A Multi-modal Data-Set for Systematic Analyses of Linguistic Ambiguities in Situated Contexts

Özge Alaçam(✉), Tobias Staron, and Wolfgang Menzel

Department of Informatics, University of Hamburg, 22527 Hamburg, Germany
alacam@informatik.uni-hamburg.de

Abstract. Human situated language processing involves the interaction of linguistic and visual processing and this cross-modal integration helps to resolve ambiguities and predict what will be revealed next in an unfolding sentence during spoken communication. However, most state-of-the-art parsing approaches rely solely on the language modality. This paper aims to introduce a multi-modal data-set addressing challenging linguistic structures and visual complexities, which state-of-the-art parsers should be able to deal with. It also briefly addresses the multi-modal parsing approach and a proof-of-concept study that shows the contribution of employing visual information during disambiguation.

1 Disambiguation and Structural Prediction

In order to achieve dynamic human-computer interaction, a better understanding of human perceptual and comprehension processes concerning multi-modal environments is one of the crucial factors that need to be taken into consideration. A considerable amount of empirical research in psycholinguistics indicated that human language processing successfully integrates available information acquired from different modalities to resolve linguistic ambiguities (i.e. syntactic, semantic or discourse) and predict what will be revealed next in the unfolding sentence [1–4]. Online disambiguation and prediction processes allow us to have more accurate and fluent conversations during spoken communication. In contrast, state-of-the-art parsing algorithms are still far away from that accuracy when it comes to challenging linguistic or visual situations. Therefore, by developing a multi-modal parser that integrates contextual (*i.e. visual knowledge*), we expect to enhance syntactic disambiguations (e.g. concerning relative clause attachments and scope ambiguities).

Tanenhaus and his colleagues' study [1] showed that visual information influences incremental thematic role disambiguation by narrowing down the possible interpretations. In their study, they focus on one of the most frequently investigated syntactic ambiguity cases in the literature, namely prepositional phrase (PP) attachment ambiguity, where different semantic interpretations are possible depending on assigning different thematic roles. It can be exemplified as *The*

J. A. Lossio-Ventura and H. Alatrista-Salas (Eds.): SIMBig 2017, CCIS 795, pp. 106–122, 2018.
https://doi.org/10.1007/978-3-319-90596-9_8

man saw the woman with a telescope, where the PP *with a telescope* can be interpreted as modifier of the seeing action (as instrument), as marked in sentence 1 below, or as a modifier of the object as possessive relation as in Sentence 2. In a multi-modal setting where the scene contains a man holding a telescope in his hand or a woman with a telescope, the visual information constrains the referential choices as well as the possible interpretations, helping the disambiguation process.

1. The man saw [the woman]$_{\text{obj}}$ [with a telescope]$_{\text{instrument}}$
2. The man saw [the woman with a telescope]$_{\text{obj}}$

Further evidence that supports this conclusion was provided by Knoeferle [3] by addressing relatively more complex scenes containing more agents and relations for both English and German. The results also indicate that the influence of visual information on language processing occurs independent from the experiment language. Furthermore, Altmann and Kamide's work [2] 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 only towards breakable objects in the scene. Some nouns may also produce expectations for certain semantic classes of verbs by activating so-called event schema knowledge [5]. Besides verbs and nouns, Berkum and his colleague's study [6] 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 [7] investigated the language - vision interaction and how it influences the interpretation of syntactically ambiguous sentences in a simple 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 psycho-linguistically 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 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 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 i.e. from visual modality, neither humans nor parsers can resolve references in syntactically ambiguous cases. They may have

only preferences. 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 in situated contexts. 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 multimodal data-set consisting of fully/temporally syntactically ambiguous sentences provided in situated contexts.

This paper is structured as follows. In Sect. 2, a data-set of various ambiguous linguistic structures and their multi-modal representations are presented. A brief description of our multi-modal parser is presented in Sect. 3, which also addresses a proof-of concept study conducted on fully ambiguous sentence structures. Section 4 summarizes the results of this work and draws conclusions.

2 Linguistic Ambiguities in Situated Contexts

Recently, a corpus of language and vision ambiguities (LAVA) in English has been released [8]. The LAVA corpus contains 237 sentences with linguistic ambiguities that can only be disambiguated using external information provided as short videos or static visual images with real world complexity. It addresses a wide range of syntactic ambiguities including prepositional 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. But our multi-modal data-set that we are introducing in this section is not limited to investigating these specific ambiguities. It can be also a useful resource in the investigation of a wide range of tasks such as object detection or relation extraction, which are more related with Computer Vision or more Cognitive Science oriented issues like the effect of perceptual/conceptual saliences.

In addition to language-specific investigations, we plan to benefit from a cross-linguistic comparison which opens up novel opportunities for studying the mechanisms of situated language processing in humans by carrying out experiments with similar stimuli in different languages. The base language of the data-set is German and English. Chinese and Turkish counterparts are being prepared as well for cross-lingual comparison. This will allow us to study the influence of different linguistic phenomena on the process of multi-modal sentence comprehension, i.e. syntactic support and constituent ordering. For example, syntactic

support is stronger for indogermanic languages like German or English in contrast to Chinese. If the available visual information facilitates syntactic parsing, the effect should be stronger in German than in Chinese. Moreover, constituent ordering can also be studied for instance by comparing the processing of relative clauses preceding their antecedent (Chinese) or following it (German or English). Because these syntactic patterns induce quite different predictions about the characteristics of the target object, they should have a strong impact on the time course of object identification.

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. Die Frau schmückt das Fenster*(n)*, das*(n)* der Mann säubert.
   The woman decorates the window that the man 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 high (*far*) and low (*near*) attachments are possible. Furthermore, the verb is semantically congruent with both NPs as well. In English, since the relative pronoun (*that*) does not have a syntactic marker, both sentences are syntactically ambiguous. 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 Fig. 1b and d).

4. Die Frau schmückt das Fenster*(n)* des Wagens*(m)*, das*(n)* der Mann säubert.
   The woman decorates the window of the car that the man cleans.
5. Die Frau schmückt das Fenster*(n)* des Zimmers*(n)*, das*(n)* der Mann säubert.
   The woman decorates the window of the room that the man cleans.

2.1 LASC Data-Set:V1

Our LASC data-set (Linguistic Ambiguities in Situated Context) is currently consisting of 206 situated contexts (*scenarios*) and 599 sentences and addresses 9 linguistically challenging cases (itemized below) concerning relative clause

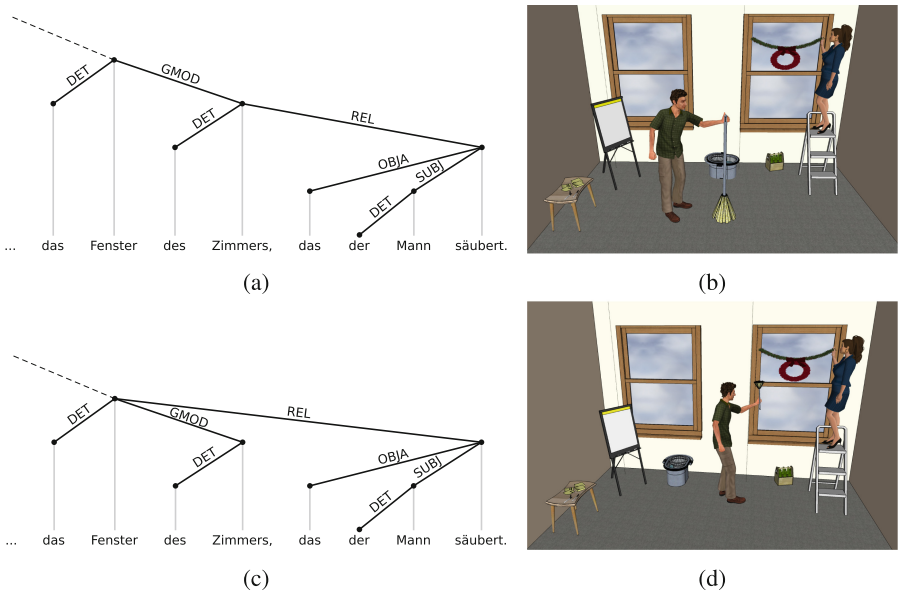

Fig. 1. (a) The first interpretation of syntactically ambiguous Sentence-5: low attachment of relative clause - syntactic gold standard annotation and (b) the corresponding visual scene. (c) the second interpretation of Sentence-5: high attachment of relative clause - syntactic gold standard annotation and (d) the corresponding visual scene.

attachments, Agent/Patient agreement[1], verb/subject agreement, and scope ambiguities for conjunctions and negations. 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. The interpretation of the sentences becomes fully unambiguous in the presence of visual stimuli. The number of scenarios and sentences for each subset is given in Table 1 and the sentences for each structure are generated by using part-of-speech templates exemplified below i.e. for Subset-1A.

– *Template:* PRO1_{nom} VP1 NP1_{acc} NP2_{gen}, $\text{WDT}^{2}{}_{\text{acc}}$ PRO2_{nom} VP2
– *Number of the lexical items:* two pronouns, 48 verbs and 48 nouns

[1] Knoeferle's sentence set [3] was used as baseline since the co-occurrence frequencies between the actions and the Agents in the sentences, as well as between the actions and the Patients, were controlled to single out the effects 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.
[2] Relative Pronoun.

Table 1. The number of scenarios and linguistic structures and the available languages for each subset (*RPA: Relative Pronoun Ambiguity).

	Type	Languages	# of Scenarios	# of Sentences
1	RPA* with a Genitive NP	DE, EN, TR	24	240
2	RP - Scope Amb.	DE, EN	24	24
3	RPA with a Dative PP	DE, EN	48	48
4	RP with an Agent/Patient ambiguity	DE	12	24
5	Negative scope	DE, EN	12	12
6	Agent-Patient agreement	DE	36	72
7	Verb-Subject agreement	DE	6	12
8	Conjunction scope	DE	9	27
9	Constituent (modifier) ordering	DE, EN	35	140
	TOTAL		206	599

Linguistic Structures

Fully Ambiguous Sentence Structures

(1) RPA with a Genitive NP *(DE, EN, TR)*

(A) *Active voice - short sentence.*
(DE) Die Frau schmückt das Fenster*(n)* des Zimmers*(n)*, das*(n)* der Mann säubert.
(EN) The woman decorates the window of the room that the man cleans.
(TR) Adam kadının dekore ettiği odanın penceresini temizliyor.
Int.1[3]*:* The man cleans the room *(low-attachment).*
Int.2: The man cleans the window *(high-attachment).*
(B) *Active voice - long sentence.* The woman with the red hair silently decorates the window of the room that the man cleans in a rush since he needs to go to a meeting soon.
(C) *Passive voice - short sentence.* The woman decorates the window of the room that is cleaned by the man.

(2) RP - Scope Ambiguities *(DE, EN)*

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: Only bananas are on the table (low-attachment).
Int.2: Both apples and bananas are on the table.

(3) RPA with a Dative PP *(DE, EN)*

(A) *Syntactically ambiguous.* 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 *(low-attachment).*
Int.2: She damages the mug *(high-attachment).*

[3] Int.=Interpretation.

(B) *Syntactically unambiguous in German.* Da befindet sich eine Flashe*(f)* auf einem Tisch*(m)*, den*(m)* sie beschädigt.
It is the bottle on the table that she damages. (She damages the table.)

(4) RPA with an Agent/Patient ambiguity *(only DE)*
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 *(DE, EN)*
Die Sängerin kauft die Jacke nicht, weil sie rot ist.
The singer does not buy 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.

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

Temporally Ambiguous Sentence Structures

(6) Agent-Patient Agreement *(only DE)* (following the data-set designed by [3])
- Die Arbeiterin*(f, nom)* kostümiert mal eben den jungen Mann*(m, acc)*.
  The (female) worker just dresses up the young man.
- Die Arbeiterin*(f, acc)* verköstigt mal eben der Astronaut*(m, nom)*.
  The (female) worker is just fed[4] *by the astronaut.*

(7) Verb-Subject Agreement *(only DE)*
- Die Sänger*(f, nom, pl)* waschen*(3rd. Pl.)* den Arzt*(m, acc, sing.)*.
  The singers wash the doctor.
- Die Sänger*(f, acc, pl)* wäscht*(3rd. Sing.)* der Offizier*(m, nom, sing.)*.
  The singers are painted (see footnote 4) *by the officer.*

(8) Conjunction Scope Ambiguities *(only DE)*
- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, acc)*.
  The singer paints the (male) officer and the (female) doctor.
- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, nom)* wäscht den Radfahrer*(m, acc)*.
  The singer paints the (male) officer and the (female) doctor washes the (male) cyclist.
- Die Sängerin*(f, nom)* bemalt den Offizier*(m, acc)* und die Ärztin*(f, acc)* besprüht der Radfahrer*(m, nom)*.
  The (female) singer paints the (male) officer and the (female) doctor is sprayed (see footnote 4) *by the (male) cyclist.*

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

(9) Constituent (Modifier) Ordering *(DE, EN)*

- Bring mir den blauen Becher vom Tresen.
  Bring me the blue mug from the counter.
- Bring mir den Becher, den blauen Becher vom Tresen.
  Bring me the mug, the blue one from the counter.

Image Construction and Semantic Annotations. The multi-modal data-set has been designed to investigate how, when and at which degree does visual complexity affect sentence comprehension and whether visual cues are still strong enough to enforce correct interpretations in such complex linguistic cases.

It should be reminded that for the computational model, we do not need visual scenes, their semantic representations are sufficient, however, the scenes 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 multi-modal data-set. 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 1250×840 resolution. Moreover, target objects and agents are located in different parts of the visual scene for each stimulus.

The objects, characters and actions in the images were annotated manually with respect to their semantic roles, similar to McCrae's approach [9], see also [10]. 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 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 2 shows one exemplary semantic annotation for the visual scene displayed in Fig. 1b. There *'die Frau'* is the Agent, who performs the decorating action, *'das Fenster'* is the Patient, the entity undergoing a change of state, caused by the action.

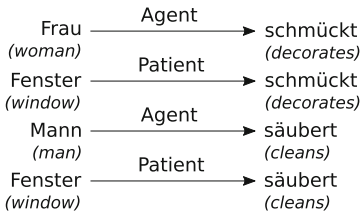

Fig. 2. One exemplary semantic annotation for the visual scene shown in Fig. 1b.

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

The current version of our LASC multi-modal data-set(v1)[6] that we constructed with the aim of studying disambiguation and structural prediction from both psycholinguistics and computational linguistics perspectives contains following items for each scenario.

- a linguistic form in various languages.
- target interpretation (*gold standard annotations*)
- possible interpretations
- a visual scene of the target interpretation in different visual complexities
- a semantic representation of the scene
- an audio file and a data file with marked onset/offsets (in msec.) of each linguistic entities in the sentence

Visual Complexity. The following figures illustrate how complexity is systematically controlled by giving one subset of the sentences as an example, namely Agent-Patient agreement (Subset-6). In the initial condition, each scenario contains three characters (one Patient, one Agent and one ambiguous Agent/Patient

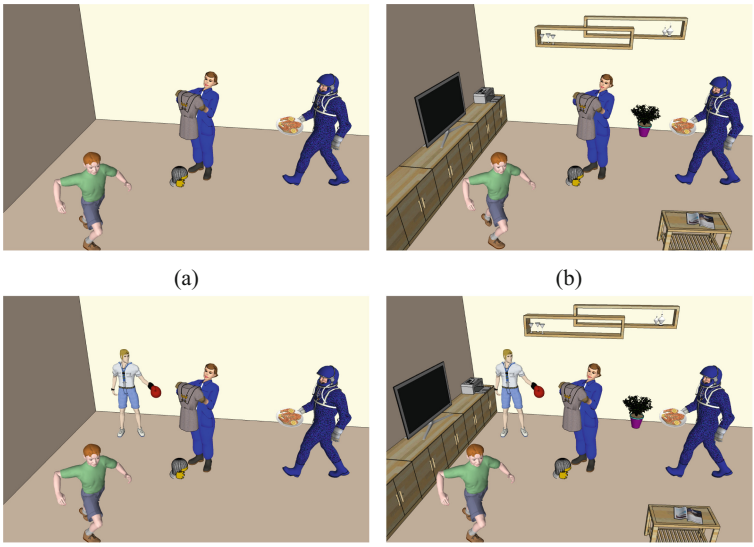

Fig. 3. (a) 3 agents in an environment with no background objects; *a Patient* (a young boy on the left), *an Agent* (an astronaut on the right) and *an ambiguous Agent/Patient character* (a female worker in the middle) (b) 3 agents in an environment with background objects, (c) 4 agents in an environment without background objects and (d) with background objects.

[6] The data-set can be accessed from https://gitlab.com/natsCML/LASC_v1.

character) and two possible actions. However, the linguistic modality in the data-set 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 without additional background objects, see Fig. 3a. This set-up resemblances Knoeferle's [3] 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 Fig. 3b. 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 Fig. 3c and d). 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.

Another example of visual complexities represented in the data-set can be given regarding the subset-9 that addresses the effect of constituent ordering. The variations in the scenes of this subset were tailored to investigate how the processing of the linguistic description of the target object (i.e. *the blue mug*) is influenced by the number of the distractor objects which can also be described with the same modifier (i.e. *blue*) or by other visual constraints like saliency and occlusion (i.e. Fig. 4b).

(a) (b)

Fig. 4. A visual scene for the sentence given as an example in Subset-9. In a simple case (a), there is only one blue mug. In a more complex scenario (b) there are two blue mugs, but the target is occluded. (Color figure online)

3 Multi-modal Parsing

As suggested by the literature discussed in Sect. 1, multi-modal integration is a key to resolve linguistic ambiguities and to anticipate what will be revealed

next in an unfolding sentence. However, most state-of-the-art parsing approaches rely solely on the language modality. One of the first systems that integrates contextual knowledge into a grammar-based parser to resolve ambiguities in German (e.g. addressing Genitive-Dative ambiguities of nouns with feminine case markings or PP attachment ambiguities) was proposed by McCrae [11]. Based on this study, Baumgärtner et al. [12] successfully realized a system integrating visual context to improve the processing of German sentences in an incremental manner leading to the only parser so far that is both incremental and cross-modal.

For the proof-of-concept study reported in the following subsection, a parser that utilizes a simple two-stage filtering approach has been developed. In contrast to the rule-based parsers [11,12], we employ a hybrid parsing approach consisting of both data-driven and grammar-based components with the aim to develop a language-independent parser that achieves state-of-the-art results in resolving linguistic ambiguities. The data-driven parser, which searches for the most plausible disambiguation of a given sentence among all possible dependency trees, is interfaced with a grammar-based component, which evaluates the most probable candidates with respect to information derived from visual input, to narrow down the hypotheses towards the most plausible representation for the sentence at hand with respect to given context.

The semantic role annotations of images elaborated in Sect. 2.1 serve as representations of contextual information. The contexts are assumed to be dynamic, i.e. the environments depicted in the visual stimuli may constantly change. So, unseen examples are not guaranteed to be represented by the training instances. Thus, additional modalities, which provide contextual information, cannot be incorporated by extracting features and adding them to the ones derived from the input sentence [17], or by learning a separate model. I the contextual information contains previously unseen information, it might not improve results or might even deteriorate parsing performances. Therefore, instead of using data-driven models for processing the context, we developed a grammar that links semantic roles and the corresponding syntactic structures. The constraints of this linking grammar work with the available information, e.g. Part-of-Speech (PoS) tags, dependency relations or syntactic labels, except for the word forms. Thus, the grammar is not lexicalized and independent of actual words. If there is a match between a semantic role and a syntactic structure, e.g. the Patient of the relative clause verb coincide with one of the possible attachments, then one linking constraint enforces that the relative clause has to be attached to that Patient. This way, it is ensured that the relative clause in Sentence-5 is attached to the *window* as the Patient of the *cleaning* action, and not to the *room*. Instead of developing a full grammar that covers all relations between every semantic role and the syntactic level, in the current version of the data-set, the content of our grammar is limited to the cases relevant with respect to the mentioned actions, agents and objects in the sentence sets. Since the images and the sentences are publicly available, any other relations with respect to user's particular interests can be extracted.

Figure 5 summarizes the parsing process. First, the data-driven RBGParser (RBG) generates the k-best candidates S_1, ..., S_k for the input sentence $x_{linguistic}$. RBG is a language independent, data-driven dependency parser and achieves state-of-the-art results [18,19]. Hill climbing is applied to approximate the optimal dependency tree, i.e. the Maximum Spanning Tree, except for cases an edge-factored model, which consists of first-order features only, is used. Then, the optimal dependency tree is determined by the Chu-Liu-Edmonds (CLE) algorithm [20]. Since we are interested in non-canonical interpretations (namely ambiguous sentences), which are usually not well represented in the training data, the best chosen parse will likely not represent the correct parse, but previous research has shown that the desired parse can more often be recovered if the k-best parses are taken into account [16]. While edge-factored models perform worse compared to higher-order models regarding 1-best parsing, their k-best results are sufficient as input for a second-stage parser, of which model can be arbitrarily complex [21]. Finding the k-best Spanning Trees based on an edge-factored model can be achieved by the algorithm of Camerini et al. [15], which extends the CLE algorithm.

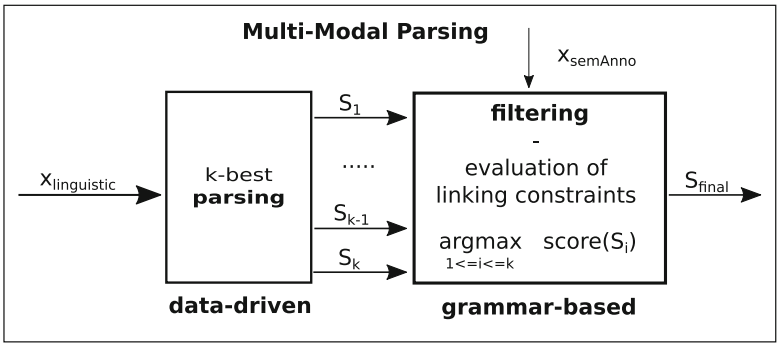

Fig. 5. Multi-modal parsing architecture - The data-driven component performs k-best parsing for input sentence $x_{linguistic}$. The grammar-based component filters all k candidates S_1, ..., S_k by evaluating constraints, which link semantic roles and the corresponding syntactic structures, with respect to the semantic annotation $x_{semAnno}$ of the visual input, the contextual information, to find the most probable solution S_{final}.

Next, all semantic role annotations $x_{semAnno}$ of the corresponding image, in other words the contextual information, are mapped onto the corresponding words of $x_{linguistic}$. For example, the *man* or the *window* from Fig. 2 are mapped onto the respective words of Sentence-5. The scoring mechanism of the grammar-based jwcdg [14], re-evaluates all candidate parses with respect to the semantic annotations using our linking grammar and excludes all candidates that violate constraints. The best remaining candidate S_{final} is the most probable solution. In case, no solution fits the contextual information, the 1-best parse of RBG

is taken as a fallback solution. To the best of our knowledge, there exists no comparable system for multi-modal broad-coverage syntactic parsing yet.

3.1 A Proof-of-Concept Study

A proof-of-concept study was conducted in order to evaluate the multi-modal approach to parsing presented in this section and to compare it to the unimodal RBG to show that a simple filtering approach is already sufficient to improve parsing results in case contextual information is provided. The study covered four subsets of fully ambiguous sentences from Sect. 2: (i) RPA with a Genitive NP *(Subset-1A, DE and EN)*[7], (ii) RP - Scope Ambiguities *(Subset-2)*, (iii) RPA with a Dative PP *(Subset-3A)* and (iv) RP with an Agent/Patient ambiguity *(Subset-4)*. For each type of ambiguity, 24 test sentences were chosen. For each sentence, the corresponding contextual information, i.e. the visual stimulus, was manually annotated as described in Subsect. 2.1. The entire test was conducted with German sentences.

In Subset (1) to (3), there are two possible antecedents for each relative clause. On the syntactic level, the relative pronoun agrees with both possible antecedents, i.e. in case, gender and number. On the semantic level, both possibilities could fill the same semantic role. In Subset (4), either the relative pronoun is the Subject and the subsequent NP is the Object or vice versa (due to the free word order of the German language) and both can fill the respective semantic roles. The first interpretation, e.g. low attachment, is the target hypothesis in one half of the test sentences and the second interpretation, e.g. the high attachment, in the other half.

Two RBG models have been trained: a full model, which exploits up to third-order local features as well as global features, and an edge-factored model. Both models have been trained on the first ≈ 98k sentences (without duplicates) of the Hamburg Dependency Treebank (HDT) part A [22]. The dependency relations of all sentences, taken from the German news website Heise Online[8], are manually annotated and the data includes word forms, gold PoS tags (from the Stuttgart-Tübingen Tagset [23]), and gold standard annotations. Instead of using the gold standard PoS tags, TurboTagger [24] is used to predict them. For tagging the training sentences, ten-way jackknifing was performed: the training set is split into ten partitions and each partition is tagged by a model trained on the other nine partitions. The test sentences were tagged by models trained on the entire training set of the respective corpus.

Table 2 shows the results for disambiguating the different types of ambiguities for both the unimodal RBG and the multi-modal filtering approach. For (1) RPA - a Genitive NP (A.), the unimodal RBG always chooses the low attachment of the relative clause, as expected due to the respective statistical distribution in the training data. Thus, it is not able to attach relative clauses

[7] See [25] for a study focused more on the experiments on this Subset regarding all three languages: German, English and Turkish.

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

Table 2. The resolution of the ambiguities for the four different types of linguistic ambiguities for unimodal parsing compared to multi-modal parsing.

	low	high	other
low attachment	12	0	0
high attachment	12	0	0

(a) RPA with a Genitive NP unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(b) RPA with a Genitive NP multi-modal RBG

	low	high	other
low attachment	12	0	0
high attachment	12	0	0

(c) RP - Scope Ambiguities unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(d) RP - Scope Ambiguities multi-modal RBG

	low	high	other
low attachment	11	0	1
high attachment	11	0	1

(e) RPA with a Dative PP unimodal RBG

	low	high	other
low attachment	12	0	0
high attachment	0	12	0

(f) RPA with a Dative PP multi-modal RBG

	s-o	s-s	o-s	o-o	s-other
subject-object	5	6	0	0	1
object-subject	5	6	0	0	1

(g) RP with an Agent/Patient ambiguity unimodal RBG

	s-o	s-s	o-s	o-o	s-other
subject-object	11	0	0	0	1
object-subject	0	10	0	0	2

(h) RP with an Agent/Patient ambiguity multi-modal RBG

correctly in case the high attachment is part of the target hypothesis. In contrast, the multi-modal filtering approach attaches all relative-clauses accordingly with respect to the target hypothesis by utilizing the contextual, i.e. visual, information as described in this section. The same behavior has been observed for (2) RPA - Scope Ambiguities and (3) RPA - a Dative PP (A.). For the latter type, the original, unimodal RBG has attached one supposed low and one supposed high attachment to completely different antecedents. The multi-modal approach avoids these errors.

For (4) RPA with an Agent/Patient ambiguity, the unimodal RBG attached all Agents and Patients respectively subjects and objects correctly, but with correct syntactic labels in only 5 out of 24 cases. Mainly, it is not able to recognize the relative pronoun as object and not able to assign the correct label to the NP following the relative pronoun either. On the other hand, the multi-modal filtering approach improve these results by labeling 11 cases correctly by assigning the correct label to the NP following the relative pronoun in 21 cases. The problem of not recognizing the relative pronoun as object remains. This problem may be

overcome by more sophisticated approaches to multi-modal parsing instead of using our simple filtering-based approach.

4 Discussion

Employing a parser that mimics some of the important mechanisms of natural language processing such as prediction and disambiguation is crucial for enabling more fluent and dynamic spoken communication. Which linguistic entity resolves the ambiguities in systematically controlled situated contexts gives us valuable information about the underlying mechanism of human language-vision interaction. In addition to reach an understanding in two endeavors, namely the cognitive aspects of language processing and technical aspects of parsing technology, a multi-modal data-set that pertains challenging 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 (see footnote 6) for temporally or fully ambiguous sentences in various languages addressing 9 different linguistic structures and 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 the performance of human subjects and of computational model regarding both disambiguation and structural predictions tasks.

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

References

1. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C.: Integration of visual and linguistic information in spoken language comprehension. Science **268**(5217), 1632 (1995)
2. Altmann, G.T., Kamide, Y.: Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition **73**(3), 247–264 (1999)
3. Knoeferle, P.S.: The role of visual scenes in spoken language comprehension: evidence from eye-tracking. Ph.D. thesis, Universitätsbibliothek (2005)
4. Ferreira, F., Foucart, A., Engelhardt, P.E.: Language processing in the visual world: effects of preview, visual complexity, and prediction. J. Mem. Lang. **69**(3), 165–182 (2013)
5. McRae, K., Hare, M., Ferretti, T., Elman, J.L.: 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, pp. 617–622 (2001)
6. Van Berkum, J.J.A., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P.: Anticipating upcoming words in discourse: evidence from ERPs and reading times. J. Exp. Psychol. Learn. Mem. Cogn. **31**(3), 443 (2005)
7. Coco, M.I., Keller, F.: The interaction of visual and linguistic saliency during syntactic ambiguity resolution. Q. J. Exp. Psychol. **68**(1), 46–74 (2015)

8. Berzak, Y., Barbu, A., Harari, D., Katz, B., Ullman, S.: Do you see what I mean? Visual resolution of linguistic ambiguities. arXiv preprint arXiv:1603.08079 (2016)
9. McCrae, P.: A computational model for the influence of cross-modal context upon syntactic parsing (2010)
10. Mayberry, M.R., Crocker, M.W., Knoeferle, P.: 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, pp. 567–572 (2006)
11. McCrae, P.: A model for the cross-modal influence of visual context upon language processing. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria, pp. 230–235 (2009)
12. Baumgärtner, C., Beuck, N., Menzel, W.: An architecture for incremental information fusion of cross-modal representations. In: IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Hamburg, Germany, pp. 498–503. IEEE (2012)
13. Beuck, N., Köhn, A., Menzel, W.: Incremental parsing and the evaluation of partial dependency analyses. In: DepLing 2011, Proceedings of the 1st International Conference on Dependency Linguistics (2011)
14. Beuck, N., Köhn, A., Menzel, W.: Predictive incremental parsing and its evaluation. In: Computational Dependency Theory. Frontiers in Artificial Intelligence and Applications, vol. 258, pp. 186–206. IOS Press (2013)
15. Camerini, P.M., Fratta, L., Maffioli, F.: The k best spanning arborescences of a network. Networks **10**(2), 91–109 (1980)
16. Charniak, E., Johnson, M.: Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 173–180. Association for Computational Linguistics, June 2005
17. Salama, A.R., Menzel, W.: Multimodal graph-based dependency parsing of natural language. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (eds.) AISI 2016. AISC, vol. 533, pp. 22–31. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48308-5_3
18. Zhang, Y., Lei, T., Barzilay, R., Jaakkola T., Globerson, A.: 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), Baltimore, Maryland, pp. 197–207. Association for Computational Linguistics (2014)
19. Lei, T., Xin, Y., Zhang, Y., Barzilay, R., Jaakkola, T.: Low-rank tensors for scoring dependency structures. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, Maryland, pp. 1381–1391. Association for Computational Linguistics, June 2014
20. Tarjan, R.E.: Finding optimum branchings. Networks **7**(1), 25–35 (1977)
21. Hall, K.: k-best spanning tree parsing. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 392–399 (2007)
22. Foth, K.A., Köhn, A., Beuck, N., Menzel, W.: Because size does matter: the Hamburg dependency treebank. In: Proceedings of the Language Resources and Evaluation Conference 2014, LREC, European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
23. Schiller, A., Teufel, S., Thielen, C.: Guidelines für das tagging deutscher textcorpora mit STTS. Universität Stuttgart und Universität Tübingen (1995)

24. Martins, A.F.T., Almeida, M.B., Smith, N.A.: Turning on the turbo: fast third-order non-projective turbo parsers. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 617–622 (2013)
25. Staron, T., Alacam, O., Menzel, W.: Incorporating contextual information for language-independent, dynamic disambiguation tasks. In: Proceedings of the 11th Language Resources and Evaluation Conference (LREC) (2018)