An LFG-type Grammar for German Based on the Topological Model

Lionel Clément Kim Gerdes Sylvain Kahane
Lionel.Clément@inria.fr kim@linguist.jussieu.fr sk@ccr.jussieu.fr
INRIA Rocquencourt
Domaine de Voluceau – BP 105
78153 Le Chesnay – France
Lattice
Université Paris 7 - case 7003
75251 Paris Cedex 05 – France

Proceedings of the LFG02 Conference
National Technical University of Athens, Athens
Miriam Butt and Tracy Holloway King (Editors)
2002
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

This paper proposes a description of German word order in a LFG-type grammar. Contrary to earlier Lexical Functional Grammar description of German, our grammar uses as c-structure the topological model. This gives us a simpler grammar, which covers (partial) VP fronting, intraposition, extraposition, auxiliary flip and all other order possibilities for a verbal dependent. Our grammar is implemented in an LFG parser.

Introduction

The aim of this article is to propose a new description of German word order in the LFG formalism. Contrary to previous attempts, we explicitly use the topological model of German as the basis of our analysis.

The topological model subdivides the sentence into a hierarchy of topological domains that are themselves composed of fields (*Vorfeld*, *Mittelfeld*, *right bracket*, ...) to which specific rules are associated (Drach 1937) (Bech 1955). It has been successfully used in HPSG (Reape 1994) (Kathol 1995) and in dependency grammars (Duchier and Debusmann 2001) (Gerdes and Kahane 2001). The present analysis is based on the description by (Gerdes and Kahane 2001), who use the hierarchy of topological domains as the only phrase structure. Translating this to LFG, we give the c-structure a topological interpretation. Therefore, contrary to earlier Dutch/German LFGs (Bresnan et al. 1982) (Zaenen and Kaplan 1995), we do not use X-bar syntax notations for our c-structure. Indeed, the topological notation corresponds better to the idea underlying LFG that the c-structure does not represent syntactic information such as subcategorization contrary to the X-bar phrase structure.

In Section 1, we propose some enrichments of the f-structure which avoid a multiplication of the phrase structure rules. The Topological model is presented in Section 2, and we present and comment the phrase structure rules of our German LFG in Section 3.

1 f-structure and dependency

We argue that word order in German (as in any natural language) depends exclusively on the syntactic links between words (the dependency tree) and the information structure (the communicative grouping of words). It has long been noted that the f-structure is closely related to classic syntactic dependency trees (Tesnière 1959) (Mel'čuk 1988). Nevertheless, there are some differences, starting with the fact that the f-structure is not in general a tree. It contains information of various types, in particular syntactic and semantic relationships. As an example consider the analysis of sentence (1) and its usual f-structure given in Figure 1.

(1) *The man seems to have slept.*

$$
\begin{bmatrix}
\text{pred} & \text{`seem}\langle\text{xcomp}\rangle\text{ subj'} \\
\text{subj} & \begin{bmatrix} \text{pred} & \text{`man'} \\ \text{gender} & \text{masc} \\ \text{number} & \text{sg} \\ \text{person} & 3 \\ \text{def} & + \end{bmatrix} \\
\text{xcomp} & \begin{bmatrix} \text{pred} & \text{`sleep}\langle\text{subj}\rangle\text{'} \\ \text{subj} & \\ \text{tense} & \text{past} \end{bmatrix} \\
\text{mode} & \text{indicative} \\
\text{tense} & \text{present}
\end{bmatrix}
$$

Figure 1: The usual f-structure for (1)

Traditionally, LFG's pred(icate) feature distinguishes between semantic arguments (carrying θ-roles) and purely syntactic arguments, by placing them inside and outside of the predicate's brackets. For instance, the subject of *seems* is not a semantic argument and is noted outside of the brackets. Yet, an argument that is only semantically and not syntactically motivated does not receive a special notation in the classical f-structure. The subject position of the infinitive, for example, is not syntactic; it encodes a more profound relation. In this work, we attempt to encode this difference explicitly by marking exclusively semantic arguments with a dot before the grammatical function. Figure 2 shows the f-structure we propose for the sentences in (1): The subject of *slept*/'sleep', which is only a semantic argument, is noted •subj.

Another problematic point concerns the status of the auxiliary: Today's customary analysis in LFG does not assign an independent predicator to the auxiliary.[1] It only verifies agreement features with the subject while adding grammatical morphemes to the main verb's feature structure. On a semantic level, we agree to treat complex verbal forms (*is sleeping*, *have slept*) in the same way as simple verbal forms (*sleeps*, *slept*); yet syntactically, they behave just like full verbs with verbal subcategorization. In German, for example, the complex verbal form *geschlafen haben* 'to have slept' has the same word order realizations as the control construction *schlafen wollen* 'to want to sleep'. Thus, in order to capture this common behavior, we introduce predicators for auxiliaries just as for control and raising verbs. We obtain a purely syntactic 'predicate' that has no semantic sub-categorization (and therefore no semantic brackets). Nevertheless, we want to indicate that the auxiliary and the participle will form one node in the semantic representation. Thus, the *xcomp* relation of the auxiliary receives a special mark, a circle, indicating that the two nodes it connects form a single semantic unit (Figure 2a,b,d). These new notations in the predicate function are completely compatible with the formalism itself, and moreover, they subsume the traditional analysis. However, our description allows distinguishing explicitly syntactic from semantic dependencies.

In our transcription of the predicate functions in Figure 2b, we denote with In Figure 2b, we indicate the predicate argument structures of the f-structure, depicting the semantic links with a dotted line and the syntactic links with a bold line; an arrow denotes the grammatical function that will be incorporated with another predicate when constructing the semantic graph 2c. Note that the subject link between 'have' and 'man' is neither syntactic nor semantic and simply ensures

[1] In (Bresnan 1982), auxiliaries still held a predicate function. In this sense, the f-structure has developed a more semantic ambition.

the percolation of the value of the syntactic subject of ‘seem’ to the semantic subject of ‘sleep’.

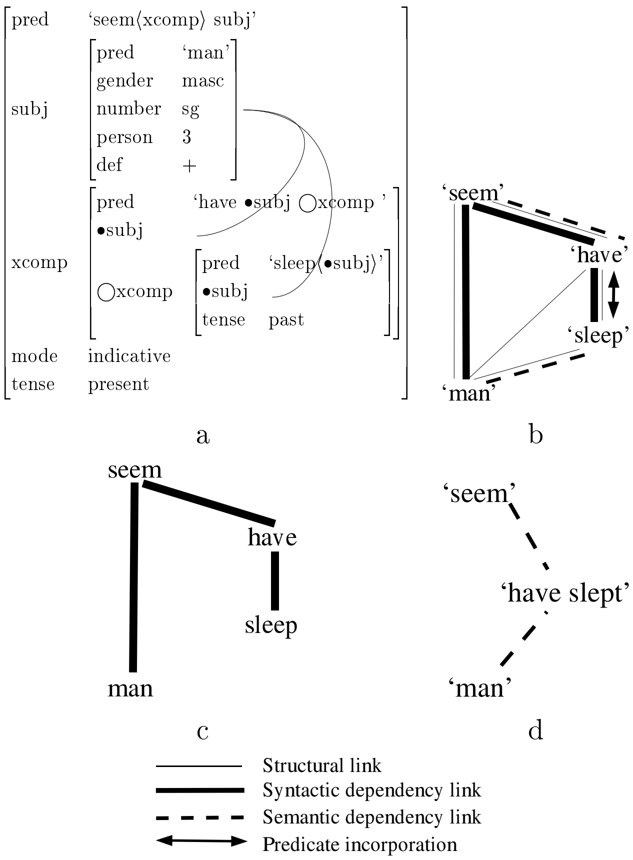


Figure 2: (a) Our f-structure for (1), (b) the underlying predicate-argument structure, and (c) the corresponding syntactic and (d) semantic structure.

2 German word order and the Topological model

Word order in German is much freer than in English. The f-structure of Figure 3, which will be our reference example, has a few dozen linearizations:

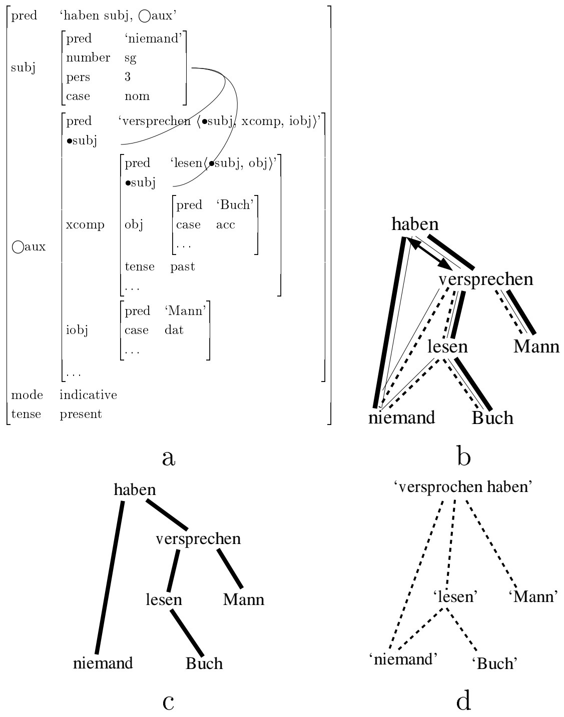


Figure 3: (a) An f-structure, (b) the graphical representation of its predicate functions, (c) the corresponding syntactic dependency tree, and (d) the semantic graph for the sentences in (2).

(2) a. *Niemand hat diesem Mann das Buch zu lesen versprochen*
nobody$_{nom}$ has this man$_{dat}$ the book$_{acc}$ to read promised
'Nobody promised this man to read the book.'

b. *Diesem Mann hat das Buch niemand zu lesen versprochen*

c. *Das Buch zu lesen hat diesem Mann niemand versprochen*

d. *Diesem Mann hat niemand versprochen, das Buch zu lesen*

e. *Diesem Mann hat, das Buch zu lesen, niemand versprochen*

f. *Zu lesen hat diesem Mann das Buch niemand versprochen*

g. *Das Buch hat niemand diesem Mann versprochen zu lesen*

2.1 The Topological model

To get all possible orders, we follow closely the classical topological model. This model supposes that a German sentence consists of a *main domain* composed of a sequence of five *fields*: *Vorfeld, left bracket, Mittelfeld, right bracket,* and *Nachfeld*. A domain is a constituent whose ordered compartments, called fields, can themselves accommodate new constituents.

Main domain

Vorfeld	left bracket	Mittelfeld	right bracket	Nachfeld

Figure 4: Fields of the main domain.

We start our description of word order with the root of the syntactic tree, whatever its semantic type (main verb, modal, or auxiliary); in our example in Figure 3, it is *hat* 'haben'. This word always takes the left bracket, which is the second position. A non-verbal syntactic dependent of the main verb, like *niemand* in our example, goes into one of the major fields, that is Vorfeld, Mittelfeld and Nachfeld. The case of a verbal syntactic dependent is more complicated: A non-finite verbal dependent of the main verb (here the past participle *versprochen* 'promised') can open two kinds of constituents:

- The first possibility is to open an *embedded domain*, consisting of three fields – Mittelfeld, right bracket, and Nachfeld. This embedded domain can accommodate all of the embedded verb's syntactic dependents.
- The second possibility is to construct a restricted phrase, called the *verb cluster*, with only one position for a verbal dependent at its left.

These two kinds of phrases must be placed in very different topological positions: an embedded domain goes to one of the major fields (just as a non verbal dependent)[2], while the verb cluster takes the right bracket. The verb in the right bracket can again have a dependent element. This dependent behaves exactly as a dependent of the main verb (in the left bracket): If it is a non verbal dependent (*diesem Mann*), it goes in one of the major fields;[3] if it is a verbal dependent (*zu lesen*),

[2]This is true for infinitives with *zu* 'to'. Bare infinitives and past participles can only create a new domain in the Vorfeld.

[3]This can be a major field of the domain of its governor, or a major field of a higher domain (2f,g). The subject, which is only a semantic dependent of the past participle, cannot join the embedded domain (except for ergative verbs and passive constructions).

either it creates an embedded domain in one of the major fields – in the Vorfeld for (2c,f), in the Mittelfeld for (2e) and in the Nachfeld for (2d,g) – or it joins its governor in the right bracket (2a,b). In this latter case, it takes the position at the left of its syntactic governor and creates a new restricted constituent, we call a *verb box*. This verb box proposes again only one position at its left reserved for a possible verbal dependent.

We show in the remaining part of this section how two other difficulties of German word order, separable verbal prefixes and auxiliary flip, can easily be integrated in the topological model.

2.2 Separable prefixes

Some verbal constructions such as *anfangen* consist of two parts: a verb, *fangen*, and a so-called separable prefix, *an*.

(3) a. *Die Schule hat um 9 Uhr angefangen*
The school$_{nom}$ has at 9 o'clock on_caught
'The school starts at 9AM.'

b. *Er fängt gleich zu schreien an*
He catches right away to shout on
'He begins to shout right away.'

The prefix *an* behaves just like a verbal dependent of *fangen*, i.e. it goes into the right bracket of the main domain.[4] When *fangen* is in the right bracket, the prefix *an* goes to its left, as a verbal dependent would do (3a). If *anfangen* has a verbal dependent, this dependent behaves as a syntactic dependent of *an* and, in particular, it can join the right bracket taken by *an* (3b).

For these reasons, we treat the verb and its prefix as two syntactic units although the writing conventions of German require joining them graphically when they are next to each other, and semantically, they clearly form an entity. This last point is taken care of in our LFG grammar by marking their syntactic link as semantic incorporation.

2.3 Auxiliary flip

Another difficulty of German verb placement is known as auxiliary flip (or *Oberfeldumstellung*).

(4) *Er wird das Buch haben lesen können*
He will the book have read can
'He will have been able to read the book.'

[4] The separable prefix cannot open an embedded domain. In case of contrast with another prefix, it can go in the Vorfeld.

Contrary to the usual order of verbal sub-categorization in the verb cluster (V_1, V_2V_1, $V_3V_2V_1$, etc.), the Oberfeldumstellung allows the auxiliaries *werden* and *haben* to place their verbal dependent to their right (V_1V_2). We handle this possibility by allowing auxiliaries to open a field for their dependents not only to their left (Oberfeld) but also to their right (Unterfeld).[5] The subsequent verbal dependent can join its governor's Oberfeld ($V_1V_3V_2$ - see example 4) or even the auxiliary's Oberfeld ($V_3V_1V_2$, *Zwischenstellung*).

We would like to stress that all the above word order rules are exclusively based on the syntactic dependencies and semantic dependencies do not intervene.

We will now present our formalization of this topological analysis in LFG.

3 Formalization

The topological model distinguishes constituents and fields. The purpose of fields is to provide a sentence position for different types of constituents, that otherwise would have to be encoded in a large number of different rewriting rules. For instance the Vorfeld can hold very different kind of constituents like nominal, verbal, adjunct, and complementizer phrases, and the same kind of constituents can go in the Mittelfeld and the Nachfeld.

We can see the constituents as boxes, one put into the other; the fields are the compartments of these boxes. In our grammar, this dichotomy is reflected in two types of labels, one for constituents, behaving in the usual way, the other for fields, passing up automatically all of the functional information. Moreover, fields do not verify the constraint that force the f-structure associated to a constituent to have a *pred* feature. In our notation, field names are preceded by underscores (following the notation of the XLFG parser).

In 3, we present the simplified phrase structure rules for our German LFG. This grammar covers only a fragment of German, leaving aside, among others, the internal structure of NPs, relative phrases, and complementizers.

[5] The Unterfeld can only be taken by an infinitive, and the past participle has to surface as an *ersatz-infinitive* if it goes into the Unterfeld. Furthermore, the governed verbs V_2 taking the Unterfeld form a closed class including modal and perception verbs (and some others like *helfen*, 'help', the causative/permissive *lassen* 'make/let' ... – *haben* 'have' itself also allows this right-placement, which suffices to explain the cases of 'double flip' giving $V_1V_2V_3$, $V_1V_2V_4V_3$).

Phrase structure rules for German:[6]

MD → _VF _LB (_MF) (_RB) (_NF)
ED → (_MF) _RB (_NF)
_VF → _XF
_VF → ED
    $(\uparrow \text{xcomp}) = \downarrow$
_XF → ADVP
    $(\uparrow \mathit{xcomp}^{*}\ \mathit{adjunct}) \ni \downarrow$
_LB → V
    $\uparrow=\downarrow$
_MF → _XF*
_RB → VC
    $\uparrow=\downarrow$
_RB → VC
    $(\uparrow \mathit{xcomp}) =\downarrow$
_NF → _XF*
_XF → ED
    $(\uparrow \mathit{xcomp}^{+}) =\downarrow$
_XF → NP
    $(\uparrow \mathit{subj}) =\downarrow$
_XF → NP
    $(\uparrow \mathit{xcomp}^{*}\ \{\mathit{obj}, \mathit{iobj}\}) =\downarrow$
VC → (_O) _H (_U)
VB → (_O) _H
_H → V
    $\uparrow=\downarrow$
_O → VB
    $(\uparrow \mathit{xcomp}^{+}) =\downarrow$[7]
_U → VC
    $(\uparrow \mathit{xcomp}) =\downarrow$
    $(\uparrow \mathit{type}) = \mathit{aux}$
    $(\downarrow \mathit{modal}) = +$
    $(\downarrow \mathit{ersatz}) = +$
    $(\downarrow \mathit{tense}) \neq$ zu-inf

Our initial symbol is the main domain (MD). In our grammar, the embedded domain (ED) takes the place of verbal phrases

[6] We suppose that cases (nom, acc, dat) are imposed in subcategorization rules in the lexicon. Nevertheless, it could be possible to consider them as grammatical rules. In such case, we would add cases in order rules (nom for subj, acc for obj, and dat for iobj). It must be remarked that some verbs have a gen subject or two accusative complements.

[7] If we left out the + sign, we could exclude the Zwischenstellung, not accepted by all German natives.

for infinitives and past participles. Other projections of verbs, which cannot take noun phrases, are the verb cluster (VC) and the verbal box (VB).

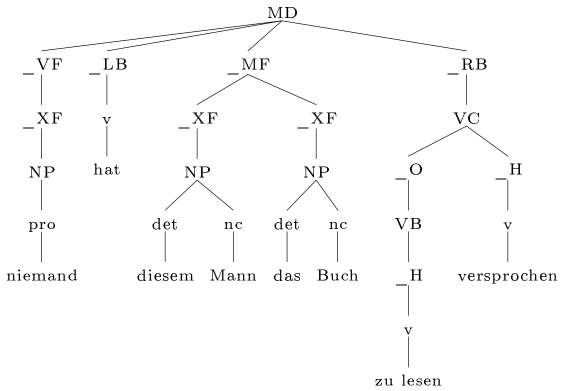
We note fields that can remain empty as optional. In order to generalize the similar behavior of the three major fields – Vorfeld (_VF), Mittelfeld (_MF) and Nachfeld (_NF) – in the LFG formalism, we have to introduce the additional label _XF, although it does not really have the status of a field.

As usual, long-distance dependencies are taken care of by functional uncertainty. We easily obtain topicalization (placement in the Vorfeld, including VP fronting) and scrambling (mixed order of complements of different verbs in the Mittelfeld).

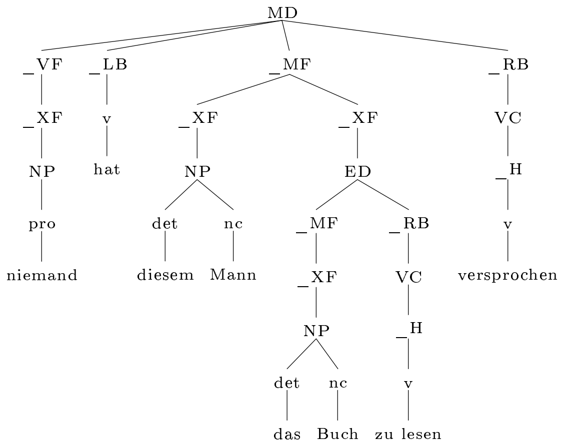
The placement of a verbal subordinative chain (a *xcomp* chain) into a topological domain obeys a simple principle: We place the verbs in descending order. The left bracket (_LB) is taken at first, then the right bracket (_RB), then the Oberfeld (_O) or the Unterfeld (_U) inside this right bracket, and so on. In the case of an embedded domain, which does not have a left bracket, the same principle applies, starting with the placement in the right bracket. Any phrase structure based formalism as LFG and HPSG does not allow expressing this generality, and our grammar thus contains two rules for the right bracket; one for the right bracket of the main domain, where the verb's f-structure gets the *xcomp* position of the main f-structure, and a second rule for the right bracket of the embedded domain, where the verb is the head of the domain.

Our analysis of the verb cluster allows all the correct orderings, in particular Oberfeldumstellung and Zwischenstellung. It is easy to exclude this last possibility, only accepted by parts of the native German speakers, by suppressing the + in the functional equation of the verb box (VB) in the rule of the Oberfeld (_O).

(a)



(b)



(c)

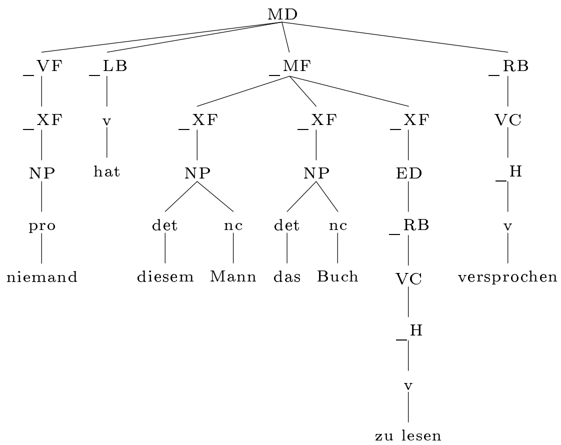


Figure 5: Three c-structures of (2a) corresponding to different topological (and communicative) structures.

Our grammar produces the desired f-structure (Figure 3a) for sentence (2a) using three different c-structures (Figure 5a,b,c). In the first analysis, *zu lesen versprochen* forms a verb cluster. This corresponds to the possibility that the two verbs form a unique prosodic pattern, with stress only on the first syllable of the first verb's radical. The second analysis reflect the fact, that embedded domains can be placed in the Mittelfeld just as in example (2e), where the verb cluster analysis is impossible. Here, the embedded domain forms an independent prosodic group. The last c-structure corresponds to the additional possibility that the embedded domain does not contain all of its verbs complements, as in (2f,g), where this is the only possible analysis. The existence of these different prosodic patterns for the same word order has been demonstrated in (Gerdes and Kahane 2002). They argue that these groupings correspond to different information structures.

Conclusion

We have shown that it is possible to create a Lexical Functional Grammar for German based on the classical topological model. We obtain a grammar of remarkable simplicity that can handle an important number of grammatical phenomena, usually considered as complex. The grammar is implemented in XLFG (Clément and Kinyon 2001), version 3.4.3, for the moment only with a toy lexicon.

In this study we are not concerned with coordination and with the order in the Mittelfeld. We believe these restrictions to be of a different nature: (Lenerz 1977) and (Choi 1999) among others have shown that the German ordering constraints for the Mittelfeld depend mostly on the information structure of the sentence. However in Dutch, a closely related language with a very similar topological structure, the word order in the Mittelfeld is constraint primarily by the syntactic position. This requires an enrichment of the LFG formalism (the *f-precedence*) proposed by (Zaenen and Kaplan 1995).

The transfer of the topological model into the LFG formalism gave us the opportunity to reexamine the theoretical status of the two principal structures of LFG:

- The f-structure can differentiate syntactic and semantic dependencies.
- The c-structure can encode word order and (prosodic and informational) groupings of words. We obtain a c-structure that is completely liberated from its functional burden inherited from X-bar syntax.

It comes out that the clear distinction of a topological level, a syntactic level, and a semantic level is just as useful for an adequate linguistic description as for an economic formalization.

We hope that this study can contribute to a convergence of various formalisms that can handle a topological description of German word order, like LFG, HPSG and dependency grammars.

References

Bech, G. 1955. *Studien über das deutsche Verbum infinitum.* Tübingen: Niemeyer.

Bresnan, J. 1982. Control and Complementation. *Linguistic Inquiry* 13(3). Also in Joan Bresnan, ed. 1982. The Mental Representation of Grammatical Relations, Chapter 5, pp. 282–390. Cambridge, MA: The MIT Press.

Bresnan, J., R. M. Kaplan, S. Peters, and A. Zaenen. 1982. Cross-serial dependencies in Dutch. *Linguistic Inquiry* 13(4):613–635. Reprinted in W. Savitch et al. (eds) The Formal Complexity of Natural Language, 286–319. Dordrecht: D. Reidel.

Choi, H.-W. 1999. *Optimizing Structure in Context - Scrambling and Information Structure.* Stanford: CSLI Publications.

Clément, L., and A. Kinyon. 2001. XLFG-an LFG parsing scheme for French. In *LFG 2001*, Hong Kong. http://www.lionel-clement.net/xlfg/.

Drach, E. 1937. *Grundgedanken der deutschen Satzlehre.* Frankfurt: Diesterweg.

Duchier, D., and R. Debusmann. 2001. Topological Dependency Trees: A Constraint-Based Account of Linear Precedence. In *ACL 2001*, Toulouse.

Gerdes, K., and S. Kahane. 2001. Word Order in German: A Formal Dependency Grammar Using a Topological Hierarchy. In *ACL 2001*, Toulouse.

Gerdes, K., and S. Kahane. 2002. Phrasing it differently. In L. Wanner (Ed.), *Works in Meaning-Text Theory in honour of Igor Mel'čuk*, Philadelphia. Benjamins.

Kathol, A. 1995. *Linearization-based German Syntax*. PhD thesis, Ohio State University.

Lenerz, J. 1977. Zur Abfolge nominaler Satzglieder im Deutschen. In *TBL Verlag Günter Narr*, Tübingen.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. New-York: SUNY Press.

Reape, M. 1994. Domain Union and Word Order Variation in German. In J. N. et al. (Ed.), *German in Head-Driven Phrase Structure Grammar*. Stanford: CSLI Lecture Notes.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Klienksieck.

Zaenen, A., and R. M. Kaplan. 1995. Formal devices for linguistic generalizations: West Germanic word order in LFG. In J. S. Cole, G. M. Green, and J. L. Morgan (Eds.), *Linguistics and Computation*, 3–27. Stanford, CA: CSLI Publications.