To what extent does sentence-internal realisation reflect discourse context? A study on word order

Sina Zarrieß **Jonas Kuhn**
Institut für maschinelle Sprachverarbeitung
University of Stuttgart, Germany
zarriesa,jonas@ims.uni-stuttgart.de

Aoife Cahill
Educational Testing Service
Princeton, NJ 08541, USA
acahill@ets.org

Abstract

We compare the impact of sentence-internal vs. sentence-external features on word order prediction in two generation settings: starting out from a discriminative surface realisation ranking model for an LFG grammar of German, we enrich the feature set with lexical chain features from the discourse context which can be robustly detected and reflect rough grammatical correlates of notions from theoretical approaches to discourse coherence. In a more controlled setting, we develop a constituent ordering classifier that is trained on a German treebank with gold coreference annotation. Surprisingly, in both settings, the sentence-external features perform poorly compared to the sentence-internal ones, and do not improve over a baseline model capturing the syntactic functions of the constituents.

1 Introduction

The task of surface realization, especially in a relatively free word order language like German, is only partially determined by hard syntactic constraints. The space of alternative realizations that are strictly speaking grammatical is typically considerable. Nevertheless, for any given choice of lexical items and prior discourse context, only a few realizations will come across as natural and will contribute to a coherent text. Hence, any NLP application involving a non-trivial generation step is confronted with the issue of soft constraints on grammatical alternatives in one way or another.

There are countless approaches to modelling these soft constraints, taking into account their interaction with various aspects of the discourse context (givenness or salience of particular referents, prior mentioning of particular concepts).

Since so many factors are involved and there is further interaction with subtle semantic and pragmatic differentiations, lexical choice, stylistics and presumably processing factors, theoretical accounts making reliable predictions for real corpus examples have for a long time proven elusive. As for German, only quite recently, a number of corpus-based studies (Filippova and Strube, 2007; Speyer, 2005; Dipper and Zinsmeister, 2009) have made some good progress towards a coherence-oriented account of at least the left edge of the German clause structure, the *Vorfeld* constituent.

What makes the technological application of theoretical insights even harder is that for most relevant factors, automatic recognition cannot be performed with high accuracy (e.g., a coreference accuracy in the 70's means there is a good deal of noise) and for the higher-level notions such as the information-structural focus, interannotator agreement on real corpus data tends to be much lower than for core-grammatical notions (Poesio and Artstein, 2005; Ritz et al., 2008).

On the other hand, many of the relevant discourse factors are reflected indirectly in properties of the sentence-internal material. Most notably, knowing the shape of referring expressions narrows down many aspects of givenness and salience of its referent; pronominal realizations indicate givenness, and in German there are even two variants of the personal pronoun (*er* and *der*) for distinguishing salience. So, if the generation task is set in such a way that the actual lexical choice, including functional categories such as determiners, is fully fixed (which is of course not always the case), one can take advantage of

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 767–776,
Avignon, France, April 23 - 27 2012.

these reflexes. This explains in part the fairly high baseline performance of n-gram language models in the surface realization task. And the effect can indeed be taken much further: the discriminative training experiments of Cahill and Riester (2009) show how effective it is to systematically take advantage of asymmetry patterns in the morphosyntactic reflexes of the discourse notion of information status (i.e., using a feature set with well-chosen purely sentence-bound features).

These observations give rise to the question: in the light of the difficulty in obtaining reliable discourse information on the one hand and the effectiveness of exploiting the reflexes of discourse in the sentence-internal material on the other – can we nevertheless expect to gain something from adding sentence-external feature information?

We propose two scenarios for adressing this question: first, we choose an approximative access to context information and relations between discourse referents – lexical reiteration of head words, combined with information about their grammatical relation and topological positioning in prior sentences. We apply these features in a rich sentence-internal surface realisation ranking model for German. Secondly, we choose a more controlled scenario: we train a constituent ordering classifier based on a feature model that captures properties of discourse referents in terms of manually annotated coreference relations. As we get the same effect in both setups – the sentence-external features do not improve over a baseline that captures basic morphosyntactic properties of the constituents – we conclude that sentence-internal realisation is actually a relatively accurate predictor of discourse context, even more accurate than information that can be obtained from coreference and lexical chain relations.

2 Related Work

In the generation literature, most works on exploiting sentence-external discourse information are set in a summarisation or content ordering framework. Barzilay and Lee (2004) propose an account for constraints on topic selection based on probabilistic content models. Barzilay and Lapata (2008) propose an entity grid model which represents the distribution of referents in a discourse for sentence ordering. Karamanis et al. (2009) use Centering-based metrics to assess coherence in an information ordering system. Clarke and Lapata (2010) have improved a sentence compression system by capturing prominence of phrases or referents in terms of lexical chain information inspired by Morris and Hirst (1991) and Centering (Grosz et al., 1995). In their system, discourse context is represented in terms of hard constraints modelling whether a certain constituent can be deleted or not.

In the linearisation or surface realisation domain, there is a considerable body of work approximating information structure in terms of sentence-internal realisation (Ringger et al., 2004; Filippova and Strube, 2009; Velldal and Oepen, 2005; Cahill et al., 2007). Cahill and Riester (2009) improve realisation ranking for German – which mainly deals with word order variation – by representing precedence patterns of constituents in terms of asymmetries in their morphosyntactic properties. As a simple example, a pattern exploited by Cahill and Riester (2009) is the tendency of definite elements tend to precede indefinites, which, on a discourse level, reflects that given entities in a sentence tend to precede new entities.

Other work on German surface realisation has highlighted the role of the initial position in the German sentence, the so-called *Vorfeld* (or "prefield"). Filippova and Strube (2007) show that once the *Vorfeld* (i.e. the constituent that precedes the finite verb) is correctly determined, the prediction of the order in the *Mittelfeld* (i.e. the constituents that follow the finite verb) is very easy. Cheung and Penn (2010) extend the approach of Filippova and Strube (2007) and augment a sentence-internal constituent ordering model with sentence-external features inspired from the entity grid model proposed by Barzilay and Lapata (2008).

3 Motivation

While there would be many ways to construe or represent discourse context (e.g. in terms of the global discourse or information structure), we concentrate on capturing local coherence through the distribution of discourse referents in a text. These discourse referents basically correspond to the constituents that our surface realisation model has to put in the right order. As the order of referents or constituents is arguably influenced by the information structure of a sentence given the previous text, our main assumption was that infor-

(1) a. Kurze Zeit später erklärte ein Anrufer bei Nachrichtenagenturen in Pakistan , **die Gruppe Gamaa** bekenne sich.
Shortly after, a caller declared at the news agencies in Pakistan, that **the group Gamaa** *avowes itself.*
b. **Diese Gruppe** wird für einen Großteil der Gewalttaten verantwortlich gemacht , die seit dreieinhalb Jahren in Ägypten verübt worden sind .
This group *is made responsible for most of the violent acts that have been committed in Egypt in the last three and a half years.*

(2) a. **Belgien** wünscht, dass sich WEU und NATO darüber einigen.
Belgium *wants that WEU and NATO agree on that.*
b. **Belgien** sieht in der NATO die beste militärische Struktur in Europa .
Belgium *sees the best military structure of Europe in the NATO.*

(3) a. **Frauen** vom Land kämpften aktiv darum , ein Staudammprojekt zu verhindern.
Women *from the countryside fighted actively to block the dam project.*
b. Auch in den Städten fänden sich immer mehr **Frauen** in Selbsthilfeorganisationen zusammen.
Also in the cities, more and more **women** *team up in self-help organisations.*

mation about the prior mentioning of a referent would be helpful for predicting the position of this referent in a sentence.

The idea that the occurence of discourse referents in a text is a central aspect of discourse structure has been systematically pursued by Centering Theory (Grosz et al., 1995). Its most important notions are related to the realisation of discourse referents (i.e. described as "centers") and the way the centers are arranged in a sequence of utterances to make this sequence a coherent discourse. Another important concept is the "ranking" of discourse referents which basically determines the prominence of a referent in a certain sentence and is driven by several factors (e.g. their grammatical function). For free word order languages like German, word order has been proposed as one of the factors that account for the ranking (Poesio et al., 2004). In a similar spirit, Morris and Hirst (1991) have proposed that chains of (related) lexical items in a text are an important indicator of text structure.

Our main hypothesis was that it is possible to exploit these intuitions from Centering Theory and the idea of lexical chains for word order prediction. Thus, we expected that it would be easier to predict the position of a referent in a sentence if we have not only given its realisation in the current utterance but also its prominence in the previous discourse. Especially, we expected this intuition to hold for cases where the morpho-syntactic realisation of a constituent does not provide many clues. This is illustrated in Examples (1) and (2) which both exemplify the reiteration of a lexical item in two subsequent sentences, (reiteration is one type of lexical chain discussed in Morris and Hirst (1991)). In Example (1), the second instance of the noun 'group' is modified by a demonstrative pronoun such that its "known" and prominent discourse status is overt in the morpho-syntactic realisation. In Example (2), both instances of "Belgium" are realised as bare proper nouns without an overt morphosyntactic clue indicating their discourse status.

Beyond the simple presence of reitered items in sequences of sentences, we expected that it would be useful to look at the position and syntactic function of the previous mentions of a discourse referent. In Example (1), the reiterated item is first introduced in an embedded sentence and realised in the *Vorfeld* in the second utterance. In terms of centering, this transition would correspond to a topic shift. In Example (2), both instances are realised in the *Vorfeld*, such that the topic of the first sentence is carried over to the next.

In Example (3), we illustrate a further type of lexical reiteration. In this case, two identical head nouns are realised in subsequent sentences, even though they refer to two different discourse referents. While this type of lexical chain is described as "reiteration without identity of referents" by Morris and Hirst (1991), it would not be captured in Centering since this is not a case of strict coreference. On the other hand, lexical chains do not capture types of reiterated discourse referents that have distinct morpho-syntactic realisations, e.g. nouns and pronouns.

Originally, we had the hypothesis that strict corefence information is more useful and accurate for word order prediction than rather loose lexical chains which conflate several types of referential and lexical relations. However, the advantage of chains, especially chains of reiteration, is that they can be easily detected in any corpus text and

that they might capture "topics" of sentences beyond the identity of referents. Thus, we started out from the idea of lexical chains and added corresponding features in a statistical ranking model for surface realisation of German (Section 4). As this strategy did not work out, we wanted to assess whether an ideal coreference annotation would be helpful at all for predicting word order. In a second experiment, we use a corpus which is manually annotated for coreference (Section 5).

4 Experiment 1: Realisation Ranking with Lexical Chains

In this Section, we present an experiment that investigates sentence-external context in a surface realisation task. The sentence-external context is represented in terms of lexical chain features and compared to sentence-internal models which are based on morphosyntactic features. The experiment thus targets a generation scenario where no coreference information is available and aims at assessing whether relatively naive context information is also useful.

4.1 System Description

We carry out our first experiment in a regeneration set-up with two components: a) a large-scale hand-crafted Lexical Functional Grammar (LFG) for German (Rohrer and Forst, 2006), used to parse and regenerate a corpus sentence, b) a stochastic ranker that selects the most appropriate regenerated sentence in context according to an underlying, linguistically motivated feature model. In contrast to fully statistical linearisation methods, our system first generates the full set of sentences that correspond to the grammatically well-formed realisations of the intermediate syntactic representation.[1] This representation is an f-structure, which underspecifies the order of constituents and, to some extent, their morphological realisation, such that the output sentences contain all possible combinations of word order permutations and morphological variants. Depending on the length and structure of the original corpus sentence, the set of regenerated sentences can be huge (see Cahill et al. (2007) for details on regenerating the German treebank TIGER).

[1]There are occasional mistakes in the grammar which sometimes lead to ungrammatical strings being generated, but this is rare.

The realisation ranking component is an SVM ranking model implemented with SVMrank, a Support Vector Machine-based learning tool (Joachims, 2006). During training, each sentence is annotated with a rank and a set of features extracted from the F-structure, its surface string and external resources (e.g. a language model). If the sentence matches the original corpus string, its rank will be highest, the assumption being that the original sentence corresponds to the optimal realisation in context. The output of generation, the top-ranked sentence, is evaluated against the original corpus sentence.

4.2 The Feature Models

As the aim of this experiment is to better understand the nature of sentence-internal features reflecting discourse context and compare them to sentence-external ones, we build several feature models which capture different aspects of the constituents in a given sentence. The sentence-internal features describe the morphosyntacic realisation of constituents, for instance their function ("subject", "object"), and can be straightforwardly extracted from the f-structure. These features are then combined into discriminative precedence features, for instance "subject-precedes-object". We implement the following types of morphosyntactic features:

- syntactic function (arguments and adjuncts)
- modification (e.g. nouns modified by relative clauses, genitive etc.)
- syntactic category (e.g. adverbs, proper nouns, phrasal arguments)
- definiteness for nouns
- number and person for nominal elements
- types of pronouns (e.g. demonstrative, reflexive)
- constituent span and number of embedded nodes in the tree

In addition, we also include language model scores in our ranking model. In Section 4.4, we report on results for several subsets of these features where "BaseSyn" refers to a model that only includes the syntactic function features and "FullMorphSyn" includes all features mentioned above.

For extracting the lexical chains, we check for any overlapping nouns in the n sentences previous to the current one being generated. We check

Rank	*Sentence and Features*
	% **Diese Gruppe** wird für einen Großteil der Gewalttaten verantwortlich gemacht.
	% This group is for a major part of the violent acts responsible made.
1	subject-<-pp-object, demonstrative-<-indefinite, overlap-<-no-overlap, overlap-in-vorfeld, lm:-7.89
	% Für einen Großteil der Gewalttaten wird **diese Gruppe** verantwortlich gemacht.
	% For a major part of the violent acts is this group responsible made.
3	pp-object-<-subject, indefinite-<-demonstrative, no-overlap-<-overlap, no-overlap-in-vorfeld, lm:-10.33
	% Verantwortlich gemacht wird **diese Gruppe** für einen Großteil der Gewalttaten.
	% Responsible made is this group for a major part of the violent acts.
3	subject-<-pp-object, demonstrative-<-indefinite, overlap-<-no-overlap, lm:-9.41

Figure 1: Made-up training example for realisation ranking with precedence features

proper and common nouns, considering full and partial overlaps as shown in Examples (1) and (2), where the (a) example is the previous sentence in the corpus. For each overlap, we record the following properties: (i) function in the previous sentence, (ii) position in the previous sentence (e.g. *Vorfeld*), (iii) distance between sentences, (iv) total number of overlaps.

These overlap features are then also combined in terms of precedence, e.g. "has_subject_overlap:3-precedes-no_overlap", meaning that in the current sentence a noun that was previously mentioned in a subject 3 sentences ago precedes a noun that was not mentioned before.

In Figure 1, we give an example of a set of generation alternatives and their (partial) feature representation for the sentence (1-b). Precedence is indicated by "<".

Basically, our sentence-external feature model is built on the intuition that lexical chains or overlaps approximate discourse status in a way which is similar to sentence-internal morphosyntactic properties. Thus, we would expect that overlaps indicate givenness, salience or prominence and that asymmetries between overlapping and non-overlapping entities are helpful in the ranking.

4.3 Data

All our models are trained on 7,039 sentences (subdivided into 1259 texts) from the TIGER Treebank of German newspaper text (Brants et al., 2002). We tune the parameters of our SVM model on a development set of 55 sentences and report the final results for our unseen test set of 240 sentences. Table 1 shows how many sentences in our training, development and test sets have at least one textually overlapping phrase in the previous 1–10 sentences.

We choose the TIGER treebank, which has no coreference annotation, since we already have a number of resources available to match the syntactic analyses produced by our grammar against the analyses in the treebank. Thus, in our regeneration system, we parse the sentences with the grammar, and choose the parsed f-structures that are compatible with the manual annotation in the TIGER treebank as is done in Cahill et al. (2007). This compatibility check eliminates noise which would be introduced by generating from incorrect parses (e.g. incorrect PP-attachments typically result in unnatural and non-equivalent surface realisations).

# Sentences in context	% Sentences with overlap		
	Training	Dev	Test
1	20.96	23.64	20.42
2	35.42	40.74	35.00
3	45.58	50.00	53.33
4	52.66	53.70	58.75
5	57.45	58.18	64.58
6	61.42	57.41	68.75
7	64.58	61.11	70.83
8	67.05	62.96	72.08
9	69.20	64.81	74.17
10	71.16	70.37	75.83

Table 1: The percentage of sentences that have at least one overlapping entity in the previous n sentences

For comparing the string chosen by the models against the original corpus sentence, we use BLEU, NIST and exact match. Exact match is a strict measure that only credits the system if it chooses the exact same string as the original corpus string. BLEU and NIST are more relaxed measures that compare the strings on the n-gram level. Finally, we report accuracy scores for the *Vorfeld* position (VF) corresponding to the percentage of sentences generated with a correct *Vorfeld*.

S_c	BLEU	NIST	Exact	VF
0	0.766	11.885	50.19	64.0
1	0.765	11.756	49.78	64.0
2	0.765	11.886	50.01	64.1
3	0.765	11.885	50.08	63.8
4	0.761	11.723	49.43	63.2
5	0.765	11.884	49.71	64.2
6	0.768	11.892	50.42	64.6
7	0.765	11.885	50.01	64.5
8	0.764	11.884	49.78	64.3
9	0.765	11.888	49.82	63.6
10	0.764	11.889	49.7	63.5

Table 2: Tenfold-crossvalidation for feature model FullMorphSyn and different context windows (S_c)

Model	BLEU	VF
Language Model	0.702	51.2
Language Model + Context $S_c = 5$	0.715	54.3
BaseSyn	0.757	62.0
BaseSyn + Context $S_c = 5$	0.760	63.0
FullMorphSyn	0.766	64.0
FullMorphSyn + Context $S_c = 5$	0.763	64.2

Table 3: Evaluation for different feature models; 'Language Model': ranking based on language model scores, 'BaseSyn': precedence between constituent functions, 'FullMorphSyn': entire set of sentence-internal features.

4.4 Results

In Table 2, we report the performance of the full sentence-internal feature model combined with context windows from zero to ten. The scores have been obtained from tenfold-crossvalidation. For none of the context windows, the model outperforms the baseline with a zero context which has no sentence-external features. In Table 3, we compare the performance of several feature models corresponding to subsets of the features used so far which are combined with sentence-external features respectively. We note that the function precedence features (i.e. the 'BaseSyn' model) are very powerful, leading to a major improvement compared to a language model. The sentence-external features lead to an improvement when combined with the language-model based ranking. However, this improvement is leveled out in the BaseSyn model.

On the one hand, the fact that the lexical chain features improve a language-model based ranking suggests these features are, to some extent, predictive for certain patterns of German word order. On the other hand, the fact that they don't improve over an informed sentence-internal baseline suggests that these patterns are equally well captured by morphosyntactic features. However, we cannot exclude the possibility that the chain features are too noisy as they conflate several types of lexical and coreferential relations. This will be adressed in the following experiment.

5 Experiment 2: Constituent Ordering with Centering-inspired Features

We now look at a simpler generation setup where we concentrate on the ordering of constituents in the German *Vorfeld* and *Mittelfeld*. This strategy has also been adopted in previous investigations of German word order: Filippova and Strube (2007) show that once the German *Vorfeld* is correctly chosen, the prediction accuracy for the *Mittelfeld* (the constituents following the finite verb) is in the 90s.

In order to eliminate noise introduced from potentially heterogeneous chain features, we look at coreference features and, again, compare them to sentence-internal morphosyntactic features. We target a generation scenario where coreference information is available. The aim is to establish an upper bound concerning the quality improvement for word order prediction by recurring to manual corefence annotation.

5.1 Data and Setup

We carry out the constituent ordering experiment on the Tüba-D/Z treebank (v5) of German newspaper articles (Telljohann et al., 2006). It comprises about 800k tokens in 45k sentences. We choose this corpus because it is not only annotated with syntactic analyses but also with coreference relations (Naumann, 2006). The syntactic annotation format differs from the TIGER treebank used in the previous experiment, for instance, it explicitely represents the *Vorfeld* and *Mittelfeld* as phrasal nodes in the tree. This format is very convenient for the extraction of constituents in the respective positions.

The Tüba-D/Z coreference annotation distinguishes several relations between discourse referents, most importantly "coreferential relation" and "anaphoric relation" where the first denotes a relation between noun phrases that refer to the same entity, and the latter refers to a link between a pronoun and a contextual antecedent, see Naumann (2006) for further detail. We expected the coreferential relation to be particularly useful, as

it cannot always be read off the morphosyntactic realisation of a noun phrase, whereas pronouns are almost always used in an anaphoric relation.

The constituent ordering model is implemented as a classifier that is given a set of constituents and predicts the constituent that is most likely to be realised in the *Vorfeld*.

The set of candidate constituents is determined from the tree of the original corpus sentence. We will assume that all constituents under a *Vorfeld* and *Mittelfeld* node can be freely reordered. Thus, we do not check whether the word order variants we look at are actually grammatical assuming that most of them are. In this sense, this experiment is close to fully statistical generation approaches. As a further simplification, we do not look at morphological generation variants of the constituents or their head verb.

The classifier is implemented with SVMrank again. In contrast to the previous experiment where we learned to rank sentences, the classifier now learns to rank constituents. The constituents have been extracted using the tool described in Bouma (2010). The final data set comprises 48.513 candidate sets of freely orderable constituents.

5.2 Centering-inspired Feature Model

To compare the discourse context model against a sentence-based model, we implemented a number of sentence-internal features that are very similar to the features used in the previous experiment. Since we extract them from the syntactic annotation instead of f-structures, some labels and feature names will be different, however, the design of the sentence-internal model is identical to the previous one in Section 4.

The sentence-external features differ in some aspects from Section 4, since we extract coreference relations of several types (see (Naumann, 2006) for the anaphoric relations annotated in the Tueba-D/Z). For each type of coreference link, we extract the following properties: (i) function of the antecedent, (ii) position of the antecedent, (iii) distance between sentences, (iv) type of relation. We also distinguish coreference links annotated for the whole phrase ("head link") and links that are annotated for an element embedded by the constituent ("contained link"). The two types are illustrated in Examples (4) and (5). Note that both cases would not have been captured in the lexical chain model since there is no lexical overlap between the realisations of the discourse referents.

	# VF	# MF
Backward Center	3.5%	5.1%
Forward Center	6.8%	6.8%
Coref Link	30.5%	23.4%

Table 4: Backward and forward centers and their positions

These types of coreference features implicitly carry the information that would also be considered in a Centering formalisation of discourse context. In addition to these, we designed features that explicitly describe centers as these might have a higher weight. In line with Clarke and Lapata (2010), we compute backward (CB) and forward centers (CF) in the following way:

1. Extract all entities from the current sentence and the previous sentence.
2. Rank the entities of the previous sentence according to their function (subject $<$ direct object $<$ indirect object ...).
3. Find the highest ranked entity in the previous sentence that has a link to an entity in the current sentence, this entity is the CB of the sentence.

In the same way, we mark entities as forward centers that are ranked highest in the current sentence and have a link to an entity in the following sentence.[2] In Table 4, we report the percentage of sentences that have backward and forward centers in the *Vorfeld* or *Mittelfeld*. While the percentage of sentences that realise a backward center is quite low, the overall proportion of sentences containing some type of coreference link is in a dimension such that the learner could definitely pick up some predictive patterns. Going by the relative frequencies, coreferential constituents have a bias towards appearing in the *Vorfeld* rather than in the *Mittelfeld*.

5.3 Results

First, we build three coreference-based constituent classifiers on their entire training set and compare them to their sentence-internal baseline. The most simple baseline records the category of

[2] In Centering, all entities in a given utterance can be seen as forward centers, however we thought that this implementation would be more useful.

(4) a. Die Rechnung geht an **die AWO**.
The bill goes to **the AWO**.
b. [Hintergrund der gegenseitigen Vorwürfe in **der Arbeiterwohlfahrt**] sind offenbar scharfe Konkurrenzen zwischen Bremern und Bremerhavenern.
Apparently, [the background of the mutual accusations at **the labour welfare***] are rivalries between people from Bremen and Bremerhaven.*

(5) a. Dies ist die Behauptung, mit der **Bremens Häfensenator** die Skeptiker davon überzeugt hat, [...].
This is the claim, which **Bremen's harbour senator** *used to convince doubters, [...].*
b. Für diese Behauptung hat **Beckmeyer** bisher keinen Nachweis geliefert. *So far,* **Beckmeyer** *has not given a prove of this claim.*

Model	VF
ConstituentLength + HeadPos	47.48%
ConstituentLength + HeadPos + Coref	51.30%
BaseSyn	54.82%
BaseSyn + Coref	56.21%
FullMorphSyn	57.24%
FullMorphSyn + Coref	57.40%

Table 5: Results from Vorfeld classification, training and evaluation on entire treebank

Model	VF
ConstituentLength + HeadPos	46.61%
ConstituentLength + HeadPos + Coref	52.23%
BaseSyn	54.63%
BaseSyn + Coref	56.67%
FullMorphSyn	55.36%
FullMorphSyn + Coref	57.93%

Table 6: Results from Vorfeld classification, training and evaluation on sentences that contain a coreference link

the constituent head and the number of words that the constituent spans. Additionally, in parallel to the experiment in Section 4, we build a "BaseSyn" model which has the syntactic function features, and a "FullMorphSyn" model which comprises the entire set of sentence-internal features. To each of these baseline, we add the coreference features. The results are reported in Table 5.

In this experiment, we find an effect of the sentence-external features over the simple sentence-internal baselines. However, in the fully spelled-out, sentence-internal model, the effect is, again, minimal. Moreover, for each baseline, we obtain higher improvements by adding further sentence-internal features than by adding sentence-external ones the accuracy of the simple baseline (47.48%) improves by 7.34 points through adding function features (the accuracy of BaseSyn is 54.82%) and by only 3.48 points through adding coreference features.

We run a second experiment in order to so see whether the better performance of the sentence-internal features is related to their coverage. We build and evaluate the same set of classifiers on the subset of sentences that contain at least one coreference link for one of its constituents (see Table 4 for the distribution of coreference links in our data). The results are given in Table 6. In this experiment, the coreference features improve over all sentence-internal baselines including the 'FullMorphSyn' model.

5.4 Discussion

The results presented in this Section consistently complete the picture that emerged from the experiments in Section 4. Even if we have high quality information about discourse context in terms of relations between referents, a non-trivial sentence-internal model for word order prediction can be hardly improved. This suggests that sentence-internal approximations of discourse context provide a fairly good way of dealing with local coherence in a linearisation task. It is also interesting that the sentence-external features improve over simple baselines, but get leveled out in rich sentence-internal feature models. From this, we conclude that the sentence-external features we implemented are to some extent predictive for word order, but that they can be covered by sentence-internal features as well.

Our second evaluation concentrating on the sentences that have coreference information shows that the better performance of the sentence-internal features is also related to their coverage. These results confirm our initial intuition that coreference information can add to the predictive power of the morpho-syntactic features in certain contexts. This positive effect disappears when sentences with and without coreferential constituents are taken together. For future work, it would be promising to investigate whether the

positive impact of coreference features can be strengthened if the coreference annotation scheme is more exhaustive, including, e.g., bridging and event anaphora.

6 Conclusion

We have carried out a number of experiments that show that sentence-internal models for word order are hardly improved by features which explicitely represent the preceding context of a sentence in terms of lexical and referential relations between discourse entities. This suggests that sentence-internal realisation implicitly carries a lot of information about discourse context. On average, the morphosyntactic properties of constituents in a text are better approximates of their discourse status than actual coreference relations.

This result feeds into a number of research questions concerning the representation of discourse and its application in generation systems. Although we should certainly not expect a computational model to achieve a perfect accuracy in the constituent ordering task – even humans only agree to a certain extent in rating word order variants (Belz and Reiter, 2006; Cahill, 2009) – the average accuracy in the 60's for prediction of *Vorfeld* occupance is still moderate. An obvious direction would be to further investigate more complex representations of discourse that take into account the relations between utterances, such as topic shifts. Moreover, it is not clear whether the effects we find for linearisation in this paper carry over to other levels of generation such as tactical generation where syntactic functions are not fully specified. In a broader perspective, our results underline the need for better formalisations of discourse that can be translated into features for large-scale applications such as generation.

Acknowledgments

This work was funded by the Collaborative Research Centre (SFB 732) at the University of Stuttgart.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, Boston,MA.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of EACL 2006*, pages 313–320, Trento, Italy.

Gerlof Bouma. 2010. Syntactic tree queries in prolog. In *Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*.

Aoife Cahill and Arndt Riester. 2009. Incorporating information status into generation ranking. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 817–825, Suntec, Singapore, August. Association for Computational Linguistics.

Aoife Cahill, Martin Forst, and Christian Rohrer. 2007. Stochastic Realisation Ranking for a Free Word Order Language. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 17–24, Saarbrücken, Germany. DFKI GmbH.

Aoife Cahill. 2009. Correlating human and automatic evaluation of a german surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 97–100, Suntec, Singapore, August. Association for Computational Linguistics.

Jackie C.K. Cheung and Gerald Penn. 2010. Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Stefanie Dipper and Heike Zinsmeister. 2009. The role of the German Vorfeld for local coherence. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten/From Form to Meaning: Processing Texts Automatically*, pages 69–79. Narr, Tübingen.

Katja Filippova and Michael Strube. 2007. The german vorfeld and local coherence. *Journal of Logic, Language and Information*, 16:465–485.

Katja Filippova and Michael Strube. 2009. Tree Linearization in English: Improving Language Model Based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 225–228, Boulder, Colorado, June. Association for Computational Linguistics.

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

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226.

Nikiforos Karamanis, Massimo Poesioand Chris Mellish, and Jon Oberlander. 2009. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1).

Jane Morris and Graeme Hirst. 1991. Lexical cohesion, the thesaurus, and the structure of text. *Computational Linguistics*, 17(1):21–225.

Karin Naumann. 2006. Manual for the annotation of in-document referential relations. Technical report, Seminar für Sprachwissenschaft, Abt. Computerlinguistik, Universität Tübingen.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*.

Massimo Poesio, Rosemary Stevenson, Barbara di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Eric K. Ringger, Michael Gamon, Robert C. Moore, David Rojas, Martine Smets, and Simon Corston-Oliver. 2004. Linguistically Informed Statistical Models of Constituent Structure for Ordering in Sentence Realization. In *Proceedings of the 2004 International Conference on Computational Linguistics*, Geneva, Switzerland.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the the 6th LREC conference*.

Christian Rohrer and Martin Forst. 2006. Improving Coverage and Parsing Quality of a Large-Scale LFG for German. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Augustin Speyer. 2005. Competing constraints on vorfeldbesetzung in german. In *Proceedings of the Constraints in Discourse Workshop*, pages 79–87.

Heike Telljohann, Erhard Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2006. Stylebook for the tübingen treebank of written german (tüba-d/z). revised version. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Erik Velldal and Stephan Oepen. 2005. Maximum entropy models for realization ranking. In *Proceedings of the 10th Machine Translation Summit*, pages 109–116, Thailand.