A Model for the Cross-Modal Influence of Visual Context upon Language Processing

Patrick McCrae
CINACS Graduate Research Group
Department of Informatics
Hamburg University
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
patrick.mccrae@informatik.uni-hamburg.de

Abstract

In this paper, we present a novel, cognitively motivated framework for modelling the cross-modal influence of visual scene context upon language processing. We illustrate how semantic relations in a knowledge representation of visual scene context can effect syntactic attachment modulations in a weighted-constraint dependency parser. In line with a central tenet of conceptual semantics, visual scene context and linguistic processing are hypothesised to interact via an intermediate, cross-modally integrated level of semantic representation. Cross-modal interaction in our model is restricted by conceptual compatibility between the concepts activated linguistically and contextually.

We apply our framework to syntactically ambiguous sentences of German and parse them in the presence of biasing visual scene contexts. The observed modulations in syntactic attachment support our two modelling hypotheses: 1) The influence of visual context upon syntactic processing is mediated by semantics. 2) The compatibility of concepts from different modalities is a suitable criterion to restrict the scope of cross-modal interaction.

Keywords

Cross-Modal Interaction, Parsing, Syntax-Semantics Interface, Context Integration, Information Fusion.

1 Introduction

The vision-language interface has become considerably more accessible to scientific enquiry with the advent of eye tracking technology. [3] showed a semantic interaction between vision and language for single-word processing with co-present visual stimuli as early as 1974. About two decades later, [13] investigated subjects' eye movements for syntactically ambiguous sentences with a particular focus on the aspect of incrementality in linguistic processing. Yet another decade later, [1] investigated whether the eye movement patterns observed as a result of the interaction between vision and language were contingent upon the visual stimulus being co-present with the linguistic stimulus.
In this paper we present a successful implementation of a framework for the integration of visual context information into the process of syntactic parsing. Starting from the review of central empirical investigations of the vision-language interface, we begin with the identification of elementary requirements for the design of a cognitively motivated framework for the cross-modal integration between vision and language. Adopting a weighted-constraint model of language processing, we outline how in our framework the interpretation of visual context constitutes an additional constraint on the cross-modally integrated semantic representation which is built up based on linguistic and contextual input.
In the following section, we provide a brief overview over selected key findings from milestone experiments at the vision-language interface to motivate the requirements for our framework. In Section 3, we outline how the components of our framework interact with each other and which procedures we employ to achieve cross-modal interaction. In Section 4, we provide experimental results from the integration of visual context into parsing for a particular class of syntactically ambiguous sentences. In Section 5, we summarise our central points and draw conclusions.

2 Milestone Investigations into the Vision-Language Interface

Cooper demonstrated that spoken word semantics influenced subjects' fixation patterns on co-present visual stimuli [3]. More specifically, Cooper was able to show that, from a selection of nine co-present visual stimuli, subjects preferably fixated those that were either direct depictions of referents denoted by the words heard or depictions of items semantically related to the words' referents. Cooper interpreted these eye movement patterns as a reflection of the on-line activation of word semantics from speech.[1]
In another milestone investigation into the vision-language interface, [13] recorded subjects' eye movements when presented with a visual scene depiction and syntactically ambiguous sentences. With their focus on eye movements in incremental sentence processing, Tanenhaus et al. concluded that "*people seek to establish reference (...) during the earliest moments of linguistic processing*". The eye movement patterns observed support their hypothesis

[1] Moreover, Cooper had the foresight that this novel methodology constituted an experimental paradigm whose "*linguistic sensitivity (...) together with its associated small latencies suggests its use as a practical new research tool for the real-time investigation of perceptual and cognitive processes*". His methodology subsequently became known as the *visual-world paradigm*.

International Conference RANLP 2009 - Borovets, Bulgaria, pages 230–235

that referentially relevant non-linguistic information *immediately* affects linguistic processing. Tanenhaus et al. further showed that eye movements and linguistic processing are tightly time-locked, which they interpret as an indication for a close and continual interaction between visual and linguistic processing.[2]
While both [3] and [13] observed anticipatory eye movements, neither of them investigated the cognitive mechanisms that drive those eye movements or ventured a hypothesis on the structure of the mental representations feeding those mechanisms. In our view, the control of the anticipatory eye movements must originate from a suitably detailed mental representation of the visual field. This mental representation must be accessible at the point in time at which the corresponding linguistic stimulus occurs. [1] examined the question of underlying mental representation by employing the *blank-screen paradigm*, a variation of the visual-world paradigm in which the visual stimulus is removed shortly before the onset of the linguistic stimulus. Given sufficiently small inter-stimulus intervals, eye movements are very similar to those obtained in the visual-world condition can be observed, even in the absence of a visual stimulus at the time of interaction. [1] concluded that the eye movements are not the result of a direct online interaction between language and the visual scene but rather result from the interaction between language and a mental representation of the visual scene.[3]

Today, there is substantial and significant empirical evidence for an online interaction between visual and linguistic processing. [1] provides solid support for the hypothesis that the cross-modal interaction between vision and language occurs at representational level. This interpretation is also in line with Jackendoff's Conceptual Structure Hypothesis [6]. Based on a wide range of linguistic and cognitive evidence, Jackendoff argues that "*there is a layer of mental representation, Conceptual Structure, at which linguistic, sensory, and motor information are mutually compatible*". While Jackendoff provides a series of subsequent refinements to this hypothesis and the model underlying it, e.g. [7], the central message remains the same, namely that cross-modal integration occurs in Conceptual Structure, a semantic level of mental representation that brings together concepts, concept instances, semantic relations, linguistic representations and is accessible to reasoning. It was our intention to include this cognitive architecture centring around an integrating level of semantic representation into our model. We hence derive the following high-level modelling requirements from these initial considerations:

R1. Visual and linguistic processing must interact continuously.

R2. The interaction between vision and language must be semantic in nature.

R3. The interaction between vision and language requires the presence of a mental representation of the visual scene rather than the physical presence of the visual scene itself.

R4. Cross-modal interaction must occur at a single representational level that encodes concepts, concept instances and semantic relations such that different modalities – be they sensory or representational in nature – are compatible with each other.

R5. The level of integrated representation must be accessible to reasoning to permit to draw elementary inferences and conclusions.

3 Framework Implementation

3.1 The Syntax-Semantics Interface and Cross-Modal Integration

Our framework implements the architecture for cross-modal integration as proposed by [9]. Core component of the architecture in [9] is WCDG, a weighted constraint dependency parser for German which provides a generic interface to incorporate additional, possibly non-linguistic, information into the parsing process [10, 5]. Constraint-based systems have the succinct advantage that in principle any modelled property of the target structure – be it linguistic or non-linguistic – can be constrained by the mere addition of appropriate further constraints. For cross-modal interaction, we use integration constraints that stipulate to what degree the cross-modally integrated semantic representation comply with the representation of visual context. A predictor component provides the contextual information to the parser.
Discussing the application of their architecture to PP-attachment, [8] proposes to achieve cross-modal integration in the parser by imposing additional context constraints upon dependencies about which the parser itself has no or only insufficient information. The additional penalty scores are provided by a predictor and are calculated based on queries to a representation of visual context through a reasoning engine. Our framework implements this approach.
Since the parser is constraint-based, a predictor influences dependency assignments by providing graded dependency vetoes. These vetoes are evaluated in the integration constraints and may further constrain the set of acceptable solutions. In our model, the predictor assigns graded penalties to semantic dependencies. These prediction scores are based on context information accessible to the predictor but unavailable to the parser.
When adopted into the cross-modally integrated semantic representation, the score on a semantic dependency affects the overall score of the syntax-semantics analysis. In our implementation, context-based prediction scores are calculated for all semantic dependencies after accessing the knowledge representation of visual context. We use the FaCT++ reasoner to do so [4]. The context representation contains instances of ontological concepts linked by thematic relations. Following [6], we assume that this kind of representation results from visual understanding.
During parsing, the parser builds up layers of syntactic and semantic representation that interface with each other. The syntax-semantics interface in our model contains correspondence rules that interlock the syntactic and semantic

[2] This finding is of particular relevance in the context of the discussion to what degree the linguistic processor acts as an encapsulated unit. The degree of interactivity of the linguistic parser clearly has a bearing on the mechanisms by which and the point in time when it can engage in interaction with information provided by other sensory or representational modalities.

[3] While beyond the scope of this paper, it should be mentioned for completeness' sake that there is some scientific debate regarding the contents and degree of detail of this mental representation as well as its actual location in memory.

levels of analysis and require consistency between them.[4] For context integration, the parser's semantic levels of representation are constrained to be consistent with the syntactic representation and the thematic relations asserted in visual context. The integration constraints stipulate just how rigidly the consistence of thematic relations be enforced between modalities. The interaction between the different levels of representation in the parser is shown schematically in Figure 1. The overall solution score comprises all levels of representation and is optimised for a minimisation of constraint violation severity.

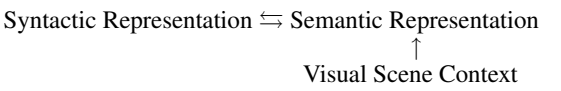

Fig. 1: *Representational interactions during the influence of visual context upon syntactic parsing in our model.*

3.2 Scoring Thematic Relations based on Visual Context

Having outlined the overall interaction of the various components in the framework we now take a closer look at the actual scoring process inside the predictor. The precondition for a thematic relation in visual context to be able to affect a thematic dependency assignment in the semantic representation is that the word in the linguistic modality map onto one or more concept instances in visual context. Only if this mapping is successful can the thematic relations in visual context provide relevant information for the dependency assignment to a given dependant-regent word pair. Consequently, the first step in cross-modal interaction must be the mapping of linguistic entities onto sets of concept instances in visual context. This process is referred to as *cross-modal matching* [2].
With WCDG, cross-modal matching is subject to a technical limitation that, at present, cannot be overcome: the predictor is invoked *prior to* the commencement of the parsing process, i.e. at a stage at which no syntactic information is available yet. As a result, the predictor can only provide dependency scores for word pairs – and not for syntactically more complex units such as phrases or clauses. To map an individual word in the input sentence to a set of concept instances in the representation of visual context, the predictor passes through the following steps (cf. Figure 2):

1. WCDG maps every surface string to a corresponding set of uniquely identified lexical entries. Each lexical entry is characterised by a unique set of lexical features. The surface string 'fragen' *ask*, e.g., can map to the infinitive lexical entry with POS tag `VVFIN` or to the finite verb form with POS tag `VVINF`.[5]

2. The predictor normalises each lexical entry to a form which, in the majority of cases, coincides with the lexical base form in the lexicon. For nouns, the normalisation typically is the nominative singular form, for verbs it is the infinitive.

3. Every normalisation activates a set of concepts in an ontology embedded in our model of Conceptual Structure. Specifically, every word activates those concepts that are lexicalised by the word's normalisation. By permitting the activation of an entire set of concepts through a single word in the linguistic input, our model robustly handles lexical ambiguity and homophony.

4. At the same time, each individual in the representation of visual context instantiates a concept from the ontology.

5. With a reasoner, the predictor determines the set of concepts instantiated in visual context that are compatible with the set of concepts activated by each word in the linguistic input.

6. A thematic relation asserted in visual context becomes relevant in cross-modal integration if the concept instances it connects instantiate concepts that have been matched to words in the input sentence.

In our framwork, dectecting a cross-modally relevant thematic relation in visual context has three effects upon the semantic level of representation in the parser:

a. The detected thematic dependency is integrated for the corresponding dependant-regent word pair. The dependency's score now contributes to the total score of the integrated representation.

b. The assignment of any other thematic dependency between the same dependant-regent word pair is penalised.

c. The assignment of any other thematic dependency originating from the same dependant or pointing to the same regent is penalised. These penalties may subsequently be overwritten if analysis of another thematic relation in visual context provides concrete positive evidence for such an additional thematic relation.

Note that b. results from the uniqueness of a thematic role assignment within any situation frame. Consequence c. reflects the assumption that the interaction between vision and language occurs in a closed-world. We hence assume that the cross-modal interaction occurs on the basis of the information available to the system at the time of interaction. Clearly, this information may be incomplete. In cases in which this information is subsequently revised, – e.g. because an additional scene participant has been detected in the visual scene – a new cross-modal interaction between vision and language based on the revised visual context representation must result.

4 Applying the Framework

To demonstrate the effectiveness of the framework, we have applied it to different classes of syntactic ambiguity phenomena in German: Genitive-Dative ambiguity of feminine nouns, subject-object ambiguities and PP attachment.

[4] Note that – since all constraints are weighted – the parser may also find solutions in which conflicts between syntactic and semantic representations occur. The parser will, however, always favour those solutions in which the overall severity of constraint violations is minimised.

[5] WCDG employs the Stuttgart-Tübingen POS tag set (STTS) [12] which is standard for German.

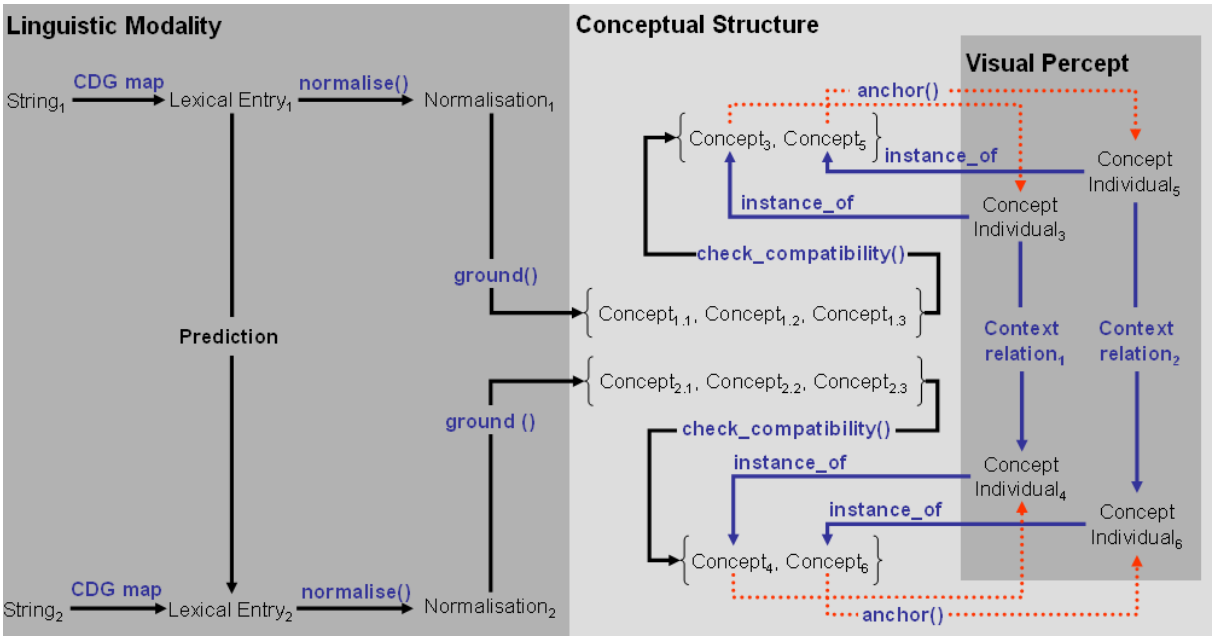

Fig. 2: *Mapping procedure in cross-modal integration.*

The latter class of ambiguity phenomena has already been envisioned as a possible application of the framework by [8].
While a detailed discussion of the application of the framework to these classes of syntactic ambiguities is beyond the scope of this paper, we now discuss how the framework processes one globally ambigous sentence that is representative of the class of German Genitive-Dative ambiguities in feminine nouns as studied by [14]. To improve comparability of the sentences, we have normalised the different introductory main clauses to 'Er weiß, dass ...' *He knows that ...* for all sentences. Consider (1) with the structural ambiguity highlighted.

(1) Er weiß, dass die Ärztin ***der Patientin*** den Leidenden präsentierte.

He knows that ...

a. Binary Situation (Genitive reading)
... the female patient's female doctor presented the male sufferer.

b. Ternary Situation (Dative reading)
... the female doctor presented the male sufferer to the female patient.

The parser's default analysis of (1) in the absence of contextual information is the Dative reading which corresponds to the syntactic structure shown in Figure 4. We can, however, modulate the semantic representation – and via the syntax-semantics interface also the syntactic analysis – by integrating a visual context that corresponds to the Genitive reading in (1) b.
The question arises how detailed such a context representation needs to be in order to be cognitively plausible and have the desired effect upon the integrated semantic representation in the parser. How, for instance, would one be able to differentiate based on the visual modality alone whether the observed scene was a 'präsentieren' *to present* or a 'zeigen' *to show* situation? In [11] Spivey argues for a level of representational detail which balances the economy of information storage against the need for access to cognitively salient information. Spivey concludes that while the mental representation of visual context need not necessarily be complete it must provide anchor points for access to information not stored in the representation via the process of active vision.
To preempt a discussion on the level of representational granularity provided by the visual modality, we have reduced the level of detail provided by the visual modality to the situation arity – i.e. the number of situation participants –, the participants' ontological category as well as their thematic role in the situation. Our model hence does not impose any restriction on how specific the ontological categorisation of participants or the situation verb needs to be. A typical context model is visualised in figure 3.
The concept instances we include need to be general enough to be attainable based on visually perceptible features. In Figure 3 we have identified the central situation verb as an instance of a generic binary situation concept and thus have intentionally underspecified the specific nature of the observed action. This visual context can be interpreted as 'I can see who is doing something to whom – even if I cannot discern exactly what it is that they are doing'.
While our framework permits to define instantiations of specific situation concepts such as 'präsentieren' *present* or 'Ärztin' *female doctor*, we think that it is cognitively questionable whether such detailed and strongly lexicalised information is really provided by the visual modality. Since our model does not rely on the one-to-one mapping of word in the linguistic modality to a concept instance in visual context we can safely model visual context with

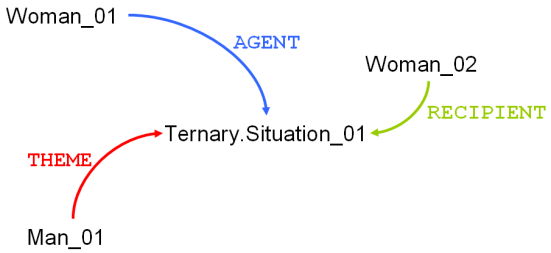

Fig. 3: *Representation of visual context for the ternary situation (Dative reading) of sentence (1).*

instances of less specific concepts. Less specific concepts are superclasses of the more specific concepts in the ontology and therefore exhibit equally or a less restrictive concept compatibility. Our results from larger evaluations show that even such rather general context information is sufficient to impart the correct situation arity to the parser's semantic representation.
The result of integrating the binary visual context into the parsing of (1) is the syntactic dependency structure in Figure 4. Integration of a visual context corresponding to the Genitive reading (1) a. has indeed succeeded in overriding the parser's default ternary situation analysis which previously was obtained in the absence of a visual context.

5 Conclusions

In this paper we have described the implementation of a framework for the cross-modal influence of visual context upon linguistic processing in a weighted constraint dependency parser. Our framework utilises semantic relations between concept instances in combination with structural constraints to achieve cross-modal interaction between visual context and syntactic analysis. In our implementation, semantic information from a parser-external knowledge representation of visual context is aligned with the semantic representation in the parser. Syntactic analysis is modulated via the syntax-semantics interface.
We have outlined that in the parser correspondence rule constraints demand the alignment between the syntactic and semantic levels of analysis and another class of constraints – the integration constraints – demand alignment of the thematic dependencies on the semantic levels with the thematic relations asserted in visual context. We found that concept instantiations related by thematic relations provide a suitable first approximation for the representation of visual context. Our account concludes with the discussion of an application of our framework to a sentence representative of an entire class of syntactic ambiguities in German. The example shows how the integration of visual context in our model permits to modulate syntactic attachment decisions.

Acknowledgements

This work has been generously funded by the German Research Foundation (DFG) as part of the CINACS project "Multimodal Representations in Communication". We would also like to express our sincere thanks to the five anonymous reviewers for their detailed and constructive feedback on an earlier version of this paper.

References

[1] G. T. M. Altmann. Language-mediated eye movements in the absence of a visual world. *Cognition*, 93:B79–B87, 2004.

[2] E. W. Bushnell. *The development of intersensory perception: comparative perspectives*, chapter A Dual-Processing Approach to Cross-Modal Matching: Implications for Development, pages 19–38. Lawrence Erlbaum Associates, 1994.

[3] R. M. Cooper. The control of eye fixation by the meaning of spoken language. a new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6:813–839, 1974.

[4] FaCT-PlusPlus Download Page. `http://code.google.com/p/factplusplus/`, 2009. Link verified: 20 April 2009.

[5] K. A. Foth. *Hybrid Methods of Natural Language Analysis*. PhD thesis, Department of Informatics, Hamburg University, Germany, 2007.

[6] R. S. Jackendoff. *Semantics and Cognition*. Cambridge, MA: MIT Press, 1983.

[7] R. S. Jackendoff. *Patterns in the Mind*. New York: Basic Books, 1994.

[8] P. McCrae. Integrating cross-modal context for PP attachment disambiguation. In *Proceedings of the 3rd International Conference on Natural Computation, Haikou (ICNC 2007)*, volume 3, pages 292–296, Los Alamitos, CA, 2007. IEEE.

[9] P. McCrae and W. Menzel. Towards a system architecture for integrating cross-modal context in syntactic disambiguation. In B. Sharp and M. Zock, editors, *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science, Madeira (NLPCS 2007)*, pages 228–237. INSTICC Press, 2007.

[10] I. Schröder. *Natural Language Parsing with Graded Constraints*. PhD thesis, Department of Informatics, Hamburg University, Germany, 2002.

[11] M. J. Spivey, D. C. Richardson, and S. A. Fitneva. *The Interface of Vision, Language, and Action*, chapter Thinking outside the brain: Spatial indices to linguistic and visual information, pages 161–189. Number 5. New York: Psychology Press, 2004.

[12] STTS Tag Set. `http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts.html`, 2009. Link verified: 24 July 2009.

[13] M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. C. Sedivy. Integration of visual and linguistic information in spoken language comprehension. *SCIENCE*, 268:1632–1634, 1995.

[14] A. van Kampen. *Syntaktische und semantische Verarbeitungsprozesse bei der Analyse strukturell mehrdeutiger Verbfinalsätze im Deutschen: Eine empirische Untersuchung*. PhD thesis, Freie Universität Berlin, 2001.

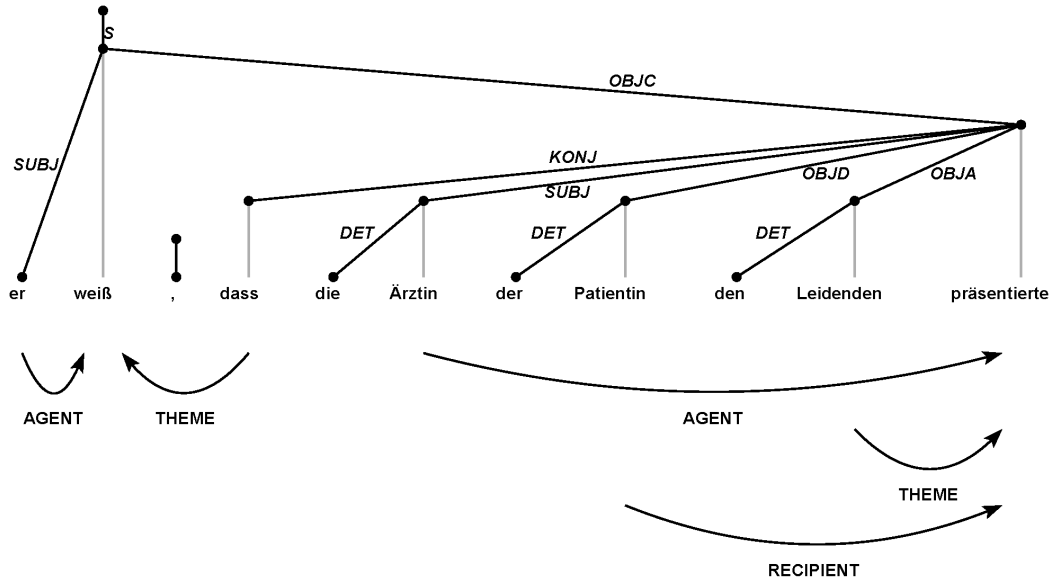

Fig. 4: *The parser's default analysis in the absence of visual context (Dative reading).*

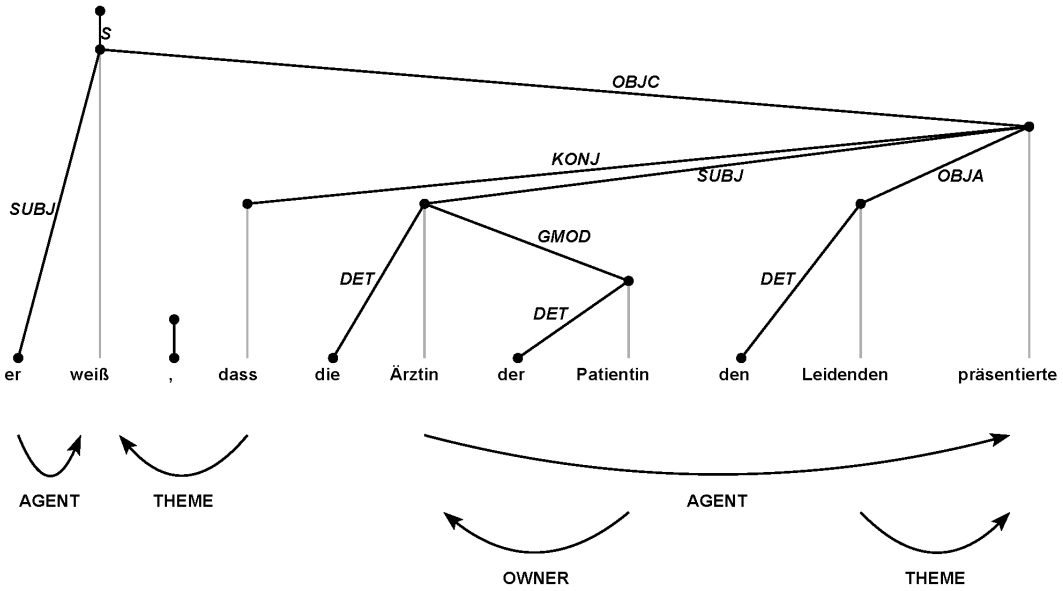

Fig. 5: *The parser's cross-modally integrated analysis with a binary visual context (Genitive reading).*