A Corpus-Based Model of Semantic Plausibility for German Bracketing Paradoxes

Corina Dima, Jianqiang Ma, Sebastian Bücking, Frauke Buscher,
Johanna Herdtfelder, Julia Lukassek, Anna Prysłopska, Erhard Hinrichs,
Daniël de Kok and Claudia Maienborn

SFB 833, Deutsches Seminar and Seminar für Sprachwissenschaft
University of Tübingen, Germany
{corina.dima,jianqiang.ma,sebastian.buecking,frauke.buscher,
johanna.herdtfelder,julia.lukassek,anna.pryslopska,
erhard.hinrichs,daniel.de-kok,claudia.maienborn}
@uni-tuebingen.de

Abstract

In this paper, we investigate German constructions composed of an adjective and a two-part nominal compound, such as *katholisches Kirchenoberhaupt* ('catholic church.leader'), focusing on two issues: (i) what are the prerequisites for semantically possible adjective-nominal compound constructions; (ii) which semantic factors determine the availability of *bracketing paradox* readings (i. e., the adjective modifies the first noun) in such constructions. We test theory-driven hypotheses using a corpus-based frequency model and evaluate the performance of the model with respect to human annotations.

1 Introduction

Nominal compounds which are modified by an adjective, as in Ex. (1), can have an *iconic* or an *anti-iconic* reading. Here, the phrase may refer to a *church leader*, who is *Catholic* (iconic reading), or to the *leader of the Catholic Church* (anti-iconic reading). The latter reading is discussed in the literature as a *bracketing paradox (BP)* because the semantic bracketing `[[adjective noun] noun]` does not iconically match the structural bracketing `[adjective [noun noun]]`, the latter of which is invariant in German. Formally explicit approaches to this phenomenon focus on systematically deriving their ambiguity ([2, 4]), but fail to satisfactorily account for the puzzling ungrammaticality of Ex. (2). While the iconic reading of this example, *four-storeyed owner of a house*, is semantically impossible, the anti-iconic reading, the *owner of a four-storeyed house*, should be fine. Surprisingly, however, this reading is considered impossible in German, according to native speaker judgments.

(1) Katholisches Kirchen*oberhaupt*
Catholic church.leader

(2) * vierstöckiger Haus*besitzer*
four-storeyed house.owner

We investigate such German constructions composed of an adjective (A) and a two-part nominal compound (N_1N_2), focusing on two issues: (i) what are the prerequisites for semantically possible A-N_1N_2 constructions; (ii) which semantic factors determine the availability of the iconic or anti-iconic reading. We analyze these constructions from a theoretical perspective (Section 2) and test the theoretical assumptions using a corpus-based frequency model (Sections 3 and 4). In a second step, the performance of the model is evaluated with respect to human annotations (Sections 5 and 6).

2 Modeling Bracketing Paradoxes

The perceived ungrammaticality of Ex. (2) suggests a prerequisite for any A-N_1N_2 construction, namely, that A-N_2 must be semantically possible (see as well [1]). We formulate this intuition as **Hypothesis 1 ($\mathbf{H_1}$)**:

$\mathbf{H_1}$: if an A-N_1N_2 is *semantically possible*, then A-N_2 is *semantically possible*.

This restriction does not make any assumptions regarding the distinction between the iconic or anti-iconic interpretation of A-N_1N_2 constructions. We hypothesize that this distinction is based on the *relative semantic plausibility* ([5],[8]) of the A-N_1 and A-N_2 constructions. We formulate this intuition as **Hypothesis 2 ($\mathbf{H_2}$)**:

$\mathbf{H_2}$: for examples where H_1 holds, the higher the *semantic plausibility* of A-N_1 *relative* to A-N_2 is, the more likely it is that A-N_1N_2 is a *bracketing paradox*.

The effects of H_2 are exemplified by Ex. (1), (3) and (4). Ex. (1) is a bracketing paradox because the semantic plausibility of the phrase *Catholic Church* is presumably greater than the one of the phrase *Catholic leader*. Ex. (3) and (4) are different: the phrase *Catholic table* in Ex. (3) is semantically impossible, and thus trivially less plausible then *Catholic prayer*. The semantic plausibilities of *Catholic company* and *Catholic leader* from Ex. (4) do not differ significantly. Therefore, Ex. (3) is not considered a bracketing paradox, whereas Ex. (4) is a borderline case which can be interpreted both iconically and anti-iconically.

(3) katholisches Tisch*gebet*
Catholic table.prayer

(4) katholisches Firmen*oberhaupt*
Catholic company.leader

3 Frequency-based Semantic Plausibility Model

We verify H_1 and H_2 using a frequency-based model derived from the 11.6 billion tokens `decow14ax` corpus [6]. The model considers the lemmatised form of the

words for computing the frequencies. For an A-N_1N_2 construction we compute the following:

- $freq_{A-N_1}$, the frequency (number of corpus occurrences) of A-N_1
- $freq_{A-N_2}$, the frequency of A-N_2
- $freq_{A-N_1N_2}$, the frequency of A-N_1N_2
- $rf_{A-N_1N_2} = \frac{freq_{A-N_1}}{freq_{A-N_2}}$, the relative frequency of A-N_1 and A-N_2.

We use the frequency of a construction in the corpus to model the notions of *semantic possibility* and *semantic plausibility* and to make judgments regarding the two hypotheses formulated in Section 2. If the frequency of a construction is higher than 0, the construction is considered *semantically possible*. We consider constructions that do not occur in the corpus to be *semantically impossible*.

The *semantic plausibility score* for an adjective-noun pair is given by the frequency count. The *relative semantic plausibility score* is the relative frequency of the two adjective-noun pairs in an A-N_1N_2 construction ($\frac{freq_{AN_1}}{freq_{AN_2}}$).

4 Testing the Hypotheses using the Frequency-based Semantic Plausibility Model

We test our two hypotheses using a dataset of 198 A-N_1N_2 constructions compiled based on the theoretical literature.

To test H_1, an A-N_1N_2 construction is considered *semantically possible* if its corpus frequency is greater than 0. The dataset contained 77 semantically possible A-N_1N_2 constructions (i. e., constructions that actually occurred in the corpus[1]). For 70 of these examples, our model predicted A-N_2 being also semantically possible, resulting in a 90.9% prediction accuracy for H_1. An interesting case is the phrase in Ex. (2), *vierstöckiger Hausbesitzer*: the full construction is considered semantically possible, because it has 10 occurrences in the corpus (as part of meta-discussions concerning its semantic impossibility). The same reasoning holds for the construction *verregnete Feriengefahr* 'rainy vacation.danger', which occurs twice in the corpus. The respective A-N_2 pairs of these constructions, however, do not occur, pointing to a logical discrepancy caused by the false initial assumption.

For H_2, the model computes a relative semantic plausibility score for each semantically possible A-N_1N_2 construction in our dataset. We identify this score with the BP-ness of an A-N_1N_2 construction. An initial inspection of the constructions with a high relative semantic plausibility score shows that these are indeed

[1]The discrepancy between the initial and the attested number of constructions can be explained as follows: on the one hand, many of the examples were ungrammatical contrasts to grammatical examples; other examples were constructed by analogy with existing examples, which of course does not imply their actual occurrence in the corpus.

constructions with an anti-iconic reading: *katholisches Kirchenoberhaupt* has a score of 1328, *europäischer Auslandsaufenthalt* ('European foreign-country.stay') has a score of 7327. In contrast, constructions with a very low score, like *verrückter Chemieprofessor* ('crazy chemistry.professor', score 0.003) or *ambulante Unfallbehandlung* ('ambulant accident.treatment', score 0.0003), clearly have no anti-iconic reading. Borderline cases include *bedrohliches Krankheitssymptom* ('menacing desease.symptom', score 4.86) and *politische Satiresendung* ('political satire.broadcast', score 1.81).

The model confirmed H_1, and thereby provided good evidence for the assumption that for all A-N_1N_2, A-N_2 must be semantically possible, irrespective of whether A can modify N_1 or not. The initial observations also suggest that the frequency model can be successfully used to test H_2. To test H_2, we annotated the set of 77 A-N_1N_2 constructions with regard to their perceived iconicity. The annotation is described in the next section.

5 Annotation

5.1 Annotation guidelines

The dataset contained those 77 A-N_1N_2 constructions that are semantically possible according to the frequency model. These items were annotated by 5 PhD students in linguistics (3 women, 2 men; native speakers of German); two of them are co-authors of this paper. The items were presented to the annotators in an Excel-spreadsheet in one of 5 randomized orders. The annotators worked independently from each other.

The annotators annotated each item according to the following two questions[2]:

Q1 Is the A-N_1N_2 construction as a whole grammatical? yes/no

Q2 Which reading is preferred? anti-iconic, iconic, equal preference

5.2 Results & discussion

Grammaticality (Q1) We categorized the 77 items according to the annotation question Q1. As a prerequisite, at least 4 (out of 5) annotators had to agree upon an answer. If there was no corresponding agreement, the item was not categorized. 73 out of 77 items are judged as grammatical. 2 items are judged as ungrammatical, and 2 items could not be categorized; these 4 examples were excluded from the evaluation of Q2. The inter-rater agreement had a Fleiss' κ value [3] of 0.45 (moderate agreement). Notably, the 2 items that were judged as ungrammatical are *vierstöckiger Hausbesitzer* and *verregnete Feriengefahr*. These are exactly those cases that are falsely classified as semantically possible by the frequency-based

[2]We also asked the annotators whether A-N_1 and A-N_2 are semantically possible. As the answers do not directly bear on H2, we will not report them here.

model as they occur in metadiscussions on their semantic impossibility (see Section 4).

Preference (Q2) We categorized the remaining 73 grammatical items according to the annotation question Q2. As a prerequisite, at least 4 (out of 5) annotators had to agree upon an answer (if all 5 annotators judged the item as grammatical), or, at least 3 (out of 4) annotators had to agree upon an answer (if only 4 annotators judged the item as grammatical). 11 items could not be categorized, as the required majority was not obtained. From the remaining 62 items, 16 were perceived to have an anti-iconic reading, while the other 46 were perceived to have an iconic reading. The inter-rater agreement had a Fleiss' κ value of 0.58 (moderate agreement). The results yield a two-way distinction between anti-iconic and iconic readings. Notably, no items were annotated as being truly ambiguous. However, for 11 items, the annotators did not agree upon a preferred reading. Among these, several examples are prototypes for bracketing paradoxes according to the theoretical literature; in fact, 4 have a tendency for being ambiguous (e. g., *politische Satiresendung*) and 2 have a tendency for being anti-iconic (e. g., *katholisches Kirchenoberhaupt*). As this result is in need of further clarification, we excluded the problematic data points from the evaluation of the frequency-based model. The remaining 62 items annotated for iconicity will be further used to train and test the frequency-based model.

6 Results of the Frequency-based Semantic Plausibility Model

This section presents the results of using the frequency-based semantic plausibility model introduced in Section 3 to predict the iconicity of A-N_1N_2 constructions. The 62 annotations which were considered to be grammatically correct (Q1) and were assigned the same preferred reading by the majority of the human annotators (Q2) are used as a dataset.

The task at hand is to predict if a particular A-N_1N_2 construction is considered a bracketing paradox or not using only the frequency information, in particular the relative semantic plausibility score which we normalize across the examples in our dataset.

We use logistic regression, a widely-used linear machine learning method, to train a prediction model. Table 1 reports the average F1 score [7] and accuracy figures obtained for 10-fold cross-validation. The model hyper-parameter (the regularization coefficient) is chosen individually for each fold, using a grid search over 10 equally spaced values in the interval $[1e-4,\ 1e+4]$. The results show that despite the imbalanced number of instances for each class, the relative semantic plausibility score is a very good predictor for the preferred interpretation of a particular construction.

Data set	F1 score	Accuracy (%)
16 BP, 46 non-BP (62 total)	0.90	95.71

Table 1: Average F1 score and accuracy at predicting if an adjective-compound construction is a bracketing paradox or not. Results for 10-fold cross-validation.

7 Conclusion

The corpus-based frequency model confirmed both H_1 and H_2. First, for all A-N_1N_2 constructions, A-N_2 must be semantically possible, irrespective of whether A can modify N_1 or not (H_1). Second, the higher the relative semantic plausibility score of A-N_1N_2 constructions, the more likely it is that the construction is a bracketing paradox (H_2). This result is based on our evaluation of the frequency-based model using human annotation as a gold standard.

Our study also pointed to issues that need to be further investigated in future work. From a theoretical perspective, it is surprising that *katholisches Kirchenoberhaupt*, the prototypical textbook example for bracketing paradoxes, received mixed ratings from the annotators. This suggests that the distinction might not necessarily be a binary one. We plan to complement our results via a rating study that elicits graded judgments for the perceived iconicity of A-N_1N_2 constructions. From the perspective of the computational modeling, we discovered some limitations of the frequency-based model. For example, in the construction *intelligenter Tierarzt* 'intelligent animal.doctor', the pair *intelligenter Arzt* is very infrequent, as the adjective 'intelligent' spells out an implied attribute of the 'doctor', whereas the locution *intelligentes Tier* is much more frequent. Thus, the model predicts that this construction should have an anti-iconic interpretation ('doctor for intelligent animals'), which is clearly wrong. Another shortcoming relates to the inability of the model to make predictions in absence of the frequency information, which resulted in analyzing only a part of the initial dataset. In order to circumvent these shortcomings, we plan to use distributional semantics models, which have the ability to integrate information about the semantics of the construction's constituents.

Acknowledgments

Financial support for the research reported in this paper was provided by the German Research Foundation (DFG) as part of the Collaborative Research Center "The Construction of Meaning" (SFB 833), projects A1 and A3. We thank the anonymous reviewers for their comments and our annotators for their support.

References

[1] Rolf Bergmann. Verregnete Feriengefahr und Deutsche Sprachwissenschaft. Zum Verhältnis von Substantivkompositum und Adjektivattribut. *Sprachwis-*

senschaft, 5(3):234–265, 1980.

[2] Markus Egg. *Flexible semantics for reinterpretation phenomena*. CSLI Publications Stanford, 2005.

[3] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[4] Richard K. Larson. Events and modification in nominals. In *Semantics and Linguistic Theory*, volume 8, pages 145–168, 1998.

[5] Angeliki Lazaridou, Eva Maria Vecchi, and Marco Baroni. Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, USA, 2013.

[6] Roland Schäfer. Processing and querying large web corpora with the COW14 architecture. *Challenges in the Management of Large Corpora (CMLC-3)*, page 28, 2015.

[7] C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 2nd edition, 1979.

[8] Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. (Linear) maps of the impossible: capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics, 2011.