Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)


# How Does Visual Complexity Influence Predictive Language Processing in a Situated Context?

**Özge Alaçam, Wolfgang Menzel** and **Tobias Staron**
Department of Informatics
University of Hamburg, Hamburg 22527, Germany
{alacam, menzel, staron}@informatik.uni-hamburg.de

## Abstract

In this study, by using a visual world paradigm, we investigate to what extent predictive language processing is affected by the visual complexity of the scene. The visual complexity is manipulated by adding irrelevant background objects or irrelevant characters (w.r.t. the spoken utterance) to the scene. The results of the eye-tracking experiment, that point out significant effects of visual complexity on situated-language processing, provide basic insights for designing cross-modal language understanding systems.

Keywords: incremental / predictive language processing; visual complexity

## 1 Predictive Language Processing

A large body of empirical evidence in psycholinguistics indicates that the interaction between linguistic and visual modalities plays a crucial role in predicting what will be revealed next in the unfolding sentence (Altmann and Kamide, 1999; Knoeferle, 2005; Tanenhaus et al., 1995).

Altmann and Kamide (1999)'s study on structural prediction has documented that listeners are able to predict complements of a verb based on its selectional constraints and immediately begin incremental parsing operations. For example, when people hear the verb "break", their attention is directed towards only breakable objects in the scene.

Focusing on prosodic cues and visual saliency, Coco and Keller's research (2015) goes deeper into the understanding of which kinds of information play role on different comprehension processes regarding situated predictive language processing. This study contains three systematic manipulations; namely (i) only the visual saliency, (ii) only the linguistic saliency and (iii) both of them together. The results point out that visual saliency narrows down the visual search space towards a target, but does not have a direct role on linguistic ambiguity resolution, while different intonational breaks favor one interpretation over the other. On the other hand no statistical interaction effect between the two modalities has been found although they complement each other and both contribute to the overall understanding of the sentence by having different roles.

Numerous studies on structural prediction have only been carried out on relatively simple visual and linguistic settings where object-action relations could be predicted relatively easily, except several studies (Ferreira et al., 2013; Coco and Keller, 2015). Therefore, to what extent this multimodal interaction occurs still needs extensive systematic investigation. As reported by Ferreira et al. (2013), when the visual context is complex, subjects have difficulties using visual information to narrow down their hypotheses about possible interpretations. Depending on the complexity of the visual environment or task, humans might choose a more passive strategy (such as waiting to have complete information about the entities referred to in the utterance) instead of anticipating upcoming information, and humans have such a preference for the cases where there is a high risk of faulty prediction.

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 fluent and efficient cross-modal and incremental parsing solutions. Better understanding of human perceptual and comprehension processes is one of the crucial factors in the realization of dynamic human-computer interaction. This papers addresses the empirical aspects of human language processing particularly focusing on the influence of different irrelevant visual components on the predictive lan-

guage processing.

## 2 Experiment

One of the well-investigated phenomena regarding the interplay between language and vision is *the garden-path effect*, which occurs when the prediction made for the upcoming sentence part does not match with the real situation, requiring re-analysis of the interpretation during online language comprehension. Subject-object ambiguities in German elicit garden-path effects due to sometimes ambiguous case marking. This phenomenon has already been observed by Knoeferle et al. (2005). In their study, two different sentence patterns are compared; unmarked word order ($\text{Subject}_{\text{ambiguous}}$ Verb $\text{Object}_{\text{accusative}}$) and marked word order ($\text{Object}_{\text{ambiguous}}$ Verb $\text{Subject}_{\text{nominative}}$). Each sentence addresses only one action between two characters and is accompanied by a scene that depicts two actions and three characters (one ambiguous AGENT/PATIENT character, one PATIENT, and one AGENT). If the preference-driven initial role assignment for the first noun phrase (before the verb) creates a conflict with the late assignment for the second one (following the verb), a reanalysis becomes necessary. Eye-tracking results show that this happens only in case of the marked word order. More interestingly, visual attention already starts to move towards the target character before the associated post-verbal noun phrase actually becomes available, i.e. while the verb is still being spoken. This clearly signals that reference resolution for the second noun phrase is not based on the observation of the phrase itself but on its prediction induced by the verb. To find out whether this effect also occurs in more complex visual environments, we compiled a set of four different visual conditions accompanied by two different versions of AGENT/PATIENT order following the same sentence patterns.

[1] Unmarked word order (SVO):
Die Arbeiterin*(f, nom)* kostümiert mal eben den jungen Mann*(m, acc)*.
*The worker just dresses up the young man.*

[2] Marked word order (OVS):
Die Arbeiterin*(f, acc)* verköstigt mal eben der Astronaut*(m, nom)*.
*The worker is just fed*[1] *by the astronaut.*

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

The spoken sentences were recorded by a male native speaker of German at a normal speech rate. We avoided unequal intonational breaks that may bias the interpretation.

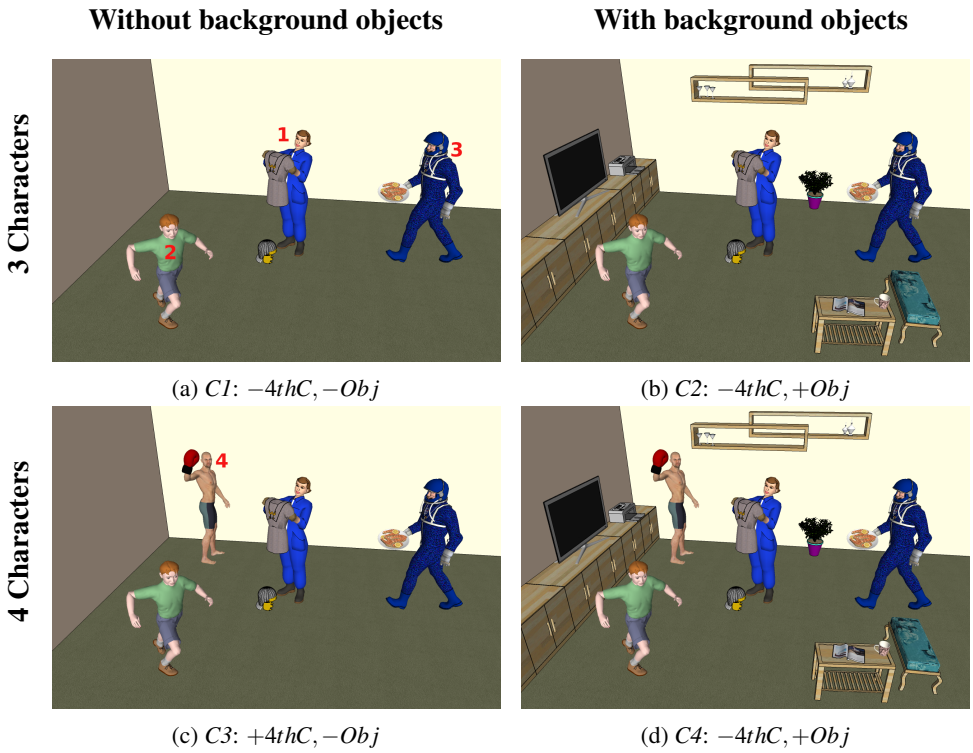
The visual context differs from those of Knoeferle et al. (2005) by a more realistic and complex background. In particular, the pictures contain more information than the sentences describe. We manipulate the visual complexity in four different conditions, see Figure 1. In the first condition (C1), the scene contains three characters (one ambiguous AGENT/PATIENT candidate, one for the PATIENT role, and one for AGENT) as in the original study. In the second condition (C2), more background objects are added, while in the third condition (C3) an additional character is included. The distractor objects and the characters have no direct thematic relation regarding the spoken sentence, therefore they should not affect the fixations on the target object particularly in the unmarked sentences. However, the additional character is acting on the ambiguous AGENT/PATIENT character, thus increasing the complexity of referential selection. Therefore, we hypothesize that having more background objects but no additional character (C2) should distract the subjects from the target not as much as condition (C3) does. Finally, C4 is a combination of C2 and C3.

In order to ensure the compatibility of the verbs and nouns with their depiction in the scenes, the scenes (with background objects and four characters) were shown to 15 participants, 4 of 40 stimuli were excluded from the stimuli set, since at least one of the actions7nouns was found not easily depictable.

27 university students participated in the experiment. The experiment was conducted in German by native speakers. Using the visual-world paradigm, a total of 32 visual displays with accompanying spoken utterances were utilized. The stimuli were displayed on an SR Eyelink 1000 Plus eye tracker with a sampling rate of 1000 Hz. The 2D visual scenes were created with the SketchUp Make Software[2] with a resolution of 1250 x 947px.

A visual scene is presented for 10 sec before the onset of the spoken sentence. The preview gives a comprehender time to encode the visual information so visual attention will be free of recognizing the objects during language processing. Then, the spoken sentence is presented accompanying the vi-

[2]`http://www.sketchup.com/` – retrieved on 10.09.2018




(a) *C1*: $-4thC, -Obj$ (b) *C2*: $-4thC, +Obj$

(c) *C3*: $+4thC, -Obj$ (d) *C4*: $-4thC, +Obj$

Figure 1: Example stimuli in four different visual complexity conditions

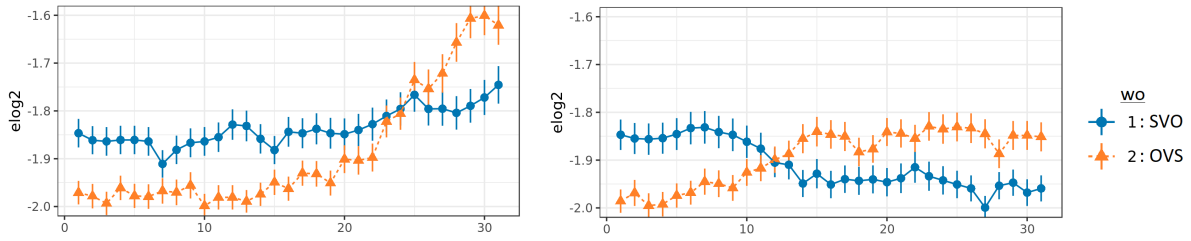
sual stimulus. Participants are asked to examine the scene carefully and attend the information given in the audio. The order of stimuli is randomized for each participant.

In this study, we focus only on the time course of fixations on the target (either AGENT or PATIENT depending on the word order) and on the competitor characters after the onset of the verb until its offset. Due to varying word length, the time window for a verb duration was normalized by stretching each individual time series to the maximum verb length observed and then reduced to 31 bins by aggregating 5 bins to one. The time window is shifted forward 200 ms in order to account for the time required to initiate eye movement (Matin et al., 1993). In total, fixation distributions of 858 trials (27 participant * 32 scenes) per character were evaluated. The fixations were coded as binomial w.r.t. whether the character is fixated or not. Fixation parameter was transformed into empirical logitbased on *population-average* estimates with weightsfollowing the reasoning discussed in (Barr, 2008; Jaeger, 2008).

## 3 Results

All analyses were carried out in R version 3.5.1. (Team, 2013) by utilizing *Lmertest*, *Lme4* and *multcomp* packages. Due to the expected curvilinear change over time, a higher-order polynomial model was chosen (Mirman, 2016; Baayen et al., 2008) to analyze the effects of word order, and visual complexities over the course of unfolding sentence. We start out with a ”base” model of fixation distribution over time with crossed-random effects of items and subjects on all orthogonal polynomial terms. Adding the visual complexity parameter (*Vis*) significantly improved model fit ($\chi^2(3) = 182.75, p < .001$), as well as the word order parameter (*Ling*), ($\chi^2(1) = 36.26, p < .001$). Finally, the full model with all interaction effects of the fixed termsprovided the best fit compared to previous models ($\chi^2(15) = 498.78, p < .001$).

As given in Table 1, the main effect of word order is significant indicating that the fixation distribution in two word order conditions differs significantly. There is also a significant effect of word order on the linear term for both characters, indicating that the slope of the fixation distribution is steeper in the OVS compared to that in the SVO order. Figure 2 shows that – when the sentence in SVO order unfolds – the look on the target objects increases, and the look on the competitor decreases (blue lines). However, in the OVS order, an increase is observed both on the target and competitor characters (orange lines).




(a) Target (b) Competitor

Figure 2: Time course of fixations on the target and competitor characters in two word order conditions

| | Target | | | Competitor | | |
|---|---|---|---|---|---|---|
| | *Est.* | *SE* | *p* | *Est.* | *SE* | *p* |
| intercept | −.19 | .02 | < .001* | .27 | .02 | < .001* |
| linear | .99 | .08 | < .001* | .75 | .09 | < .001* |
| quadratic | .15 | .08 | .07 | −.61 | .09 | < .001* |

Table 1: Estimated parameters for the word order parameter on the target and competitor characters

| | *Linear* | | | *Quadratic* | | |
|---|---|---|---|---|---|---|
| **Contrasts** | *Est.* | *SE* | *p* | *Est.* | *SE* | *p* |
| **1 vs. 2** | −.48 | .06 | < .001* | −.51 | .06 | < 0.001* |
| **1 vs. 3** | −.34 | .06 | < .001* | −.19 | .06 | < 0.001* |
| **1 vs. 4** | −.48 | .06 | < .001* | −.25 | .06 | < 0.001* |
| **2 vs. 3** | −.15 | .06 | < .05* | .32 | .06 | > .05 |
| **3 vs. 4** | −.14 | .06 | < .05* | −.06 | .06 | > 0.05 |

Table 2: Multiple comparisons on the target character regarding four visual complexity levels

As summarized in Table 2, the first three contrasts on slope (linear term) and curvature (quadratic term) of the fixation parameter over time show that the slope of the *C1* is always steeper with significantly more fixations compared to the other three complexities. Figure 3 shows that the look on the target character starts around timestamp 21 (out of 31), directly after half of the verb has been uttered. On the other hand, this reaction is significantly slower for the other conditions. When background objects are added (*C2*), less fixation is observed on the target characters. The inclusion of the 4th character *C3* also causes a decrease of the look on the target compared to the first condition. The same pattern is observed between *C1* and *C4*.

The difference between *C2* and *C3* indicates that although the curvatures of the fixation distributions display similar pattern (quadratic term is $> 0.05$), the slope terms are significantly different indicating that adding 4th character results in overall slower increase on the fixation to the target compared to adding background objects *C2*. Finally, 3rd and 4th condition also show a difference only in the slope

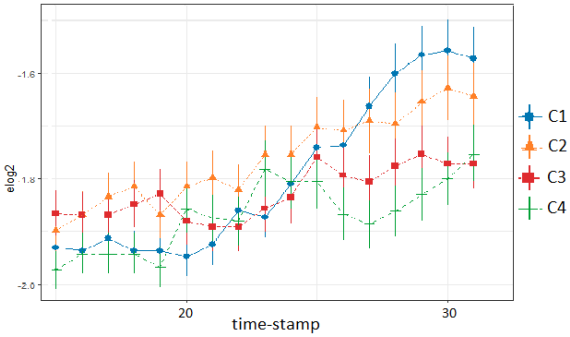



Figure 3: Time course of fixations on the target character for four visual complexity levels

term but not in the curvature.

Further analysis on the interaction between the word order and the visual complexity parameters indicated that while the fixation distributions for the sentences in different word order conditions significantly differed within the first ($Estimate = -0.133, SE = 0.013, p < 0.001$) and the second ($Estimate = -0.062, SE = 0.012, p < 0.001$) complexity levels, , it, however, seems that this difference between the unmarked and marked sentence forms starts to fade with the inclusion of additional character (non-significant differences in SVO and OVS fixations for the 3rd and 4th visual complexity levels).

## 4 Discussion and Conclusion

The results replicate the findings previously reported in the literature that states the participants are garden-pathed when they hear a sentence in a OVS order, in which the expected order of the thematic roles are reversed (Knoeferle, 2005). This is in line with the NVN strategy, which states there is a tendency to assume that the first argument of a sentence is a proto-agent and the second is a proto-patient (Ferreira, 2003). Moreover, in OVS order one can see a late increase on the target. This result is also in line with the surprisal effect theory that

states that less predictable items are fixated longer (Levy, 2008; Hale, 2001). However, when we look at the comparisons in more detail for each visual complexity level separately, it seems that while the difference between the unmarked and marked sentences is preserved for the simple visual complexity conditions, no clear sign of prediction of the upcoming sentence part is observed when the complexity is increased. Similar to the findings by Coco and Keller (2015), this result implies that visual complexity may not have direct role on thematic role assignment, however, it does definitely have an effect on target identification.

Furthermore, our results also support an interactive processing architecture that claims visual information influences the processing of syntactic linguistic information (MacDonald and Seidenberg, 2006), in our case even when they are irrelevant to sentence context. Regarding visual complexity, although none of the manipulations directly has an association with the entities mentioned in the sentence, the results still reveal that the looks on the target are affected by the complexity of the environment, probably due to visual search despite given 10s preview. In the C1 condition, people look at the target object more compared to other conditions. The overall fixation rate decreases when the complexity increases (also confirming that having 4th character distracts more compared to having background objects).

Although our investigations into this area are still ongoing, the results could be a useful aid for developing models for cross-modal NLP that aim to account for visual complexity. The most interesting outcome for NLP implementation is to exhibit the varying effect of different and irrelevant visual objects on language processing. This highlights the fact that while contributing visual information into language processing, the effect of different visual components should be treated differently with respect to not just their direct relevance but also possible interference even though they do not have a direct relation to the linguistic input.

In a further study, we aim to address how our context-integrated parser reacts to those visual manipulations and whether thematic role assignments are affected by irrelevant and varying visual components.

## Acknowledgments

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

## References

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

Baayen. R, Davidson. D J, and Bates. D M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412.

Bailey. K, and Ferreira. F. 2007. The processing of filled pause disfluencies in the visual world. *Eye movements: A window on mind and brain*, 485–500.

Knoeferle. P S. 2005. The role of visual scenes in spoken language comprehension: Evidence from eye-tracking The role of visual scenes in spoken language comprehension: Evidence from eye-tracking. Universitätsbibliothek.

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

Ferreira. F. 2003. The misinterpretation of noncanonical sentences. *Cognitive psychology*, 47(2):164–203.

Van Berkum. J J, Brown. C M., Zwitserlood. P., Kooijman. V., and Hagoort. P. 2005. Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Ferreira. F., Foucart. A., and Engelhardt. P E. 2013. Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, 69(3):165–182.

Matin. E., Shao. K C. and Boff. K R. 1993. Saccadic overhead: Information-processing time with and without saccades. *Perception & psychophysics*, 53(4):372–380.

Barr. D J. 2008. Analyzing 'visual world' eye-tracking data using multilevel logistic regression. *Journal of memory and language*, 59(4):457–474.

Jaeger. T F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models *Journal of memory and language*, 59(4):434–446.

Meng. M., and Bader. M. 2000. Mode of disambiguation and garden-path strength: An investigation of subject-object ambiguities in German. *Language and Speech*, 43(1):43–74.

Mirman. D. 2016. Growth curve analysis and visualization using R *CRC Press*.

Snedeker. J., and Trueswell. J. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context *Journal of Memory and language*, 48(1) 103–130.

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

Team. R C. 2013. R: A language and environment for statistical computing

Bates. D. 2005. Fitting linear mixed models in R. *R news*, 5(1): 27–30.

Hale. J. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8.

Levy. R. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

MacDonald. M C., and Seidenberg. M S. 2006. *Handbook of psycholinguistics*.

McRae. K., Hare. M., Ferretti. T., and Elman. J L. 2001. Activating verbs from typical agents, patients, instruments, and locations via event schemas. *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, 617–622.

Quené. H. and Van den Bergh. H. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4):413–425.