Towards a Systematic Analysis of Linguistic and Visual Complexity in Disambiguation and Structural Prediction

Özge Alaçam , Tobias Staron and **Wolfgang Menzel**
Department of Informatics
University of Hamburg
alacam, staron, menzel@informatik.uni-hamburg.de

Abstract

Situated language processing in humans involves the interaction of linguistic and visual processing and this cross-modal integration helps resolving ambiguities and predicting what will be revealed next in an unfolding sentence. However, most state-of-the-art parsing approaches rely solely on the language modality. This paper aims to introduce a new multi-modal data-set (containing sentences and respective images and audio files) addressing challenging linguistic and visual complexities, which state-of-the-art parsers should be able to cope with. It also briefly addresses a proof-of-concept study that shows the contribution of employing external visual information during disambiguation.

1 Disambiguation and Structural Predictions

A better understanding of human perceptual and comprehension processes concerning multi-modal environments is one of the crucial factors for realizing dynamic human-computer interaction. A large body of empirical evidence in psycholinguistics suggests that human language processing successfully integrates available information acquired from different modalities in order to resolve linguistic ambiguities (i.e. syntactic, semantic or discourse) and predict what will be revealed next in the unfolding sentence (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Knoeferle, 2005). During spoken communication, online disambiguation and prediction processes allow us to have more accurate and fluent conversations. In contrast, state-of-the-art parsing algorithms are still far away from that accuracy and fluency when it comes to challenging linguistic or visual situations. Therefore, by developing a cross-modal parser to exploit visual knowledge, we expect to enhance syntactic disambiguation, e.g. concerning relative clause attachments and various scope ambiguities.

One of the most frequently investigated syntactic ambiguity cases is the prepositional phrase (PP) attachment ambiguity, where different semantic interpretations are possible depending on assigning different thematic roles (Tanenhaus et al., 1995). A well-known example is the imperative sentence: "put the apple on the towel in the box", where the PP "on the towel" can be interpreted as modifier of an apple (as location of the apple), as marked in 1 below, or as goal location as in 2.

[1] put [the apple on the towel]$_{\text{obj}}$ [in the box]$_{\text{goal}}$

[2] put [the apple]$_{\text{obj}}$ [on the towel in the box]$_{\text{goal}}$

The re-analysis of the interpretation during online language comprehension is termed as garden-path example. In a multi-modal setting where the scene contains an empty towel or an apple on a towel, the visual information constrains the referential choices as well as the possible interpretations, helping the disambiguation process.

Tanenhaus and his colleagues' study (1995) showed that visual information influences incremental thematic role disambiguation by narrowing down the possible interpretations. Further evidence that supports this conclusion was provided by Knoeferle (2005) by addressing relatively more complex scenes containing more agents and relations for both English and German. The results also indicated that this influence occurs independent from the experiment language. Fur-

thermore, Altmann and Kamide (1999)'s study has documented that listeners are able to predict complements of a verb based on its selectional constraints. For example, when people hear the verb 'break', their attention is directed towards only breakable objects in the scene. Some nouns may also produce expectations for certain semantic classes of verbs by activating so-called event schema knowledge (McRae et al., 2001). Beside verbs and nouns, Van Berkum et al. (2005)'s study also showed the effect of syntactic gender cues for Dutch in the anticipation of the upcoming words. Similar to German, pre-nominal adjectives as well as nouns are gender-marked in Dutch and the gender of the adjective has to agree with the gender of the noun. Their results showed that the human language processing system uses the gender cue, when it becomes available, to predict the target object if its gender is different than the gender of the other objects in the environment. They interpreted this as evidence for the incremental nature of the human language system, which can predict the upcoming words and immediately begin incremental parsing operations. In a more recent work, Coco and Keller (2015) investigated the language - vision interaction and how it influences the interpretation of syntactically ambiguous sentences in a simple but real-world setting. Their study provided further evidence that visual and linguistic information influences the interpretation of a sentence at different points during online processing. The aforementioned empirical studies provided insights regarding psycholinguistically plausible parsing. However, those studies were limited to simple (written) linguistic or visual stimuli where object-action relations could be predicted relatively easily.

Based on the prior research, our project focuses also on studying underlying mechanisms of human cross-modal language processing of incrementally revealed utterances with accompanying visual scenes, with the aim of using the empirically gained insights to develop a psycholinguistically plausible cross-modal and incremental syntactic parser which can be implemented e.g. on a service robot. A parser that processes only linguistic information is expected to be able to successfully handle syntactically unambiguous cases by using linguistic constraints or statistical methods. However, without external information from visual modality, neither humans nor parsers can resolve references in syntactically ambiguous cases. They may have preferences but the accuracy of the preferences are bounded by chance. On the other hand, humans naturally use external information from other modalities for disambiguation when available. Incorporating this feature, cross-modal parsers may also resolve those ambiguities and reach correct interpretations of the visually depicted events. Therefore, a better understanding of human language processing concerning cross-model environments is one of the crucial factors in the realization of dynamic human-computer interaction. Furthermore, comparing the performance of the computational model with human performance (e. g. whether ambiguities were resolved correctly, at which point of a spoken utterance a correct resolution was achieved, how many changes were made before reaching the correct thematic role assignment) also provides valuable information about the plausibility and the effectiveness of the proposed parsing architecture. Constructing a data-set that contains challenging linguistic and visual cases and complex multi-modal settings, where state-of-the-art parsers often fail, are fundamental towards achieving this ultimate goal. In this paper, we aim to introduce a multi-modal data-set consisting of garden-path (fully/temporally syntactically ambiguous) sentences.

This paper is structured as follows. In section 2, a data-set of ambiguous German sentences and their multi-modal representations are presented. A brief description of our cross-modal parser is presented in Section 3. Section 3 also addresses a test run conducted on fully ambiguous sentence structures. Section 4 summarizes the results of this work and draws conclusions

2 Linguistic and Visual Complexities

Recently, a corpus of language and vision ambiguities (LAVA) in English has been released (Berzak et al., 2016). LAVA corpus contains 237 sentences with linguistic ambiguities that can only be disambiguated using external visual information provided as short videos or static visual images with real world complexity. It addresses a wide range of syntactic ambiguities including prepositional phrase or verb phrase attachments and ambiguities in the interpretation of conjunctions. However, this corpus does not take linguistically challenging cases like relative clause attachments

or scope ambiguities, which may also give valuable insights understanding the underlying mechanisms of cross-modal interactions, into account. To our knowledge, the reference resolution concerning these linguistic cases and the effect of linguistic complexity in visually disambiguated situations have been scarcely investigated. Our multimodal data-set consists of challenging linguistic cases in German (itemized below), which becomes fully unambiguous in the presence of visual stimuli. Our main question from the psycholinguistic point of view is whether the presence of linguistic ambiguity and the linguistic complexity affect the processing of multi-modal stimuli. On the other hand, from the computational perspective, we focus on whether and to what extent visual information is useful for the disambiguation and structural prediction processes in order to develop more fluent and accurate computational parsing.

German has three grammatical genders, namely each noun is either feminine*(f)*, masculine*(m)*, or neuter*(n)*. In a sentence that contains a relative clause attachment, the gender of the relative pronoun has to be the same as the gender of its antecedent. Sentence [3] illustrates an example, which contains a relative clause licensing the NP.

[3] Sie schmückt das Fenster*(n)*, das*(n)* er säubert. *(She decorates the window that he cleans.)*

In Sentence [4], the NP is modified by an additional NP, i.e. a genitive object. In this case, since the gender of the relative pronoun matches only the first NP, it is clear that *the window* is being cleaned, not *the car*. However, due to ambiguous German case-marking, if the genders of the nouns of both NPs are the same, as in sentence [5], both far and near attachments are possible. Furthermore, the verb is semantically congruent with both NP and PP as well. Correct reference resolution can not be achieved based on linguistic information alone. On the other hand, having access to visual information eliminates other interpretations and it favors only one assuming there will be no ambiguity in the visual modality (see Figure 1 and 2).

[4] Sie schmückt das Fenster*(n)* des Wagens*(m)*, das*(n)* er säubert. *(She decorates the window of the car that he cleans.)*

[5] Sie schmückt das Fenster*(n)* des Zimmers*(n)*, das*(n)* er säubert. *(She decorates the window of the room that he cleans.)*

Our data-set is currently consisting of 191 sentences[1] and addresses 8 linguistically challenging cases concerning relative clause attachments, agent/patient agreement, verb/subject agreement, and scope ambiguities for conjunctions and negations. The sentence sets for each structure are generated by using part-of-speech templates given in Table 1. Parsers often have problems with correct reference resolution for such linguistic expressions because they usually attach the relative clause to a nearest option with respect to statistical distributions in their training data or explicitly stated rules.

Knoeferle's (2005) sentence set was used as baseline since the co-occurrence frequencies between the action and the Agent in the sentence, as well as between the action and the Patient, were controlled to single out the effect of semantic associations or preferences during parsing operations. For a syntactic parser, this may seem irrelevant, however in order to develop a comparable experimental setup for human comprehension, this parameter needs to be taken into account.

Fully Ambiguous Sentence Structures

[1] **RPA[2] - a Genitive NP**
Sie schmückt das Fenster*(n)* des Zimmers*(n)*, das er säubere.
She decorates the window of the room that he cleans.
Int.1[3]: He cleans the room *(near-attachment)*.
Int.2: He cleans the window *(far-attachment)*.

[2] **RPA - Scope Ambiguities**
Ich sehe Äpfel*(pl)* und Bananen*(pl)*, die*(pl)* auf dem Tisch liegen.
I see apples and bananas that lie on the table.
Int.1: Both apples and bananas are on the table.
Int.2: Only bananas are on the table.

[3] **RPA - a Dative PP**
Da befindet sich ein Becher*(m)* auf einem Tisch*(m)*, den*(m)* sie beschädigt.
It is the mug on the table that she damages.
Int.1: She damages the table *(near-attachment)*.
Int.2: She damages the mug *(far-attachment)*.

[1] The short-term goal is to increase the sample size to 450 sentences.
[2] Relative Pronoun Agreement
[3] Int.=Interpretation

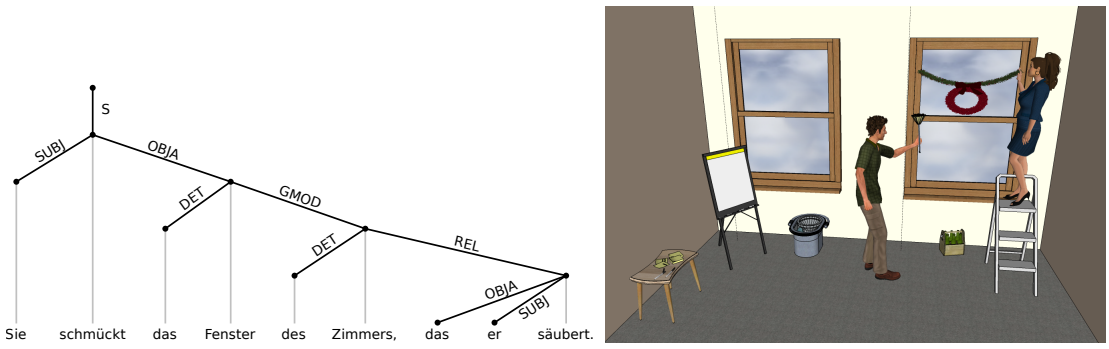

Figure 1: First interpretation of syntactically ambiguous sentence [5]: near attachment of relative clause - syntactic gold standard annotation and visual scene.

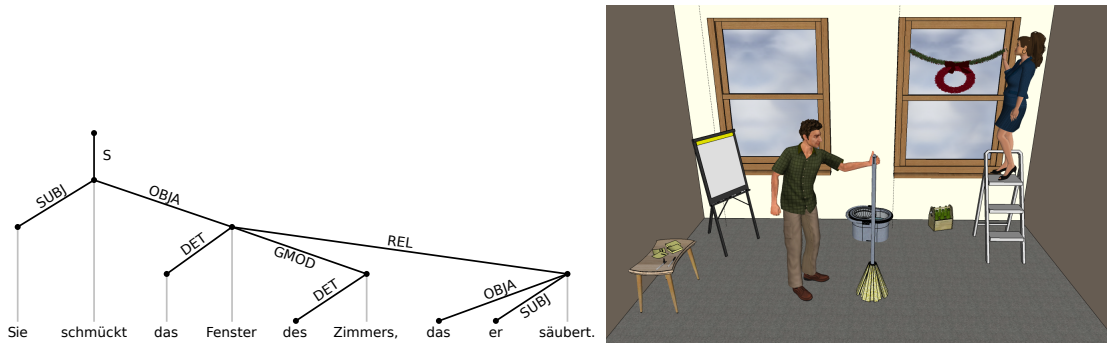

Figure 2: Second interpretation of syntactically ambiguous sentence [5]: far attachment of relative clause - syntactic gold standard annotation and visual scene.

[4] **RPA with an agent/patient ambiguity**
Da ist eine Japanerin*(f)*, die*(f, $RP_{nom/acc}$)* die Putzfrau*(f)* soeben attackiert.
There is a Japanese, who(m) the cleaning lady attacks.
Int.1: The cleaning lady attacks the Japanese woman.
Int.2: The Japanese woman attacks the cleaning lady.

[5] **Negative Scope Ambiguities**
Die Sängerin kauft die Jacke nicht, weil sie rot ist.
The singer does not wear the coat because it is red.
Int.1: The singer does not buy the coat because of its color.
Int.2: The singer actually buys the coat but not because it is red.

All the sentence structures for the fully ambiguous set (except negative scope sentences) presented above can be also transformed to temporally ambiguous sentence structures by changing the noun in either of the NPs (or PPs) with another noun that has an article in different gender. Below, three additional types of temporal ambiguities, which are convenient for the investigation of how/when structural prediction mechanisms are employed during parsing process are presented.

Temporally Ambiguous Sentence Structures

[6] **Agent-Patient Agreement** (following the dataset designed by Knoeferle (2005))

- Die Arbeiterin kostümiert mal eben den jungen Mann.
  The worker(f) just dresses up the joung man(m).
- Die Arbeiterin verköstigt mal eben der Astronaut.
  The worker(f) is just fed[4] *by the astronaut(m).*

[7] **Verb-Subject Agreement**

- Die Sänger waschen den Arzt.
  The singers wash the doctor(m).

[4]The original German sentence is in active voice in OVS word order.

- Die Sänger wäscht der Offizier.
  The singers are painted [4] *by the officer(m).*

[8] **Conjunction Scope Ambiguities**

- Die Sängerin bemalt den Offizier und die Ärztin.
  The singer(f) paints the officer(m) and the doctor(f).
- Die Sängerin bemalt den Offizier und die Ärztin wäscht den Radfahrer.
  The singer(f) paints the officer(m) and the doctor(f) washes the cyclist(m).
- Die Sängerin bemalt den Offizier und die Ärztin besprüht der Radfahrer.
  The singer(f) paints the officer(m) and the doctor(f) is sprayed[4] *by the cyclist(m).*

2.1 Image Construction and Visual Complexity

Besides the effect of linguistic complexity, the data-set was designed to be used in the investigation of the following research questions: how, when and at which degree does visual complexity affect sentence comprehension and are visual cues in such a complex linguistic case still strong enough to enhance correct interpretation.

The 2D visual scenes were created with the SketchUp Make Software[5] and all 3D objects were exported from the original SketchUp 3D Warehouse. The images were set to 1250 x 840 resolution. Moreover, target objects and agents are located in different parts of the visual scene for each stimulus. It should be reminded that for the computational model, we do not need visual depictions, their semantic representations are sufficient, however the visual depictions are crucial to conduct comparable experimental studies with human subjects. Furthermore, an automatic extraction of semantic roles from the images is another task that we are aiming for. That is the reason why not just semantic representations but the images themselves are integral part of our data-set.

The following figures illustrate how complexity is systematically controlled on one of the cases in the data-set, namely Agent-Patient agreement. In the initial/original case, each scenario contains three characters (one Patient, one Agent and one ambiguous Agent/Patient character) and two possible actions. On the other hand each sentence addresses only one action and two characters, see sentences in [6]. For each scenario, four different complexity levels were designed. In the first condition, a visual scene contains three characters in an environment, where there is no additional background object, see Figure 3. This set-up resemblances Knoeferle's (2005) images and provides a baseline to compare our results with previous research. The images in the second condition also contain three characters, but in an environment with noninteracting distractor objects, see Figure-4. In the last two conditions, a fourth character in an Agent role, who acts on the ambiguous character is added to the scene. While the images in the third condition do not have additional objects , the images in the fourth condition are in a cluttered environment as in the condition 2 (see Figure 5 and Figure 6). It should be noted that background objects and the fourth character do not have any semantic association with the actions mentioned in the sentences. Besides, visual complexities can be further diversified, e.g. by adding another patient character to the scene or by adding semantically congruent distractor objects.

Figure 3: 3 agents in an environment with no background objects; *a Patient* (a young boy on the left), *an Agent* (an astronaut on the right) and *an ambiguous Agent/Patient character* (a female worker in the middle).

2.2 Semantic Annotations

The objects, characters and actions in the images were annotated manually with respect to their semantic roles, similar to McCrae's approach (McCrae, 2010), see also Mayberry et al. (2006). Semantic roles are used to establish a relation between semantic and syntactic levels as an important part of modeling the cross-modal interaction. Semantic roles are linguistic abstractions to distinguish and classify the different functions of

[5] http://http://www.sketchup.com/ - retrieved on 03.08.2016

Ambiguity Types	Template	# of unique items	# of sample
1– RPA with a Genitive NP	$PRO1_{nom}$ VP1 $NP1_{acc}$ $NP2_{gen}$, WDT^*_{acc} $PRO2_{nom}$ VP2	Pro(2), $NP_{acc/gen}$(48), VP(48)	24
2– RPA Scope Ambiguities	PRO_{nom} VP1 $NP1_{nom,pl.}$ $NP1_{nom,pl.}$, $WDT_{acc,pl.}$ VP2 PP1	Pro(3), VP(36), $NP_{acc/dat}$(72)	24
3– RPA with a Dative-PP	$NP_{it-cleft}$ VP1 $NP1_{nom}$ $NP2_{dat}$, WDT_{dat} $PRO_{3rd-sing.}$ ADV VP2	NPs(44), VP(23), ADV(24)	20
4– RPA Ambiguous Gender Case Marking	EX V_{aux} $NP1_{nom}$ WDT_{nom} $NP2_{acc}$ ADV VP1 EX V_{aux} $NP1_{nom}$, WDT_{acc} $NP2_{acc}$ ADV VP1	NPs(30), VP(20), ADV(12)	24
5– Negative Scope Ambiguities	$NP1_{nom}$ VP1 $NP2_{acc}$ NEG, Conj. PROnom ADJ VP2	NPs(6), VP(6), ADJ(12), ADV(6)	12
6– Agent–Patient Agreement (all in 3rd P. Sing.)	$NP1_{nom}$ VP $NP2_{acc}$ $N1_{acc}$ V $N2_{nom}$	NPs(37), VP(48), ADV(6)	48
7– Verb–Subject Agreement	$NP1_{nom\text{-}3rd\ Pl.}$ VP $_{3rd\ Pl.}$ $NP2_{acc\text{-}3rd\ Sing.}$ $NP1_{acc\text{-}3rd\ Pl.}$ V $_{3rd\ Sing.}$ $NP2_{nom\text{-}3rd\ Sing.}$	NPs(3), VP(6), ADV(6)	12
8– Conjunction Scope Ambiguities (all in 3rd P. Sing.)	$NP1_{nom}$ VP1 $NP2_{acc}$ Conj. $NP3_{acc}$ $N1_{nom}$ VP1 $NP2_{acc}$ Conj. $NP3_{nom}$ VP2 $NP4_{acc}$ $NP1_{nom}$ VP1 $NP2_{acc}$ Conj. $NP3_{acc}$ VP2 $NP4_{nom}$	NPs(32), VP(27), ADV(6)	27
TOTAL			191

Table 1: POS templates, the number of sentences for each ambiguity case, and the number of unique items in each POS category (*Relative Pronoun)

Figure 4: 3 agents in an environment with background objects.

Figure 5: 4 agents in an environment with no background objects.

Figure 6: 4 agents in an environment with background objects.

the action in an utterance, in other words they are a useful tool to specify *"who did what to whom"*. The most common set of semantic roles includes Agent, Theme, Patient, Instrument, Location, Goal and Path. Figure 7 shows one exemplary semantic annotation for the visual scene displayed in Figure 1. There *"Sie"* is the Agent, who performs the decorating action, *"das Fenster"* is the Patient, the entity undergoing a change of state, caused by the action.

To wrap-up, the current version of our multimodal data-set in German that we constructed with the aim of studying disambiguation and structural prediction from both psycholinguistics and computational linguistics perspectives contains fol-

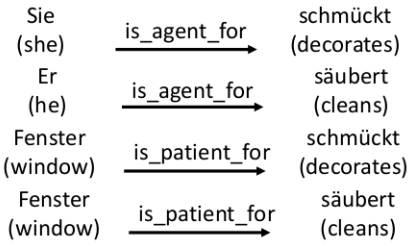

Figure 7: One exemplary semantic annotation for the visual scene shown in Figure 1.

lowing items for each scenario in the data-set [6]

- a linguistic form and a sentence in German with its English translation
- gold standard annotations
- possible interpretations
- a target interpretation
- a visual depiction of the target interpretation in four different visual complexities
- a semantic representation of the visual depiction of the target interpretation
- an audio file and a data file with marked onset/offsets (in msec.) of each linguistic entities in the sentence

3 Cross-modal Parsing

As suggested by the literature mentioned in Section 1, cross-modal integration facilitates to resolve ambiguities and predict what will be revealed next in an unfolding sentence. However, most state-of-the-art parsing approaches rely solely on the language modality. McCrae (2009) proposed a system for the integration of contextual knowledge into a rule-based syntactic and semantic parser to resolve ambiguities in German, e.g. Genitive-Dative ambiguity of feminine nouns or PP attachment ambiguities. Baumgärtner et al. (2012) extended that system by adding incremental processing capabilities leading to the only cross-modal and incremental syntactic parser so far. In their study of visually guided natural language processing, Baumgärtner et al. (2012) propose a computational model that successfully integrates visual context to improve the processing of sentences of German, and semantic information derived from language input that is used to guide the parser to find the correct referent in the description of visual context.

However, in contrast to those rule-based parsers (McCrae, 2009; Baumgärtner et al., 2012), we employ statistical parsing with the aim to achieve state-of-the-art results and developing a language-independent parser. To realize cross-modality, we interface the data-driven parser (RBGParser, Zhang et al. (2014)), which is utilized to search for the most plausible disambiguation of a given sentence among all possible dependency trees, with a rule-based component (jwcdg, Beuck et al. (2011)), which evaluates possible analyses produced by RBG with respect to the visual knowledge. This contextual information guides the parsing process and narrows down the hypotheses towards the most plausible representation for a given sentence.

Another approach that could have been used is to train a parser on combined linguistic and visual features (Salama and Menzel, 2016). However, due to lack of available data to train the parser with, RBG is not dedicated to process the contextual information in our approach. Instead, we embed a constraint-based component that is able to evaluate a dependency tree based on symbolic knowledge, i. e. the semantic role annotations. jwcdg is utilized to link the semantic roles that the visual scenes are annotated with and the syntactic level of RBG. For example, the Agent of an active sentence is supposed to be its Subject. Instead of developing a full grammar that covers all relations between every semantic role and the syntactic level, our grammar covers the cases relevant with respect to our test data and has been developed for German only. But, our grammar will be extended to further cases during the remainder of this project. Also, we plan to extended it to English, Turkish and Chinese. To the best of our knowledge, there exists no comparable system for cross-modal broad-coverage syntactic parsing yet. Since we aim to introduce the corpus of fully/temporally ambiguous sentences in German, more technical aspects of the current parser have been left out of scope here.

3.1 A Test Run

This section presents the results of our proof-of-concept test run, where the performance of our developed cross-modal parser has been tested and compared with the performance of the original

[6] The data-set can be accessed from https://gitlab.com/natsCML/SIMBig2017

RBG model in order to see whether the contextual information improves parsing results.

Our task for the computational model in this test run is to assign thematic roles correctly with respect to the visual depiction of the event. Therefore, the disambiguation task was performed by the cross-modal parser on fully ambiguous sentence (see Table 1, Type [1 - 4], 108 sentences in total). For each sentence, the corresponding visual stimulus has been manually annotated as described in Subsection 2.2.

The RBG models had been trained on the first 100k sentences of the Hamburg Dependency Treebank (HDT) (Foth et al., 2014) part A, a German corpus that is freely available for research purposes. All sentences, which are from the German news website Heise Online[7], are manually dependecy-annotated. TurboTagger[8] is used to predict the PoS tags, from the tag set of Schiller et al. (1995), instead of using the gold standard ones.

RPA-Genitive (Type [1]) case involves 24 sentences. In one half of it, the relative clause is attached to the first NP, *far attachment*, and in the other half, it is attached to genitive object, *near attachment*. The original RBG was not able to attach relative clauses correctly in all 12 cases of far attachment, while there was no wrong attachment in the case of near-attachments as expected due to the respective statistical distribution in the training data. In contrast, our cross-modal parser was able to attach all relative-clauses correctly by utilizing external contextual information.

RPA-Scope (Type[2]) is also consisting of 24 sentences; in one half, relative clause is attached to both NPs (wide scope), while it is attached to only the closest NP in the rest (narrow-scope). A similar pattern in the parsing results as in the previous case was observed. While the original RBG was not able to make any correct attachment for the wide-scope cases, our model correctly attached all relative clauses.

RPA-Dative (Type [3]) set contains 20 sentences; one half is far-attached and the other half is near-attached. The previous pattern was again observed in this case. While the original RBG was blind to far-attachments, our parser was able to disambiguate the sentences by using external cues.

In case of RC-gender, RBG attached all agents and patients correctly but with wrong syntactic labels in 20 out of 40 cases. Our cross-modal parser improved those results by labeling only 10 agent/patients wrongly. The performance of this is expected to be improved by fine-tuning of the semantic annotations employed during parsing operations.

[7] https://www.heise.de

[8] TurboTagger is distributed together with TurboParser (Martins et al., 2013)

4 Discussion

Which linguistic entity resolves the ambiguities under different ambiguity and complexity conditions by humans gives us valuable information about the underlying mechanism of language-vision interaction in a situated setting, enabling us to improve a psycho-linguistically plausible parser. However, for designing such a parser, in addition to reach an understanding in two endeavors, namely the cognitive aspects of language processing and technical aspects of parsing technology, the multi-modal data-set that pertains very challenging garden-path (fully or temporally ambiguous) cases for both areas in a systematic way needs to be designed carefully. This paper addresses this bridging component.

Here we introduce a multi-modal set[6] for ambiguous German sentences addressing 8 different linguistic and four different visual complexities. Furthermore, the contribution of the external information in parsing operations was shown by a proof-of concept study. Further studies will address the comparison between performance of human subjects and computational model on both disambiguation and structural predictions tasks concerning the entire data-set.

Acknowledgments

This research was funded by the German Research Foundation (DFG) in project “Crossmodal Learning”, TRR-169.

References

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73(3):247–264.

Christopher Baumgärtner, Niels Beuck, and Wolfgang Menzel. 2012. An architecture for incremental information fusion of cross-modal representations. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*. IEEE, Hamburg, Germany, pages 498–503.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2016. Do you see what i mean? visual resolution of linguistic ambiguities. *arXiv preprint arXiv:1603.08079* .

Niels Beuck, Arne Köhn, and Wolfgang Menzel. 2011. Incremental parsing and the evaluation of partial dependency analyses. In *DepLing 2011, Proceedings of the 1st International Conference on Dependency Linguistics*.

Moreno I Coco and Frank Keller. 2015. The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology* 68(1):46–74.

Kilian A. Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because size does matter: The Hamburg Dependency Treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Language Resources and Evaluation Conference 2014*. LREC, European Language Resources Association (ELRA), Reykjavik, Iceland.

Pia Stefanie Knoeferle. 2005. *The role of visual scenes in spoken language comprehension: Evidence from eye-tracking*. Ph.D. thesis, Universitätsbibliothek.

André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pages 617–622.

MR Mayberry, Matthew W Crocker, and Pia Knoeferle. 2006. A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In *Proceedings of the 28th annual conference of the Cognitive Science Society*. pages 567–572.

Patrick McCrae. 2009. A model for the cross-modal influence of visual context upon language procesing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2009)*. Borovets, Bulgaria, pages 230–235.

Patrick McCrae. 2010. A computational model for the influence of cross-modal context upon syntactic parsing .

Ken McRae, Mary Hare, Todd Ferretti, and Jeffrey L Elman. 2001. Activating verbs from typical agents, patients, instruments, and locations via event schemas. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Erlbaum Mahwah, NJ, pages 617–622.

Amr Rekaby Salama and Wolfgang Menzel. 2016. Multimodal graph-based dependency parsing of natural language. In *International Conference on Advanced Intelligent Systems and Informatics*. Springer International Publishing, pages 22–31.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das tagging deutscher textcorpora mit STTS. *Universität Stuttgart und Universität Tübingen* .

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268(5217):1632.

Jos JA Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(3):443.

Yuan Zhang, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Amir Globerson. 2014. Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pages 197–207.