennisspielerin boxt hier den Sträfling.* (Literally: The tennis player(ambiguous) boxes here the convict(object)) 'The tennis player is boxing the convict.' in the subject-verb-object sequence (Taken from [54])

Altogether we used a set of 24 items, each consisting of four sentences and two pictures, for a total number of 96 test-runs. For each experimental run a representation (see Figure 67) of the scene in the picture was loaded by the connecting component. Example sentences for this description would be *Die Tennisspielerin boxt hier den Sträfling.* (Literally: The tennis player(ambiguous) boxes here the convict(object).) 'The tennis player is boxing the convict.' for the subject-verb-object condition and *Die Tennisspielerin kämmt hier der Flötist.* (Literally:The tennis player(ambiguous) combs here the flutist(subject).) 'The tennis player is combed by the flutist.' for the object-verb-subject condition. The role of the tennis player is initially ambiguous as the word alone can refer to the agent role of the boxing action as well as to the theme role of the combing action. After onset of the verb, the system has enough information to infer which role is the correct one.

WCDG was started with the grammar of German as well as the constraint set for assigning thematic roles described in [68]. Parsing was performed in incremental mode.

$$\begin{aligned}
\text{Tennis.Player.f\_001} &\xrightarrow{\text{is\_THEME\_for}} \text{Etw.Kaemmen\_001} \\
\text{Tennis.Player.f\_001} &\xrightarrow{\text{is\_AGENT\_for}} \text{Etw.Boxen\_001} \\
\text{Flutist.m\_001} &\xrightarrow{\text{is\_AGENT\_for}} \text{Etw.Kaemmen\_001} \\
\text{Convict\_001} &\xrightarrow{\text{is\_THEME\_for}} \text{Etw.Boxen\_001}
\end{aligned}$$

Figure 67: Representation of the picture in Figure 66

Results and Discussion

Metric	**Value**
average first correct	0.47
first correct applicable	268
average first final	0.48
first final applicable	268
mean edits per utterance	0.07
edit overhead	0.02
correctness	0.53

Table 4: Results for Experiment 1

The correct referent was first found correctly after roughly half of the input sentence (average first correct: 0.47) which is as expected, considering that this value is computed for each referent of the sentence, whether the word is already part of the sentence or not.
The computed value for average first final is only slightly higher (0.48) showing that the referent initially found for a word was almost always the correct one. This is also confirmed by the mean edits per utterance of 0.07 and the edit overhead of 0.02 showing that the system seldom switched referents, and that almost all of these switches were necessary to find the correct referent. The system assigned the correct referent in half of the processing steps shown by the correctness of 0.53.

6.2.1.2 Experiment 2 : Finding the correct referent in a simple context with incremental prediction

In this experiment we investigate if and when the system chooses the intended set of referents for a given sentence, based on linguistic cues. We added constraints connecting linguistic analysis and visual information, combining both knowledge sources to predict upcoming referents. Our expectation is that, during incremental parsing, the system will find referents for parts of the sentence even before the corresponding word is presented, thus finding the correct set of referents at an earlier time-point than in Experiment 1.

Materials and design

The experimental setup and materials used were the same as in Experiment 1 with one exception: in addition to the constraints of German and for thematic roles, we included a set of visual constraints. Two types of constraints were used. The first type consists of constraints that incorporate visual information about thematic roles in the context into the processing of the language parser. One example of such a constraint can be found in Figure 68. Similar constraints have been devised for other thematic relations (THEME, OWNER, INSTRUMENT, COMITATIVE) between referents. These constraints operate by penalizing any semantic edge between two words for which visual referents were found, whenever the referents are not connected by the relation specified in the constraint. Thus, for the example in Figure 68, any AGENT edge between two words is subject to a penalty when the referent of the modifying word is not described as the acting character for the action of the modified word in the context model. The other type of constraints is responsible for the task of integrating the information from the language channel to predict upcoming referents for words that are not yet part of a sentence fragment. An example is the constraint shown in Figure 69. Whenever the language parser predicts an AGENT relationship for referents of two words in the sentence, the connecting component increases the likelihood of scenes of reference that contain such a relationship between entities.
A crucial effect for the prediction of referents for upcoming words is the capability of the system to use the presented mechanism in order to find referents for virtual nodes

```
X!AGNT : ’Connect Agent correct ’ : visual : 0.5 : (visual) :
X.label = AGENT &
hasReferent(X@id) &
hasReferent(X^id)
->
visual(X@id, X^id, AGENT) > 0.9 ;
```

Figure 68: Constraint expressing that an AGENT dependency between words having referents requires visual evidence in the context

of the dependency structure. Whenever language input and visual information for the sentence fragment at hand will result in the prediction of a relationship for a referent of a word yet to be received, constraints like the one in Figure 69 will have the effect of assigning a possible referent to the virtual node the semantic edge is connected to.

```
X!AGNT : ’Search referent for AGENT’ : visual : 0.5 : (visual) :
X.label = AGENT
->
visualChange(X@id, X^id, 5.2, 1, AGENT);
```

Figure 69: Constraint expressing that an AGENT dependency between words is an indicator that referents for these words are in an AGENT relation

Results and Discussion

The constraints used resulted in a lowered average first correct (0.36) referent for a word compared to the results of experiment 1, indicating that the correct referent was found even before the corresponding word was received by the language parser. We can find one example of this behavior when processing the sentence *Die Joggerin verhext mal eben den Doktor*(The jogger(ambiguous) bewitches just now the doctor(object).)’The jogger is bewitching the doctor.’ (see Figure 70) After finding a referent **Jogger.f_001** for the word *Jogger*(jogger). In this case, the system infers that the agent edge of the dependency tree predicts another, not yet named, visual entity (**Etw.verhexen_001**(bewitching something)).

The question if this prediction of an upcoming referent is a good one is at least partly

Metric	Value Experiment 1	Value Experiment 2
average first correct	0.47	0.36
first correct applicable	268	268
average first final	0.48	0.40
first final applicable	268	268
mean edits per utterance	0.07	0.38
edit overhead	0.02	0.20
correctness	0.53	0.61

Table 5: Results for Experiments 1 and 2

answered by the computed value of average first final. Although the difference between first correct and first final is considerable larger compared to results of experiment 1, showing that the first correct referent found for a word was changed to an incorrect one more often (also indicated by the higher mean edits per utterance) the first final ratio is still smaller than in Experiment 1 and even smaller than first correct in Experiment 1. The reason for the drop in both values as well as for the difference between first final and first correct is the prediction of referents before word onset. Using prediction of referents the system finds the correct referent earlier (explaining the drop in both values) but is also more likely to change this to an incorrect one (by predicting wrong referents due to intermediate parsing decisions), which will be corrected at word onset (which results in a discrepancy between first final and first correct). Another impact of the added constraint is the increase in correctness (0.61) due to the fact that correct referents are found earlier than in Experiment 1.

6.2.1.3 Experiment 3 : Finding the correct referent in complex context

In this experiment we investigate if and when the system chooses the intended set of referents for a given sentence using a context that is complex compared to the contextual descriptions used in Experiment 1 and 2. The visual representation for this experiment contains a large set of possible referents, while the individual sub-scenes (i.e. sets of entities that are related) vary only slightly, making the choice of the correct referent difficult. As we use the same set of constraints as in Experiment 1, which does not support linking the results of language parsing to the visual representation, we expect a large number of incorrect referents for a given sentence.

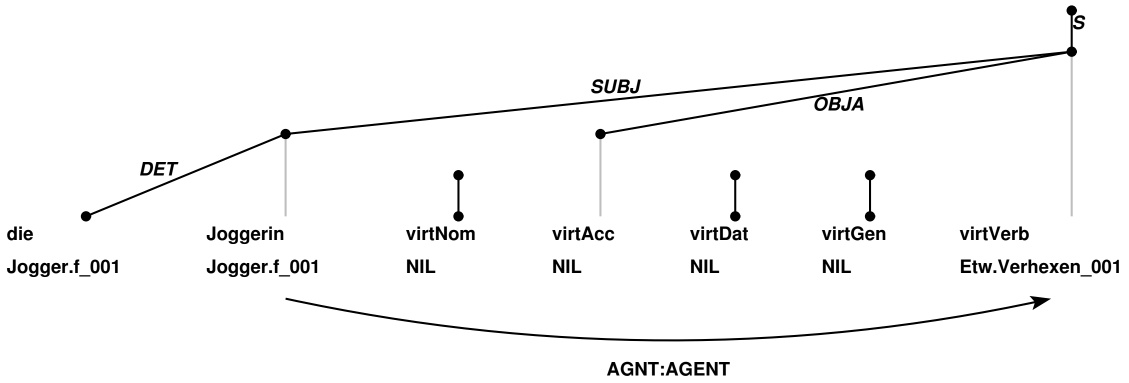


Figure 70: Dependency tree for the sentence fragment *Die Joggerin* (The jogger). Although the word *verhext*(bewitches) is not yet part of the sentence, the correct referent **Etw.verhexen_001**(bewitching something) is already predicted

Materials and design

We used the same experimental setup as in Experiment 1. The contextual representation differed by including all descriptions used in the preceding experiment into one representation. Therefore, a word of the input sentence could be potentially grounded in more than one referent. Given the context in Figure 71 and the sentence *Die Prinzessin wäscht offensichtlich den Pirat.* (Literally: The princess(ambiguous) washes apparently the pirate(object).) 'The princess is apparently washing the pirate.' several possible choices for the selection of referents are possible when using only the words of the sentence as guiding information. The verb of the sentence and each of the nouns can refer to several visual entities. For instance, the word *Prinzessin* (Princess) has two obvious referents as two princesses are present in the representation. Actually the whole number of possible referents is much higher as we do not connect the words of the sentence directly to entities named by these words. As we have pointed out, the system is able to connect words via the conceptual hierarchy to entities of other types than the one denoted by the word. Thus, the word *Prinzessin* might refer to all of the 72 persons described as these are semantically connected via the concept **human being**. This also applies to the verb of the sentence which can refer to

Pirate.m_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Waschen_001
Fencer.m_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Malen_001
Princess_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Malen_001
Princess_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Waschen_001
Fencer.m_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Malen_002
Pirate.m_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Waschen_002
Princess_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Waschen_002
Princess_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Malen_002

Figure 71: A subset of the context used in Experiment 3, describing six characters as participants of four actions

several actions in the context that are connected by the concept **Situation.Concept**.

Results and Discussion

Using a context containing ambiguous referents for each word did not result in a significant change in the values of first correct(0.47) and first final(0.49) compared to Experiment 1, showing that if the system finds the correct referent at all, it does so at roughly the same parsing steps as in Experiment 1. The analysis of the experiment also shows a significant drop in the scores for first correct applicable(133) and first final applicable(133) that is evidence for a large number of incorrectly assigned referents. This result is also evident in the correctness which drops significantly (0.26) compared to Experiment 1.

The low number of edits per utterance(0.04) is due to the fact that without correcting information from the language parser, the system seldom changes its assignment of referents, no matter if they are correct or not. This also shows in the edit overhead(0.54): In most cases, the system carries out one assignment. Therefore the edit overhead for a single word/referent-combination will be one (when the single assignment is incorrect) or zero (if correct). As the number of correct assignments and

Metric	Value Experiment 1	Value Experiment 3
average first correct	0.47	0.47
first correct applicable	268	133
average first final	0.48	0.49
first final applicable	268	133
mean edits per utterance	0.07	0.04
edit overhead	0.02	0.54
correctness	0.53	0.26

Table 6: Results for Experiments 1 and 3

incorrect assignments are evenly distributed in our result set, slightly more than half (0.54) of all assignments are necessary ones.

We will discuss these results given the example sentence *Die Prinzessin wäscht offensichtlich der Pirat*(The princess(ambiguous) apparently washes the pirate(subject)) 'The princess is apparently washed by the pirate'. Although the sentence is parsed correctly by assigning the AGENT-role to the pirate and the THEME-role to the princess (see Figure 72), the referents found are incorrect given the context in Figure 71. Of the two princesses described in the context, the one being the THEME of the washing action (**Princess_002**) should be selected as this choice would be in line with suggestions made by the parser. Due to the lack of linking constraints, the system instead chooses **Princess_001** which contradicts the language analysis as it is the AGENT of an action.

6.2.1.4 Experiment 4 : Finding the correct referent in a complex context with incremental prediction

In this experiment we investigate if and when the system finds the intended set of referents for a given sentence, based on linguistic cues using the same complex context as in Experiment 3. In this case we add linguistic constraints that incorporate the results of the language parser into the choice of a referent for a given word. We expect a significant improvement in the performance compared to Experiment 3.

Materials and design

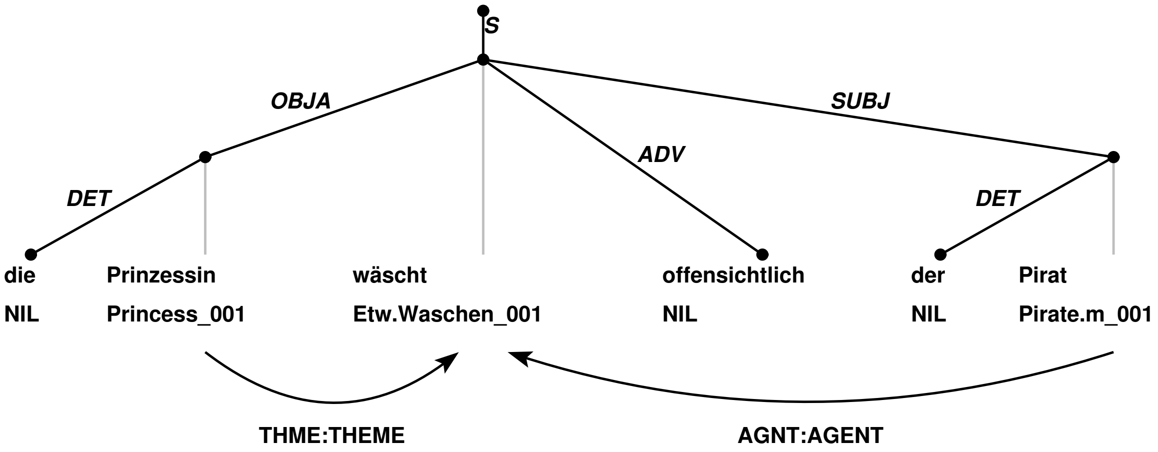


Figure 72: Dependency tree for the sentence *Die Prinzessin wäscht offensichtlich der Pirat*(The princess(ambiguous) apparently washes the pirate(subject)) 'The princess is apparently washed by the pirate'

The experimental design was the same as in Experiment 2: We used three sets of constraints (German, thematic roles, and visual) as input to the parser, which was started in incremental mode. The materials were the same that we used in Experiment 3: the visual context contained descriptions of all 24 pictures in one representation.

Results and Discussion

Metric	Value Experiment 3	Value Experiment 4
average first correct	0.47	0.48
first correct applicable	133	247
average first final	0.49	0.57
first final applicable	133	247
mean edits per utterance	0.04	0.31
edit overhead	0.54	0.44
correctness	0.26	0.42

Table 7: Results for Experiments 3 and 4

The constraints did not improve the value for average first correct, indicating that the first time the correct referent was assigned did not change under the influence of linguistic input. Average first final is even worse than in Experiment 3, which means that the final decision for a correct referent was even later than without linguistic information. What is crucial are the number of times these values were applicable (i.e. how often the correct word/referent combination was found at all (for first correct applicable) or was the result at the end of parsing (for first final applicable)). Using linguistic constraints, the correct referent was found (first correct applicable) and was the final decision (first final applicable) 247 times which is a considerable improvement compared to 133 in Experiment 3. One example of this behavior can be observed in Figures 73 and 74: although the system initially decides that **Devil.f_001** is the correct referent for the word *Teufelin*(female devil), as soon as the parser changes its analysis that the initial noun phrase (in Figure 73) being an AGENT by assigning the THEME role (Figure 74), the Connector assigns **Devil.f_002** as the correct referent. As the sentence is initially ambiguous with regard to its thematic roles, the change of the analysis from AGENT to THEME is made by the language parser lately during processing (e.g. when the second noun phrase *der Clown*(by the clown) is available).

This also explains why the value for average first final is worse than in the previous

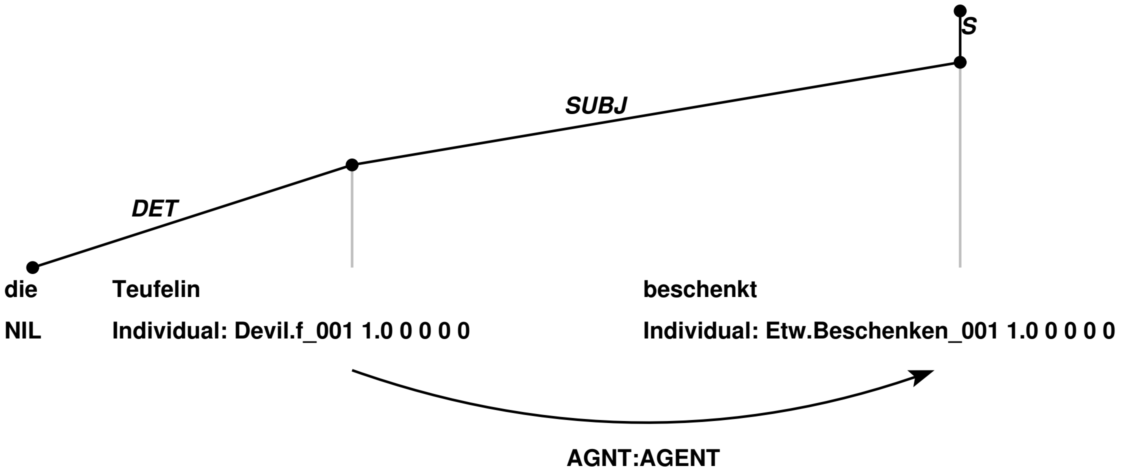

S
SUBJ
DET
die
Teufelin
beschenkt
NIL
Individual: Devil.f_001 1.0 0 0 0 0
Individual: Etw.Beschenken_001 1.0 0 0 0 0
AGNT:AGENT

Figure 73: Dependency tree for the sentence fragment *Die Teufelin beschenkt* (Literally: The devil(ambiguous) gifts) 'The devil is gifted by'

experiment: in all cases where the system finds the referent only due to linguistic cues provided by constraints of the WCDG-grammar, the system uses this information to switch its initially wrong decision for a referent to a correct one during later processing steps. Without constraints, the system sticks to its incorrect referent which results in a lowered applicable scores (as a wrongly assigned referent is not counted in this value) but actually improves the average scores (i.e. a decision for the correct referent

late during processing increases average first final, an incorrect final decision has no influence at all as it is not included in the calculation).
The improvement due to linguistic input also affects the overall number of parsing steps in which the system found the correct referent for a word which can be seen in the value of correctness (correctness 0.42 compared to 0.26 in Experiment 3). The number of edits per utterance increased to 0.31 as the information provided by the added constraints results in more changes of decisions, while the edit overhead(0.44) dropped due to the fact, that we have a higher percentage of necessary decisions (e.g. those for a correct referent) than in Experiment 3.
Our conclusion is that the system can find correct referents in an ambiguous context when integrating results of language parsing.

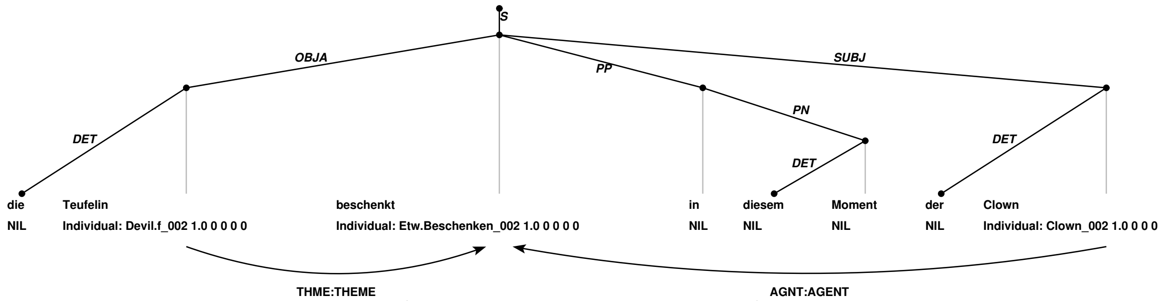

S
OBJA
PP
SUBJ
PN
DET
DET
DET
die
NIL
Teufelin
Individual: Devil.f_002 1.0 0 0 0 0
beschenkt
Individual: Etw.Beschenken_002 1.0 0 0 0 0
in
NIL
diesem
NIL
Moment
NIL
der
NIL
Clown
Individual: Clown_002 1.0 0 0 0 0
THME:THEME
AGNT:AGENT

Figure 74: Dependency tree for the sentence *Die Teufelin beschenkt in diesem Moment der Clown* (Literally: The devil(ambiguous) gifts at this moment the clown(subject)) 'The devil is gifted by the clown'

6.2.2 Influencing Attention

6.2.2.1 Experiment 5 : Changing saliency by incrementally receiving language information

In this experiment we investigate the systems ability to increase saliency in a picture region that is addressed in the corresponding sentence at a specific processing step. The purpose of this experiment is twofold: Firstly, we want to investigate if the system is indeed capable to switch its focus (which is the region of highest saliency in the picture) depending on language input. Also we want to compare the evolution of saliency with the changes of visual focus observed in humans. As we do not predict upcoming parts of the sentence in this experimental setup, our expectation is that the attentional focus will change at word onset.

Materials and design

We used the same materials as in Experiment 1: The context model described the pictures taken from [54] showing agent, patient and ambiguous characters engaging in actions. We also included a constraint that increases saliency for the last word of the sentence fragment at each parsing step.

Results and Discussion

Figures 75 and 76 show the evolution of saliency for the regions containing the three characters under SVO- and OVS-conditions. As the time course of processing (i.e. how much time is required for a specific incremental step) differs between experimental runs, we chose to normalize the time dimension such that processing of any sentence is divided into twelve steps. Figure 75 shows that in both conditions, saliency for the regions of agent and patient drops when the parser receives the first noun phrase (steps 3-5) which refers to the ambiguous character. At the onset of the second noun phrase (at timepoint 8), which is the first time either the patient (in SVO sentences) or agent (in OVS sentences) is referred to by the language input, saliency for the corresponding region increases. A similar observation can be made with regard to the development of saliency for the ambiguous character (Figure 76): saliency for this character is only increased while the parser processes a sentence fragment that ends with the first noun phrase (3-5).

Our conclusion is, that the system increases saliency for the correct region but only when a word referring to such a region is already part of the sentence fragment processed. Without inclusion of prediction of referents, the system does not exhibit the behavior of human attentional switches to focus on yet unmentioned parts of the picture which has been observed in [54].

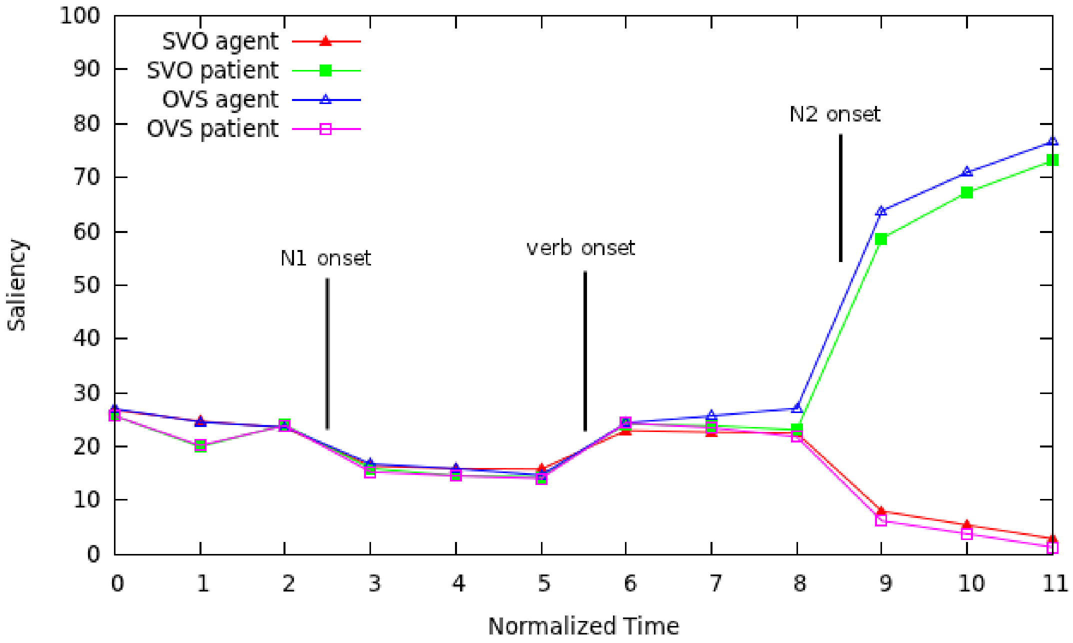


Figure 75: Development of saliency in Experiment 5 for agent and patient in SVO and OVS sentences

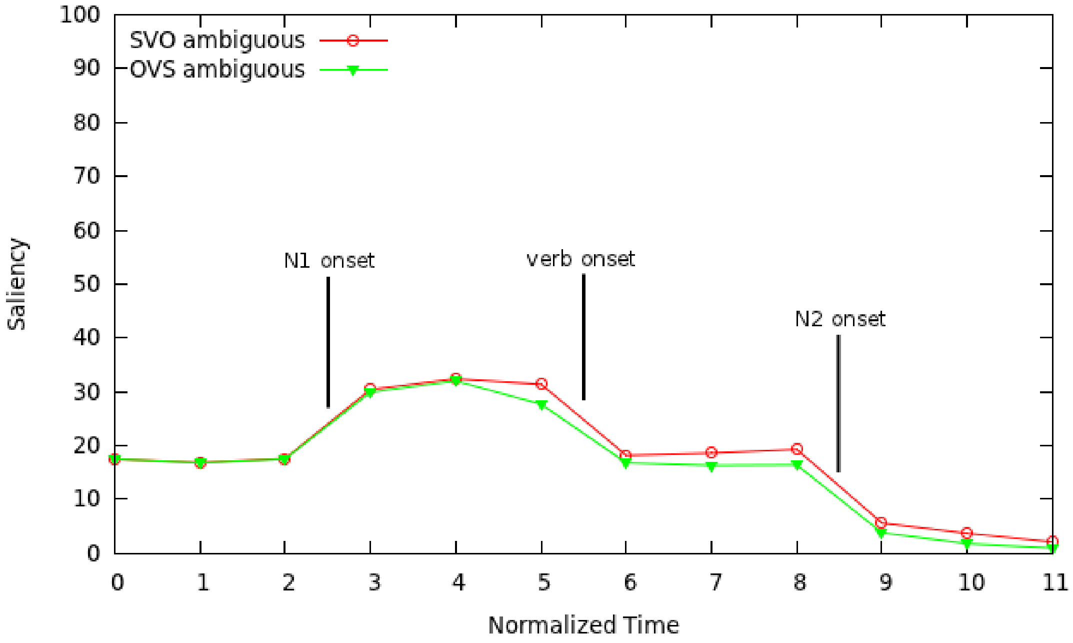


Figure 76: Development of saliency in Experiment 5 for the ambiguous character in SVO and OVS sentences

6.2.2.2 Experiment 6 : Changing saliency by predicting upcoming referents

In this experiment, which was presented in [9], we investigate the effect of merging language and visual information on the prediction of upcoming referents for uncompleted utterances. In [54] it was shown that the bi-directional interaction of visual and contextual information in the human cognitive system will result in prediction of missing role fillers whenever a person hears a part of the sentence. The task of this experiment is to find out whether the integration of top-down information from the language channel, visual information from our context representation and bottom-up information from the attentional model can result in predictions of missing information similar to the effects found in human reference resolution. Contrary to Experiment 5, the effect on the attentional model does not only depend on the language information already received, but will also integrate the predictive elements we already used in our setup of Experiment 2.

Materials and design

The experimental setup was the same as in Experiment 5. In addition we integrated the constraints that assign referents to words not yet part of the sentence by predicting thematic edges early. For each of the thematic edges (AGENT, THEME/PATIENT, INSTRUMENT, OWNER and COMITATIVE), one constraint is integrated.

Results and Discussion

Figures 77 and 78 show the development of saliency in the regions of interest for the depicted characters as incremental parsing progresses. The saliency before the onset of the first noun (steps 0-2) relies only on bottom-up cues provided by the picture. After the first noun has been added to the sentence fragment (steps 3-5), the saliency for the agent as well as the patient decreases, because saliency is increased for the ambiguous character in the picture, as denoted by the first noun of the sentence. After verb onset (steps 6-8) the saliency of the character not yet addressed (i.e. the patient for SVO-sentences, the agent for OVS-sentences) already increases, although the second noun is not processed until steps 9-11. Saliency for the ambiguous character drops significantly at this point.

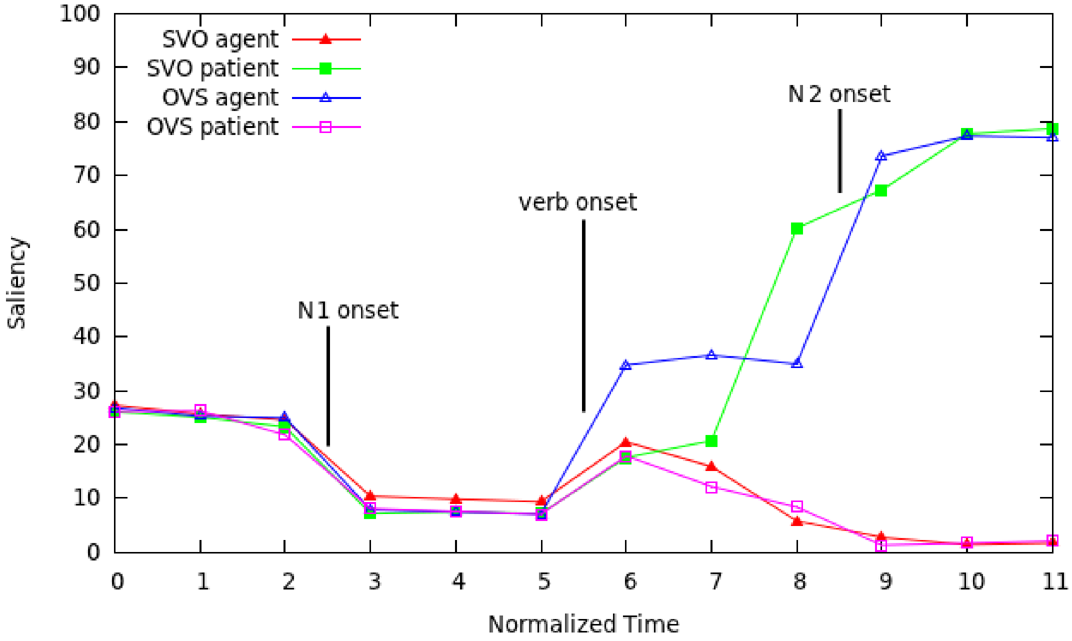


Figure 77: Development of saliency in Experiment 6 for agent and patient in SVO and OVS sentences

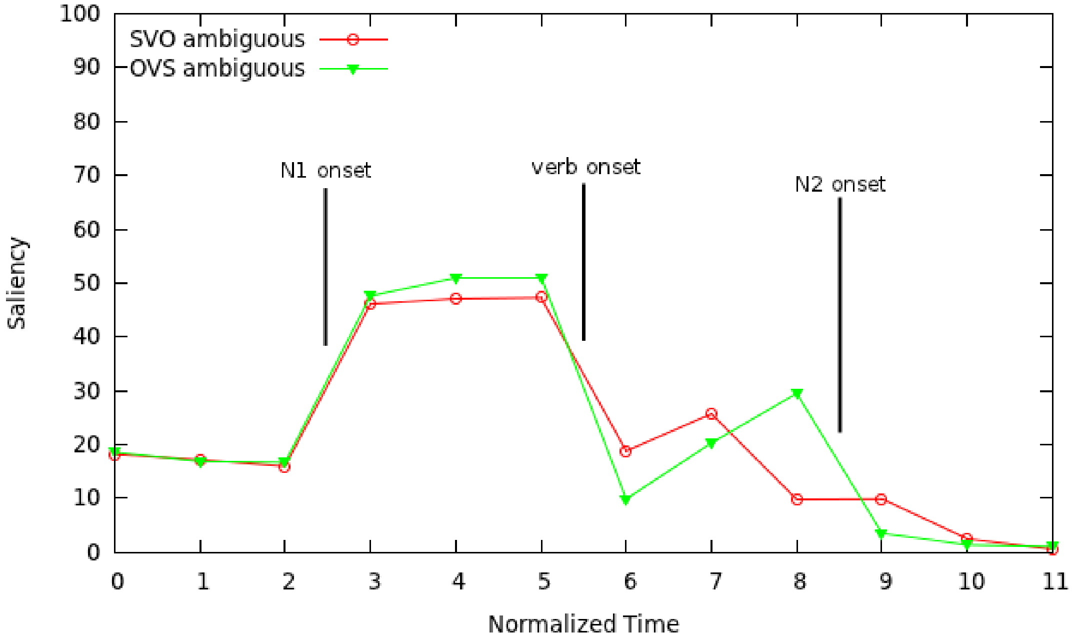


Figure 78: Development of saliency in Experiment 6 for the ambiguous character in SVO and OVS sentences

This development is in agreement with findings about human eye-movements in [55]. A closer inspection of the parsed sentences shows indeed that this development is caused by the fact, that the first noun and the verb enable the system to assign the correct roles to both characters, even if the second noun is not yet part of the sentence fragment.

6.2.2.3 Experiment 7 : Comparing the effects of hard and soft contextual integration on the attentional model

In Experiment 6, the results of any of the three subsystems (model of attention, language parser and contextual representation) were tightly locked with the outcome of processes in the other two models by the integration constraints. The fact that we need this dependency between the information gained from each subsystem is obvious when we compare Experiment 6 with Experiment 5 where we omitted these constraints, which results in a system behavior not in line with effects of human reference resolution.

The following experiment is designed to investigate which degree of coupling between the different types of information is needed in order to model the desired effects. We therefore change the weights of the integrating constraints between different experimental runs. We expect that these changes influence the development of saliency. A higher weight of the integrating constraints should result in a reduced influence of visual evidence. This should result in the system making wrong parsing decisions (by attaching dependency edges not in line with visual evidence), which in turn results in wrong predictions of upcoming referents and therefore leads to changes in saliency that differ from those observed when the coupling is stronger.

Materials and design

For this experiment we repeated Experiment 6 two times: the two tests differed in the adjustments of the weights of the integrating constraints. As described in Section 4, the weight of the integrating constraints has an effect on language parsing as constraints with a lower weight are more likely to influence the attachment of dependency edges. We choose two different weights for the constraints: 0.4 and 0.95.

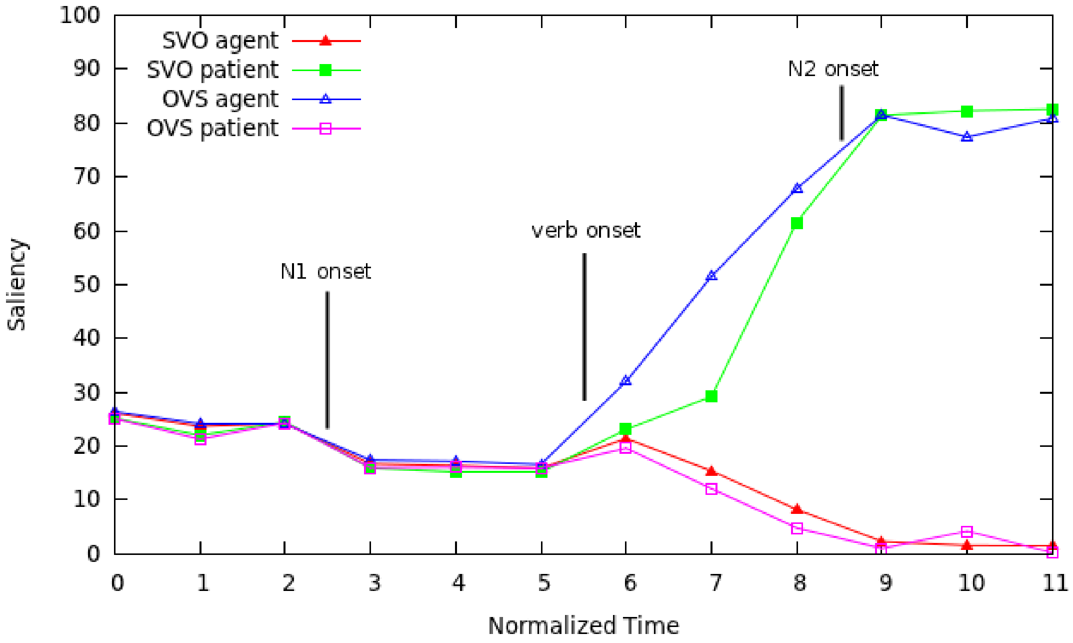


Figure 79: Development of saliency for a constraint weight of 0.4

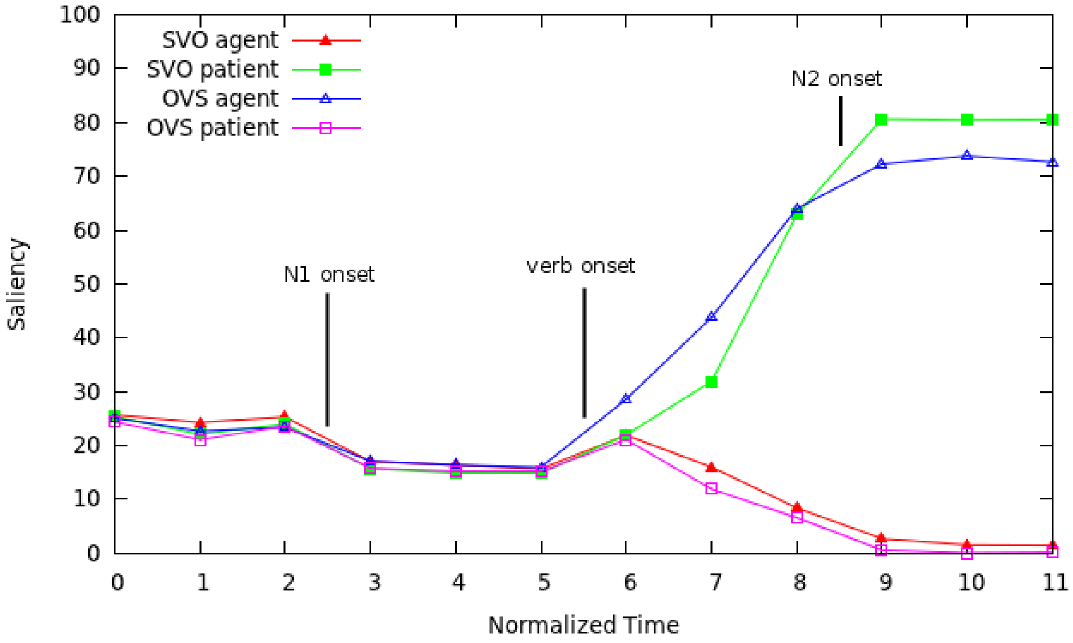


Figure 80: Development of saliency for a constraint weight of 0.95

Results and Discussion

Figures 79 and 80 show the development of saliency for each of the used weights for AGENT and THEME character-region. Contrary to our expectations, the curves are almost the same for the two experimental conditions, indicating that the system is able to make correct predictions about upcoming referents early during processing even when the integration of visual information has a low priority due to the high weight of the integrating constraints.
After taking a closer look at the output, we explain this phenomenon by assuming that as the lower weighted (and thus more influential) constraints improve parsing at early incremental processing steps, as the parser needs a lower number of steps to parse the sentence and each individual step takes less time than in the case of higher weights which result in incorrect intermediate parsing decisions. Unfortunately, this difference in parsing steps and parsing time also counteracts the effects on saliency that we predicted.

To give an example of this effect, we compare the two cases with regard to parsing results of the sentence *Die Aerobic-Trainerin verwarnt offensichtlich der Astronaut.* (Literally: The aerobics-trainer(ambiguous) warns obviously the astronaut(AGENT)) 'The aerobics-trainer is obviously warned by the astronaut'. Figure 81 shows the earliest parsing step at which the lower weighted constraint influences the system to make the correct prediction that the **Astronaut_002**(astronaut) is an upcoming referent of the sentence. This prediction is due to the fact that visual evidence is used to assign the correct role of THEME to the aerobics-trainer. In case of the higher

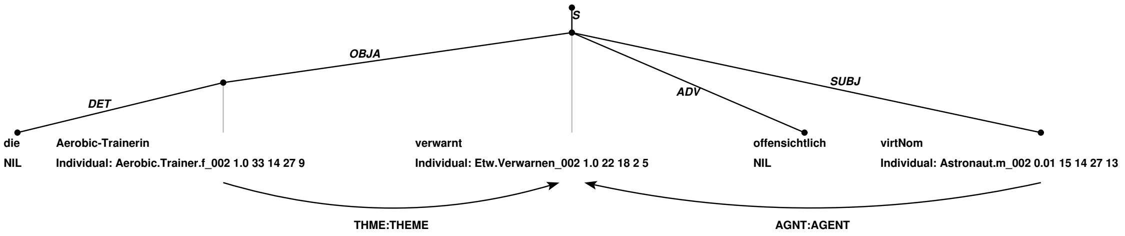

S
OBJA
DET
ADV
SUBJ
die
NIL
Aerobic-Trainerin
Individual: Aerobic.Trainer.f_002 1.0 33 14 27 9
verwarnt
Individual: Etw.Verwarnen_002 1.0 22 18 2 5
offensichtlich
NIL
virtNom
Individual: Astronaut.m_002 0.01 15 14 27 13
THME:THEME
AGNT:AGENT

Figure 81: Dependency structure for the sentence *Die Aerobic-Trainerin verwarnt offensichtlich der Astronaut.*(Literally: The aerobics-trainer(ambiguous) warns obviously the astronaut(AGENT))'The aerobics-trainer is obviously warned by the astronaut' with hard integration of context

weighted constraints (see Figure 82) the system assigns the incorrect role of AGENT because visual evidence does not overrule the influence of language. As soon as the language information results in the parser making the correct role assignments in a later parsing step, the system also predicts the correct upcoming referent. As the parser reattaches a large number of dependency edges for this late re-evaluation a lot of time is spent during these steps after the correct prediction has already been made. Therefore saliency is adjusted correctly for a long time which results in a similar development when comparing Figure 79 and 80.

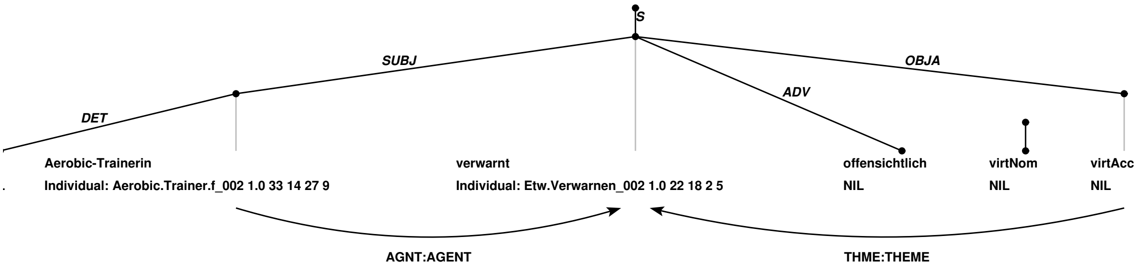


Figure 82: Dependency structure for the sentence *Die Aerobic-Trainerin verwarnt offensichtlich der Astronaut.*(Literally: The aerobics-trainer(ambiguous) warns obviously the astronaut(AGENT))'The aerobics-trainer is obviously warned by the astronaut' with soft integration of context

6.2.3 Influencing Language Parsing

6.2.3.1 Experiment 8 : Changing the attachment of prepositional phrases with non-incremental parsing in a simple context

This experiment investigates the influence of context on the processing of sentences containing prepositional phrases. The sentences used were taken from the SALSA-corpus (see [84]) . Each sentence contains parts that would be interpreted as a PP-edge in the corresponding dependency tree. For each sentence, we present two distinct scene descriptions: one containing a visual context that suggests an attachment of the prepositional phrase that is in line with the semantic interpretation of the sentence and one containing information that would indicate a wrong attachment. Our goal in conducting this experiment is twofold: on the one hand, we want to show that integrating contextual descriptions will be beneficial when resolving the complex problem of PP-attachment. Secondly this experiment will explore the system dependency on the visual descriptions. Our expectation is that the system will choose the attachment of the prepositional phrase that is in line with the visual descriptions if contextual information strongly favors this interpretation.

COMITATIVE

$$\text{Idiom\_01} \xrightarrow{\text{is\_COMITATIVE\_for}} \text{German\_01}$$

INSTRUMENT

$$\text{Idiom\_01} \xrightarrow{\text{is\_INSTRUMENT\_for}} \text{Etw.Sprechen\_01}$$

Figure 83: Representation of an idiom (**Idiom_01**) being either the INSTRUMENT of the speaking action (**Etw.Sprechen_01**) or the COMITATIVE of German (**German_01**)

Materials and design

Ten sentences have been have been constructed. For each sentence two different contextual descriptions existed: in one case, the context would suggest an interpretation of the sentence in line with the annotation found in the SALSA-corpus. In the other case, visual context would be evidence for the wrong interpretation. Contexts represented parts of the scene described in the sentence by either an instrument or a comitative relationship between an object and an action (see Figure 83). Thus, all in all, twenty experimental runs were conducted.

We used the same constraints as in Experiment 2 to integrate the contextual information into language processing. One example constraint is shown in Figure 84, which requires an instrument edge between nodes of the dependency tree whenever the referents for the words are represented in an instrument relation in the context. A similar constraint was used for the comitative relation.

Results and Discussion

We found that the system was able to integrate the context to influence the attachment of edges in the dependency tree in all sentences used. The effect of the integration

```
X!INST, Y:INST : 'Connect INSTRUMENT correct 1' : visual : 0.5 :
(visual) :
X.label = INSTRUMENT &
->
visual(X@id, X^id, INSTRUMENT) > 0.9 ;
```

Figure 84: Constraint expressing that an INSTRUMENT dependency between words having referents requires visual evidence in the context

was that the analysis changed from its independent interpretation (see Figure 85 for an example of the system choosing the instrument interpretation) to one in line with the information representation of the visual evidence (see Figure 86), given that the linguistic model and the visual evidence are not consistent. The outcome of the experiment for all sentences was an analysis in line with visually perceivable information. This result is not surprising as we choose a low (0.5) weight for the integrating constraints in Figure 84. As the value of an analysis is the product of its violated constraints, any analysis not consistent with visual evidence (and therefore violating the integrating constraint) must be scored at least twice as high as an interpretation in line with context information in order to be considered a viable alternative. As the attachment of prepositional phrases of a sentence and the related semantic interpretation of thematic roles in the scene are highly ambiguous when derived solely from linguistic evidence, the difference between the WCDG-score of the two different results tends to be small. An integrating constraint with a low weight will therefore have a high impact. Thus, the results of sentence parsing in this experiment are always in line with the representation of contextual information.

6.2.3.2 Experiment 9 : Changing the attachment of prepositional phrases during incremental parsing in a simple context

In this experiment (presented in [10]) we use sentences containing an ambiguity with regard to the attachment of prepositional phrases. In contrast to Experiment 8, we conduct the processing in incremental mode while integrating the contextual information late during processing. Our goal is to investigate the systems ability to

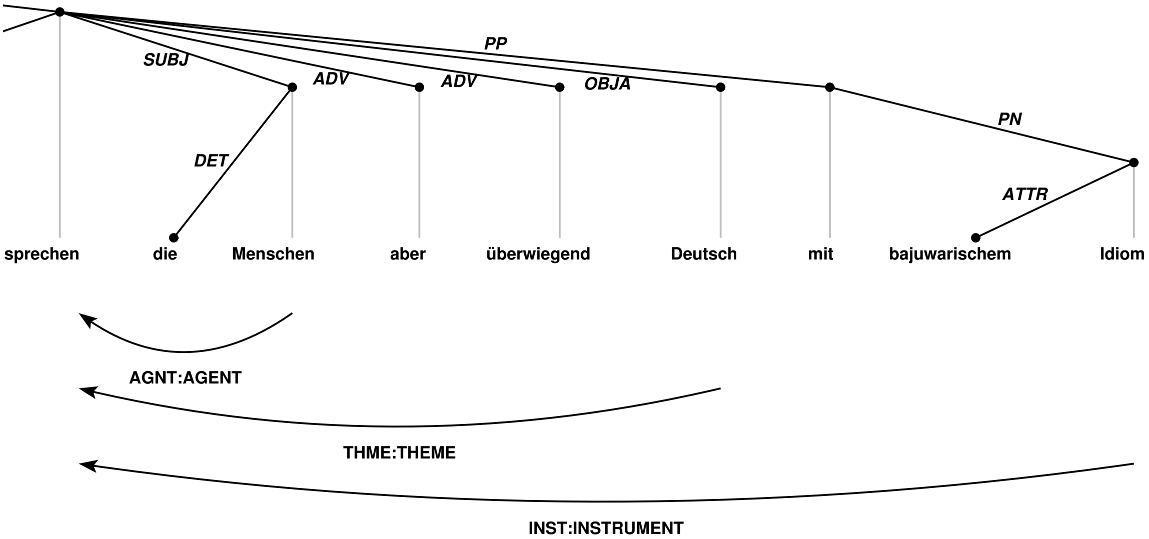


Figure 85: Dependency tree for the sentence *Die Region ist offiziell zweisprachig, im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom*(The region is officially bilingual, but in everyday life people mainly speak German with a Bavarian idiom) without integrating any context

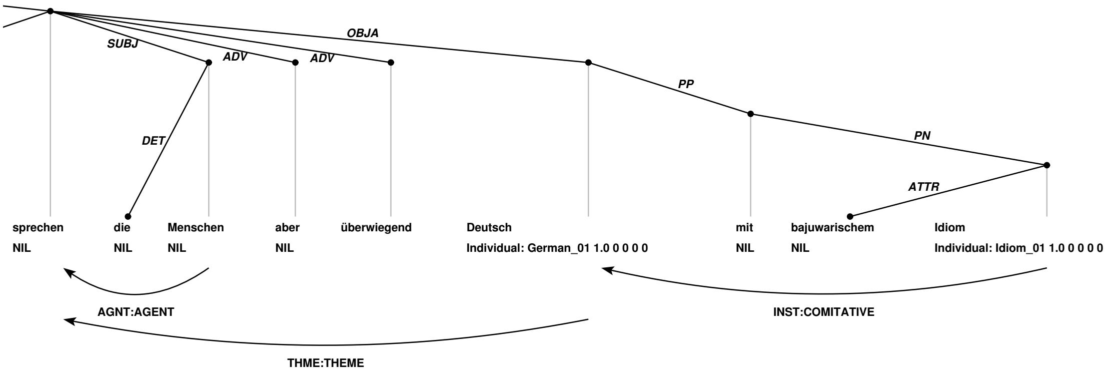


Figure 86: Dependency tree for the sentence *Die Region ist offiziell zweisprachig, im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom*(The region is officially bilingual, but in everyday life people mainly speak German with a Bavarian idiom) integrating the comitative context

re-evaluate attachments of phrases when the contextual description provides evidence contradicting previously made parsing decisions. We expect the system to prefer an attachment of the phrase in line with the represented visual description.

Materials and design

For this experiment, we used the same materials as in Experiment 8. The capability of the system to act upon newly introduced visual context is tested by integrating the visual representation late during processing of a sentence. WCDG was started in incremental mode. From previously conducted experiments without context integration, we knew that the parser prefers the comitative interpretation for each sentence when being processed in incremental mode. We therefore introduced the contradictory instrument context late during processing. The timepoint of integration differs for each sentence, as the sentences are of different length and the context is integrated only after the language parser already has decided for a specific attachment of the prepositional phrase.

Results and Discussion

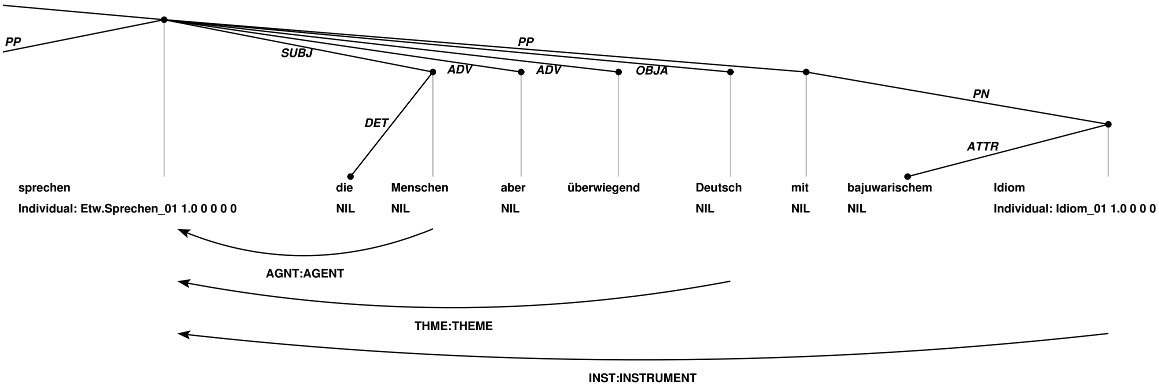


Figure 87: Dependency structure for a fragment of the sentence *Die Region ist offiziell zweisprachig, im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom*(The region is officially bilingual, but in everyday life people mainly speak German with a bavarian idiom) integrating the instrument context of Figure 83

Before the representation of visual context influences processing, the parser chooses the comitative reading as the most plausible one. Introducing an instrument context (see Figure 83) late during processing, the parser will re-evaluate its initial interpretation, reattaching semantic as well as syntactic edges of the dependency tree to conform with the visual evidence for an alternative reading (see Figure 87). This re-evaluation of previously made parsing decisions shows that the system can react to newly introduced context (the source of which could be any change in the visual environment).

6.2.4 Use-cases

In this section we test our system on tasks in a changing virtual environment. The goal of these tests is to show the systems performance in a field that resembles real-world environments with moving objects and a shifting point of view. The tests are carried out using the 3d-environment described in Section 5. We use scenes with a wide range of different objects and sentences describing a subset of them, including their spatial relationships. Dynamic changes in the visual modality are modeled by changing the users viewpoint and by moving the objects to different locations in the environment.

6.2.4.1 Use-case 1 : Evaluating the resolution of reference in a dynamically changing environment

In this use-case we apply our system to scenes of unmoving objects. The scenes contain several objects of a specific class making the link between the visual context and a natural language description a complex task as words of the sentence can refer to several visual entities. Our goal is to show that, depending on the dynamically changing environment and ambiguity inherent in the sentence, the system will produce different results for each sentence/scene combination. Each scene is observed from two different positions. The change of the viewpoint always results in a change of spatial relationships between objects.

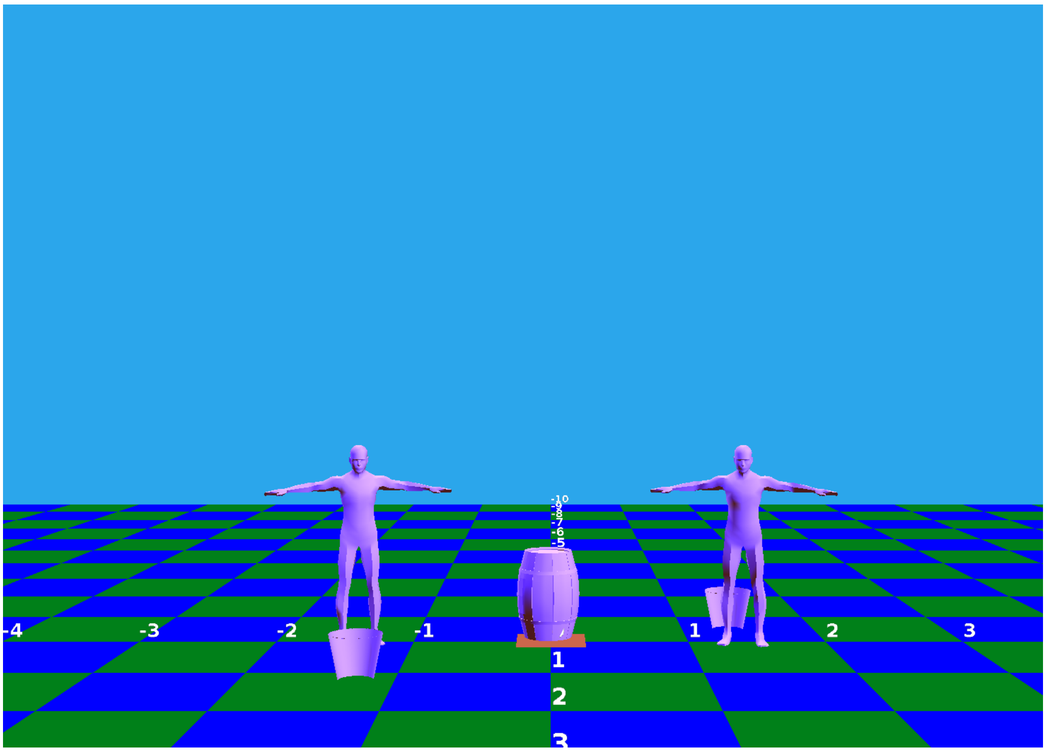


Figure 88: Scene of two persons, two buckets and a barrel: one person is standing in front of a bucket and one standing behind another bucket

bucket_1	$\xrightarrow{\text{has\_LOCATION\_vor}}$	Human_1
bucket_1	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	barrel_1
bucket_2	$\xrightarrow{\text{has\_LOCATION\_hinter}}$	Human_2
bucket_2	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	barrel_1
Human_1	$\xrightarrow{\text{has\_LOCATION\_hinter}}$	bucket_1
Human_1	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	barrel_1
Human_2	$\xrightarrow{\text{has\_LOCATION\_vor}}$	bucket_2
Human_2	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	barrel_1
barrel_1	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	bucket_1
barrel_1	$\xrightarrow{\text{has\_LOCATION\_rechts\_von}}$	Human_1
barrel_1	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	Human_2
barrel_1	$\xrightarrow{\text{has\_LOCATION\_links\_von}}$	bucket_2

Figure 89: Representation of the 3D-scene in Figure 88

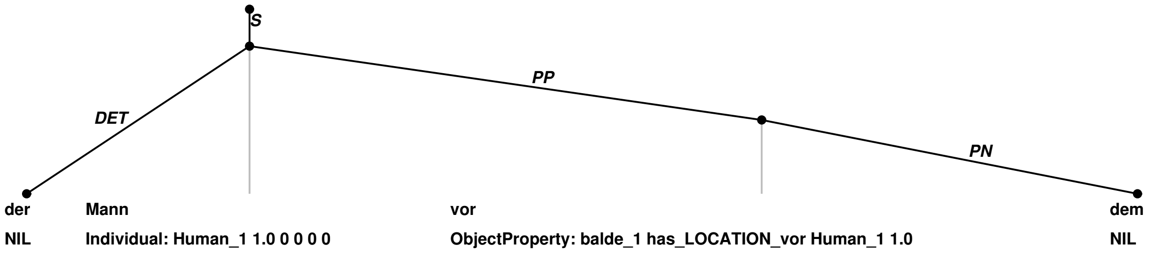

S
DET
PP
PN
der
NIL
Mann
Individual: Human_1 1.0 0 0 0 0
vor
ObjectProperty: balde_1 has_LOCATION_vor Human_1 1.0
dem
NIL

Figure 90: Dependency tree for the sentence fragment *Der Mann vor dem*(The man in front of the)

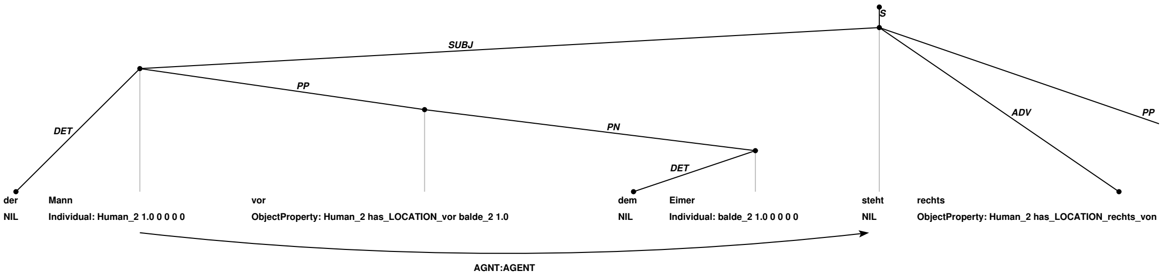

S
SUBJ
PP
DET
PN
DET
ADV
PP
der
NIL
Mann
Individual: Human_2 1.0 0 0 0 0
vor
ObjectProperty: Human_2 has_LOCATION_vor balde_2 1.0
dem
NIL
Eimer
Individual: balde_2 1.0 0 0 0 0
steht
NIL
rechts
ObjectProperty: Human_2 has_LOCATION_rechts_von
AGNT:AGENT

Figure 91: Dependency tree for the sentence fragment *Der Mann vor dem Eimer steht rechts*(The man in front of the bucket is standing to the right)

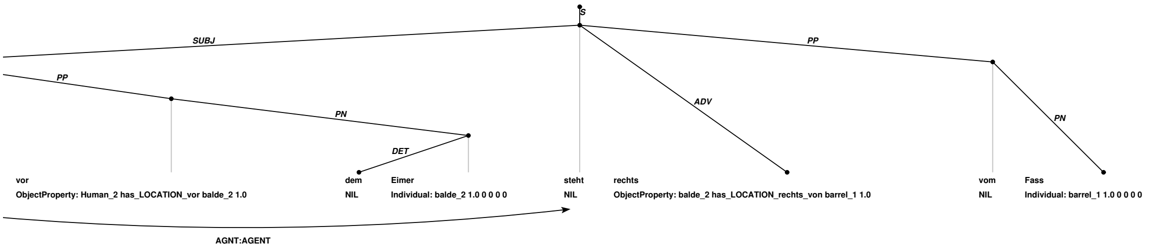

S
SUBJ
PP
PP
PN
DET
ADV
PN
vor
ObjectProperty: Human_2 has_LOCATION_vor balde_2 1.0
dem
NIL
Eimer
Individual: balde_2 1.0 0 0 0 0
steht
NIL
rechts
ObjectProperty: balde_2 has_LOCATION_rechts_von barrel_1 1.0
vom
NIL
Fass
Individual: barrel_1 1.0 0 0 0 0
AGNT:AGENT

Figure 92: Dependency tree for the sentence *Der Mann vor dem Eimer steht rechts vom Fass*(The man in front of the bucket is standing to the right of the barrel)

One example of this behavior is the sentence *Der Mann vor dem Eimer steht rechts von dem Fass*(The man in front of the bucket is standing to the right of the barrel) applied to the 3d-scene in Figure 88. The scene consists of one barrel, two persons and two buckets. Therefore finding the correct referents by using only word information will result in the system having the choice between several visual entities. As each class of object in the representation (see Figure 89) is denoted once in the test sentence, the objects can be used to construct 4 different scenes of reference.

Furthermore, those parts of the sentence referring to observed spatial relationships also allow 4 different combinations (as the words *vor*(in front of) can refer to a man in front of a bucket or a bucket in front of a man and *rechts*(to the right) can refer to the barrel to the right of a man or the man to the right of the barrel), resulting in 16 possible scenes of reference.

In order to influence language parsing depending on the object placement in the visual field, we used constraints that connect spatial relations with the attachment of PP-edges in the dependency tree of our parser. Using information integrated by these constraints, and by preferring scenes of semantically related entities the system decides for one specific scene in favor of the others.

The development of reference resolution depends on the results of incremental processing. In Figure 90, the system chooses **Human_1** (which is located on the left of the scene) as referent for the word *Mann*(man) and sticks to this decision after receiving the word *vor*(in front of). After parsing progresses to the word *Eimer*(bucket) the system re-evaluates its choice and switches to the other person (**Human_2**) presented in the scene (see Figure 91). The reason for this change is a visual constraint that influences reference resolution depending on PP-attachments suggested by WCDG. As the parser infers a syntactic relation between the words *Mann*(man), *vor*(in front of) and *Eimer*(bucket), the Connector chooses referents for these words that are in a corresponding relationship in the contextual representation. The system then sticks to this decision and finds correct referents for the words *rechts*(to the right) and *Fass*(barrel) (see Figure 92).

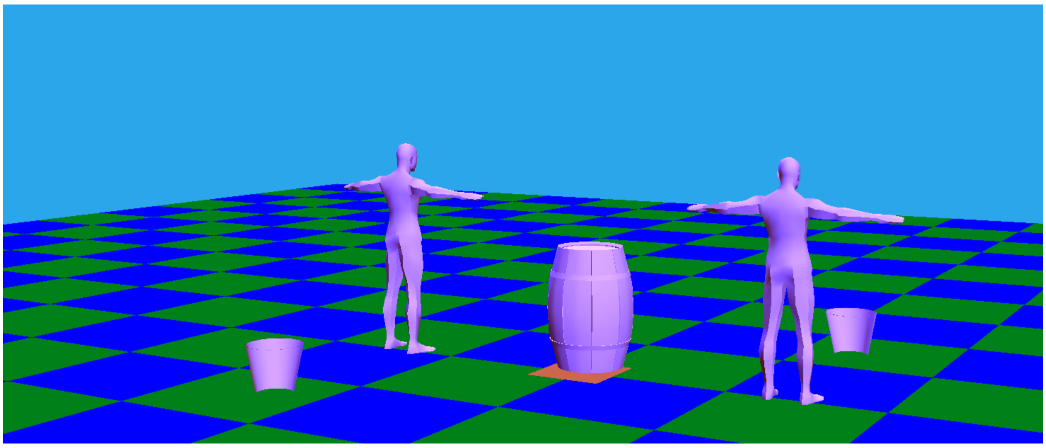


Figure 93: The same scene as in Figure 88 from a different viewpoint

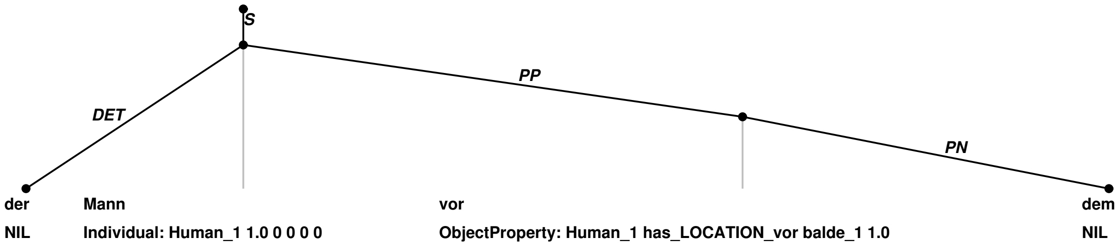

S
DET
PP
PN
der
NIL
Mann
Individual: Human_1 1.0 0 0 0 0
vor
ObjectProperty: Human_1 has_LOCATION_vor balde_1 1.0
dem
NIL

Figure 94: Dependency tree for the sentence fragment *Der Mann vor dem*(The man in front of the)

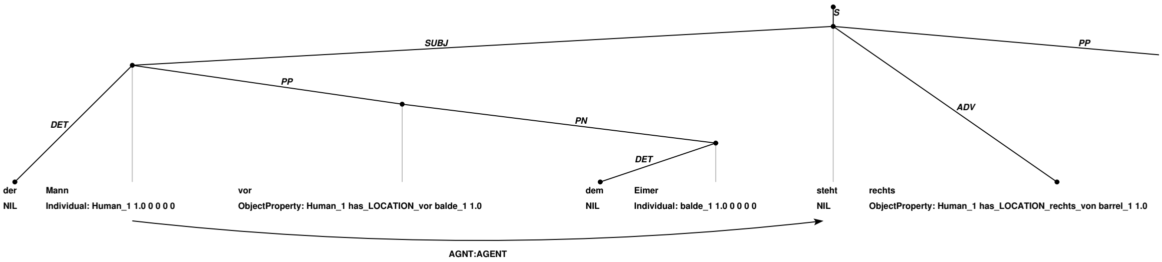

S
SUBJ
PP
ADV
DET
PP
PN
DET
der
NIL
Mann
Individual: Human_1 1.0 0 0 0 0
vor
ObjectProperty: Human_1 has_LOCATION_vor balde_1 1.0
dem
NIL
Eimer
Individual: balde_1 1.0 0 0 0 0
steht
NIL
rechts
ObjectProperty: Human_1 has_LOCATION_rechts_von barrel_1 1.0
AGNT:AGENT

Figure 95: Dependency tree for the sentence *Der Mann vor dem Eimer steht rechts vom Fass*(The man in front of the bucket is standing to the right of the barrel)

Changing the point of view (see Figure 93) the same sentence used on the same object configuration results in a different set of referents. Consistent with the spatial relationships between the objects, seen from the changed angle of view, the system chooses **Human_1** as the correct referent (Figure 94) and sticks to this decision in later parsing steps (see Figure 95) as it is standing in front of the bucket seen from this point of view.

An example for a more sophisticated reference resolution is the sentence *Der Polizist steht vor dem Laster links von der Palme*(The policeman is standing in front of the truck, to the left of the palm tree) used to describe the scene in Figure 96. In contrast to the previous example, the sentence is already ambiguous in itself due to its two possible attachments of the prepositional phrase *links von der Palme*(to the left of the palm tree). This phrase can either refer to the policeman or the truck being located to the left of the palm tree.

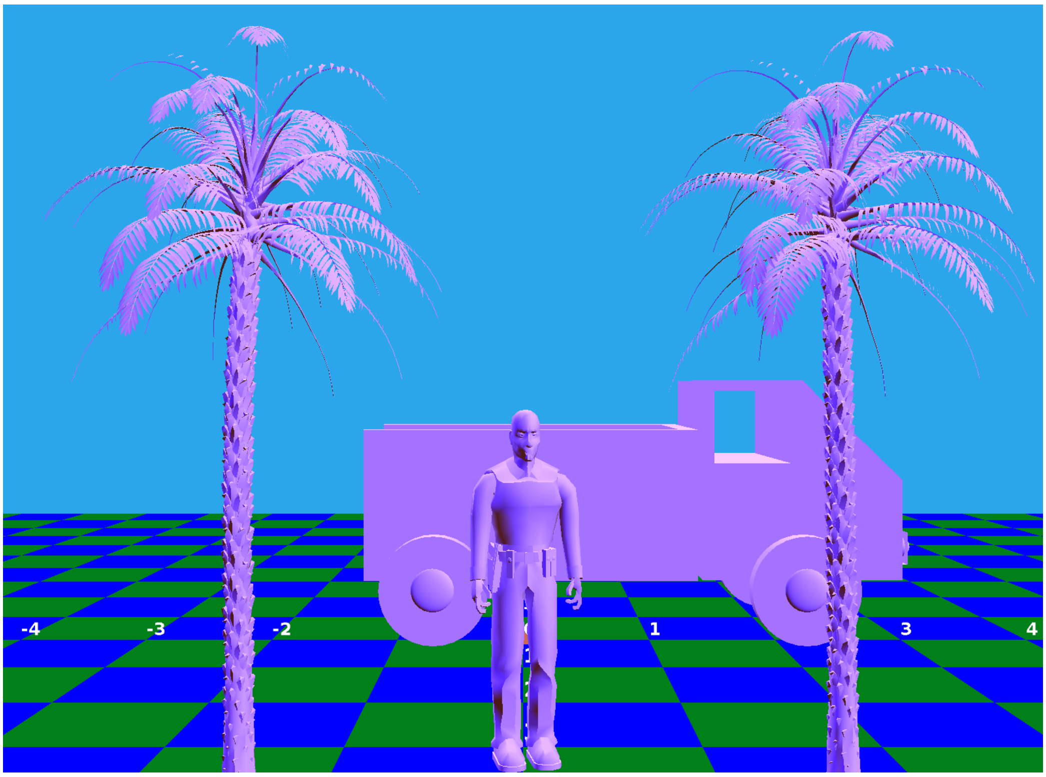


Figure 96: Picture of a 3d-environment containing a policeman, two palm trees and a truck

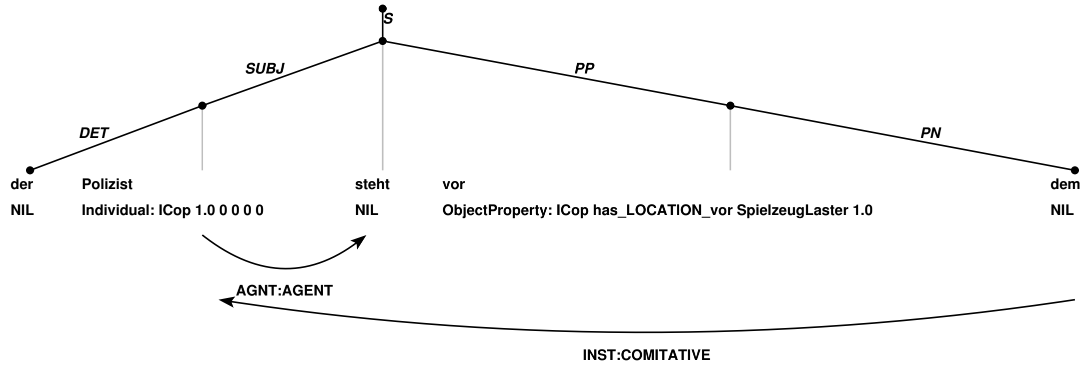

S
SUBJ
PP
DET
PN
der
NIL
Polizist
Individual: ICop 1.0 0 0 0 0
steht
NIL
vor
ObjectProperty: ICop has_LOCATION_vor SpielzeugLaster 1.0
dem
NIL
AGNT:AGENT
INST:COMITATIVE

Figure 97: Dependency tree for the sentence fragment *Der Polizist steht vor dem*(The policeman is standing in front of the)

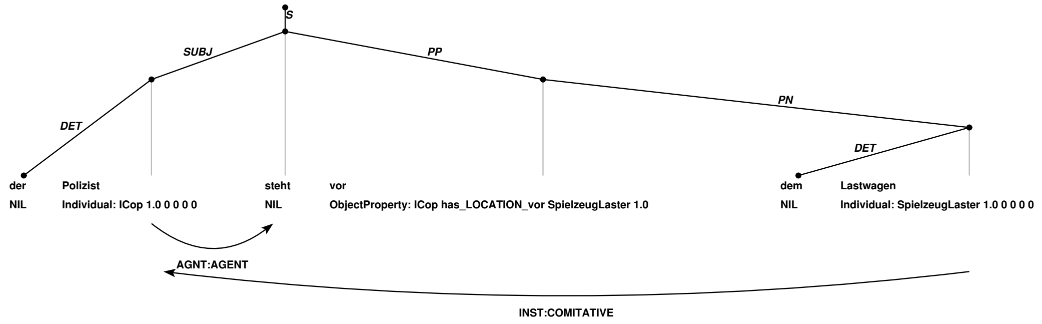

S
SUBJ
PP
DET
PN
DET
der
NIL
Polizist
Individual: ICop 1.0 0 0 0 0
steht
NIL
vor
ObjectProperty: ICop has_LOCATION_vor SpielzeugLaster 1.0
dem
NIL
Lastwagen
Individual: SpielzeugLaster 1.0 0 0 0 0
AGNT:AGENT
INST:COMITATIVE

Figure 98: Dependency tree for the sentence fragment *Der Polizist steht vor dem Lastwagen*(The policeman is standing in front of the truck)

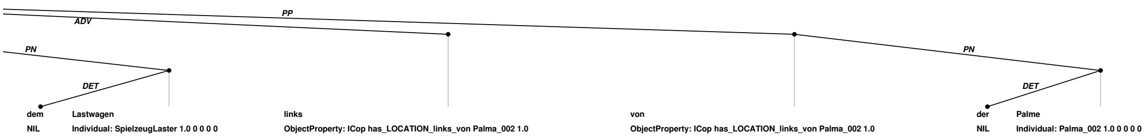

PP
ADV
PN
DET
PN
DET
dem
NIL
Lastwagen
Individual: SpielzeugLaster 1.0 0 0 0 0
links
ObjectProperty: ICop has_LOCATION_links_von Palma_002 1.0
von
ObjectProperty: ICop has_LOCATION_links_von Palma_002 1.0
der
NIL
Palme
Individual: Palma_002 1.0 0 0 0 0

Figure 99: Dependency tree for the sentence fragment *Der Polizist steht vor dem Lastwagen links von der Palme*(The policeman is standing in front of the truck to the left of the palm tree)

In this case, the system finds the correct referents up to the word *Lastwagen*(truck) (see Figure 98). This is not surprising for the words referring to observable objects(*Polizist*(Policemen) and *Lastwagen*(Truck)) have only one possible referent in the context. The correct spatial relationship for the word *vor*(in front of) is found as soon as it is part of the sentence fragment (Figure 97), although the context contains several such relations. The reason for this fact is that the system already has chosen the policeman as one referent and therefore links the word *vor* to the relation connected to the **ICop**, which is **ICop has_LOCATION_vor SpielzeugLaster**.
After receiving the words *links von*(to the left) the system incorrectly chooses a relation between the left palm tree and the policemen as referent by inferring that any relation that fits to the words and is connected to already found referents is a viable one. This mistake is corrected after finding the correct referent for the word *Palme*(palm tree) in Figure 99. This re-evaluation of incorrectly assigned referents is only possible due to the information provided by our integrating constraints. The effect of these constraints is an increase of the score of any scene containing a palm tree that is in a left(Cop, Palm) relationship due to the attachment of the prepositional phrase to the verb of the sentence. This attachment indicates that the policeman is standing to the left of the palm tree and not the truck, which only applies to **Palma_002**. After this correct referent is found the system consequently reassigns correct spatial relationships for the words *links von*. We can see in this example how information from words of the sentence, semantic relations and the results of language parsing are used in conjunction to find correct referents for a sentence and how this reference resolution is changed as soon as new language information is received.
When we modify the point of view in the scene (see Figure 100), the same sentence used on the same set of objects will bring forth a different result. As before, the integration of visual context will result in choosing **Palma_002** as the correct referent for the word *Palme*(palm tree) and **Truck** for the word *Lastwagen*(Truck) (see Figure 101). But in this case, the phrase *links von der Palme*(to the left of the palm tree) is attached to *Lastwagen*(truck) as the **Truck** is standing to the left of the **Palma_002**. What is important with regard to the referents found is, that the system includes the **Cop** in the set of visual entities although the relationship **vor**(in front of) is not applicable to **Truck** and **Cop**.
What is evident in this example is the systems ability to adapt its interpretation of language input to adhere to even minor changes in what is visually observable, such

as a slight change in position. Furthermore, the set of applicable referents is chosen even when parts of the sentence do not fit what is observed in the context, as long as the misfitting set of entities is the best choice for the fragment of the language input. This shows that the system is able to cope with conflicting information from the modalities while still being able to find meaningful referents

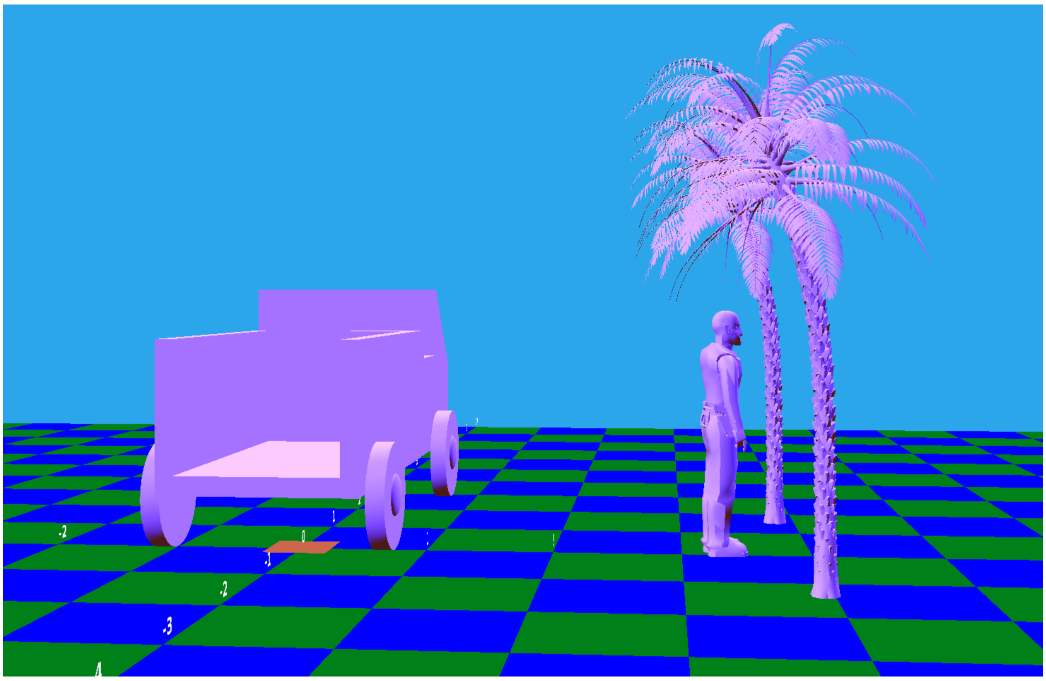


Figure 100: The same 3d-environment as in 96, but from a different position

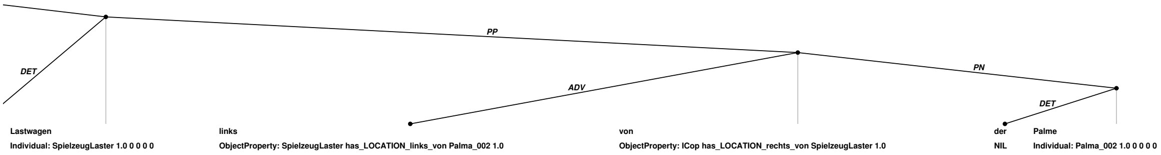

PP
DET
ADV
PN
DET
Lastwagen
Individual: SpielzeugLaster 1.0 0 0 0 0
links
ObjectProperty: SpielzeugLaster has_LOCATION_links_von Palma_002 1.0
von
ObjectProperty: ICop has_LOCATION_rechts_von SpielzeugLaster 1.0
der
NIL
Palme
Individual: Palma_002 1.0 0 0 0 0

Figure 101: Dependency tree for the sentence fragment *Der Polizist steht vor dem Lastwagen links von der Palme*(The policeman is standing in front of the truck to the left of the palm tree) after changing the point of view

6.2.4.2 Usecase 2 : Dynamically changing the environment as incremental parsing proceeds

In our second use-case we test the system in an environment consisting of objects that change their positions. At a crucial point during incremental parsing (when the system already decided which referents are the best fit for a sentence fragment depending on the position of objects), we change the positions of objects in the visual field. This results in different spatial relationships between them, forcing the system to re-evaluate its choice of referents.

As in the previous test, we use sentences and 3d-scenes. Instead of walking through the environment, we dynamically change the spatial configuration of objects at specific

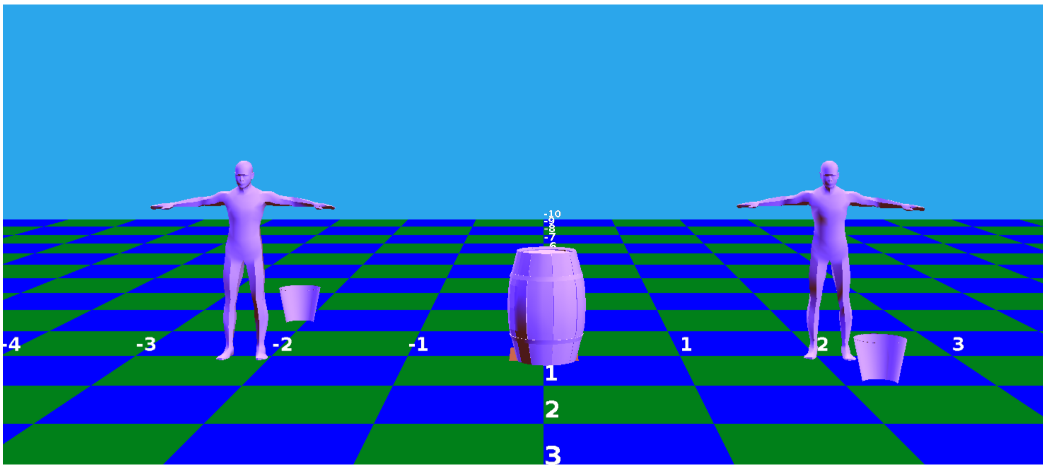


Figure 102: The same 3d-environment as in 88 with both buckets having changed their positions)

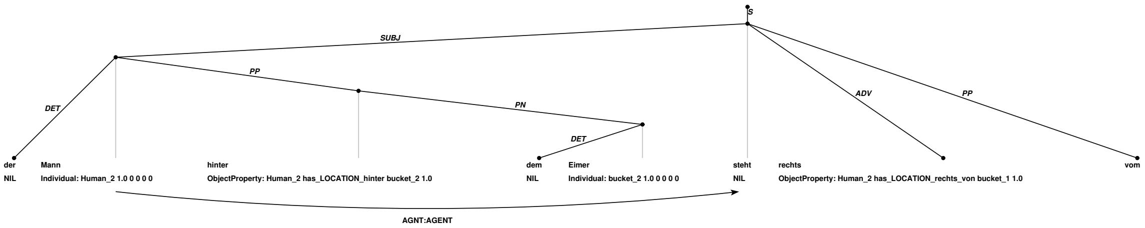

S
SUBJ
PP
PN
DET
DET
ADV
PP
der
NIL
Mann
Individual: Human_2 1.0 0 0 0 0
hinter
ObjectProperty: Human_2 has_LOCATION_hinter bucket_2 1.0
dem
NIL
Eimer
Individual: bucket_2 1.0 0 0 0 0
steht
NIL
rechts
ObjectProperty: Human_2 has_LOCATION_rechts_von bucket_1 1.0
vom
AGNT:AGENT

Figure 103: Dependency tree for the sentence fragment *Der Mann hinter dem Eimer steht rechts vom*(The man behind the bucket is standing to the right of)

points of parsing the sentence. In each picture, some types of objects are present more than once. We changed the environment at onset of a specific word of the sentence during incremental processing.

Using the example sentence *Der Mann hinter dem Eimer steht rechts vom Fass* (The man behind the bucket is standing to the right of the barrel), the system will choose one of the persons (**Human_1**) after processing the sentence fragment *Der Mann hinter dem Eimer steht rechts*(The man behind the bucket is standing to the right). The scene changes after onset of the word *rechts*(to the right) of the sentence by moving both buckets (compare Figures 88 and 102). This forces the system to re-evaluate its first choice of referential subset after the scene is changed to the one in Figure 103 assigning another referent to the word *Mann*(man) which is the right handed person **Human_2**. It persists in this choice unto the end of the sentence and chooses the barrel as the referent for the final word.

Testing the system on the dynamic development in the scene involving the policemen the truck and the two palm trees (see Figure 96) using again the sentence *Der Polizist steht vor dem Laster links von der Palme*(The policeman is standing in front of the truck, to the left of the palm tree) we observe the behavior of the system when we move an object after onset of the word *links*(left) by positioning the palm tree(**Palma_02**) to the right of the truck (see Picture 104). After this change in spatial relationships, the system finds the correct tree after onset of the word *Palme*(palm tree) (see Figure 106). Furthermore, the change in the trees position has an effect on language processing with regard to the attachment of the prepositional phrase *links von der Palme*(to the left of the palm tree): The phrase is attached to the verb of the sentence (see Figure 105) at the onset of the word *links*(left).

This interpretation is changed (see Figure 106) as soon as the word *Palme*(palm tree) is received due to the integration of the knowledge that the moved object is standing to the right of the truck in the 3d-scene. The language information is used to find correct referents (including the choice of the correct palm) and this set of referents is used to influence language processing such that the attachment of the prepositional phrase is corrected to be in line with spatial relationships in the visual context.

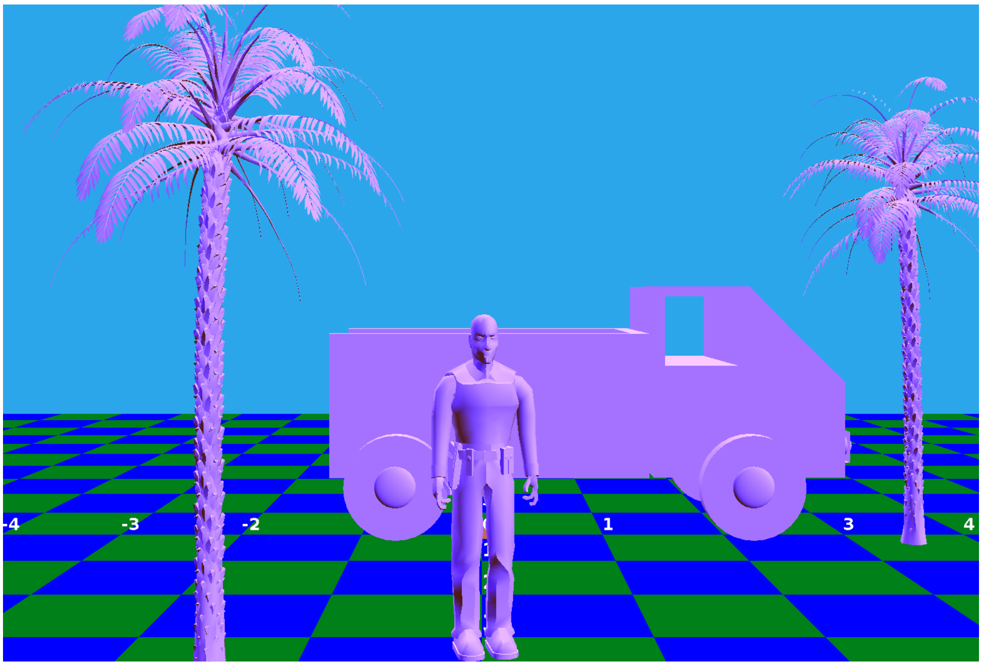


Figure 104: The same 3d-environment as in Figure 96 with one palm tree having moved its position to stand to the right of the truck

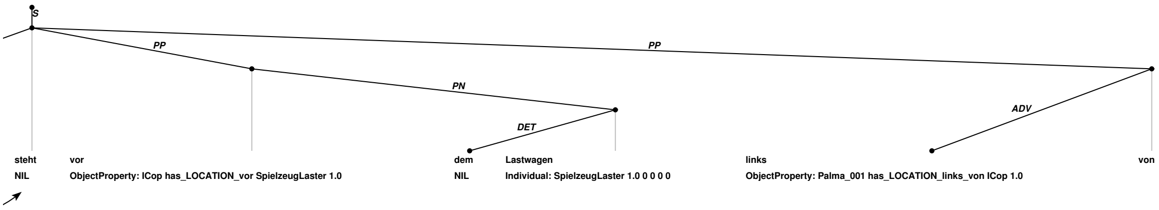

S
PP
PP
PN
DET
ADV
steht
NIL
vor
ObjectProperty: ICop has_LOCATION_vor SpielzeugLaster 1.0
dem
NIL
Lastwagen
Individual: SpielzeugLaster 1.0 0 0 0 0
links
ObjectProperty: Palma_001 has_LOCATION_links_von ICop 1.0
von

Figure 105: Dependency tree for the sentence fragment *Der Polizist steht vor dem Lastwagen links von*(The policeman is standing in front of the truck to the left of)

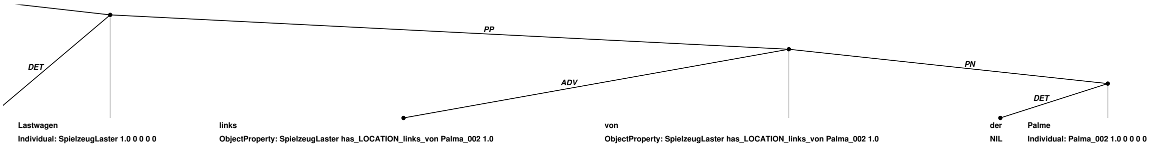

DET
PP
ADV
PN
DET
Lastwagen
Individual: SpielzeugLaster 1.0 0 0 0 0
links
ObjectProperty: SpielzeugLaster has_LOCATION_links_von Palma_002 1.0
von
ObjectProperty: SpielzeugLaster has_LOCATION_links_von Palma_002 1.0
der
NIL
Palme
Individual: Palma_002 1.0 0 0 0 0

Figure 106: Dependency tree for the sentence fragment *Der Polizist steht vor dem Lastwagen links von der Palme*(The policeman is standing in front of the truck to the left of the palm tree)

7 Future Work

One further area of research using the model described in this thesis is the application of highly dynamic scenes which are changing constantly either due to interactions with the environment triggered by the same agent that is also the source of the natural language input or by external events due to other agents. We have seen a toned down version of this approach in Section 6.2.4 when we investigated reference resolution in the 3d-environment. A promising application of our approach is the use in complex environments such as computer games (see [34] for one approach to integrate situated language understanding into such an environment).

The model has yet to be tested with complex scenes and sentences. Although we investigated this already in the experimental section by creating scenes for sentences from German newspaper articles (which already are complex, especially when compared to the linguistic input used in completely artificial experimental setups), this line of experiments can be pursued further by increasing the complexity of the contextual information available. To conduct a test in such an environment a corpus would have to be created that confronts human participants with situations where they have to describe their environment referring to objects and situations. A thorough approach to investigating the effects of the model would have to include scenes and descriptions about actions conducted by participants found in the visually observable environment.

Another interesting field of research which can be beneficial to further improve and test our model comes from the embodied mind thesis (see [5]) which holds that all aspects of cognition, including the attentional effects and the effect of language processing on them, is in large part dependent on the workings of the human body. For one example on how action intention can influence visual attention during search see [11], where errors of test subjects increased depending on whether they were told to either grasp or point at specific objects. At the moment our system is completely neglecting these effects. In our opinion the model is suitable for these kind of influences as it uses constraints which increase or decrease the saliency of visual entities depending on linguistic cues. A set of constraints which changes saliency depending on the level of embodiment found in the input instructions by means of natural language is imaginable.

Another idea is the integration of information that is not directly extractable from observing things in the visual field, but related to the visual information due to knowledge provided on the intermediate level of representation. We already did that with regard to the conceptual relatedness between objects but a way to go one step further would be the integration of affordance based information [32]. This means that objects in certain configurations, involving certain agents are affording some kinds of actions. For instance a cup standing on the table in front of a person affords the action of “grabbing” it. The integration of information about actions related to objects could help to overcome difficulties inherent in our approach of linking two modalities: as we used an intermediate level of semantic representations, including thematic roles of participants in actions, linking the descriptions of scenes contained in a sentence to what is visually observable is dependent on the link between actions found in the context and verbs referring to these actions. As the extraction of actions from visual stimuli is highly complicated, affordance based information might be a means to provide related action-information to the system.

The system as it is presently implemented suffers from the shortcoming that the bottom-up model does neither influence how high level representations of visual context are created, nor does it influence in any way the decision process for choosing the original subset of attention due to top-down cues such as natural language. The processes of bottom-up and top-down attention are merged solely on the level of saliency generation. It might be advantageous to integrate bottom-up cues also on a higher level of processing.

One effect of this integration would be the language dependent choice of a sub-scene of the available context. Although the linguistic modality points out specific visual entities during every step of incremental parsing, every other lower graded scene is still available to processing. High saliency of specific areas in the visual field could influence the assessment of any scene by drawing the focus of attention to elements located in these areas. Thus an alternative scene would be chosen despite not being the scene of reference chosen by the component for top-down influence of language.
the language channel.

8 Summary

We have proposed a detailed model for the integration of contextual information from the visual channel with a natural language processing component. We model processing of the two modalities by a broad coverage parser for unrestricted German input and a model for bottom-up visual attention. The system connects the two types of information on two levels: the first one is the level of visual representation, where visual entities are described in abstract terms of a formalism based on description logics. The second level of integration is a saliency map which shows the regions of a picture most likely to attract the attention of an observer. The rules governing the interaction of language phenomena with what is present in the visual field are specified as constraints.

Experimental results showed that the system is able to identify correct referents in visual contexts of varying complexity by combining a wide range of linguistic cues. When a sentence unfolds incrementally over time, the connection of parts of the sentence and parts of the context is changing as new linguistic information comes up. Although the system sometimes fails to find the correct referents for an utterance at the beginning of a sentence it will gradually improve its choice as the sentence unfolds, often finding the correct referent at a later time.

Given appropriate constraints that model the rules governing the interaction of language phenomena and visual knowledge, we have found that the system performs similar to humans on tasks such as identifying participants of a description that fit into a thematic role of an action denoted by a verb. This effect relies heavily on the prediction of future referents of a sentence depending on what is present in the visual field. The system thus goes beyond simple reference resolution due to linguistic input by entangling visual and linguistic information such that the information from both channels is not merely linked when present, but the system is actively searching possible future referents by making its predictions.

Furthermore, the integration of visual knowledge is able to solve language parsing problems with certain linguistic phenomena such as the attachment of prepositional phrases. The use-cases demonstrated one application of the model in a virtual environment which dynamically changes over time. The system identifies the correct objects and entities and changes its judgment depending on position and perspective of the observer. Even changing the spatial configuration of the objects in view has an effect on reference resolution and language processing.

9 References

[1] Paul D. Allopenna, James S. Magnuson, Michael K. Tanenhaus. 1998. Tracking the Time Course of spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, 38:419–439.

[2] Gerry T.M. Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73:247–264.

[3] Gerry T.M. Altmann and Jelena Mirkovic. 2009. Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, 33:583–609.

[4] Gerry T.M. Altmann and Yuki Kamide. 2009. Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, 111(1):55–71.

[5] Michael L. Anderson. 2003. Embodied Cognition: A field guide. *Artificial Intelligence*, 149:91-130.

[6] Richard Anderson, Fernanda Ferreira, John M. Henderson. 2011. I see what you're saying: the integration of complex speech and scenes during language comprehension. *Acta Psychologica.*

[7] Jennifer E. Arnold. 1998. Reference Form and Discourse Patterns. *Stanford*, Ph.D. Thesis.

[8] Jennifer E. Arnold. 2001. The Effect of Thematic Roles on Pronoun Use and Frequency of Reference Continuation. *Discourse Processes*, 31(2):137–162.

[9] Christopher Baumgärtner and Wolfgang Menzel. 2012. Integrating a Model for Visual Attention into a System for Natural Language Parsing *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2012* , 24–33 .

[10] Christopher Baumgärtner, Niels Beuck and Wolfgang Menzel. 2012. An Architecture for Incremental Information Fusion of Cross-Modal Representations *Proceedings of the Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE* , 498–503 .

[11] Harold Bekkering, Sebastian F.W. Neggers. 2002. Visual Search is Modulated by Action Intentions *Psychological Science*, 13:370–374.

[12] Niels Beuck, Arne Koehn, and Wolfgang Menzel 2011. Incremental parsing and the evaluation of partial dependency analyses *In Proceedings of the 1st International Conference on Dependency Linguistics, DepLing-2011*, 290–299.

[13] Blanchard, E., Harzallah, M., Briand, H., and Kuntz, P. 2005. A Typology Of Ontology-Based Semantic Measures. *Proceedings of EMOI-INTEROP*

[14] Peter Burt and Ted Adelson 1983. The Laplacian Pyramid as a Compact Image Code *IEEE Trans. Communications*, 9(4):532—540.

[15] Craig G. Chambers, James S. Magnuson, Michael K. Tanenhaus. 2004. Actions and Affordances in Syntactic Ambiguity Resolution. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30(3):687-696.

[16] Champollion, Lucas, Sauerland, Uli. 2010. Move and accommodate: A solution to Haddock's puzzle. *Empirical Issues in Syntax and Semantics*, 8, Olivier Bonami and Patricia Cabredo Hofherr (eds.).

[17] E. Colin Cherry 1953. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25:975–979

[18] Roger M. Cooper 1974. A New Methodology for the Real-Time Investigation of Speech and Perception, Memory, and Language Processing *Cognitive Psychology*, 6(1):84–107.

[19] Kenny R. Coventry, Dermot Lynott, Angelo Cangelosi, Lynn Monrouxe, Dan Joyce, Daniel C. Richardson. 2010. Spatial language, visual attention, and perceptual simulation. *Brain and Language*, 112(3):202–213.

[20] George S. Cree and Ken McRae. 2003. Analyzing the Factors Underlying the Structure and Computation of the Meaning of Chipmunk, Cherry, Chisel, Cheese, and Cello (and Many Other Such Concrete Nouns) *Journal of Experimental Psychology: General*, 132(2):163–201.

[21] Delphine Dahan, James S. Magnuson, Michael K. Tanenhaus, Ellen M. Hogan. 2001. Subcategorical mismatches and the time course of lwxical access: Evidence for lexical competition. *Language and Cognitive Processe*, 16(5/6):507–534.

[22] Frederick J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171-176.

[23] Gautam Das, Rudolf Fleischer, Leszek Gasieniec, Dimitrios Gunopulos, Juha Kärkkäinen. 1997. Episode Matching. *Proceedings of the CPM 97*, 1264:12-27.

[24] Robert Desimone, John Duncan 1995. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222.

[25] David DeVault, Mathew Stone 2003. Domain inference in incremental interpretation. *ICOS 4: Workshop on Inference in Computational Semantics.*, 18:193–222.

[26] Kathleen M. Eberhard, Michael J. Spivey-Knowlton, Julie C.Sedivy, Michael K. Tanenhaus. 1995. Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. *Journal of Psycholinguistic Research*, 24:409–436.

[27] Howard E. Egeth, Steven Yantis. 1997. Visual attention: Control, representation, and time course. *Annual Review of Psychology*, 48: 269-297.

[28] Jacob Feldman. 2003. The Simplicity Principle in Human Concept Learning. *Current Directions in Psychological Science, American Psychological Society*, 227-232.

[29] Jerry A. Fodor. 1983. Modularity of Mind.

[30] Kilian Foth. 1999. Transformationsbasiertes Constraint-Parsing *Diplomarbeit Universität Hamburg.*

[31] Kilian Foth. 2006. Hybrid Methods Of Natural Language Analysis *PhD Thesis Universität Hamburg.*

[32] James J. Gibson. 2004. The Ecological Approach to Visual Perception. *Boston: Houghton Mifflin*

[33] Peter Gorniak and Deb Roy. 2004. Grounded Semantic Composition for Visual Scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

[34] Peter Gorniak, Jeff Orkin, and Deb Roy. 2005. Speaking with your Sidekick: Understanding Situated Speech in Computer Role Playing Games. *Proceedings of Artificial Intelligence and Digital Entertainment*

[35] Peter Gorniak, Jeff Orkin, and Deb Roy. 2006. Speech, Space and Purpose: Situated Language Understanding in Computer Games. *Twenty-eighth Annual Meeting of the Cognitive Science Society Workshop on Computer Games*

[36] Thomas Gruber. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal Human-Computer Studies*, 43(5-6):907–928.

[37] Nicholas J. Haddock. 1988. Incremental Semantics and Interactive Syntactic Processing *PhD University of Edinburgh*

[38] Nicholas J. Haddock. 1989. Computational models of incremental semantic interpretation. *Language and Cognitive Processes*, 4(3):337–368.

[39] R.W. Hamming. 1950. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29:147–160.

[40] Graeme Hirst , David St-Onge. 1998. Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms. *WordNet: An electronic lexical database*, 305–332.

[41] Evgenia Hristova, Severina Georgieva, Maurice Grinberg. 2011. Top-Down Influences on Eye-Movements during Painting Perception: The Effect of Task and Titles *Lecture Notes in Computer Science*, 6456/2011:104-115.

[42] Albert S. Huang , Stefanie Tellex , Abraham Bachrach , Thomas Kollar , Deb Roy , Nicholas Roy. 2010. Natural Language Command of an Autonomous Micro-air Vehicle. *International Conference on Intelligent Robots and Systems.*, 2010:2663-2669.

[43] Falk Huettig, Gerry T.M. Altmann. 2005. Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition* 96:23–32.

[44] Falk Huettig, James McQueen. 2007. The tug of war between phonological, semantic and shape information in language mediated visual search *Journal of Memory and Language* 57:460–482.

[45] Falk Huettig, Christian N.L. Olivers, Robert J. Hartsuiker. 2010. Looking, language and memory: Bridging research from the visual world and visual search paradigms. *Acta Psycholgica.*

[46] Falk Huettig, Joost Rommers, Antje S. Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica.*

[47] Laurent Itti. 2000. Models of Bottom-Up and Top-Down Visual Attention. *California Institute of Technology*, Ph.D. Thesis.

[48] Jay Jiang, David Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceeding of International Conference Research on Computational Linguistics*, 24–26

[49] Elsi Kaiser, John C. Trueswell 2004. The role of discourse context in the processing of a felxible word-order language. *Cognition* Cognition 94(2), 113–147.

[50] Elsi Kaiser, John C. Trueswell 2008. Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes* 23(5):709-748.

[51] Elsi Kaiser, Jeffrey T. Runner, Rachel S. Sussman, Michael K. Tanenhaus. 2009. Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition* 112:50-80.

[52] Yuki Kamide, Christoph Scheepers, Gerry T.M. Altmann. 2003. Integration of Syntactic and Semantic Information in Predictive Processing: Cross-Liguistic Eveidence from German and English *Journal of Psycholinguistic Research* 32(1) 37–55.

[53] John D. Kelleher , Fintan J. Costello . 2009. Applying computational models of spatial prepositions to visually situated dialog *Computational Linguistics* 35(2):271–306.

[54] Pia Knoeferle. 2005. The Role of Visual Scenes in Spoken Language Comprehension: Evidence from Eye-Tracking. *PhD thesis Universität des Saarlandes.*

[55] Pia Knoeferle, Mathew W. Crocker, Christoph Scheepers, Martin J. Pickering. 2005. The influence of the immediate visual context on incremental thematic role-assignment evidence from eye-movements in depicted events. *Cognition*, 95:95–127.

[56] Pia Knoeferle, Mathew W. Crocker. 2006. The Coordinated Interplay of Scene, Utterance, and World Knowledge: Evidence From Eye Tracking. *Cognitive Science*, 30(3):481–592.

[57] Pia Knöferle, Mathew W. Crocker. 2007. The influence of recent scene events on spoken comprehension: evidence from eye movements. *Journal of Memory and Language*, 57(2):519-543.

[58] Christof Koch and Shimon Ullman. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4(4):219–227

[59] Geert-jan M. Kruijff , Pierre Lison , Trevor Benjamin , Henrik Jacobsson , Nick Hawes. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. *Proceedings of the Symposium on Language and Robots*, 55–64

[60] Stephen Laurence and Eric Margolis. 1999 Concepts and Cognitive Science. *In E. Margolis & S. Laurence (eds.) Concepts: Core Readings, Bradford Books/MIT Press*, 3–81.

[61] Claudia Leacock, Martin Chodorow. 1998 Combining Local Context and WordNet Similarity for Word Sense Identification. *An Electronic Lexical Database In WordNet: A Lexical Reference System and its Application*, 265–283.

[62] Vladimir Levenshtein. 1966 Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10: 707–10.

[63] Audie G. Leventhal. 1991 The Neural basis of visual function. *CRC Press*

[64] Dekang Lin. 1998 An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conference on Machine Learning*, 133–138.

[65] Sebastian Löbner 2002 Understanding Semantics *London: Arnold*

[66] William D. Marslen-Wilson. 1987. Functional parallelism in spoken word-recognition. *Cognition*, 25:71–102.

[67] Marshall Mayberry, Mathew W. Crocker, Pia Knoeferle. 2009. Learning to Attend: A Connectionist Model of Situated Language Comprehension. *Cognitive Science*, 33:449-496.

[68] Patrick McCrae. 2009. A model for the cross-modal influence of visual context upon language processing. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 230–235.

[69] Wolfgang Menzel. 2009. Towards radically incremental parsing of natural language. *Current Issues in Linguistic Theory*, 309:41–56.

[70] Daniel Mirman, James S. Magnuson. 2009. Dynamics of activation of sematically similar concepts durong spoken word recognition. *Memory & Cognition*, 37(7):1026–1039.

[71] Gonzalo Navarro . 2001. A Guided Tour to Approximate String Matching . *ACM Computing Surveys*, 33(1):31–88.

[72] Saul B. Needleman, Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453

[73] W3C-World Wide Web Consortium. 2004. OWL Reference, 10.02.2004. *http://www.w3.org/TR/2002/REC-owl-ref-20040210.*

[74] Michael I. Posner. 1980. Orienting of attention. *Quat. J. Exper. Psych*, 32:2-25.

[75] Michael I. Posner, Steven E. Petersen. 1990. The attention system of the human brain. *Annual Review of Neuroscience*, 13:25-–42..

[76] D. R. Powell, L. Allison and T. I. Dix. 1999. A versatile divide and conquer technique for optimal string alignment. *Information Processing Letters*, 70(3):127–139.

[77] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner 1989. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems Management and Cybernetics*, 19:17–30.

[78] Keith Rayner and Arnold D. Well 1996. Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin and Review*, 3(4):504–509.

[79] Keith Rayner, Xingshan Li and Barbara J. Juhasz 2005. The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin and Review*, 12(6):1089–1093.

[80] Terry Regier and Laura Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130:273–298.

[81] Hilke Reckman, Jeff Orkin and Deb Roy. 2011. Extracting aspects of determiner meaning from dialogue in a virtual world environment. *Proceedings of the Ninth International Conference on Computational Semantics (IWCS)*, 245–254.

[82] Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI'95)*, 1:448-453.

[83] Jeffrey N. Rouder and Roger Ratcliff. 2004. Comparing Categorization Models. *Current Directions in Psychological Science*, 15:9–13.

[84] SALSA Corpus Homepage. 2012. http://www.coli.uni-saarland.de/projects/salsa/page.php?id=index. *Link verified: 19.09.2012*

[85] David Sankoff, Joseph Kruskal. 1983. Time Warps, String Edits, and Macromolecules - The Theory and Practice of Sequence Comparison. *Addison-Wesley, Reading, MA.*

[86] Mathias Scheutz, Kathleen Eberhard, Virgil Andronache. 2004. A Real-time Robotic Model of Human Reference Resolution using Visual Constraints. *Connection Science Journal*, 16(3):145–167.

[87] David Schlangen, Timo Baumann, Michaela Atterer 2009. Incremental Reference resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. *Proceedings of SigDial 2009*, 30–37.

[88] William Schuler 2003. Using model-theoretic semantic interpretation to guide statistical parsing and word recogntion in a spoken language interface. *Proceedings if the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, 529–536

[89] William Schuler, Stephen Wu, Lane Schwartz 2009. A framework for fast incremental interpretation during speech decoding. *Computational Linguitics*, 35(3):313–343.

[90] Ingo Schröder. 2002. Natural Language Parsing with Graded Constraints. *PhD Thesis Universität Hamburg.*

[91] Julie C. Sedivy, Michael K. Tanenhaus, Craig G. Chambers, Gregory N. Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71:109–147.

[92] Karen O. Solomon, Douglas L. Medin, Elizabeth Lynch. 1999. Concepts do more than categorize. *Trends in Cognitive Science*, 3(3):99–105.

[93] Michael J. Spivey-Knowlton, John C Trueswell, Michael Tanenhaus. 1993. Context effects in syntactic ambiguity resolution: discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology*, 37:276–309.

[94] Jesse Snedeker, John C. Trueswell 2004. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, 49:238–299.

[95] Alexander Siebert, David Schlangen 2008. A simple method for resolution of definite reference in a shared visual context. *Proceeding SIGdial '08 Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 84-87.

[96] Mark Steedman 1987. Combinatory grammars and parasitic gaps. *Natural Language and Linguistic Theory*, 5:403–439.

[97] Elizabeth A. Styles 1997. The Psychology of Attention. *Psychology Press*

[98] Elizabeth A. Styles 2005. Attention, Preception and Memory. An Integrated Introduction. *Psychology Press*

[99] Michael Sussna 1993. Word sense disambiguation for free-text indexing using a massive semantic network. *CIKM '93 Proceedings of the second international conference on Information and knowledge management* 67–74

[100] Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, Julie Sedivy. 1995. Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science In Science*, 268(5217):1632–1634.

[101] Michael K. Tanenhaus, Kathleen M. Eberhard, Julie Sedivy. 1996. Using eye-movements to study spoken language comprehension: Evidence for visually mediatedincremental interpretation. *In T. Inui & J.L. McClelland (Eds.). Attention & Performance XVI: Information integration in perception and communication*, 457–478.

[102] Christopher Phillip Town. 2004. Ontology based visual information processing. *University of Cambridge (Trinity College)*, Ph.D. Thesis.

[103] Robert A. Wagner, Michael J. Fischer. 1974. The String-to-String Correction Problem. *Journal of the ACM.* 21(1):168–173

[104] Terry Winograd. 1973. A Procedural Model of Language Understanding. *Computer models of thought and language.*

[105] Zhibiao Wu, Martha Palmer. 1994. Verb semantics and lexical selection. *32nd Annual Meeting of the Association for Computational Linguistics* 133–138.

[106] Stephen Wu , Lane Schwartz , William Schuler. 2008. Exploiting Referential Context in Spoken Language Interfaces for Data-Poor Domains. *Proceedings of the 2008 International Conference on Intelligent User Interfaces* 285–292.

10 Appendix

10.1 List of Publications

Christopher Baumgärtner and Wolfgang Menzel. 2012. Integrating a Model for Visual Attention into a System for Natural Language Parsing *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2012* , 24–33 .

Christopher Baumgärtner, Niels Beuck and Wolfgang Menzel. 2012. An Architecture for Incremental Information Fusion of Cross-Modal Representations *Proceedings of the Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE* , 498–503 .

10.2 Sentences

10.2.1 AGENT-PATIENT-Ambiguous Sentences

(Taken from [54])
01a Die Prinzessin wäscht offensichtlich den Pirat.
01a Die Prinzessin malt offensichtlich der Fechter.
01b Die Prinzessin malt offensichtlich den Fechter.
01b Die Prinzessin wäscht offensichtlich der Pirat.
02a Die Amazone erdolcht gerade den Mechaniker.
02a Die Amazone besprüht gerade der Fußballspieler.
02b Die Amazone besprüht gerade den Fußballspieler.
02b Die Amazone erdolcht gerade der Mechaniker.
03a Die Krankenschwester schubst in diesem Moment den Sportler.
03a Die Krankenschwester fönt in diesem Moment der Priester.
03b Die Krankenschwester fönt in diesem Moment den Priester.
03b Die Krankenschwester schubst in diesem Moment der Sportler.
04a Die Journalistin fesselt in diesem Moment den Matrosen.
04a Die Journalistin füttert in diesem Moment der Oberarzt.
04b Die Journalistin füttert in diesem Moment den Oberarzt.
04b Die Journalistin fesselt in diesem Moment der Matrose.

05a Die Bauarbeiterin attackiert offensichtlich den Cellist.
05a Die Bauarbeiterin interviewt offensichtlich der Golfer.
05b Die Bauarbeiterin interviewt offensichtlich den Golfer.
05b Die Bauarbeiterin attackiert offensichtlich der Cellist.
06a Die Teufelin beschenkt in diesem Moment den Clown.
06a Die Teufelin skizziert in diesem Moment der Koch.
06b Die Teufelin skizziert in diesem Moment den Koch.
06b Die Teufelin beschenkt in diesem Moment der Clown.
07a Die Putzfrau bewirft soeben den Kellner.
07a Die Putzfrau ohrfeigt soeben der Ritter.
07b Die Putzfrau ohrfeigt soeben den Ritter.
07b Die Putzfrau bewirft soeben der Kellner.
08a Die Schlittschuhläuferin schrubbt mal eben den Detektiv.
08a Die Schlittschuhläuferin stupst mal eben der Zauberer.
08b Die Schlittschuhläuferin stupst mal eben den Zauberer.
08b Die Schlittschuhläuferin schrubbt mal eben der Detektiv.
09a Das Dienstmädchen parfümiert in diesem Moment den Henker.
09a Das Dienstmädchen bandagiert in diesem Moment der Trommler.
09b Das Dienstmädchen bandagiert in diesem Moment den Trommler.
09b Das Dienstmädchen parfümiert in diesem Moment der Henker.
10a Die Tennisspielerin boxt hier den Sträfling.
10a Die Tennisspielerin kämmt hier der Flötist.
10b Die Tennisspielerin kämmt hier den Flötist.
10b Die Tennisspielerin boxt hier der Sträfling.
11a Die Meerjungfrau krönt gerade den Student.
11a Die Meerjungfrau zupft gerade der Soldat.
11b Die Meerjungfrau zupft gerade den Soldat.
11b Die Meerjungfrau krönt gerade der Student.
12a Die Nonne impft gerade den Schülerlotsen.
12a Die Nonne zwickt gerade der Klarinettist.
12b Die Nonne zwickt gerade den Klarinettist.
12b Die Nonne impft gerade der Schülerlotse.
13a Die Oma kratzt soeben den Bogenschützen.
13a Die Oma filmt soeben der Saxophonist.

13b Die Oma filmt soeben den Saxophonist.

13b Die Oma kratzt soeben der Bogenschütze.

14a Die Fee bürstet hier den Gangster.

14a Die Fee bespritzt hier der Tourist.

14b Die Fee bespritzt hier den Tourist.

14b Die Fee bürstet hier der Gangster.

15a Die Joggerin verhext mal eben den Doktor.

15a Die Joggerin frottiert mal eben der König.

15b Die Joggerin frottiert mal eben den König.

15b Die Joggerin verhext mal eben der Doktor.

16a Die Cheerleaderin verprügelt offensichtlich den Pagen.

16a Die Cheerleaderin vergiftet offensichtlich der Angler.

16b Die Cheerleaderin vergiftet offensichtlich den Angler.

16b Die Cheerleaderin verprügelt offensichtlich der Page.

17a Die Braut verhaut gerade den Pfadfinder.

17a Die Braut verbrüht gerade der Postbote.

17b Die Braut verbrüht gerade den Postboten.

17b Die Braut verhaut gerade der Pfadfinder.

18a Die Stewardess pudert soeben den Leichtathlet.

18a Die Stewardess rempelt soeben der Wanderer.

18b Die Stewardess rempelt soeben den Wanderer.

18b Die Stewardess pudert soeben der Leichtathlet.

19a Die Hexe bestrahlt hier den Zeitungsverkäufer.

19a Die Hexe bestiehlt hier der Strassenkehrer.

19b Die Hexe bestiehlt hier den Strassenkehrer.

19b Die Hexe bestrahlt hier der Zeitungsverkäufer.

20a Die Japanerin beschmiert mal eben den Kameramann.

20a Die Japanerin bekränzt mal eben der Ordnunghüter.

20b Die Japanerin bekränzt mal eben den Ordnungshüter.

20b Die Japanerin beschmiert mal eben der Kameramann.

21a Die Rollstuhlfahrerin kostümiert hier den Schiedsrichter.

21a Die Rollstuhlfahrerin besoldet hier der Chinese.

21b Die Rollstuhlfahrerin besoldet hier den Chinesen.

21b Die Rollstuhlfahrerin kostümiert hier der Schiedsrichter.

22a Die Geschäftsfrau umgürtet mal eben den Klempner.
22a Die Geschäftsfrau verköstigt mal eben der Imker.
22b Die Geschäftsfrau verköstigt mal eben den Imker.
22b Die Geschäftsfrau umgürtet mal eben der Klempner.
23a Die Badenixe maskiert soeben den Skifahrer.
23a Die Badenixe entlohnt soeben der Musketier.
23b Die Badenixe entlohnt soeben den Musketier.
23b Die Badenixe maskiert soeben der Skifahrer.
24a Die Aerobic-Trainerin verwarnt offensichtlich den Astronaut.
24a Die Aerobic-Trainerin bekocht offensichtlich der Handwerker.
24b Die Aerobic-Trainerin bekocht offensichtlich den Handwerker.
24b Die Aerobic-Trainerin verwarnt offensichtlich der Astronaut.

10.2.2 Prepositional-Phrase-Ambiguous Sentences from the SALSA Corpus

s3025 Dort griff die Polizei unter Gewaltanwendung einzelne Demonstranten heraus, wobei Tritte und Schläge mit dem Knüppel von Polizisten beobachtet wurden.
s3277 Nach Darstellung der Nippon-Firma hatten Gewerkschaftsvertreter Bezahlung für die Zeit verlangt, die sie während früherer Streiks in Verhandlungen mit der Unternehmensleitung verbrachten.
s3839 Staatschef Soares argumentiert, daß die Regierung nicht einfach Gesetze mit früheren Laufbahnzusagen kurzfristig ändern könne.
s5459 Im Wahlkampf 1992 trat Rabin mit dem Versprechen an, "in neun Monaten" Frieden mit Israels arabischen Nachbarn zu schaffen.
s7177 Insgesamt werden Braunkohlemeiler mit zusammen 8500 Megawatt (MW) abgeschaltet.
s7650 Die ganze Nacht über landeten auf dem internationalen Ben-Gurion-Flughafen Flugzeuge mit Trauergästen.
s17512 Die höchsten Renditen erwirtschafteten Agenturen in Hauptgeschäftslagen von Orten mit 10 000 bis 500 000 Einwohnern.
s19569 Um Familien mit Kindern nicht zusätzlich zu belasten, wird eine Neuordnung des Familienlastenausgleichs und des Steuerrechts erwartet.
s23135 Vorrangig erwirbt der Pool Strom zu höheren Preisen (Einspeisevergütung) aus regenerativen Quellen (Wind, Wasser, Sonne, Biogas) und Anlagen mit Kraft-

Wärme-Kopplung.
s28600 "Wir richten uns nicht nach Müller und Schulz", wehrt er Vergleiche mit der Konkurrenz ab.
s31611 Die Region ist offiziell zweisprachig, im Alltag sprechen die Menschen aber überwiegend Deutsch mit bajuwarischem Idiom.

10.3 Pictures

(Taken from [54])

10.3.1 AGENT-PATIENT-Ambiguous Pictures

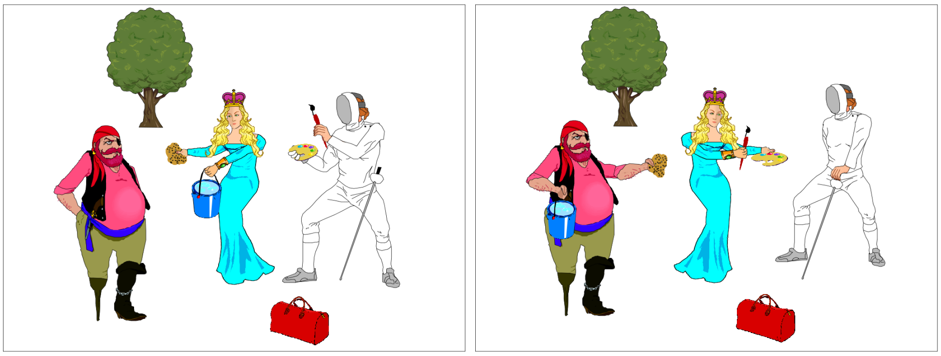


Figure 107: Scene 01

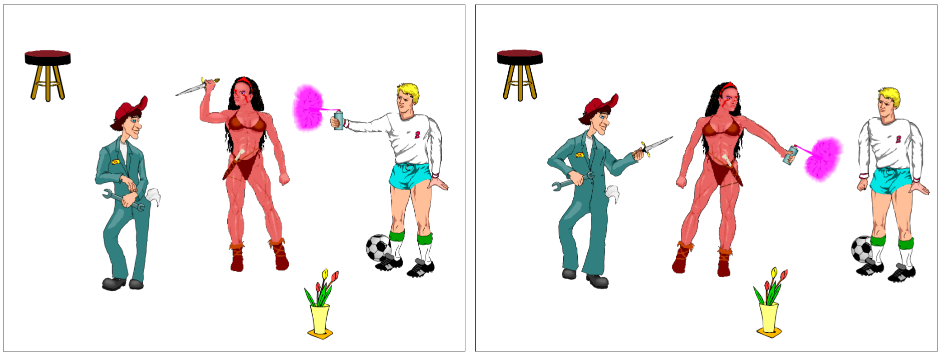


Figure 108: Scene 02

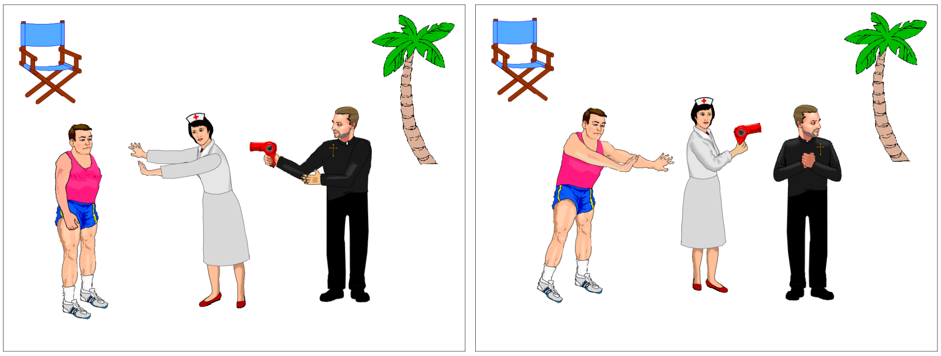


Figure 109: Scene 03

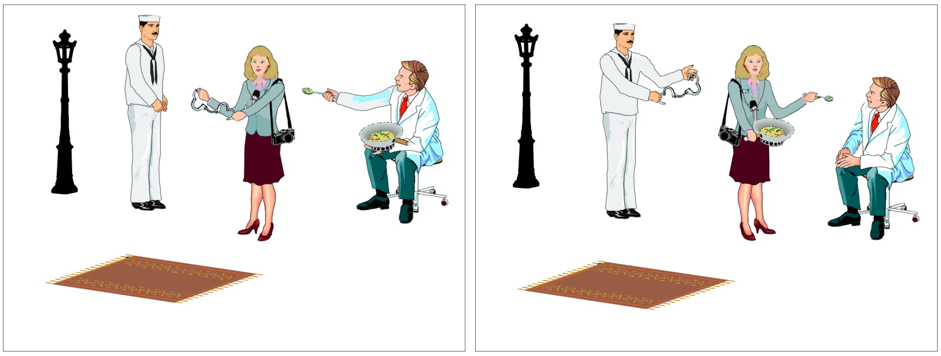


Figure 110: Scene 04

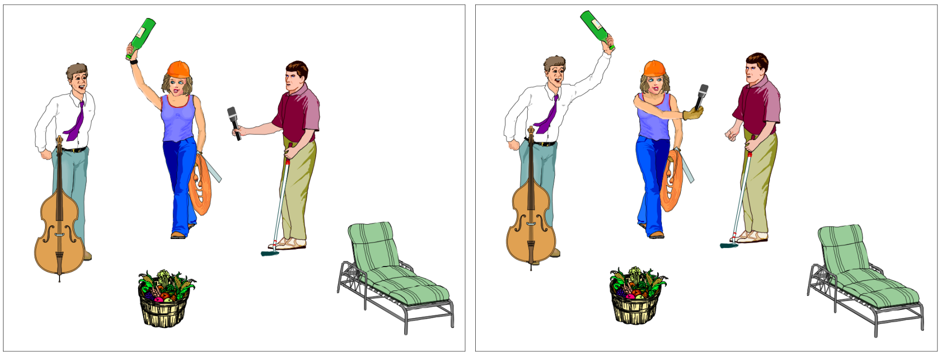


Figure 111: Scene 05

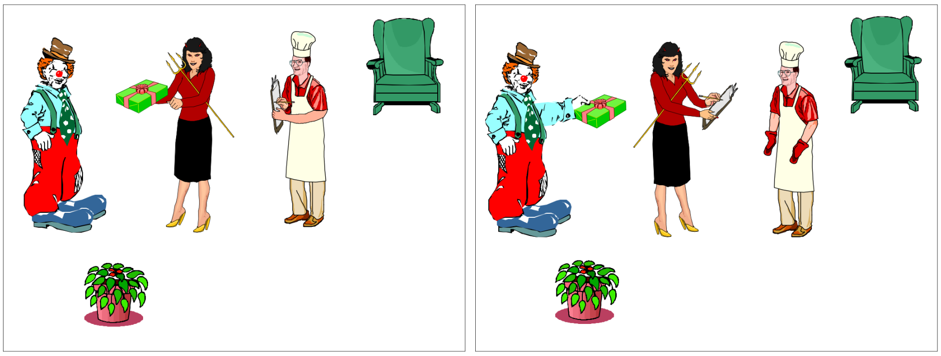


Figure 112: Scene 06

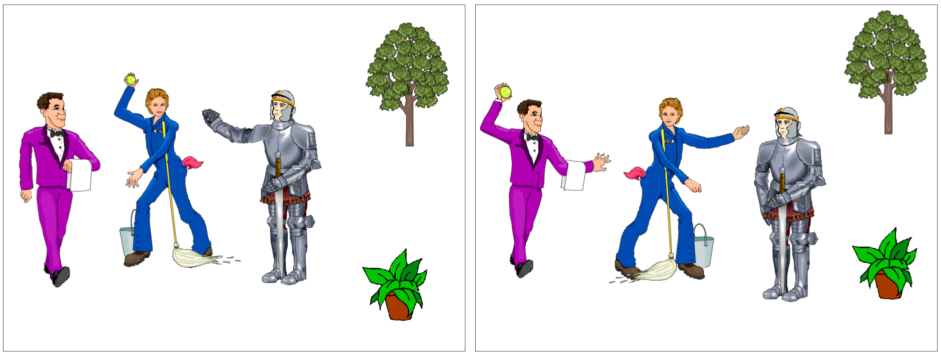


Figure 113: Scene 07

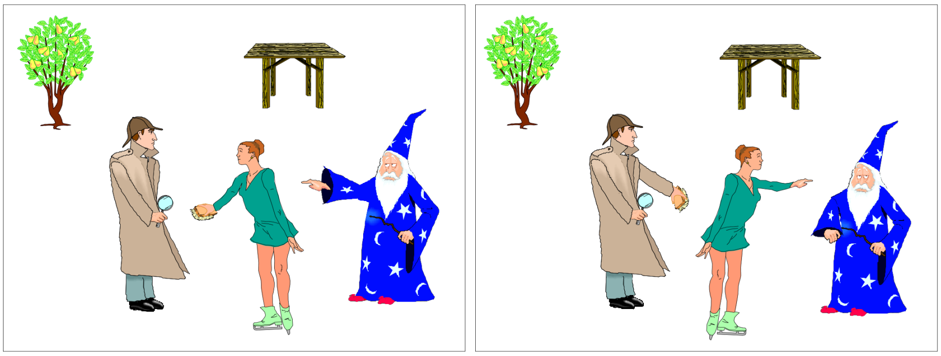


Figure 114: Scene 08

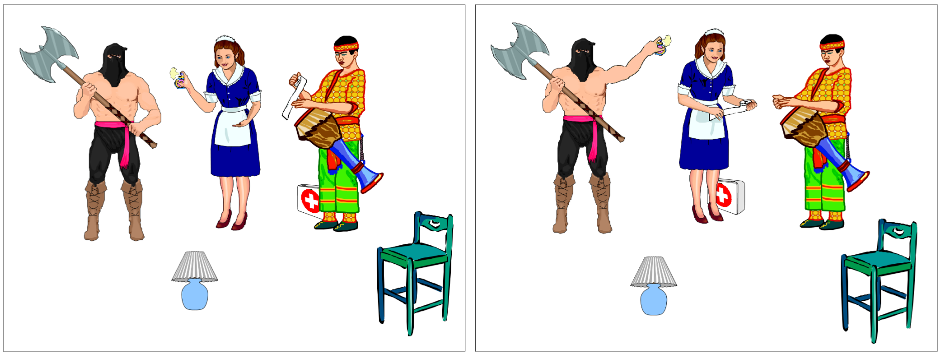


Figure 115: Scene 09

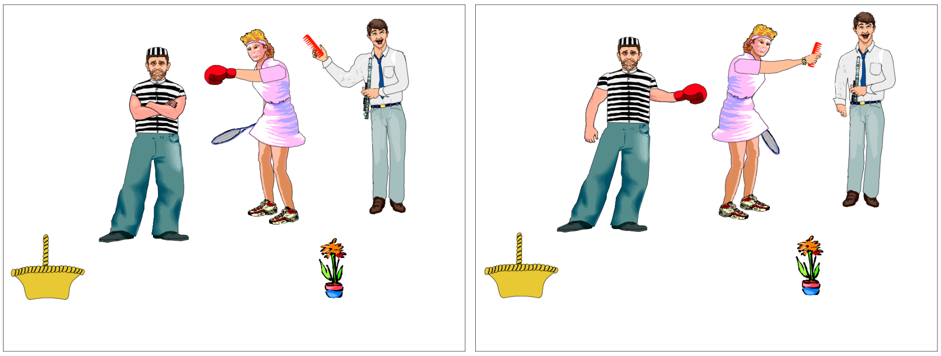


Figure 116: Scene 10

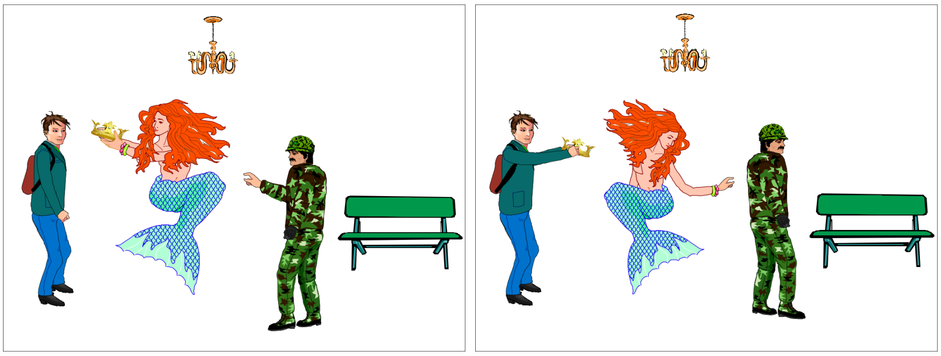


Figure 117: Scene 11

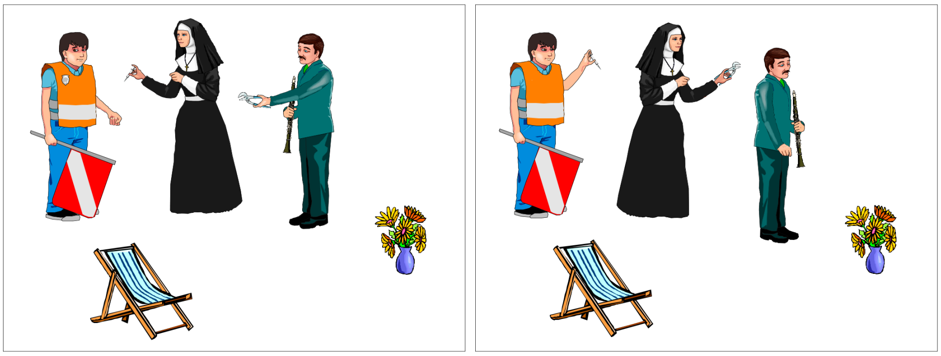


Figure 118: Scene 12

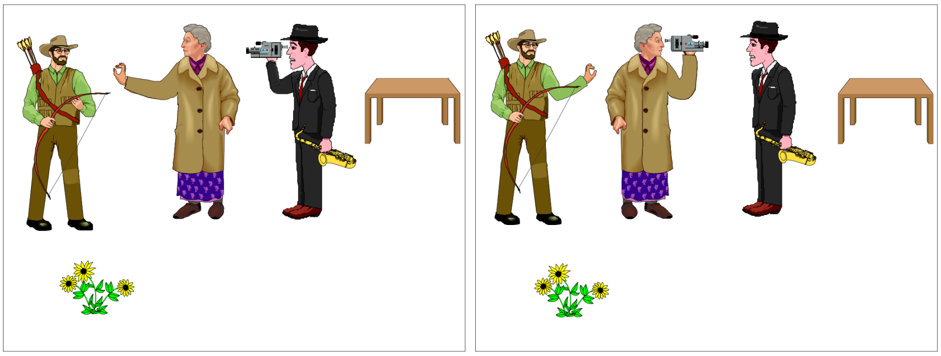


Figure 119: Scene 13

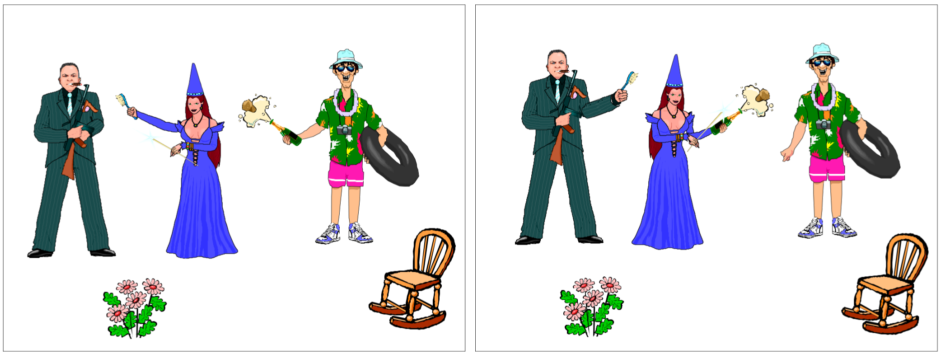


Figure 120: Scene 14

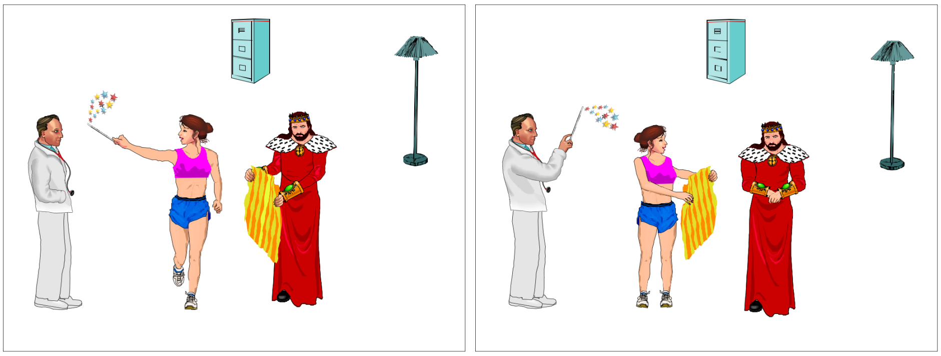


Figure 121: Scene 15

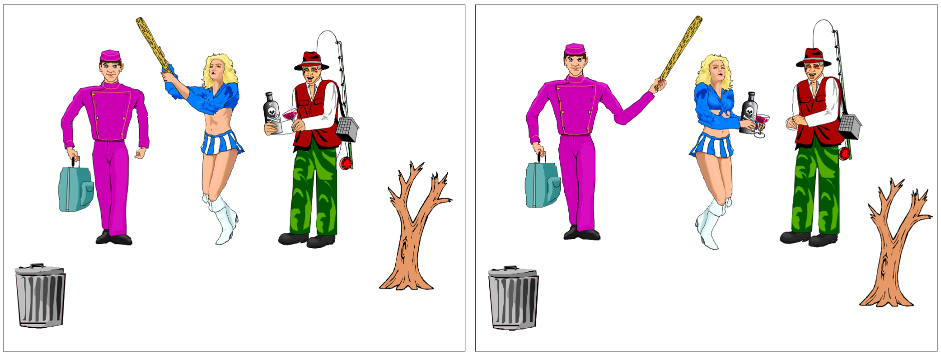


Figure 122: Scene 16

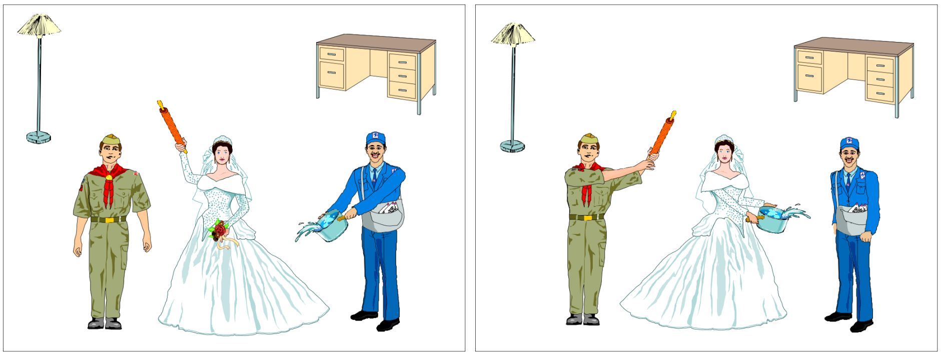


Figure 123: Scene 17

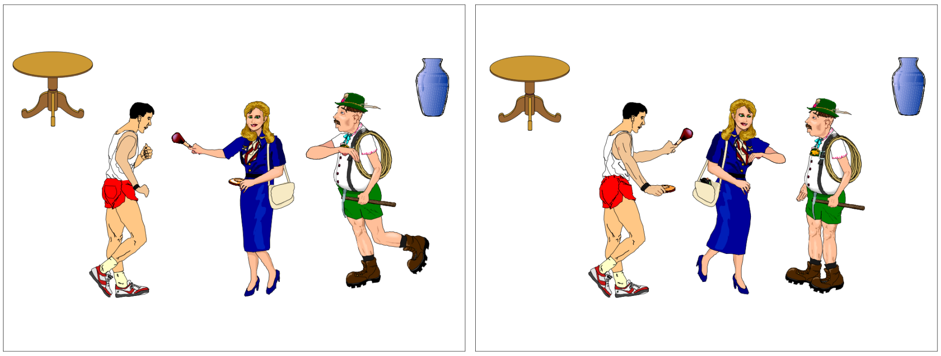


Figure 124: Scene 18

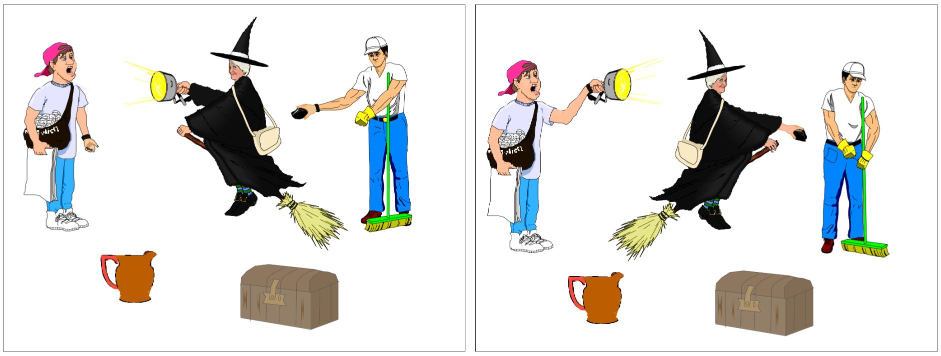


Figure 125: Scene 19

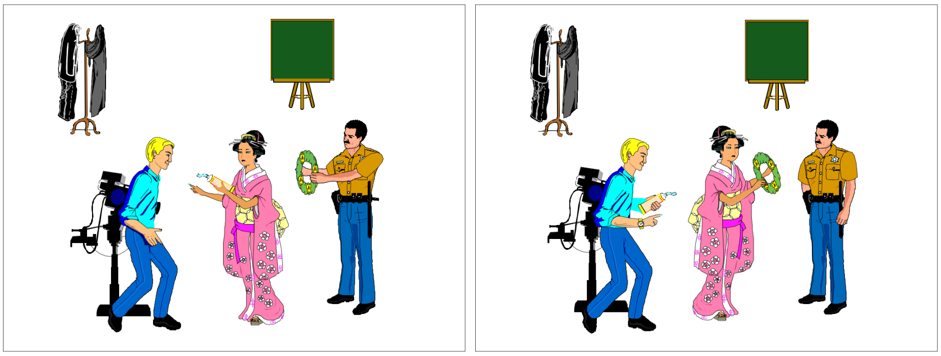


Figure 126: Scene 20

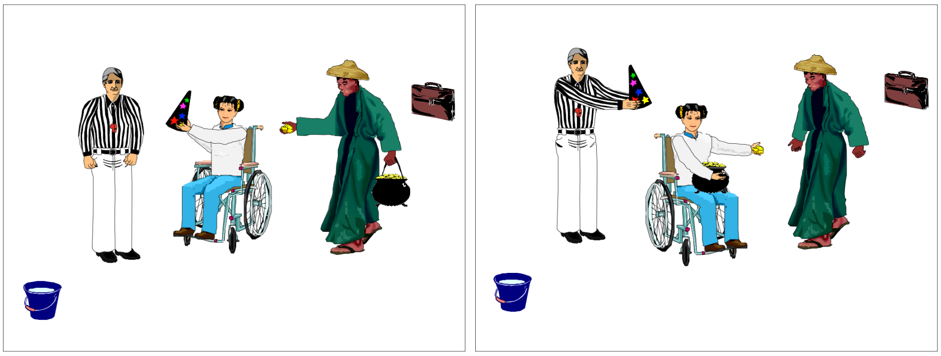


Figure 127: Scene 21

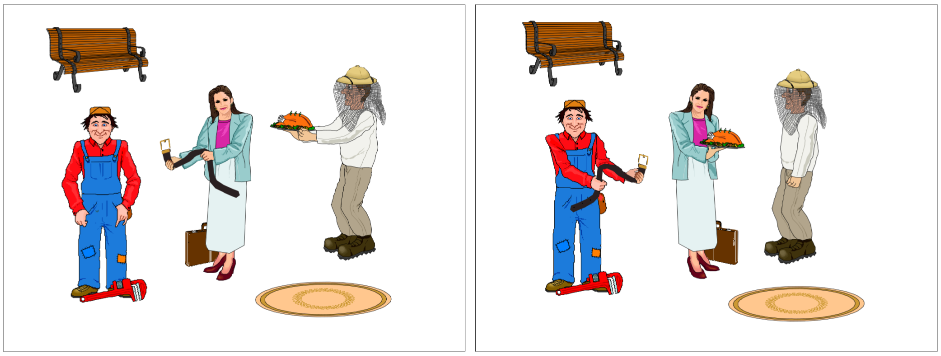


Figure 128: Scene 22

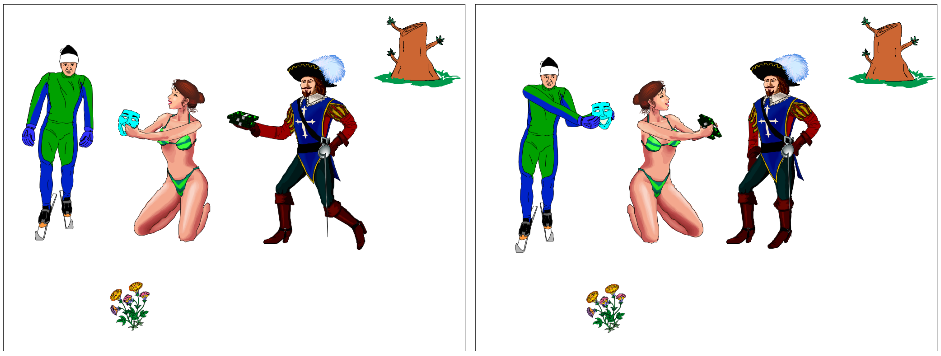


Figure 129: Scene 23

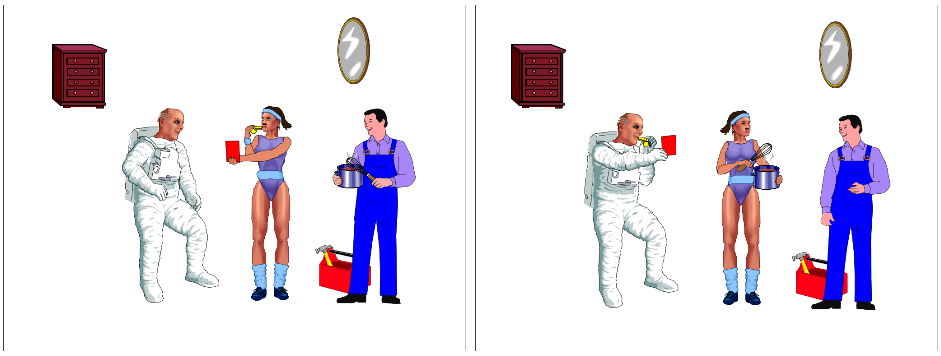


Figure 130: Scene 24

10.4 Context Models

01a:

Pirate.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Waschen_001

Fencer.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Malen_001

Princess_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Malen_001

Princess_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Waschen_001

01b:

Fencer.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Malen_002

Pirate.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Waschen_002

Princess_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Waschen_002

Princess_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Malen_002

02a:

Football.Player.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Besprühen_001

Mechanic.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Erdolchen_001

Amazon_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Erdolchen_001

Amazon_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Besprühen_001

02b:

Mechanic.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Erdolchen_002

Amazon_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Besprühen_002

Amazon_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Erdolchen_002

Football.Player.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Besprühen_002

03a:

Nurse.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Schubsen_001

Nurse.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Fönen_001

Athlete.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Schubsen_001

Priest_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Fönen_001

03b:

Athlete.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Schubsen_002

Priest_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Fönen_002

Nurse.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Schubsen_002

Nurse.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Fönen_002

04a:

Senior.Physician.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Füttern_001
Sailor.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Fesseln_001
Journalist.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Fesseln_001
Journalist.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Füttern_001

04b:

Senior.Physician.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Füttern_002
Journalist.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Füttern_002
Journalist.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Fesseln_002
Sailor.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Fesseln_002

05a:

Golfer.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Interviewen_001
Construction.Worker.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Interviewen_001
Construction.Worker.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Attackieren_001
Cellist.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Attackieren_001

05b:

Cellist.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Attackieren_002

Construction.Worker.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Interviewen_002

Construction.Worker.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Attackieren_002

Golfer.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Interviewen_002

06a:

Chef.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Skizzieren_001

Clown_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Beschenken_001

Devil.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Beschenken_001

Devil.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Skizzieren_001

06b:

Chef.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Skizzieren_002

Clown_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Beschenken_002

Devil.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Skizzieren_002

Devil.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Beschenken_002

07a:

Knight_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Ohrfeigen_001

Cleaner.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Ohrfeigen_001

Cleaner.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bewerfen_001

Waiter_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bewerfen_001

07b:

Cleaner.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bewerfen_002

Cleaner.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Ohrfeigen_002

Waiter_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bewerfen_002

Knight_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Ohrfeigen_002

08a:

Skater.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Schrubben_001

Skater.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Stupsen_001

Wizard.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Stupsen_001

Detective.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Schrubben_001

08b:

Skater.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Schrubben_002

Skater.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Stupsen_002

Detective.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Schrubben_002

Wizard.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Stupsen_002

09a:

Drummer.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bandagieren_001

Maid_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Parfümieren_001

Maid_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bandagieren_001

Hangman_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Parfümieren_001

09b:

Hangman_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Parfümieren_002

Maid_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Parfümieren_002

Maid_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bandagieren_002

Drummer.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bandagieren_002

10a:

Tennis.Player.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Kaemmen_001

Tennis.Player.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Boxen_001

Flutist.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Kaemmen_001

Convict_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Boxen_001

10b:

Tennis.Player.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Boxen_002

Tennis.Player.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Kaemmen_002

Flutist.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Kaemmen_002

Convict_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Boxen_002

11a:

Student.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Krönen_001

Soldier.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Zupfen_001

Mermaid_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Krönen_001

Mermaid_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Zupfen_001

11b:

Mermaid_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Krönen_002

Mermaid_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Zupfen_002

Student.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Krönen_002

Soldier.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Zupfen_002

12a:

Crossing.Guard.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Impfen_001

Clarinetist.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Zwicken_001

Nun_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Impfen_001

Nun_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Zwicken_001

12b:

Nun_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Impfen_002

Nun_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Zwicken_002

Crossing.Guard.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Impfen_002

Clarinetist.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Zwicken_002

13a:

Saxophonist.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Filmen_001

Archer.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Kratzen_001

Grandmother_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Kratzen_001

Grandmother_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Filmen_001

13b:

Archer.m_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Kratzen_002
Grandmother_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Kratzen_002
Grandmother_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Filmen_002
Saxophonist.m_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Filmen_002

14a:

Gangster.m_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Bürsten_001
Tourist.m_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Bespritzen_001
Fairy_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Bespritzen_001
Fairy_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Bürsten_001

14b:

Tourist.m_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Bespritzen_002
Gangster.m_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Bürsten_002
Fairy_002	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Bespritzen_002
Fairy_002	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Bürsten_002

15a:

King_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Frottieren_001
Medical.Doctor.m_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Verhexen_001
Jogger.f_001	$\xrightarrow{\text{is\_AGENT\_for}}$	Etw.Verhexen_001
Jogger.f_001	$\xrightarrow{\text{is\_THEME\_for}}$	Etw.Frottieren_001

15b:

Medical.Doctor.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verhexen_002

King_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Frottieren_002

Jogger.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verhexen_002

Jogger.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Frottieren_002

16a:

Page.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verprügeln_001

Angler.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Vergiften_001

Cheerleader.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Vergiften_001

Cheerleader.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verprügeln_001

16b:

Cheerleader.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verprügeln_002

Cheerleader.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Vergiften_002

Page.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verprügeln_002

Angler.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Vergiften_002

17a:

Boy.Scout_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verhauen_001

Bride_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verbrühen_001

Bride_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verhauen_001

Postman_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verbrühen_001

17b:

Postman_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verbrühen_002

Bride_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verhauen_002

Bride_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verbrühen_002

Boy.Scout_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verhauen_002

18a:

Stewardess.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Pudern_001

Stewardess.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Rempeln_001

Athlete.m_003 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Pudern_001

Hiker.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Rempeln_001

18b:

Stewardess.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Rempeln_002

Stewardess.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Pudern_002

Athlete.m_004 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Pudern_002

Hiker.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Rempeln_002

19a:

Street.Sweeper.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bestehlen_001

Newspaper.Seller.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bestrahlen_001

Witch_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bestrahlen_001

Witch_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bestehlen_001

19b:

Witch_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bestehlen_002

Witch_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bestrahlen_002

Newspaper.Seller.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bestrahlen_002

Street.Sweeper.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bestehlen_002

20a:

Japanese.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Beschmieren_001

Japanese.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bekränzen_001

Camera.Operator.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Beschmieren_001

Police.Man_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bekränzen_001

20b:

Police.Man_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bekränzen_002

Japanese.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Beschmieren_002

Japanese.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bekränzen_002

Camera.Operator.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Beschmieren_002

21a:

Wheelchair.User.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Besolden_001

Wheelchair.User.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Kostümieren_001

Chinese.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Besolden_001

Referee.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Kostümieren_001

21b:

Wheelchair.User.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Kostümieren_002

Wheelchair.User.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Besolden_002

Referee.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Kostümieren_002

Chinese.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Besolden_002

22a:

Beekeeper.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verkoestigen_001

Plumber.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Umgürten_001

Business.Woman_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Umgürten_001

Business.Woman_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verkoestigen_001

22b:

Business.Woman_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verkoestigen_002

Business.Woman_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Umgürten_002

Beekeeper.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verkoestigen_002

Plumber.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Umgürten_002

23a:

Musketeer_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Entlohnen_001

Bathing.Beauty_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Entlohnen_001

Bathing.Beauty_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Maskieren_001

Skier.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Maskieren_001

23b:

Bathing.Beauty_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Maskieren_002

Bathing.Beauty_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Entlohnen_002

Skier.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Maskieren_002

Musketeer_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Entlohnen_002

24a:

Craftsmen.m_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bekochen_001

Aerobic.Trainer.f_001 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verwarnen_001

Aerobic.Trainer.f_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bekochen_001

Astronaut.m_001 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verwarnen_001

24b:

Craftsmen.m_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Bekochen_002

Aerobic.Trainer.f_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Bekochen_002

Aerobic.Trainer.f_002 $\xrightarrow{\text{is\_THEME\_for}}$ Etw.Verwarnen_002

Astronaut.m_002 $\xrightarrow{\text{is\_AGENT\_for}}$ Etw.Verwarnen_002

10.5 The Asserted T-Box Class Hierarchy

Thing
- → Entity.Concept
  - → Abstract
    - → Address
    - → Application
    - → Article
    - → Clarity
    - → Cogeneration
    - → Comment
    - → Commitment
    - → Comparison
    - → Compensation
    - → Competition
    - → Contract
    - → Courage
    - → Damage
    - → Danger
    - → Decision
    - → Demand
    - → Electricity
    - → Every.Day.Life
    - → Excursion
    - → Extrapolation
    - → Favour
    - → Flight
    - → Force
      - → Hit
      - → Kick
    - → Fun
    - → 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
    - → Job
    - → Language
      - → English
      - → French
      - → German
    - → Law
      - → Tax.Law

  - → Leadership
  - → Lexicalisation
  - → Location
  - → Market
  - → Megawatt
  - → Money
    - → Debts
    - → Entrance.Fee
  - → Month
  - → Mood
  - → Murder
  - → Negotiation
  - → Night
  - → Patience
  - → Payment
  - → Peace
  - → Place
  - → Police
  - → Price
  - → Prize
  - → Promise
  - → Proof
  - → Representative
  - → Request
  - → Respect
  - → Restructure
  - → Return
  - → Saving
  - → Sleep
  - → Song
  - → Source
  - → Statement
  - → Storm
  - → Story
  - → Strike
  - → Success
  - → Task
  - → Technology
  - → Time
  - → Warning
  - → Weekday
    - → Friday
    - → Monday
    - → Saturday
    - → Sunday
    - → Thursday
    - → Tuesday
    - → Wednesday
  - → Work

- → Concrete
  - → Human.m.f
    - → Acquaintance.m.f
      - → Acquaintance.f
    - → Artist.m.f
      - → Acrobat.m.f
      - → Choreographer.m.f
      - → Dancer.m.f
      - → Musician.m.f
        - → Violinist.m.f
      - → Sculptor.m.f

  - → Athlete.m.f
  - → Chess.Player.m.f
  - → Confidant
    - → Confidante
  - → Cousin.m.f
  - → Enemy.m.f
    - → Enemy.m
  - → Friend.m.f
  - → Guest
  - → Inhabitant.m.f
  - → Liberator.m.f
    - → Liberator.m
  - → Lifeguard.m.f
    - → Lifeguard.m
  - → Martyr.m.f
    - → Martyr.f
    - → Martyr.m
  - → Member
  - → Movie.Star
  - → Murderer.m.f
  - → Novice.m.f
  - → Offspring
    - → Daughter
    - → Grandson
    - → Son
  - → Opponent.m.f
  - → Patient.m.f
  - → Reader.m.f
  - → Rival.m.f
  - → Ruffian.m.f
  - → Sister
  - → Sister-offspring
    - → Niece
  - → Tourist.m.f
  - → Trainee.m.f
  - → Visitor.m.f
- → Physical.Object
  - → Aeroplane
  - → Airport
  - → Award
  - → Basket
  - → Bier
  - → Bouquet
  - → Cape
  - → City
  - → Clock
  - → Coast
  - → Drug
  - → Envelope
  - → Flat
  - → Gas
  - → Highwater
  - → House
  - → Newspaper
  - → Plan_Concept
  - → Plant
  - → Ring_Concept
  - → Robe

            - → Score
            - → Stick
            - → Street
            - → Sun
            - → Table
            - → Umbrella
            - → Water
            - → Wind
    - → Entity.Feature
        - → Age
            - → Old
            - → Young
        - → Personal.Pronoun
            - → He
            - → It
            - → They
                - → They.f
                - → They.m
                - → They.mixed
- → Helper.Concept
    - → Lexicalised.Concept
    - → Participant
        - → AGENT
        - → RECIPIENT
        - → THEME
        - → THEME_THEME
    - → Situation
- → Meta.Data
    - → Natural.Gender
        - → Female
            - → Abbess
            - → Acquaintance.f
            - → Conductress
            - → Daughter
            - → Maid
            - → Niece
            - → She
            - → Sister
            - → They.f
        - → Male
            - → Construction.Worker.m
            - → Grandson
            - → He
            - → Liberator.m
            - → Lifeguard.m
            - → Physical.Object
                - → Aeroplane
                - → Airport
                - → Award
                - → Basket
                - → Bier
                - → Bouquet
                - → Cape
                - → City
                - → Clock
                - → Coast
                - → Drug
                - → Envelope
                - → Flat
                - → Gas
                - → Highwater

```
|   |   |   |   → House
|   |   |   |   → Newspaper
|   |   |   |   → Plan_Concept
|   |   |   |   → Plant
|   |   |   |   → Ring_Concept
|   |   |   |   → Robe
|   |   |   |   → Score
|   |   |   |   → Stick
|   |   |   |   → Street
|   |   |   |   → Sun
|   |   |   |   → Table
|   |   |   |   → Umbrella
|   |   |   |   → Water
|   |   |   |   → Wind
|   |   |   → Son
|   |   |   → They.m
|   |   → Mixed
|   |   |   → They.mixed
|   |   → Neuter
|   |   |   → It
|   |   → Personal.Pronoun
|   |   |   → He
|   |   |   → It
|   |   |   → They
|   |   |   |   → They.f
|   |   |   |   → They.m
|   |   |   |   → They.mixed
|   → Number
|   |   → Personal.Pronoun
|   |   |   → He
|   |   |   → It
|   |   |   → They
|   |   |   |   → They.f
|   |   |   |   → They.m
|   |   |   |   → They.mixed
|   |   → Plural
|   |   |   → They
|   |   |   |   → They.f
|   |   |   |   → They.m
|   |   |   |   → They.mixed
|   |   → Singular
|   |   |   → He
|   |   |   → It
|   |   |   → She
→ Situation.Concept
|   → Binary.Situation
|   |   → Takes.AGENT.RECIPIENT
|   |   |   → Jmd.Trauen
|   |   → Takes.AGENT.THEME
|   |   |   → Etw.Ablehnen
|   |   |   → Etw.Abnehmen
|   |   |   → Etw.Abschalten
|   |   |   → Etw.Abschlagen
|   |   |   → Etw.Abwehren
|   |   |   → Etw.Aendern
|   |   |   → Etw.Anbieten
|   |   |   → Etw.Andrehen
|   |   |   → Etw.Anhaengen
|   |   |   → Etw.Ankuendigen
```

→ Etw.Anrichten
→ Etw.Antreten
→ Etw.Argumentieren
→ Etw.Aufschwatzen
→ Etw.Aufspueren
→ Etw.Aufsuchen
→ Etw.Aushaendigen
→ Etw.Bedienen
→ Etw.Begleichen
→ Etw.Behalten
→ Etw.Bekommen
→ Etw.Belasten
→ Etw.Benoetigen
→ Etw.Benutzen
→ Etw.Beobachten
→ Etw.Beschreiben
→ Etw.Bezahlen
→ Etw.Bitten
→ Etw.Brauchen
→ Etw.Bringen
→ Etw.Daempfen
→ Etw.Draengen
→ Etw.Einschaerfen
→ Etw.Empfehlen
→ Etw.Entreissen
→ Etw.Erlassen
→ Etw.Erstatten
→ Etw.Erwarten
→ Etw.Erwerben
→ Etw.Erwirtschaften
→ Etw.Faelschen
→ Etw.Feiern
→ Etw.Finden
→ Etw.Fordern
→ Etw.Fragen
→ Etw.Geben
→ Etw.Geniessen
→ Etw.Goennen
→ Etw.Greifen
→ Etw.Halten
→ Etw.Herausgreifen
→ Etw.Hoeren
→ Etw.Holen
→ Etw.Kaufen
→ Etw.Kennen
→ Etw.Landen
→ Etw.Leihen
→ Etw.Lieben
→ Etw.Liefern
→ Etw.Machen
→ Etw.Malen
→ Etw.Melden
→ Etw.Missgoennen
→ Etw.Moegen
→ Etw.Nehmen
→ Etw.Nennen
→ Etw.Nutzen
→ Etw.Praesentieren
→ Etw.Reichen
→ Etw.Richten
→ Etw.Rufen
→ Etw.Schaffen

      - → Etw.Schenken
      - → Etw.Schicken
      - → Etw.Schildern
      - → Etw.Schulden
      - → Etw.Sehen
      - → Etw.Sein
      - → Etw.Senden
      - → Etw.Sprechen
      - → Etw.Spueren
      - → Etw.Stornieren
      - → Etw.Suchen
      - → Etw.Tragen
      - → Etw.Treffen
      - → Etw.Treten
      - → Etw.Trinken
      - → Etw.Uebergeben
      - → Etw.Uebermitteln
      - → Etw.Uebertragen
      - → Etw.Verbieten
      - → Etw.Verbringen
      - → Etw.Verdanken
      - → Etw.Verderben
      - → Etw.Verdienen
      - → Etw.Verkaufen
      - → Etw.Verlangen
      - → Etw.Verlieren
      - → Etw.Vermissen
      - → Etw.Vermitteln
      - → Etw.Versagen
      - → Etw.Verschweigen
      - → Etw.Versorgen
      - → Etw.Verstecken
      - → Etw.Vertreten
      - → Etw.Verweigern
      - → Etw.Vorluegen
      - → Etw.Vorsingen
      - → Etw.Vorstellen
      - → Etw.Waehlen
      - → Etw.Wissen
      - → Etw.Wuerdigen
      - → Etw.Zeigen
      - → Etw.Zurueckweisen
      - → Etw.Zutrauen
      - → Etw.lesen
      - → Fuer.Etw.Sorgen
      - → Jmd.Auffordern
      - → Jmd.Beschuldigen
      - → Jmd.Bitten
      - → Jmd.Richten
      - → Jmd.Verdaechtigen
  - → Ternary.Situation
    - → Takes.AGENT.RECIPIENT.THEME
      - → Jmd.Etw.Ablehnen
      - → Jmd.Etw.Abnehmen
      - → Jmd.Etw.Abschlagen
      - → Jmd.Etw.Anbieten
      - → Jmd.Etw.Andrehen
      - → Jmd.Etw.Anhaengen
      - → Jmd.Etw.Ankuendigen
      - → Jmd.Etw.Aufschwatzen
      - → Jmd.Etw.Aushaendigen
      - → Jmd.Etw.Begleichen

→ Jmd.Etw.Behalten
→ Jmd.Etw.Bekommen
→ Jmd.Etw.Benoetigen
→ Jmd.Etw.Benutzen
→ Jmd.Etw.Beschreiben
→ Jmd.Etw.Bezahlen
→ Jmd.Etw.Brauchen
→ Jmd.Etw.Bringen
→ Jmd.Etw.Einschaerfen
→ Jmd.Etw.Empfehlen
→ Jmd.Etw.Entreissen
→ Jmd.Etw.Erlassen
→ Jmd.Etw.Erstatten
→ Jmd.Etw.Faelschen
→ Jmd.Etw.Feiern
→ Jmd.Etw.Finden
→ Jmd.Etw.Geben
→ Jmd.Etw.Geniessen
→ Jmd.Etw.Goennen
→ Jmd.Etw.Greifen
→ Jmd.Etw.Herausgreifen
→ Jmd.Etw.Hoeren
→ Jmd.Etw.Holen
→ Jmd.Etw.Kaufen
→ Jmd.Etw.Kennen
→ Jmd.Etw.Leihen
→ Jmd.Etw.Lesen
→ Jmd.Etw.Lieben
→ Jmd.Etw.Liefern
→ Jmd.Etw.Machen
→ Jmd.Etw.Melden
→ Jmd.Etw.Missgoennen
→ Jmd.Etw.Moegen
→ Jmd.Etw.Nehmen
→ Jmd.Etw.Nennen
→ Jmd.Etw.Nutzen
→ Jmd.Etw.Praesentieren
→ Jmd.Etw.Reichen
→ Jmd.Etw.Rufen
→ Jmd.Etw.Schenken
→ Jmd.Etw.Schicken
→ Jmd.Etw.Schildern
→ Jmd.Etw.Schulden
→ Jmd.Etw.Sehen
→ Jmd.Etw.Sein
→ Jmd.Etw.Senden
→ Jmd.Etw.Stehlen
→ Jmd.Etw.Stornieren
→ Jmd.Etw.Suchen
→ Jmd.Etw.Tragen
→ Jmd.Etw.Uebergeben
→ Jmd.Etw.Uebertragen
→ Jmd.Etw.Verbieten
→ Jmd.Etw.Verdanken
→ Jmd.Etw.Verderben
→ Jmd.Etw.Verdienen
→ Jmd.Etw.Verkaufen
→ Jmd.Etw.Verlieren
→ Jmd.Etw.Vermissen
→ Jmd.Etw.Vermitteln
→ Jmd.Etw.Versagen
→ Jmd.Etw.Verschweigen

```
|         |         |         → Jmd.Etw.Verstecken
|         |         |         → Jmd.Etw.Vertreten
|         |         |         → Jmd.Etw.Verweigern
|         |         |         → Jmd.Etw.Vorluegen
|         |         |         → Jmd.Etw.Vorsingen
|         |         |         → Jmd.Etw.Vorstellen
|         |         |         → Jmd.Etw.Waehlen
|         |         |         → Jmd.Etw.Wuerdigen
|         |         |         → Jmd.Etw.Zeigen
|         |         |         → Jmd.Etw.Zutrauen
|         |
|         |         → Takes.AGENT.THEME.THEME
|         |         |         → Jmd.Etw.Fragen
|
|         → Unary.Situation
```

Index

Index