Ad Hoc Compounds for Stance Detection

Qi Yu[1,2], **Fabian Schlotterbeck**[3], **Hening Wang**[3], **Naomi Reichmann**[1], **Britta Stolterfoht**[3], **Regine Eckardt**[1,2], **Miriam Butt**[1,2]
[1]Department of Linguistics, University of Konstanz
[2]Cluster of Excellence "The Politics of Inequality", University of Konstanz
[3]Department of Modern Languages, University of Tübingen
firstname.lastname@{uni-konstanz, uni-tuebingen}.de

Abstract

In this paper we focus on a subclass of multi-word expressions, namely compound formation in German. The automatic detection of compounds is a known problem and we argue that its resolution should be given more urgency in light of a new role we uncovered with respect to ad hoc compound formation: the systematic expression of attitudinal meaning and its potential importance for the down-stream NLP task of stance detection. We demonstrate that ad hoc compounds in German indeed systematically express attitudinal meaning by adducing corpus linguistic and psycholinguistic experimental data. However, an investigation of state-of-the-art dependency parsers and Universal Dependency treebanks shows that German compounds are parsed and annotated very unevenly, so that currently one cannot reliably identify or access ad hoc compounds with attitudinal meaning in texts. Moreover, we report initial experiments with large language models underlining the challenges in capturing attitudinal meanings conveyed by ad hoc compounds. We consequently suggest a systematized way of annotating (and thereby also parsing) ad hoc compounds that is based on positive experiences from within the multilingual ParGram grammar development effort.

Keywords: ad hoc compounds, attitudinal meaning, stance detection, German, universal dependencies, psycholinguistic validation, large language models

1. Introduction

The automatic detection of compounds is known to be a difficult problem for Natural Language Processing (NLP) (Constant et al., 2017; Baldwin and Kim, 2010), particularly in a language like German which uses compounding as a central strategy for novel word formation. In this paper we present research showing that novel, ad hoc compound formations in German can be used strategically to convey attitudinal meaning, thus making them an interesting area of research from the overall perspective of stance detection (Mohammad et al., 2016; Schiller et al., 2021) and adding urgency to finding reliable ways of automatically detecting compounds, and in particular, novel compound formations in a language. We adduce evidence that combines insights from theoretical linguistic analysis, corpus linguistic investigations and psycholinguistic experimentation to show that a subset of ad hoc compounds in German, termed *enigmatic compounds* (ECs; Wildgen, 1981) are indeed systematically used to convey attitudinal meaning and are therefore of inherent interest for the NLP task of stance detection.

The types of compounds falling under the rubric of ECs are illustrated in (1)–(3). We noted the use of such compounds for the expression of stance as part of a larger project investigating the framing of politically charged issues across several German newspapers. We have marked the extra expressive meaning carried by these ad hoc compound formations as attitudinal meaning (AM) in the examples.

(1) Flüchtlinge wollen Österreich meiden und
refugees want Austria avoid and
lieber in Merkel-Land einreisen.
rather in Merkel Country travel.into
'Refugees want to avoid Austria and instead enter Merkel-Country.'
AM: The German refugee crisis is Merkel's fault.
(SOURCE: Facebook)

(2) Jede 5. China-Maske ist unbrauchbar
every fifth China mask is unusable
'Every fifth China-mask is unusable'
AM: China is notorious for low-quality products.
(SOURCE: BILD, 2020-05-03)

(3) Neue Stelle für Kopftuch-Praktikantin
new position for hijab intern
'New position for hijab-intern'
AM: Religious practices of Muslims often cause trouble for others.
(SOURCE: BILD, 2016-08-25)

Intended but deliberately masked meanings of speakers such as the AMs above are known to play a crucial role in political communication (Beaver and Stanley, 2018). Our data indicate that ECs are a useful rhetorical device for speakers/authors to implicitly convey attitudinal meaning. In particular, we observed that ECs can be employed as so-called *dog-whistles* (Henderson and McCready, 2019), whereby their use – at least for a certain time span – speaks to a certain subgroup and con-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@*LREC-COLING-2024*, pages 231–242. May 25, 2024.

veys a meaning that is on the surface rather vague, but decodable as to its hidden meaning by that subgroup. This seems particularly interesting, as ad hoc compounds are instances of innovated language and thus, dog whistles and pejorative uses in expressing attitudinal meaning clearly cannot rest on conventional lexical meanings alone. This makes an automatized stance detection task challenging yet interesting.

We consequently examine how compounds are currently treated in available dependency parsers and Universal Dependencies (UD) treebanks (de Marneffe et al., 2021; Nivre et al., 2016) for German. We find that the current treatment is uneven. We also explored the potentially greater capabilities of current large language models (LLMs) with respect to detecting attitudinal meaning in ECs, but found that while the results from LLMs may provide an explanation for substantial variation in our experimental data, they do not easily capture the effect of our experimental manipulation involving ECs. We therefore provide suggestions for a systematic UD annotation for compounds that is based on the multilingual ParGram grammar development experience (Butt et al., 1999; Sulger et al., 2013) so as to allow for a more successful learning.

This paper is structured as follows: in section 2 we provide some background on the current state-of-the-art. We follow this in section 3.1 with the results of a corpus study of three different newspapers, which yielded indications that more conservative leaning newspapers used ad hoc compounds to trigger attitudinal meaning more than other newspapers. However, our results are most robust for the conservative tabloid *BILD*, which is also known for an editorial policy that prefers the use of pictures coupled with short, expressive texts. The greater use of ad hoc compounds could also therefore just be a matter of newspaper writing style. To test the perception of attitudinal meaning in compounds, we therefore designed and executed an experiment that sought to establish the stance triggering effect of ECs using psycholinguistic methods. This is described in section 3.3, following a discussion of how the semantics of ECs are hypothesized to come about in section 3.2. In section 4, we report on our attempts to use current LLMs to simulate our experimental results. We did not find any indication that these models can capture the central contrasts observed in the experimental outcomes. In section 5, we combine the insights from the corpus and psycholinguistic results to formulate recommendations for the systematic annotation of compounds in corpora. Section 6 concludes the paper.

2. Background

2.1. Evaluative Language

Evaluative language is of interest for a range of NLP tasks, perhaps currently most prominent among the sentiment analysis (Pang and Lee, 2008; Taboada et al., 2011), but also hate speech detection (Davidson et al., 2017) and stance detection (Mohammad et al., 2016; Schiller et al., 2021). Sentiment analysis and stance detection are closely related tasks but differ in their overall goals. Sentiment analysis is concerned with identifying whether a given text, sentence or passage overall can be classified as being positive, negative or neutral. This has generally involved a bag-of-words approach, where the internal structure of the text is not considered and the target has generally been reviews or statements about movies, books, objects or persons. More recently, approaches to sentiment analysis have become more nuanced in that the classification aims at *aspect based* (what aspect is the sentiment targeted at, e.g., the acting or the plot?) or *target based* (what is the precise target of the sentiment, e.g. an iPhone or the ear phones that came with the iPhone?) sentiment analysis (Alturayeif et al., 2023).

Stance detection is informed by the Stance Triangle defined by Du Bois (2007), by which the author of a text is taken to want to influence or align the recipient/reader of the text with his/her beliefs. The difference between sentiment analysis and stance detection is that in sentiment analysis the object of study are texts expressing a given sentiment, prototypically reviews. In these the author articulates their opinion to an audience, but is not necessarily seeking to align the audience with their own views. Given that our overall interest lies in determining how issues are framed (Chong and Druckman, 2007), we are interested in stance detection as a subtask for determining the overall framing of a narrative or text. As far as we have been able to determine, no previous work on stance detection has attempted to include information from compounds in a focused manner, though Li and Caragea (2019) note as part of their stance detection error analysis that it would be useful to separate the individual components used in hashtags such as as *#VoteGOP* or *#NoHilary*, as found in the SemEval-2016 dataset developed specifically for stance detection (Mohammad et al., 2016).

Stance detection includes identifying instances of subjective language (Wiebe et al., 2004). Subjective language can be detected on the basis of linguistically informed lexicon and/or construction based information (Biber and Finegan, 1989; Biber and Conrad, 2019; Taboada et al., 2011), or it can be detected by machine learning on the basis of annotated data (Alturayeif et al., 2023). Our data

is German, for which an automatic annotation tool for subjective language already exists (El-Assady et al., 2016, 2019). This tool provides POS-tagging and syntactic parsing of a given text along with a systematic identification of linguistic cues for subjective language such as the annotation of various modals or German discourse particles (Zimmermann, 2011). However, the tool does not include a facility for the automatic detection of ECs.

2.2. Annotation and Automatized Detection of Compounds

The compounds in (1)–(3) each contain a hyphen. However, German compounds generally do not contain a hyphen. One could hypothesize that ad hoc compounds in particular are marked with a hyphen, but our data also contains instances of ad hoc formations such as *Asylprügler* 'asylum beater' and *Migrantenschreck* 'migrant scare' that have been written without a hyphen. Nevertheless, the inclusion of a hyphen provides a potentially important clue for the automatic identification of at least a subset of compounds and one that could be picked up on easily. In surveying existing dependency parsers and treebanks annotated according to the Univeral Dependencies (UD) scheme (de Marneffe and Manning, 2008; de Marneffe et al., 2021; Nivre et al., 2016), we found that only the Stanza toolkit (Qi et al., 2020) could reliably identify German compounds characterized by a hyphen. The sample of other dependency parsers for German that we tried were not reliable in the identification of compounds, with most merely labeling them with the POS-tag of NN for common nouns, as shown in Figure 1 for the Mate parser (Björkelund et al., 2010),[1] where both of the compounds *Flüchtlingsorganisation* 'refugee organisation' and *Asyl-Verschärfungen* 'asylum restrictions' are tagged as NN. The same is true for spaCy,[2] ParZu (Sennrich et al., 2009)[3] and a German dependency parser[4] based on the MaltParser framework[5], as well as the very high quality morphological analyzer SMOR (Schmid et al., 2004). An investigation of UD treebanks for German collected at the INESS website[6] yielded much the same result. See also the reports and conclusions in Baldwin et al. (2023).

A morphological analyzer can be integrated as part of a dependency parser and so we set out to test SMOR for our purposes. We worked with this system because it has been designed especially to deal with the productive word formation possibilities in German, including ad hoc compounds occurring in newspaper texts. However, a pilot study with respect to our data showed that while SMOR is indeed able to identify ad hoc compounds successfully, the uneven nature of the overall results means that quite a bit of manual postprocessing would be required to obtain a useable data set. For example, the ad hoc compound *Pegida-Anhänger* 'Pegida follower/supporter' could not be analyzed at all, while the lexically established word *Bezirksamt* 'district office' was incorrectly analyzed. Instead of the correct split into the morphemes *Bezirk+s+amt* (the *s* is a so-called linking element that appears for phonological reasons), the word was split into *Bezirk+Samt* 'district velvet' as one of the three most likely results.

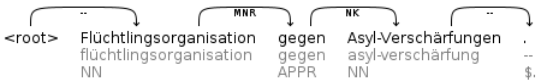

Figure 1: Sample Mate parse

Thus, the challenges posed by automatic compound detection (Constant et al., 2017; Baldwin and Kim, 2010) continue to be a problem, and one that – we argue – gains more urgency given our findings. Given that ECs express attitudinal meaning and as such can provide an important linguistic cue for stance detection, search for these cues should be operationalized.

3. Enigmatic Compounds

In this section, we combine results from a corpus linguistic study and a psycholinguistic experiment to show that ECs can be used systematically to express attitudinal meaning. We first present results from a corpus study that demonstrates a systematic use of ECs to express a negative stance in newspapers (section 3.1). We then discuss how ad hoc compounds invite such attitudinal meaning from a theoretical linguistic aspect (section 3.2), and report a psycholinguistic experiment (section 3.3) to confirm that ECs are indeed a systematic part of language use. All data and code resulting from this work are publicly available at: https://github.com/qi-yu/enigmatic-compounds.

3.1. Corpus Study

Our corpus study was conducted as part of a larger investigation into the framing of the Syrian refugee crisis by German newspapers in the time span of

[1] https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools/
[2] https://spacy.io/
[3] https://github.com/rsennrich/ParZu
[4] https://pub.cl.uzh.ch/users/siclemat/lehre/ecl1/ud-de-hunpos-maltparser/html/
[5] http://www.maltparser.org/
[6] https://clarino.uib.no/iness-prod/treebanks#

2014–2018. We chose the three German newspapers with the highest circulation rates (IVW, 2023): *BILD*, *Frankfurter Allgemeine Zeitung* (FAZ) and *Süddeutsche Zeitung* (SZ). These three newspapers cover a representative range of political leanings within the German media landscape, with *BILD* being the most conservative on the political spectrum, the SZ the most left leaning, and the FAZ also leaning towards the conservative end. Moreover, they also build a diverse sample of different styles, with BILD characterized as a tabloid newspaper whereas FAZ and SZ contain high quality, in-depth reporting. Examples (4)–(6) illustrate the different styles: they are headlines from articles reporting on the same event and published around the same time.

(4) **BILD**, 2014-09-29:
Folter-Skandal in deutschen Asylbewerberheimen: "Die Wachleute schlagen und treten uns"
'Torture-scandal in German asylum seekers' accommodations: "The guards beat and kick us"'

(5) **FAZ**, 2014-09-30:
Misshandlung von Asylbewerbern: Sicherheitsleute werden überprüft
'Mistreatment of asylum seekers: security guards undergo checks'

(6) **SZ**, 2014-09-30:
Ermittlungen nach Misshandlungsverdacht in drei Flüchtlingsheimen
'Investigations into suspected mistreatment in three refugee accommodations'

As part of this investigation, we noticed that compounds seemed to be used to express a negative stance towards refugees and the handling of the crisis by the government (see, e.g., *Folter-Skandal* 'torture-scandal' in (4)). A more in-depth investigation of this phenomenon was hampered by the difficulty of automatically detecting compounds. We therefore decided to experiment with training a language model on the basis of annotated data. The best performing model was a logistic regression model that resulted in a value of 0.68 for F1.

Given these unsatisfactory results, we asked ourselves whether it was indeed necessary to detect these compounds. As we report on in the following sections, the result of our investigations has established that ECs indeed have the potential for providing important information for stance detection. Efforts should be redoubled so as to be able to operationalize ECs for stance detection.

Our data set consisted of a total of 23,889 articles. Given the necessity for manual annotation of the compounds (since automatic detection is a challenge), we considered only the articles' headlines for our study. We manually identified 19,353 referential/neutral ad hoc compounds and 828 ECs in these headlines. We structured our resulting data set into pieces of information as follows: the target compound, the sentence in which it appeared, the year it was released, the newspaper source, and the annotation (`0` = referential, `1` = enigmatic). We categorized the compounds as enigmatic if they met the following two criteria:

(i) the compound carries an attitudinal meaning;

(ii) the compound is an innovative, ad hoc formation and is thus not established in a recognized dictionary or lexicon of German.

To validate the application of criterion (ii), the German dictionary *Duden*[7] as well as the online dictionary *Digitales Wörterbuch der deutschen Sprache*[8] were consulted. For instance, based on these criteria, the compound *Karajan-Schüler* 'Karajan student' was defined as referential (neutral), as it does not seem to express an additional evaluative meaning; only its literal meaning is being transmitted. In contrast, the compound *Flüchtlings-Tsunami* 'refugee tsunami' was categorized as enigmatic, as it does not only refer to a large amount of refugees, but it also carries an additional AM to the effect that refugees are overwhelming the transit and host countries.

Our overall results of the annotation per newspaper are given in Table 1. They show that *BILD* uses by far the most ECs. We furthermore sampled the top most ECs per newspaper per year and found that *BILD* predominantly used these in contexts of discussing security or issues of criminality, whereas the FAZ and the SZ placed a greater emphasis on problems of capacity and the rights of individual refugees. For example, with compounds such as *Asylprügler* 'asylum beater', *Migrantenschreck* 'migrant scare', and *Amok-Afrikaner* 'amok African', *BILD* focuses on criminality related to the refugees in Germany through the use of ECs. This is in line with the hostile reporting style previously observed for tabloid newspapers (see Innes, 2010; Kleinsteuber and Thomass, 2007).

Newspaper	#Enigmatic	#Neutral
BILD	726	10,059
FAZ	58	5,525
SZ	44	3769

Table 1: Total number of enigmatic and neutral compounds in newspaper headlines.

Whether or not the ECs are employed as attention-getters as part of *BILD*'s sensationalist

[7] https://www.duden.de
[8] https://www.dwds.de

writing style (see Greussing and Boomgaarden, 2017) becomes irrelevant in the face of their extensive use by *BILD* in combination with the negative attitudinal meanings triggered by these ECs: they are a significant contributing factor to the overall articulated stance towards a topic.

3.2. Compound Meaning

Compounds have a range of interpretational possibilities because their meanings are not compositional. Earlier theoretical linguistic studies on compound meaning share the common assumption that there is some covert, meaning-decisive *semantic relation* $\mathcal{R}$ between the constituents of a compound:

(7) Let C_1C_2 be a compound where $[\![C_1]\!] = m_1$ and $[\![C_2]\!] = m_2$.
Then: $[\![C_1C_2]\!] = \mathcal{R}(m_1, m_2)$

Levi (1978) and Fanselow (1981) propose taxonomies of semantic relations that play a role in *ad hoc* compound interpretation, and Meyer (1993), Ryder (1994) and Benczes (2009) propose different assumptions on how the semantic relations in (7) are derived. In the simplest case, *ad hoc* compounds serve as abbreviations for phrases, as in *Karajan-Schüler* 'Karajan Student' which is equivalent to *Schüler von Karajan* 'student of Karajan'. In (1)-(3); however, there is clearly an attitudinal meaning, an extra meaning dimension that is not found in the equivalent non-compound phrase. Consider, for example, *China-Maske* 'China mask' in the context in (2): it has a negative attitudinal meaning that is not conveyed by the compositional alternative phrase *chinesische Maske* 'Chinese mask'.

Sassoon (2011) opens an avenue towards an explanation of attitudinal enrichment in ECs. The author summarizes comparative studies in the conceptual structure of nouns and adjectives: nouns denote similarity-based concepts with a prototype structure (Murphy, 2002), whereas adjectives denote rule-based properties (Kennedy, 1999). The distinction is backed up by converging evidence from neurolinguistics, patholinguistics and language acquisition. Sassoon's proposal predicts that the modifier in ECs (*China-* in *China-Maske*) contributes to a similarity-based concept. This happens, plausibly, by adding a further dimension in which exemplars must match the prototype. Specifically, similarity-based categorization rests on prototypical values that can be attributed to this dimension. In our example the similarity-based categorization invites a comparison to typical 'products from China', which provides a hook for the accommodation of an interpretation including negative expectations about products from China. The corresponding adjective in a phrasal alternative ('Chinese mask'), in contrast, adds a simple categorical property 'be Chinese' (yes/no). Sassoon thus predicts that the processing of modifiers does not trigger novel stereotypes and should not provide an entry-point for attitudinal meaning.

We were interested in this prediction as it also provides a systematic way of testing whether the attitudinal meaning associated with ECs we found as part of the corpus study in section 3.1 is a general, systematic part of language use or whether it is perhaps attributable to the particular corpus. If the attitudinal meaning associated with ECs is found to be a systematic part of language, it provides another argument for taking ECs seriously as part of the overall task of stance detection. We describe the psycholinguistic experiment we set up to test Sassoon's prediction in section 3.3.

3.3. Experiment

3.3.1. Methods

Materials and Design We manually selected 21 text snippets from newspapers and social media which contain ECs along the lines of (1)–(3) that trigger negative AM according to our own intuitions. We restricted ourselves to negative AMs in our experiment as these were more prevalent in the corpus study. To test for the AM-triggering effects of ECs, three variants were created from each snippet. Table 2 provides examples of such snippets (translated into English). The three variants were:

(i) COMPOUND: original text snippet with the EC.

(ii) PHRASAL: EC substituted by a corresponding phrasal construction.

(iii) NEUTRAL: EC substituted by a corresponding noun that is attitudinally neutral.

The PHRASAL condition controls for truth-conditional information, as it conveys the same truth-conditional information as the COMPOUND condition but in a pragmatically unmarked phrasal expression, not an ad hoc compound. The condition NEUTRAL is intended as a baseline: though there is no stylistic difference in terms of innovative language use between PHRASAL and NEUTRAL, these two conditions differ in their information load, as the modifier part of the PHRASAL (and COMPOUND) condition provides extra information that is not necessary for reference resolution but can be inferred from the prejacent context (see Table 2). Comparing the PHRASAL and the NEUTRAL condition thus allows us to examine whether the addressees' perception of the attitudinal strength is affected by such additional but in principle unnecessary information while keeping the style constant. With these three conditions, we test the following two hypotheses:

(i) COMPOUND VS. PHRASAL (different style, same information load): compounding amplifies the perceived attitudinal strength;

(ii) PHRASAL VS. NEUTRAL (same style, different information load): the additional information that is not necessary for reference resolution amplifies the perceived attitudinal strength.

The items were distributed over 3 lists using a Latin square. 24 stylistically similar text snippets were added to each list as fillers. For each item, participants rated its attitudinal strength by answering question (8) on a 7-point Likert-scale.

(8) *How does the author talk about _____?*
1=positive ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 ◯7 *7=negative*

As our overall interest is in the framing of politically charged discourse, we also collected the political leaning of each participant by asking question (9) at the end of the experiment. This allows us to further control whether participants' perception of attitudinal strength is affected by their political leaning:

(9) *In politics, people often use "left" and "right" to denote political leanings. Where would you place your own political leaning?*
1=left ◯1 ◯2 ◯3 ◯4 ◯5 ◯6 ◯7 *7=right*

Participants The participant recruitment and data collection was carried out online via Prolific.[9] 212 German native speakers, identified through Prolific's demographic prescreening function, took part in the study (103 female, 102 male, 7 other genders; mean age = 26.52 years, $SD = 8.10$ years). The experiment was carried out anonymously and voluntarily. Each participant received a compensation of £8.50 per hour, a fair rate suggested by Prolific.

3.3.2. Results

Figure 2 shows the rating distributions of each condition. Overall, all items were rated rather negatively, with more negative ratings for COMPOUND than PHRASAL and PHRASAL than NEUTRAL conditions. We fitted a *cumulative link model* (CLM) with random effects using the R package *ordinal* (Christensen, 2018) to test these differences statistically. CLM is a variant of logistic regression generalized to multinomial ordinal dependent variables. A CLM models the probability, $P(Y \leq j)$, that an ordinal response variable Y is less than or equal to a specific category $j \in \{1, \ldots, J\}$ $(J \geq 2)$ according to the equation below, where θ_j is the intercept of level j, $\mathbf{x}$ is a vector of predictors, and $\boldsymbol{\beta_j}$ is a vector of coefficients:

$$\text{logit}(P(Y \leq j)) = \log \frac{P(Y \leq j)}{P(Y > j)} = \theta_j - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}$$

In our initial model, we predicted participants' *ratings* using *condition* and participants' *political leaning* as well as their interactions. For the predictor *condition*, PHRASAL is set as reference level (cf. hypotheses above). For the predictor *political leaning*, we mapped the seven original levels (see (9) above) to three aggregated levels in order to ease the model interpretation: 1-3 = LEFT, 4 = NEUTRAL, 5-7 = RIGHT. We used dummy encoding to code the three levels. Random intercepts and random slopes were fitted for *items* and *participants*, as likelihood ratio tests showed that they improved the model fit. In a following model selection step based on likelihood ratio tests, the predictor *political leaning* and the interaction term were removed as they were not significant in improving the model fit (likelihood ratio test without interaction: $\chi^2(2) = 0.384$, $p = 0.826$; likelihood ratio test with interaction: $\chi^2(6) = 2.004$, $p = 0.919$).

Our final model showed a significant difference between COMPOUND and the reference level PHRASAL. Compared to PHRASAL, COMPOUND led to a significant decrease in the logit of ratings in lower (i.e., more positive) categories (COMPOUND VS. PHRASAL: $\beta = 0.526, SE = 0.152, p < 0.001$). No significant difference between NEUTRAL and PHRASAL was found (NEUTRAL VS. PHRASAL: $\beta = -0.272, SE = 0.176, p = 0.123$).

3.3.3. Discussion

The result of our experiment with a large population is in line with the corpus study. The significant decrease of the likelihood of positive ratings indicates that the authors' negative attitudes are perceived as more pronounced when ECs are used instead of the PHRASAL counterpart. The difference in information load between PHRASAL and NEUTRAL condition did not show significant influence on the participants' perception of attitudinal strength. Furthermore, the non-significant effect of *political leaning* as well as the non-significant interaction between *political leaning* and *condition* show that the increased perception of attitudinal meaning in ECs is general part of how language works, rather than being domain or population specific.

4. Simulations with Large Language Models (LLMs)

Recent advances of LLMs have underscored their remarkable utility across a wide variety of NLP

[9] https://www.prolific.co

COMPOUND	PHRASAL	NEUTRAL
The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **China-masks** are unusable for medical purposes.	The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **Chinese masks** are unusable for medical purposes.	The federal government purchased more than 108 million masks from China for German clinics and medical practices. However, about 10 percent of these **masks** are unusable for medical purposes.
The big **refugee-mistake**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.	The big **mistake about refugees**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.	The big **mistake**: no labor market miracle has been brought by refugees. Unfortunately, most of the newcomers were not Syrian doctors and engineers.

Table 2: Example stimuli (translated into English from German). The variation between different conditions are marked in bold.

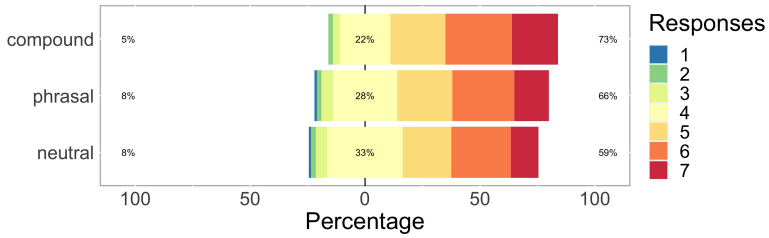

Figure 2: Distribution of participants' ratings by condition.

tasks (e.g., Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023). However, the challenges associated with compound detection, particularly in identifying the associated attitudinal meanings of certain types of compounds like ECs remain significant. An avenue worth exploring is whether current LLMs encounter comparable challenges in this domain, particularly within the context of our psycholinguistic experiment. Recent work similar in spirit focused on human-likeness of LLMs' linguistic performance, e.g., testing language models on different syntactic phenomena,(Wilcox et al., 2018, 2020; Futrell et al., 2019; Arehalli et al., 2022) semantic judgements (e.g., Levy et al., 2017; Kauf et al., 2023), and on subtle pragmatic phenomena like irony or compliance with Gricean maxims (Hu et al., 2023; Tsvilodub et al., 2023).

We conducted experiments testing two of the latest versions of ChatGPT, namely GPT-4 and GPT-3.5-turbo (Achiam et al., 2023), employing various temperature settings. We designed a prompt that closely simulates the task employed in the experiment, and fed experimental items from the previous psycholinguistic experiment with human participants to these LLMs. Among these configurations, the one utilizing GPT-4 with a temperature set to 0 yielded the best results. Overall, we found that the best LLM captured a significant portion of the observed by-item variance in our experimental results ($R^2 = .48$, $p < .001$; see Fig. 3). Contrary to our experimental results, however, at the condition level, there was no indication of any alignment with human data ($R^2 = .43$, $p = .55$).

Our current LLM simulations thus provide initial evidence that these models currently have difficulty picking up cues for AMs conveyed by ECs. While further analyses (e.g., of the involved contextual embeddings or attention patterns) or future LLMs may provide a closer match between human ratings and modeling results, the current lack of effect was observed concurrently with the models' ability of capture substantial variation in other dimensions of our experimental results. This further highlights the specific subtleties and challenges involved in the detection and interpretation of ECs.

5. Recommendations and Outlook

We have now established that ECs systematically convey attitudinal meaning which can provide information for the NLP task of stance detection. We have also established that the current state of the art with respect to dependency parsers and UD treebank representations does not facilitate the automatic detection and identification of ECs. We furthermore showed that LLMs also struggle with the identification of EC contributions that are natural for humans, despite their otherwise impressive capabilities.

In this section, we propose that the UD community adopt a uniform approach towards the annotation of compounds. A systematic and uniform approach towards annotation will be able to result

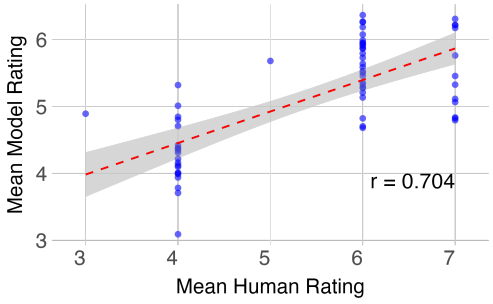

Figure 3: By-item correlation between participants' ratings from our experiment and LLM simulations.

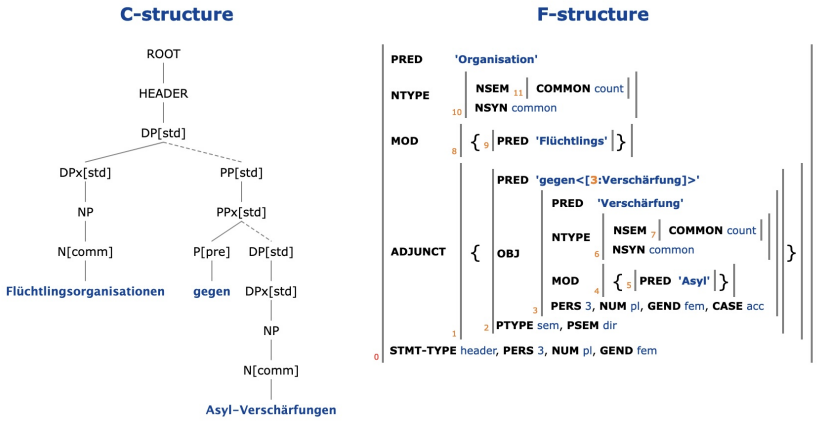

Figure 4: LFG analysis of ad hoc compounds.

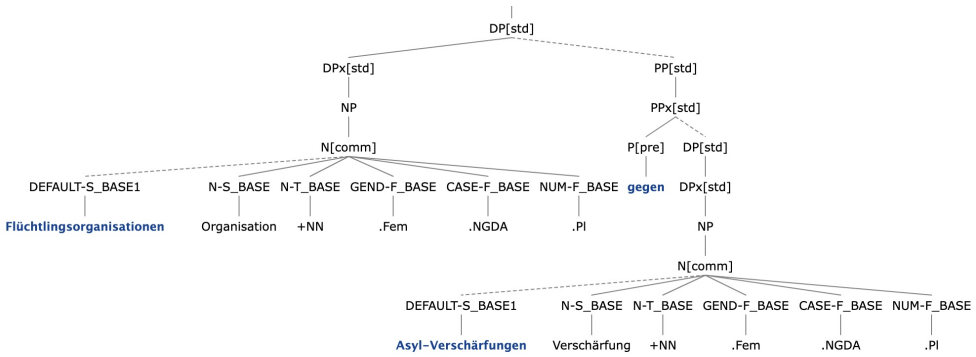

Figure 5: LFG analysis of ad hoc compounds with morphological analysis by DMOR.

in better down-stream machine learning and thus better results with respect to dependency parsers. Concretely we recommend adopting the approach deployed within the multilingual ParGram grammar development effort (Butt et al., 1999; Sulger et al., 2013). This is illustrated in Figures 4 and 5 from the German ParGram grammar (Dipper, 2003). The grammar is hosted on the INESS XLE website and can be used interactively.[10] The German ParGram grammar is based on Lexical Functional Grammar (LFG; Dalrymple, 2001), which has a context-free phrase structure part (the *c-structure*) and a dependency part (the *f-structure*). A c-structure of the compounds in question are simply tagged as common nouns (`N[comm]`). However, as shown in Figure 5, the German grammar also contains a finite-state morphological analyzer (DMOR, a precursor of SMOR; Schiller, 1994) and if one uses the built-in facility to look into the morphological analysis, one can see that the morphological analyzer separates out the parts of the compound into a base noun (the head noun) and the modifier, with the modifier then being flagged as such in the

[10] https://xle.uni-konstanz.de/iness/xle-web

dependency analysis at f-structure (Figure 4). We propose a UD annotation of the following form: a separation out of the head noun from the modifier, with the modifier being identified clearly as such in the dependency analysis. The curly brackets in the f-structure denote a set. This indicates that this attribute may have more than one value. Translating this into UD, we would assume that a head noun can have more than one modifier, all of which would be represented as sisters (at the same level) in the dependency graph.

However, a systematic annotation scheme only provides us with part of the necessary information for the detection of ad hoc compounds. Another part will necessarily involve the consultation of existing dictionaries, as was done as part of our corpus study (section 3.1). This type of lexical information can be further supplemented by lists of nouns and likely combinations, as was done in Schulte im Walde and Borgwaldt (2015). We propose that the data set we gleaned from the German newspaper study could be used in this way: one can compile an initial list of compounds for any given domain, identify the parts (i.e., heads and modifiers) of the compounds, and use the combined list of heads and modifiers as a seed list. This seed list can be then fed into models calculating clusters of lexically similar words for the identification of further ad hoc compounds. We leave this approach for exploration in further research.

6. Conclusion

We have presented a study of German ad hoc compounds that establishes that a subset of these compounds, dubbed *enigmatic compounds*, is systematically used to convey extra attitudinal meaning. We showed this via a combination of theoretical linguistic analysis, a corpus study and a psycholinguistic experiment. We also showed that the extra attitudinal meaning was predominantly used to express a negative stance in the newspapers and thus see enigmatic compounds as providing an important source of information for the end user NLP task of stance detection. A survey of existing dependency parsers and treebanks for German showed an uneven treatment for the annotation of German compounds and we therefore proposed a systematic annotation scheme that is based on the existing multilingual ParGram grammar development experience. We believe that a systematic annotation combined with lexical resources of the type developed in this paper will help ameliorate the challenge of automatized compound detection.

7. Acknowledgements

This project was funded by the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under Germany's Excellence Strategy – EXC-2035/1 – 390681379.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35:5113–5144.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova. 2023. Universals of linguistic idiosyncrasy in multilingual computational linguistics. *Dagstuhl Reports*, 13(4):22–70. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing (2nd ed.)*. Chapman and Hall/CRC.

David Beaver and Jason Stanley. 2018. Toward a non-ideal philosophy of language. *Graduate Faculty Philosophy Journal*, 39(2):503–547.

Reka Benczes. 2009. What motivates the production and use of metaphorical and metonymical compounds. *Cognitive approaches to English: Fundamental, methodological, interdisciplinary and applied aspects*, pages 49–69.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press, Cambridge. 2nd edition.

Douglas Biber and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, 9(1):93–124.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China. Coling 2010 Organizing Committee.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran & Associates, Inc.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications, Stanford.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:1–113.

Rune Haubo B. Christensen. 2018. Cumulative link models for ordinal regression with the R package *ordinal*. *Journal of Statistical Software*.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Mary Dalrymple. 2001. Lexical Functional Grammar. *Syntax and Semantics*, 34.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive-language. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 512–515.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stefanie Dipper. 2003. Implementing and documenting large-scale grammars – German LFG. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, 9(1).

John W. Du Bois. 2007. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, pages 139–182. John Benjamins, Amsterdam.

Mennatallah El-Assady, Valentin Gold, Annette Hautli-Janisz, Wolfgang Jentner, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2016. VisArgue - a visual text analytics framework for the study of deliberative communication. In *Proceedings of The International Conference on the Advances in Computational Analysis of Political Text (PolText2016)*, pages 31–36.

Mennatallah El-Assady, Wolfgang Jentner, Fabian Sperrle, Rita Sevastjanova, Annette Hautli-Janisz, Miriam Butt, and Daniel Keim. 2019. lingvis.io - a linguistic visual analytics framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Florence, Italy. Association for Computational Linguistics.

Gisbert Fanselow. 1981. Zur Syntax und Semantik der Nominalkomposition: ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen. *Linguistische Arbeiten*, 107.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Esther Greussing and Hajo G. Boomgaarden. 2017. Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of ethnic and migration studies*, 43(11):1749–1774.

Robert Henderson and Elin McCready. 2019. Dogwhistles and the at-issue/non-at-issue distinction. In *Secondary content*, pages 222–245. Brill.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A

fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Alexandria J. Innes. 2010. When the threatened become the threat: The construction of asylum seekers in British media narratives. *International Relations*, 24(4):456–477.

IVW. 2023. Ranking der auflagenstärksten überregionalen Tageszeitungen in Deutschland im 2. Quartal 2023.

Carina Kauf, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan S. She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11).

Christopher Kennedy. 1999. *Projecting the adjective: the syntax and semantics of gradability and comparison*. Garland.

Hans J. Kleinsteuber and Barbara Thomass. 2007. The German media landscape. *European Media Governance: National and regional dimensions*, pages 111–123.

Judith N. Levi. 1978. *The syntax and semantics of complex nominals*. Academic Press, New York.

Joseph Patrick Levy, John Bullinaria, and Samantha McCormick. 2017. Semantic vector evaluation and human performance on a new vocabulary MCQ test. In *Proceedings of the Annual Conference of the Cognitive Science Society: CogSci 2017 London: "Computational Foundations of Cognition"*. Cognitive Science Society.

Yingjie Li and Cornelia Caragea. 2019. Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Ralf Meyer. 1993. Compound comprehension in isolation and in context. In *Compound Comprehension in Isolation and in Context*. Max Niemeyer Verlag, Tübingen.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Gregory Murphy. 2002. *The Big Book of Concepts*. MIT Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mary Ellen Ryder. 1994. *Ordered chaos: The interpretation of English noun-noun compounds*, volume 123. University of California Press, Berkeley.

Galit Weidmann Sassoon. 2011. Adjectival versus nominal categorization processes: The rule vs. similarity hypothesis. *Belgian Journal of Linguistics*, 25:104–147.

Anne Schiller. 1994. DMOR - User's Guide. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *Künstliche Intelligenz*, 35:329–341.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sabine Schulte im Walde and Susanne Borgwaldt. 2015. Association norms for German noun compounds and their constituents. *Behavior Research Methods*, 47(4):1199–1221.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M. Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoglu, I Wayan Arka, and Meladel Mistica. 2013. ParGramBank: The ParGram parallel treebank. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 550–560, Sofia. Association for Computational Linguistics.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Polina Tsvilodub, Michael Franke, Robert Hawkins, and Noah D. Goodman. 2023. Overinformative question answering by humans and machines. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Wolfgang Wildgen. 1981. *Makroprozesse bei der Verwendung nominaler ad hoc-Komposita im Deutschen*. Linguistic Agency University of Trier.

Malte Zimmermann. 2011. Discourse particles. In Paul Portner, Claudia Maienborn, and Klaus von Heusinger, editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 2011–2038. De Gruyter Mouton, Berlin.