Improving Text Fluency by Reordering of Constituents

Katja Filippova and **Michael Strube**
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/nlp

Abstract

We present a method for improving local coherence in German with implications for automatically as well as for human-generated texts. We demonstrate that local coherence crucially depends on which constituent occupies the initial position in a sentence. We provide statistical evidence based on a corpus investigation and on results of an experiment with human judges to support our hypothesis. Additionally, we implement our findings in a generation module for determining the *Vorfeld* constituent automatically.

1 Introduction

Multi-document summarization extracts important sentences from different input documents and joins them together in one output document. Obviously, this procedure may not lead to well-written summaries as they may lack coherence. Even if the extracted sentences exhibit some coherence on the entity level, they cannot present the information in the right word order thus leading to difficult to read sentences.

In this paper we propose a method for improving local coherence of German texts by making transitions between sentences smoother. We show that the fluency of a transition from one sentence to the next one depends on which constituent occupies the initial position of the next one. This work is done within a project on automatic text-to-text biography generation which proceeds as follows: Given a number of documents about a certain person and a keyword query as input, first, the sentences which are relevant to the user are found; second, a coherent text is generated from them. The tasks performed during the generation phase, when selected sentences are being put together, concern the order of sentences (global coherence) as well as the order of constituents within a sentence and pronominalization (local coherence). In this paper we investigate the tasks constituent order and pronominalization.

Other applications which could benefit from our method are text summarization, machine translation, or any other application whose output consists of more than one sentence. Moreover, as we will demonstrate, simple rules can improve the fluency of a text produced by human writers.

Unlike some other approaches investigating the relation between information structure and word order, our scope is not limited to noun phrases only, but also includes adverbs and discourse connectives. Because of that we deliberately decided not to formalize our approach within such well-established frameworks as, for example, Centering (Grosz et al., 1995; Prince, 1999). The modifications needed for such formalization would require extending the notion of the (backward-, forward-looking) center not just to constituents other than NPs but also to propositions and would lead to a loss of conceptual simplicity of this framework.

The remainder of the paper is organized as follows: Having outlined related work on generation (Section 2) and on information structure (Section 3), we first motivate and present our approach (Section 4), then we introduce our data whose analysis provides statistical evidence for our hypothesis (Section 5). The results of an experiment with human judges which also confirm the claims concerning the functions of the VF and an application to generation are presented in Section 6 and Section 7 respectively.

2 Related Work

Recent papers on local coherence have suggested algorithms for ordering discourse units like sentences or clauses while phrase ordering within a sentence has not received as much attention. Barzilay et al. (2002) consider the task of sentence ordering within a multi-document summarization approach and experiment with majority and chronological ordering. Lapata (2003) infers constraints on sentence order from a corpus of domain specific texts and approaches the problem in a probabilistic manner. Karamanis et al. (2004) assume a set of clauses as the input and compute a metric for text structuring which utilizes the Centering perspective on coherence. Since all these studies concern English, the question of phrase or word ordering does not play an important role there. The German language, allowing for word order variations, introduces another challenge for generating locally coherent texts.

Kruijff et al. (2001) combine the Prague School and the systemic-functional frameworks and recognize the importance of the information structure for word order variation. They propose an approach to characterizing word order which can be equally well applied to different languages, no matter whether the word order is driven pragmatically or syntactically. In their study, they consider English, Czech and German and demonstrate that in each case the word order can be determined by so called *communicative dynamism* (Firbas, 1974) as well as by the language specific *systemic ordering* (Sgall et al., 1986). Generally, communicative dynamism prescribes that explicitly or implicitly given entities (termed context-bound) precede new information and systemic ordering describes the canonical order in a clause which in case of German corresponds to the following:

Actor < TemporalLocative < SpaceLocative < Means < Addressee < Patient < Source < Destination < Purpose

The authors apply their algorithm to English and Czech software instruction manuals and note that it can be applied to German as well.

3 Background on Information Structure

Due to divergencies in terminology, information structure is notoriously difficult to talk about (see Levinson (1983, p.x)). Therefore, given that the sentence topic is what our proposal relies on, it is a matter of necessity to provide an operative definition and clearly express similarities and differences to existing approaches before presenting our idea.

In general, there are two views on *topic*: as what the sentence is about, and as the measure of salience of an entity. The former has its origin in the work of Strawson (1964); an extreme example of the latter is Givon (1983) whose topic is very similar to the notion of the backward-looking center in the Centering model (Grosz et al., 1995). We adhere to the first view and define topic based on the pragmatic relation of aboutness only, thus excluding the discourse status of the referent from our definition. The topic is the referent the proposition is about, or more precisely, the referent the speaker assumes to be a center of current interest. Consequently, we do not subscribe to the view that the element about which the information is provided always occupies sentence initial position. On the contrary, like Reinhart (1981), we think that topiclike elements may and do appear on other positions as well. The role of the sentence initial element, on the other hand, is more similar to the role of 'real' topics as described by Chafe (1976) in that they 'are not so much "what the sentence is about" as "the frame within which the sentence holds"'. Splitting these two functions makes our approach different from Vallduví's (1990). For him, 'by starting a sentence with link speakers indicate to hearers that the focus must go to that address, and enter the information under its label.' (Vallduví, 1990, p.59). Although in many cases his link and our topic coincide, we find it unintuitive and improbable that dates or discourse connectives are the addresses where the new information is attached.

Apart from that, we distinguish between what has been introduced by Chafe (1987) as active, accessible and inactive referents. Topic and activeness correlate in that the most easily processed sentences are those whose topic referents are active in the discourse (Lambrecht, 1994, p.165).

Our approach is similar to the one of Kruijff et al. (2001) in that we also consider the relation between word order and information structure but differs from it in several respects. Firstly, Kruijff et al. (2001) concentrate on how to generate not just a grammatical but acceptable ordering whereas we focus on how to determine not just a grammatical and acceptable ordering but the one that makes the transition as smooth as possible. Secondly, we extend context-bound information to accessible and treat context-bound NPs, temporal expressions and discourse connectives in the same way. The fact that absolute temporal expressions are perfectly acceptable in the beginning of a sentence (Heidolph et al., 1981, ch.4) which can not be explained in terms of given and new information is noticed by Kruijff-Korbayová et al. (2002) where locations or temporal expressions are treated as the theme or point of departure of the clause. Extending given information to accessible makes it possible to treat these cases uniformly.

4 Our Hypothesis

Because of the fixed verb position, the German V-second clause is divided into two parts. The part preceding the finite verb, "prefield", or *Vorfeld (VF)* usually contains only one constituent, and the part between the finite verb or complementizer and the verbal elements at the end of the clause, "middle field", or *Mittelfeld (MF)* incorporates the rest. In (1) and all the following examples the VF is indicated by italics:

(1) *Marie Curie* wurde am 7. November
Marie Curie was on the 7th November
1867 in Warschau geboren.
1867 in Warsaw born.

'Marie Curie was born in Warsaw on the 7th of November 1867'

The problem of constituent ordering in German can be reformulated then: Which constituent is to be placed in the VF? What should be the order of constituents in the MF?

Concerning the VF, the following claims are made: the VF, being a cognitively prominent position (Gernsbacher & Hargreaves, 1988), has two major[1] functions: Whenever the topic of a sentence needs to be established, it is placed into the VF. Otherwise, if the topic has already been established and is still activated in the mind of the reader, it should be pronominalized and there is no need for it to occupy the VF (see Frey (2004) for recent research on the topic position in German). In this case the VF is the position responsible for a smooth transition from the previous sentence to the current one. The smoothness or fluency of transitions is ensured by placing a constituent in the VF which helps linking the introduced sentence to the representation readers have already built in their mind.

The best candidate for the VF is to be selected from the set of accessible elements. These are entities accessible due to the preceding context, e.g. repeated mentions or anaphoric elements, inferentially accessible constituents (bridging anaphora). Temporal expressions – absolute, *am 23. Mai 1900 (on the 24th of May)*, or relative, *Im gleichen Jahr (in the same year)* – belong to this group because of the relevance of the time scale for the biography genre. For newspaper texts locations (e.g. *Berlin, Sankt-Petersburg*) and other named entities (e.g. *SPD, Merkel, SAP*) are expected to be as readily accessible as temporal expressions here. Discourse connectives count as accessible constituents as well: They establish a relation between the proposition expressed in the current sentence and propositions expressed earlier in the discourse. *So (so), anschliessend (finally), dabei (in doing so)* are examples of such connectives but not *weil (because), obwohl (although)* which link two clauses within one sentence. Proadverbials, e.g. *damit (with that), darüber (about that)* are also included in this group.

The first impression might be that it is inconsistent to unify such diverse phenomena as discourse connectives and noun phrases. This impression may change if we distinguish between structural connectives and discourse adverbials and consider the latter as anaphora (Webber et al., 2003). From this point of view the fact that an adverbial connective, e.g. *sonst* (otherwise), and an inferrable NP, e.g. *die Familie* (the family) following a discourse where the parents are mentioned, are both treated as accessible elements, should not be surprising because both of them are instances of anaphora (bridging anaphora in the latter case).

To sum up, we identify the topic in the sentence, which is the address for new information, we also find *other* linking or framing elements and in case of the topic being activated place the best candidate from the linking list to the VF. We hypothesize that this strategy provides smoother transitions than reserving the VF for the topic. The rest of the paper provides evidence from different sources which confirm our hypothesis.

[1] Other elements, such as contrastive topics, are encountered in the VF as well but considerably less frequent.

5 Data

5.1 Preprocessing

The data we investigate is a collection of biography texts from the German version of Wikipedia[2]. The data is homogeneous in the sense that it contains all biographies under the Wikipedia category of physicists.

Fully automatic preprocessing in our system comprises the following stages: First, a list of people of a certain Wikipedia category is taken and for every person an article is extracted. The text is purged from Wiki tags and comments, the information on subtitles and paragraph structure is preserved. Second, sentence boundaries are identified with a Perl CPAN module[3] whose performance we improved by extending the list of abbreviations and modifying the output format. Next, the sentences are split into tokens. The TnT tagger (Brants, 2000) and the TreeTagger (Schmid, 1997) are used for tagging and lemmatizing. Finally, the texts are parsed with the CDG dependency parser (Foth & Menzel, 2006). Thus, the text is split on three levels: paragraphs, sentences and tokens, and morphological and syntactic information is provided.

A publicly available list of about 300 discourse connectives was downloaded from the Internet site of the Institute of the German Language[4] (Institut für Deutsche Sprache, Mannheim) and slightly extended. These are identified in the texts and annotated automatically as well. Named entities are classified according to their type using information from Wikipedia: *person*, *location*, *organization* or *undefined*. Given the peculiarity of our corpus, we are able to identify all mentions of the biographee in the text by simple string matching. We also annotate different types of referring expressions (*first*, *last*, *full name*) and resolve anaphora by linking personal pronouns to the biographee provided that they match in number and gender.

Temporal expressions (both relative and absolute) and VFs are identified automatically by a set of patterns. VFs, for example, are determined as the part of the sentence standing before the root verb.

5.2 Corpus Analysis

We analyzed 370 texts with an average length of 17 sentences, 6521 sentences in total. 2857 of them mentioned the biographee (with the name or with a personal pronoun) and hence were of interest for us. Whenever such a sentence opens a new section in an article, we assume that the topic should be explicitly established, therefore a concrete reference to the person is needed and the referring expression should be placed in the VF, no matter what its syntactic function is. Whenever a sentence is preceded by one or several sentences which already are about the biographee, we assume the person to be activated in the mind of the reader. In such a case a pronominal reference should be used, and the preferred position for it is the MF. Examples (2) and (3) should make the point clearer:

(2) a Familie und frühe Jahre
Familiy and early years
’Family and early years’

b *Marie Curie* wurde am 7. November
Marie Curie was on 7th November
1867 als Maria Salomea Sklodowska in
1867 as Maria Salomea Sklodowska in
Warschau geboren.
Warsaw born.
’Marie Curie was born in Warsaw on the 7th of November 1867 as Maria Salomea Sklodowska.’

(3) a Zusammen mit ihrem Mann Pierre
Together with her husband Pierre
Curie und dem Physiker Antoine Henri
Curie and the physicist Antoine Henri
Becquerel erhielt sie 1903 den
Becquerel received she 1903 the
Nobelpreis für Physik.
Nobel prize in physics.
’Together with her husband Pierre Curie and the physicist Antoine Henri Becquerel, she received the Nobel prize in physics in 1903.’

b *Acht Jahre später* wurde ihr der
Eight years later was her the
Nobelpreis für Chemie verliehen.
Nobel prize in chemistry given.
’Eight years later, the Nobel prize in chemistry was given to her’

[2] http://de.wikipedia.org

[3] http://search.cpan.org/˜holsten/Lingua-DE-Sentence-0.07/Sentence.pm

[4] http://hypermedia.ids-mannheim.de/index.html

	pronoun	name	conn.	temp.expr.
VF	680	953	359	1358
MF	2177	602	1013	1355
Total	2857	1555	1372	2713

Table 1: Distribution of expressions according to their position

Following a title, (2b) opens the biography which is devoted to Marie Curie. The topic is established by placing the full name reference to the VF. Pronominalization and placing the constituent in the MF are deprecated. In (3) the situation is different: The biographee is already activated in the mind of the hearer, and in (3b) there is a better candidate for the VF – a temporal expression.

Considering sentences with a reference to the biographee, it was of interest to us to see which constituents usually occupy the VF. Table 1 shows the distribution of the expressions referring to the biographee (pronominal and non-pronominal), temporal expressions, and connectives with respect to their position in a sentence. The results clearly indicate that the VF is not a preferred position for pronouns, whereas non-pronominal reference may appear in the VF about one and a half times as often as in MF. Unfortunately, some connectives are ambiguous and can mark relations between clauses of one sentence as well as relations between sentences. In the future we plan to improve the annotation and rule out all instances of intrasentential connectives. The fact that temporal expressions appear in the VF as often as in the MF does not support our hypothesis so far. In order to find out which candidate is more preferrable, we performed an experiment with human judges.

6 Experiment

In order to verify our hypotheses on text fluency, we performed an experiment with human judges, all native speakers of German who were presented with 24 short text fragments from our corpus. Each fragment had two possible continuations which were identical in all aspects but for the word order. The order of the two alternative sentences as well as the order of the fragments was generated randomly. The judges were asked to choose from the two variants the one which continues the preceding text in the most fluent way or choose nothing in case of both continuations sound equally fluent.

(4) a Nach seiner Kriegsteilnahme am
After his War participation in the
Ersten Weltkrieg folgte er
First World War followed he
Berufungen nach Jena, Stuttgart,
invitations to Jena, Stuttgart,
Breslau und Zürich.
Wroclaw and Zürich.

’Having taken part in the First World War, he accepted invitations from Jena, Stuttgart, Wroclaw and Zürich’

b′ *Dort* belegte er den Lehrstuhl für
There hold he the chair for
Theoretische Physik, den vor ihm
theoretical physics, which before him
bereits Albert Einstein und Max von
already Albert Einstein and Max von
Laue inne hatten.
Laue had.

b′′ *Er* belegte dort den Lehrstuhl für
He hold there the chair for
Theoretische Physik, den vor ihm
theoretical physics, which before him
bereits Albert Einstein und Max von
already Albert Einstein and Max von
Laue inne hatten.
Laue had.

’He hold there the chair of theoretical physics, which was before him occupied by Albert Einstein and Max von Laue’

Sentences (4b′) and (4b′′) have the same propositional content and differ only in what stands in the VF: the proadverbial *there* or the personal pronoun *he*. If our hypothesis is right, then the judges would choose (4b′) more often than (4b′′).

The purpose of the experiment was twofold: to check, first, whether in cases where topic establishing is necessary (e.g. example (2)), the VF is the preferrable position for the topic. Second, whether an established topic occupying the VF makes the transition to the sentence smoother, or there are better candidates for this position (example (4)).

18 human judges (9 female and 9 male) took part in the experiment. The statistical significance of our results was computed using χ^2 test on the

	inferrable	temp.expr.	connective	proadverbial	total
pronoun	– + + +	∘∘ + +	– + + +	– + – – –	17
name	+		+		2

Table 2: Results of the experiment with human judges

$p = 0.01$ level or below. It turned out that the preference for a certain variant was significant if it was chosen by at least 15 judges.

6.1 Topic-establishing Sentences

We selected three section initial sentences which mention the biographee because such sentences open a new discourse topic (this is explicitly marked by using section titles) and therefore require non-pronominal reference to the person. Three pairs – a sentence and a propositionally equivalent variant of it – were presented to the judges. Example (2) is one of such fragments. In these three fragments the judges had a choice of what to place into the VF: an absolute temporal expression, an NP with a reference to a previously mentioned and therefore accessible person, and an inferentially accessible NP or a name reference to the biographee. In all three cases the biographee was preferred over other candidates for the VF position, and in two of the cases the difference was significant. This finding alone is in accordance with the well-known correlation between topics, subjects and sentence initial position and does not have a dramatic impact on coherence.

6.2 Sentences with the Established Topic

The second part of the experiment concerned sentences where the biographee is established as the topic due to the immediately preceding context (like (3a,b) and (4a,b)). From the 19 test pairs of this kind, seventeen contained a pronominal reference, and in two other pairs the biographee was referred to with the last name. For these examples, constituents of the following kinds were supposed to be better candidates for the VF: *inferrable constituents* (5 fragments), *temporal expressions* (4), *discourse connectives* (5), or *proadverbials* (5). Here we distinguish between connectives which have a distinct semantic meaning (e.g. temporal or additive), these are labeled as *discourse connectives*, from *proadverbials* (*dabei*, *darüber*) whose meaning is usually context-dependent.

Syntactic function was expected to play a minor role for the choice of the best constituent for the VF. This parameter was set in favour of the activated referent: in all sentences the syntactic role of the biographee is subject.

For *every* pair it turned out that the majority of judges preferred accessible constituents over activated subjects. In five cases, the judges preferred the modified version over the original sentence, i.e. the sentence from the Wikipedia article, because they found the modified fragment sound more fluent. A plus (+) in Table 2 stands for cases where the difference in preferences is significant on the $p = 0.01$ level, a circle (∘) for significance on the $p = 0.05$ level, a minus (–) for non-significant preference.

Interestingly, for both examples with a non-pronominal reference to the biographee the connective as well as the accessible constituent were preferred significantly more often. This brings us to the conclusion that for a fluent transition the established topic should not be placed into the VF no matter what its surface or syntactic realization is. The last two test sentence pairs let the reader choose between, first, a temporal expression and an accessible constituent; second, a temporal expression and a proadverbial. For the former case, no difference in preferences was found; for the latter, the proadverbial was picked significantly more frequent than the temporal expression.

Obviously, in order to rank candidates of different kinds more subtle experiments need to be performed: Form of the expression, semantics of connectives, and degree of accessibility should be taken into account. So far, it can only be stated that, concerning candidates for the VF, the established topic follows any of the listed above.

7 Implications for Generation

In this section we present an application of our findings to the automatic identification of the best candidate for the VF. This can be considered a first step towards the automatic generation of phrase and word order. We split our 370 articles corpus into training and testing sets and selected parsed sentences which mention a biographee. Thus we obtained 3080 and 616 sentences for training and

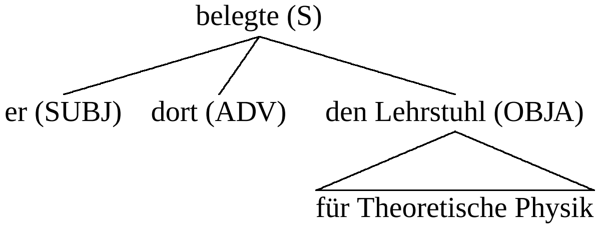

Figure 1: Essential part of example (4)

testing respectively. The number of candidate constituents for a sentence ranged from 1 to 8 being 4 on average. Consider the example (4b$''$) again: *Er belegte dort den Lehrstuhl für Theoretische Physik, den vor ihm bereits Albert Einstein und Max von Laue inne hatten.* For our purposes we may ignore the structure dependent on the OBJA *Lehrstuhl* and consider only the nodes dependent on the root verb (Figure 1). In this example there are three candidates which can occupy the VF because there are three constituents dependent on the main verb.

Using maximum entropy learning which has been successfully applied to a number of NLP tasks, including word order generation (Ratnaparkhi, 2000), we trained a binary classifier which for every constituent estimated the probability of it being in the VF. The three feature vectors for Figure 1 are presented in Table 3. The first seven features apply to any candidate, these are the word immediately dependent on the verb (DEP.WORD), the non-auxiliary root verb (VERB), the lexical head of the dependent constituent (LEX.HEAD), part of speech (POS), syntactic function (SYNT), maximal depth (DEP) and the length (LEN) of the constituent. If the constituent is a named entity, a temporal expression or a connective then this is expressed as TYPE. If it is a person, then it is marked whether it refers to the biographee (ROLE) and the type of the referring expression is given (REF.EXPR). For a temporal expression, REF.EXPR expresses whether it is an absolute or a relative one. The last line gives values of the temporal expression from example (2b) – *am 7. November 1867*. From all candidates for one sentence, the one with the highest probability was chosen as the best candidate. The results were evaluated against the original ordering. Note, that with this setting contextual information is totally absent, and inferrable constituents can not be identified.

From the 616 test instances the algorithm made a mistake in 211 cases, thus the accuracy is about 65%. Having analysed the first 100 errors, we summarize our observations in Table 4. In 17 cases the algorithm preferred a temporal expression over a pronoun which occupied the VF in the original Wikipedia article. This counts as a mistake although, as the experiment has demonstrated, human judges find text more coherent provided there is a temporal expression and not a pronoun in the VF. Likewise, the fact that 8 connectives were classified falsely does imply that the generated order would make the text less coherent than the original. Apart from that, name references may have been used in topic established sentences, which means that some of the 11 mistakes might not be errors, just as it is in the case of pronouns.

Wikipedia	MaxEnt	
pron	temp	17
pron	conn	8
name	temp	11
XP	pron	22

Table 4: Types of errors with their frequency

In 22 cases a pronominal reference to the biographee was chosen instead of a NP, PP or a sub-clause (labeled XP in the table) which were accessible due to the preceding context. By extending the list of features and taking the context into account we expect to improve the results significantly. Whereas temporal expressions, NEs and connectives can be identified relatively easily, identifying inferrable NPs is a much harder task. A straightforward way to measure inferrability is by means of string matching but, obviously, this method would work for the most trivial cases only. Measuring semantic relatedness (using GermaNet (Gurevych, 2005) or Wikipedia (Strube & Ponzetto, 2006)) could offer a more intelligent way of finding accessible referents.

8 Conclusions

Corpus investigation as well as experiments on constituent reordering confirmed our claims concerning the role of the VF: In most cases, it is either the topic establishing position, or the position for accessible constituents. In line with the hypothesis, human judges find transitions between sentences smoother when the VF is occupied by accessible elements, and not by topics, no matter what their discourse status is. The first results on automatic phrase ordering motivate fur-

DEP.WORD	VERB	LEX.HEAD	POS	SYNT	DEP	LEN	TYPE	ROLE	REF.EXPR.
[er]	belegte	er	pper	subj	d=0	l=1	pers	biogr	re=pron
[dort]	belegte	dort	adv	adv	d=0	l=1			
[lehrstuhl]	belegte	lehrstuhl	nn	obja	d=7	l=18			
[am]	geboren	november	card	pp	d=3	l=5	temp		re=abs

Table 3: Vectors for the three constituents in Figure 1 and the temporal expression from example (2b)

ther research in this direction. In the future we would like to automatically generate word order for whole sentences. The ultimate goal is to apply the method to generating coherent biographies.

Acknowledgments: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a KTF grant. We would also like to thank our judges.

References

Barzilay, Regina, Noemie Elhadad & Kathleen R. McKeown (2002). Inferring strategies for sentence ordering. *Journal of Artificial Intelligence Research*, 17:35–55.

Brants, Thorsten (2000). TnT – A statistical Part-of-Speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing,* Seattle, Wash., 29 April – 4 May 2000, pp. 224–231.

Chafe, Wallace (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles Li (Ed.), *Subject and Topic*, pp. 25–55. New York: Academic Press.

Chafe, Wallace (1987). Cognitive constraints on information flow. In Russell S. Tomlin (Ed.), *Coherence and Grounding in Discourse*, pp. 21–52. Amsterdam, The Netherlands: John Benjamins.

Firbas, Jan (1974). Some aspects of the Czechoslovak approach to problems of functional sentence prespective. In F. Daneš (Ed.), *Papers on Functional Sentence Perspective*, pp. 11–37. Prague: Academia.

Foth, Kilian & Wolfgang Menzel (2006). Robust parsing: More with less. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 25–32.

Frey, Werner (2004). A medial topic position for German. *Linguistische Berichte*, 198:153–190.

Gernsbacher, Morton A. & David J. Hargreaves (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27:699–717.

Givon, Talmy (1983). Topic continuity in spoken English. In T. Givon (Ed.), *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. Amsterdam, Philadelphia: John Benjamins.

Grosz, Barbara J., Aravind K. Joshi & Scott Weinstein (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Gurevych, Iryna (2005). Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing,* Jeju Island, South Korea, 11-13 October, 2005, pp. 767–778.

Heidolph, Karl Erich, Walter Flämig & Wolfgang Motsch (1981). *Grundzüge einer deutschen Grammatik*. Berlin: Akademie-Verlag.

Karamanis, Nikiforos, Massimo Poesio, Chris Mellish & Jon Oberlander (2004). Evaluating Centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 392–393.

Kruijff, Geert-Jan M., Ivana Kruijff-Korbayová, John Bateman & Elke Teich (2001). Linear Order as higher-level decision: Information Structure in strategic and tactical generation. In *8th European Workshop on Natural Language Generation,* Toulouse, France, July 6-7 2001, pp. 74–83.

Kruijff-Korbayová, Ivana, Geert-Jan Kruijff & John Bateman (2002). Generation of appropriate word order. In K. van Deemter & R. Kibble (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pp. 193–222. Stanford: CSLI.

Lambrecht, Knud (1994). *Information Structure and Sentence Form*. Cambridge University Press.

Lapata, Maria (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pp. 545–552.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Prince, Ellen F. (1999). How not to mark topics: 'Topicalization' in English and Yiddish. *Texas Linguistics Forum*.

Ratnaparkhi, Adwait (2000). Trainable methods for surface natural language generation. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics,* Seattle, Wash., 29 April – 3 May, 2000, pp. 194–201.

Reinhart, Tanya (1981). Pragmatics and linguistics. An analysis of sentence topics. *Philosphica*, 27(1):53–93.

Schmid, Helmut (1997). Probabilistic part-of-speech tagging using decision trees. In Daniel Jones & Harold Somers (Eds.), *New Methods in Language Processing*, pp. 154–164. London, UK: UCL Press.

Sgall, Peter, Eva Hajičová & Jarmila Panevová (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: D. Reidel.

Strawson, Peter F. (1964). Identifying reference and truth-values. In D. Steinberg & L. Jacobovits (Eds.), *Semantics*, pp. 86–99. Cambridge: Cambridge University Press.

Strube, Michael & Simone Paolo Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, Mass., 16–20 July 2006. To appear.

Vallduví, Enric (1990). *The Informational Component*, (Ph.D. thesis). Philadelphia, Penn.: University of Pennsylvania, Department of Linguistics.

Webber, Bonnie, Matthew Stone, Aravind Joshi & Alistair Knott (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–588.