

# The Role of the German Vorfeld for Local Coherence: A Pilot Study\*

Stefanie Dipper<sup>1</sup> and Heike Zinsmeister<sup>2</sup>

<sup>1</sup> Institute of Linguistics  
Bochum University  
dipper@linguistics.rub.de

<sup>2</sup> Institute of Linguistics  
Konstanz University  
Heike.Zinsmeister@uni-konstanz.de

**Abstract.** This paper investigates the contribution of the German Vorfeld to local coherence. We report on the annotation of a corpus of parliament debates with a small set of coarse-grained labels, marking the functions of the Vorfeld constituents. The labels encode referential and discourse relations as well as non-relational functions. We achieve inter-annotator agreement of  $\kappa = 0.66$ . Based on the annotations, we investigate different features and feature correlations that could be of use for automatic text processing. Finally, we perform an experiment, consisting of an insertion task, to assess the individual impact of different types of Vorfeld on local coherence.

**Key words:** German Vorfeld, local coherence, discourse annotation

## 1 Introduction

A text is said to be coherent if it is easy to read and understand. *Global coherence* is achieved on different levels of the text: (i) on the content level, topics are shared across the sentences (Halliday and Hasan 1976); (ii) on the level of information structure, the focus of attention is shifted smoothly in course of the text (Grosz and Sidner 1986); (iii) finally on the logical level, discourse relations mediate between sentences and other parts of the text (Mann and Thompson 1988).

*Local coherence*, the correlate of global coherence at sentence level, is determined by the smoothness of the transition from one sentence to the next. (i) Topic continuity manifests itself in chains of coreferent entities. (ii) The way these entities are realized in a sequence of sentences (position, grammatical role, choice of determiner, etc.) relates to their salience, which, in turn, determines the reader's focus of attention when shifting from one sentence to the next. (iii) Discourse relations become visible in the occurrence of connectives such as conjunctions, adverbials, or other fixed expressions.

---

\* We would like to thank three anonymous reviewers of a previous version of this paper for valuable comments as well as our student annotators. The research reported in this paper was partly financed by Europäischer Sozialfonds in Baden-Württemberg.

Models of coherence refer to textual, cohesive means to approximate coherence. These means include (i) referential relations, which approximate shared topics (Barzilay and Lapata 2008, Filippova and Strube 2007b, Elsner and Charniak 2008), (ii) referential relations in combination with grammatical role assignment or information status, which approximate attentional focus shift (Grosz *et al.* 1995, Strube and Hahn 1999, Barzilay and Lapata 2008), and (iii) discourse connectives, which are explicit markers of discourse relations (Stede and Umbach 1998, Knott and Dale 1994, Prasad *et al.* 2008).

In general, the beginning of a sentence seems to be a naturally distinguished position for relating the sentence to its preceding context. For instance, it is well-known that old information, which takes up information from the prior context, tends to occur early in the sentence. In a language such as German, the first obligatory position in a declarative sentence, the *Vorfeld* (“pre-field”) is not restricted to a specific grammatical function, such as the subject. For instance, in the sequence *Max ist krank. Deshalb wird er im Bett bleiben.* (‘Max is sick. Therefore, he will stay in bed.’), these positions are occupied by the subject and a discourse connective, respectively. The sequence is locally coherent, due to two textual means. First, there is a coreference link between the sentences (*Max, er*), and second, the connective (*deshalb* ‘therefore’) signals a cause-consequence relation between the sentences. Due to its flexibility, the *Vorfeld* is a highly suitable position for all kinds of coherence-inducing elements. Hence, our working hypothesis is that the majority of the *Vorfeld* constituents should be related to the prior context.

In the present study, we investigate the contribution of the *Vorfeld* constituent to the emergence of local coherence. We define a classification of functions—some contributing to local coherence, some not—and report on an annotation experiment in which 113 argumentative texts were annotated with coreference and discourse relations, among others (Sec. 3). The annotated texts serve as an empirical base for investigating (i) what kind of relations occur in the *Vorfeld* position in spoken monologue texts; (ii) whether the different kinds can be distinguished by means of textual properties; and (iii) to what extent they contribute to local coherence (Sec. 4). The last issue is supplemented by a human insertion experiment in which a sentence that has been extracted from a text has to be re-inserted in the remaining text. Accuracy and ease of insertion is matched with the local *Vorfeld* contexts of the extracted sentence (Sec. 5).

## 2 Related Work

The classical centering approach (Grosz *et al.* 1995) models local coherence in terms of sentence transition types that describe whether the focus of attention continues, is about to change or is changed. The focus of attention is defined on the basis of “centers”—referential expressions—and a salience hierarchy of the involved expressions in terms of grammatical roles; transition types take into account cross-sentential pairs of salient expressions and whether they are referentially related. A similar approach but more robust is the entity-

grid model (Barzilay and Lapata 2008, Filippova and Strube 2007a). It takes all kinds of equivalence sets and referential chains into account and combines them with grammatical-role information. For each referential entity and each sentence, it is recorded whether the entity occurs in that sentence or not. The entity-grid model learns patterns of occurrences. Functional centering (Strube and Hahn 1999) modifies the classical centering approach by employing information status—which characterizes information as old/new, familiar/unfamiliar (Prince 1981)—for determining salience of referential expressions. The importance of information status is also emphasized in the coreference-inspired model of local coherence by Elsner and Charniak (2008).

Discourse connectives indicate discourse relations (Prasad *et al.* 2008), e.g., *daher*, *deswegen* ‘therefore’ indicate a cause function (Stede and Umbach 1998), and strengthen local coherence. Kibble and Power (2004) combine constraints on centering and discourse connectives, among others, to model local coherence. A summary of discourse annotation projects dealing with discourse relations that are explicitly marked as well as implicit ones is given in Stede *et al.* (2007).

In contrast to previous studies we focus on the information provided by the Vorfeld constituent. In the case of discourse relations, we consider relations explicitly marked by discourse connectives as well as implicit relations, provided they are indicated by the Vorfeld constituent as such.

The exceptional status of the Vorfeld constituent is also acknowledged in the corpus-based study of Filippova and Strube (2006), which served as the empirical base for a two-step generation model in which one classifier is trained specifically to pick the Vorfeld constituent (Filippova and Strube 2007b). Among a number of general features, they use grammatical functions, coreferential information as well as textual hints on information status. They do not include discourse relations. Another recent corpus study is Speyer (2005, 2007). He finds that the initial position is usually occupied by brand-new elements or scene-setting elements. Contrastive topics are less preferred, and salient centers—which he equals with non-contrastive topics—are even lower on the preference scale. His findings imply that the most preferred occupants of the Vorfeld are *not* related to the previous context but have a sentence-internal function only.

Most models of local coherence provide algorithms to order discourse units like sentences or clauses. A recent overview of centering approaches and sentence ordering is provided in Karamanis *et al.* (2009). Chen *et al.* (2007) use such a model to insert new information into existing documents. Elsner and Charniak (2008) adopt this task to test the quality of their coherence model.

### 3 The Corpus and Its Annotations

To investigate the role of the Vorfeld constituents, we created and annotated a corpus of selected debates from the European Parliament. The corpus served three purposes: First, we wanted to know the ratio of the sentences that are related to the previous context by virtue of some cohesive item located in the Vorfeld position. Speyer (2007) showed that in his corpus, which consisted of

texts from different genres, almost half of the Vorfelds were *not* related to the previous context. However, it is well known that discourse structure depends on the text type and genre; see, e.g., Berzlánovich *et al.* (2008), and the figures that Speyer (2007) presents for the genres contained in his corpus. Hence, we were interested in the ratios that would show up in our corpus. Our working hypothesis was that the majority of the Vorfeld constituents should be related to the prior context.

Second, we wanted to investigate the types of expressions that occur in the German Vorfeld, be they cohesive or non-cohesive, and search for correlations between morpho-syntactic and discourse properties.

Third, we wanted to examine the influence of different types of cohesive expressions on coherence; for this, we designed an experiment that we present in Sec. 5.

### 3.1 The Corpus

The texts that we included in our corpus are part of the Europarl corpus (Koehn 2005). The Europarl corpus consists of protocols of debates in the European Parliament, both in the original language, as delivered by the speaker, as well as in translations into ten other languages, as delivered by the translation services of the European Union. In the Europarl corpus, individual contributions (“turns”) are marked by SGML elements, along with the names, parties and languages of the respective speakers. As our basis, we selected all contributions whose original language is German (including Austrian German).

We transformed the SGML representation into XML. After tokenization, we applied the German chunker of the TreeTagger (Schmid 1994) to the corpus. Based on the chunk and part-of-speech annotations, we used a heuristics to automatically determine the location of the Vorfeld for each sentence, if any: Our script reads in the analyses of the TreeTagger and searches for the first occurrence of a finite verb, which marks the Vorfeld boundary. If the first word of the current sentence is a subjunction, the finite verb is included in the Vorfeld; otherwise it is excluded from the Vorfeld; leading conjunctions are skipped.

For the annotation task, we isolated medium-sized turns, consisting of 15–20 sentences. This was done (i) to guarantee that the turns contained a sufficient number of Vorfeld sentences, and, hence, would realize a coherent discourse, (ii) to guarantee that the turns would be long enough to allow us to look for cross-sentential patterns, and (iii), at the same time, to avoid turns that are too lengthy and (maybe) more difficult to annotate than shorter ones; also, some long turns contain written reports that are read out by the speakers.

The turns were presented to the annotators without further context information.

### 3.2 Annotation Tagset

We aimed at a tagset that would be rather easy to apply. Hence, we decided not to use a detailed tagset as, e.g., provided by the classical set of RST re-

lations (Rhetorical Structure Theory, Mann and Thompson 1988). Moreover, we decided not to differentiate between discourse relations at the semantic or pragmatic level, since this distinction often poses problems for non-expert annotators. Instead, we opted for coarse-grained labels that are intuitively accessible (cf. Marcu and Echiabi 2002). We finally adopted a small set of coherence labels that can be grouped into five groups. In the following, we describe each group; annotated examples are provided in the Appendix.

*Reference relations: coreference, bridging, reference to the global theme* We distinguish between real coreference, and indirect relations or bridging, as in *I am reading [an interesting paper]. [The author] claims that . . .* Sometimes, the Vorfeld constituent refers to the (implicit) general global topic of the parliament's session. Coreference and bridging relations are prime examples cohesive means.

*Discourse relations: cause, result, continue, contrast, text* The Vorfeld can be occupied by a constituent that indicates a discourse relation to the preceding context. The labels *cause* and *result* are complementary relations: they are designed to cope with any kind of motivational relation (semantic or pragmatic) that holds between two statements, one representing the “premise”, the other the “conclusion”. The label *cause* indicates that the current sentence represents the premise. The label *result* is the complementary relation: it marks sentences that represent the conclusion. A clear indicator of a causal relations is, e.g., the phrase *der Grund (hierfür) ist, dass . . .* ‘the reason (for this) is’; characteristic indicators of result relations are the discourse connectives *daher, deswegen* ‘therefore’.

The *continue* and *contrast* relations are opposed to each other in a similar way: the first one, *continue*, deals with sentences that continue the discourse in the same vein as before, e.g., by elaborating a topic or by drawing comparisons; the second one, *contrast*, introduces a contrastive sentence, which deviates from the previous discourse. The *text* relation is for references to text structure.

A discourse relation is only annotated if the respective indicator is located in the Vorfeld position. The discourse relation, however, holds between the *entire* current sentence and the underlined preceding segment.

Like reference relations, discourse relations are cohesive means.

*Situational relations: deictic, address* The label *deictic* marks references to the situation of the speaking event. Usually it is used for the pronouns *ich, wir* ‘I, we’. The label *address* marks those Vorfeld constituents that are used to address the audience, e.g., the chair of the parliament.

Vorfeld deictics are very frequent in our type of corpus. They relate the current sentence to the external situation, hence they could be considered as non-cohesive (at the textual level). However, successive use of deictics could be interpreted as an instance of a coreference relation, i.e. it would be a cohesive means.

*Internal functions* Some Vorfeld constituents do not refer to the previous context but to (some segment within) the current sentence. Often, these constituents are instances of frame-setting topics, setting the context of interpretation for the current sentence (Jacobs 2001).

*Expletive Function* This label is reserved for the so-called *Vorfeld-es* (Pütz 1986): under certain circumstances, the pronoun *es* ‘it’ occupies the Vorfeld position without having the status of an argument.

The function of the *Vorfeld-es* is not entirely clear but it is often said to be associated with presentational sentences: similar to the *there is*-construction in English, expletive *es* would serve as a means to place all other constituents behind the finite verb, which is the canonical position of new information. In our context, we decided to analyze expletive *es* as non-cohesive.

### 3.3 Annotation Process

The corpus was annotated by 15 students of a seminar about information and text structure and by two paid student assistants. They had a short training period with sample texts, followed by a round of discussion.

For the annotation, the tool MMAX2 was used.<sup>3</sup> The Vorfeld constituents were marked in advance and highlighted, to ease the annotation task. Manual annotation consisted of picking the correct label among a set of 11 predefined features. In case of coreference relations, the antecedent had to be marked. Similarly, in case of discourse relations, the discourse segment that represents the “sister node” or “antecedent” of the annotated relation (*causal*, *result*, etc.) was marked.

On average, the annotation took about 10min/text. In total, the corpus currently consists of 113 annotated texts, average length is 17.2 sentences. We computed inter-annotator agreement on the basis of 18 texts and found  $\kappa = 0.66$ , which is an acceptable value for a task like ours (Artstein and Poesio 2008).

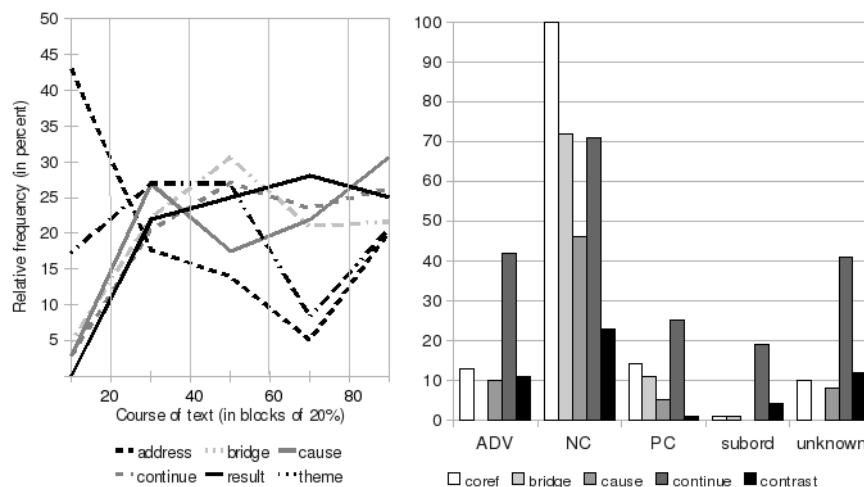
## 4 Corpus Exploration

The corpus consists of 1,940 sentences in total. 86% of the sentences feature a Vorfeld (that could be recognized automatically). Ignoring the first sentence of each text, 91% have a Vorfeld. Among these Vorfeld sentences, 45.1% have a Vorfeld that is clearly related to the previous context (via a reference relation or discourse relation). 21.9% contain a reference to the situation, which could be special instances of coreference relations (see discussion in Sec. 3). The Vorfeld “types” occur with the frequencies displayed in the table below.<sup>4</sup>

Related Vorfeld	Deictic Vorfeld	Unrelated Vorfeld
22.8% Reference	21.9% Situational	4.8% Expletives
22.3% Discourse		28.2% Internal
45.1% Total	21.9% Total	33.0% Total

<sup>3</sup> <http://mmax2.sourceforge.net/>, accessed 2009, April 10.

<sup>4</sup> Situational relations: 19.5% deictic, 2.4% address; Reference relations: 16.0% coreference, 5.0% bridging, 1.8% global theme; Discourse relations: 11.9% continue, 4.0% cause, 3.0% contrast, 1.9% result, 1.5% text.



**Fig. 1.** Distribution of selected Vorfeld functions across the text (left), and selected correlations between morpho-syntactic categories and Vorfeld functions (right)

We decided to focus on *related* Vorfeld types for this study. To get a first impression of the data, we started by looking at the distribution of the different Vorfeld types within each text. The chart in Fig. 1 displays for selected Vorfeld types the relative number of occurrences and their relative text positions. Relative numbers are computed separately for each type. Relative positions are accumulated in blocks of 20%. Vorfeld types that show a rather even distribution across the entire text have been omitted from the chart.

As can be expected, address relations usually occur at the beginning and the end of a turn. The majority of the discourse relations show similar distributions; interestingly, the peak of the result relation occurs rather late. This could be a characteristic of argumentative text such as the Europarl debates.

Another feature that we looked at is the distance between relational Vorfelds and their “antecedents”. The medians of the distances show that discourse relations typically relate to the immediate preceding sentence or, less frequently, to the penultimate sentence. Coreference relations tolerate longer distances to their antecedents than discourse relations do. Still, coreference relations have a distance median of 1, but show large variation. The bridging relation has a distance median of 2.

A further feature is the correlation between Vorfeld types and morpho-syntactic categories, cf. Fig. 1.<sup>5</sup> Not surprisingly, there is a positive correlation between reference relations and noun chunks, similarly between deictic and noun

<sup>5</sup> “ADV”: adverb; “NC”: noun chunk, “PC”: prepositional chunk; “subord”: subordinate clause; “unknown” means that the category could not be determined automatically.

chunks (not displayed). Another positive correlation can be observed with discourse relations and adverbs. In contrast to ordinary reference relations, bridging relations show a positive correlation both with noun chunks and prepositional chunks.<sup>6</sup> This shows that Vorfeld constituents that stand in a bridging relation are often not realized as a prominent grammatical function but mainly seem to serve as a bridge between the current sentence and the prior context. Finally, sentence internal functions also correlate with prepositional chunks; it is still to be confirmed whether these are frame-setting elements.

## 5 Experiment: Insertion Task

As mentioned in Sec. 3, our third goal was to examine different types of cohesive expressions according to their influence on coherence. Our hypothesis was that all types of coherence relations link sentences together, but to a different degree. To validate this hypothesis, we designed an experiment consisting of an insertion task (cf. Elsner & Charniak 2008):

*Procedure* From each turn of our annotated corpus, an arbitrary sentence was extracted. Turn-initial sentences and sentences without Vorfeld marking were skipped. The extracted sentence, followed by the text block consisting of the remaining sentences, was presented to the annotators, who had to guess the original location of the extracted sentence.

*Results* Annotators correctly marked the location in 54% of the texts (53 in total). If we also consider locations as correct that differ from the actual position by just one sentence, accuracy increases to 73%. A closer look at the results reveals that incorrect insertions are located more frequently in front of the actual position than behind it.

Next, we were interested in the question whether the relations that occur in the immediate contexts of the extraction sites are relevant. That is, we investigated whether there are contexts that facilitate or complicate the localization task. According to the annotators themselves, the preceding context was most helpful. According to the results of the experiment, we observe the following tendencies: (i) If the dislocated sentence has a Vorfeld that is referentially related to the previous context, insertion accuracy improves. (ii) Similarly, if the sentence following the dislocated one has a Vorfeld with an internal function, accuracy goes up. (iii) Finally, if the dislocated sentence is the first one of a paragraph, accuracy improves as well.—For preceding sentences, no clear picture emerges from our experiment. Likewise, negative effects do not show up as clearly as positive effects.

Based on the results of the insertion experiment, we could hypothesize that referential relations expressed in the Vorfeld have a stronger contribution to local coherence than discourse relations. However, to be able to really interpret the results and assess the impact of the context, we certainly need more annotated data and have to scrutinize individual cases.

<sup>6</sup> Annotators have been instructed to mark reference relations also in cases where the embedded NP rather than the entire PC is related.



## 6 Conclusion

Coherence is an important issue for all NLP tasks that depend on text generation, such as text summarization or machine translation. It is a crucial step to generate sentences that fit smoothly into the local context, which is defined as the transition of one sentence to the next one.

Contrary to our working hypothesis, the majority of Vorfeld constituents in our corpus is not related to the prior context (by a relation reference or discourse).

A further interesting finding is that most discourse relations connect adjacent sentences. This insight could facilitate the task of discourse analysis considerably. Possibly, this result is related to the fact that we took only relations into account that were indicated by the Vorfeld constituent.

Comparing real coreference and bridging relations, we found that coreference relations are predominantly established by noun chunks while bridging relations are often instantiated by prepositional chunks, which correspond to less-focused grammatical functions (such as prepositional objects or adjuncts). Potentially, this shows that bridging elements rather serve to connect two sentences than to provide for the next focus of attention.

In our next steps, we will explore the Vorfeld types that have not been in our focus so far: unrelated and situation-deictic Vorfeld functions. In addition, we want to find out whether related and non-related Vorfeld constituents can be distinguished automatically, and whether they can be classified into meaningful subclasses. The corpus study indicates that features such as (morpho-)syntactic category, relative position in the text, and distance to the antecedent could be employed in this classification task. The training of a system would require a large database to get a larger number of significant results.

## Appendix: Example Sentences

In the examples, the Vorfeld constituents are marked in boldface. In the case of reference and discourse relations, the segment in the preceding context that is referred to by the Vorfeld constituent is also marked (underlined). Usually, just the semantic head of the antecedent is underlined, as in Ex. 4; in case of propositional antecedents, suitable fragments or entire sentences are marked, cf. Ex. 3 and 6.<sup>7</sup>

*Example 1. Expletive:*

G: **Es** sind alle Versuche gestartet worden, wir müssen jetzt zu Entscheidungen kommen.

E: ‘We have tried everything (lit: **It** has been tried everything); now we need to make decisions.’

<sup>7</sup> The line labeled “E” presents an English translation that is based on the original translations from Europarl. We used the tool OPUS (<http://urd.let.rug.nl/tiedeman/OPUS>) to retrieve the English translations.

*Example 2. Deictic:*

- G: **Ich** würde mir eine bessere Verständigung mit dem britischen Außenministerium wünschen.
- E: ‘**I** would like to see a closer understanding between the British Foreign Office and ourselves.’

*Example 3. Coreference:*

- G: Nun haben aber die Juristen gesprochen — ich muss zugeben, dass ich nicht verstanden habe, was Herr Lehne gesagt hat, aber das liegt vielleicht an mir. **Die Überlegungen von Frau Kaufmann und von Herrn Lehne** will ich jetzt nicht bewerten.
- E: ‘Now, though, the lawyers have spoken, and I have to concede that I did not understand what Mr Lehne said, but perhaps that is my fault. I do not, right now, want to weigh up the pros and cons of **Mrs Kaufmann’s and Mr Lehne’s arguments.**’

*Example 4. Bridging:*

- G: Deshalb sollte es in punkto Handel und Investitionen ein attraktiver Partner für die Union sein. **Dieses Potenzial** wird von Unternehmen der Europäischen Union nicht voll ausgeschöpft, die eine starke Präferenz für China zu haben scheinen.
- E: ‘Therefore, it ought to be an attractive partner for the Union where trade and investment are concerned. **This potential** is not fully exploited by European Union companies, which seem to have a strong preference for China.’

*Example 5. Theme: (The parliament debates an action plan for animal protection)*

- G: **Der nun diskutierte Aktionsplan** ist sicherlich ein weiterer wichtiger Schritt in die richtige Richtung.
- E: ‘**The Action Plan now under discussion** is certainly a further important step in the right direction.’

*Example 6. Contrast:*

- G: Bereits im Jahr 2003 hat die EU bekanntlich ihrer Besorgnis über Hunde-, Stier- und Hahnenkämpfe Ausdruck verliehen, was auch erfreulicherweise im vorliegenden Dokument seinen Niederschlag gefunden hat. **Seltsamerweise** wurde allerdings die Fuchsjagd vergessen.
- E: ‘As we know, the EU expressed its concern about dog, bull and cock fighting back in 2003 and I am pleased to note that this is also reflected in the present document. **Strangely**, however, fox-hunting has been overlooked.’

*Example 7. Text:*

- G: **Der zweite Punkt** betrifft die berühmten 500 Millionen.
- E: ‘**My second point** is that of the celebrated 500 million.’

*Example 8. Internal:*

- G: **Seit Anbeginn des bewussten Tierschutzes infolge der zunehmend technisierten Viehzucht im 19. Jahrhundert** hat sich bekanntlich einiges getan.
- E: ‘As we know, a fair amount has happened **since animal protection as a concept was born as a result of increasingly mechanised animal breeding in the 19th century.**’

## Bibliography

- Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.
- Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, **34**(1), 1–34.
- Berzlánovich, I., Egg, M., and Redeker, G. (2008). Coherence structure and lexical cohesion in expository and persuasive texts. In A. Benz, P. Kühnlein, and M. Stede, editors, *Proceedings of the Workshop on Constraints in Discourse III*, pages 19–26.
- Chen, E., Snyder, B., and Barzilay, R. (2007). Incremental text structuring with online hierarchical ranking. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 83–91.
- Elsner, M. and Charniak, E. (2008). Coreference-inspired coherence modeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and Human Language Technologies (ACL-HLT)*, pages 41–44, Columbus, Ohio.
- Filippova, K. and Strube, M. (2006). Improving text fluency by reordering of constituents. In *Proceedings of the ESSLLI Workshop on Modelling Coherence for Generation and Dialogue Systems*, pages 9–16, Málaga.
- Filippova, K. and Strube, M. (2007a). Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*.
- Filippova, K. and Strube, M. (2007b). Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 320–327, Prague, Czech Republic.
- Grosz, B. and Sidner, C. (1986). Attentions, intentions and the structure of discourse. *Computational Linguistics*, **12**, 175–204.
- Grosz, B., Joshi, A., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**, 203–225.
- Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Longman, London.
- Jacobs, J. (2001). The dimensions of topic–comment. *Linguistics*, **39**, 641–681.
- Karamanis, N., Mellish, C., Poesio, M., and Oberlander, J. (2009). Evaluating centering for information ordering using corpora. *Computational Linguistics*, **35**(1), 29–46.
- Kibble, R. and Power, R. (2004). Optimizing referential coherence in text generation. *Computational Linguistics*, **30**(4), 401–416.
- Knott, A. and Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, **18**(1), 35–62.
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*.

- Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, **8**(3), 243–281.
- Marcu, D. and Echiabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 368–375, Philadelphia, PA.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.
- Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, New York.
- Pütz, H. (1986). *Über die Syntax der Pronominalform 'es' im modernen Deutsch*. Number 3 in *Studien zur deutschen Grammatik*. Tübingen: Narr, 2 edition.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- Speyer, A. (2005). Competing constraints on Vorfeldbesetzung in German. In *Proceedings of the Constraints in Discourse Workshop*, pages 79–87, Dortmund.
- Speyer, A. (2007). Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, **26**, 83–115.
- Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*, pages 1238–1242, Montreal, Quebec, Canada.
- Stede, M., Wiebe, J., Hajičová, E., Reese, B., Teufel, S., Webber, B., and Wilson, T. (2007). Discourse annotation working group report. In *Proceedings of the Linguistic Annotation Workshop (LAW) at ACL*, pages 191–196, Prague.
- Strube, M. and Hahn, U. (1999). Functional Centering — grounding referential coherence in information structures. *Computational Linguistics*, **25**, 309–344.