Annotating Unrestricted German Text

Wojciech Skut, Thorsten Brants, Brigitte Krenn, Hans Uszkoreit
Universität des Saarlandes, Computational Linguistics
D-66041 Saarbrücken, Germany
{skut,brants,krenn,uszkoreit}@coli.uni-sb.de

In Proceedings der 6. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft, 1997, Heidelberg, Germany

Abstract

This paper discusses the development of an annotation scheme for unrestricted German text. We argue for a uniform representation format based on argument structure but allowing us to recover other kinds of representations. We also discuss several methodological issues and the analysis of some phenomena.

The presented annotation format has been successfully tested in corpus annotation.

1 Corpora for Data-Driven NLP

An important pardigm shift is currently taking place in linguistics and language technology. Purely introspective research focussing on a limited number of isolated phenomena is being replaced by a more data-driven view of language. The growing importance of stochastic methods opens new avenues for dealing with the wealth of phenomena found in real texts, especially phenomena requiring a model of preferences or degrees of grammaticality.

This new research paradigm requires very large corpora annotated with different kinds of linguistic information. Since the main objective here is rich, transparent and consistent annotation rather than putting forward hypotheses or explanatory claims, the following requirements are often stressed:

descriptivity: phenomena should be described rather than explained as explanatory mechanisms can be derived (induced) from the data.

data-drivenness: the formalism should provide means for representing all types of grammatical constructions occurring in the corpus[1].

theory-neutrality: the annotation format should not be influenced by theory-internal considerations. However, annotations should contain enough information to permit the extraction of theory-specific representations.

In addition, the architecture of the annotation scheme should make it easy to refine the information encoded, both in width (adding new description levels) and depth (refining existing representations). Thus a structured, multi-stratal organisation of the representation formalism is desirable.

The representations themselves have to be easy to determine on the basis of simple empirical tests, which is crucial for the consistency and a reasonable speed of annotation.

2 Why Tectogrammatical Structure?

In the data-driven approach, the choice of a particular representation formalism is an engineering problem rather than a matter of 'adequacy'. More important is the theory-independence and reusability of linguistic knowledge, i.e., the recoverability of theory/application specific representations, which in the area of NL syntax fall into two classes:

Phenogrammatical structure: the structure reflecting surface order, e.g. *constituent structure* or topological models of surface syntax, cf. (Ahrenberg, 1990), (Reape, 1994)).

Tectogrammatical representations: predicate-argument structures reflecting lexical argument structure and providing a guide for assembling meanings. This level is present in almost every theory: D-structure (GB), f-structure (LFG) or argument structure (HPSG). A theory based mainly on tectogrammatical notions is dependency grammar, cf. (Tesnière, 1959).

As annotating both structures separately presents substantial effort, it is better to recover constituent structure automatically from an argument structure treebank, or vice versa. Both alternatives are discussed in the following sections.

2.1 Annotating Constituent Structure

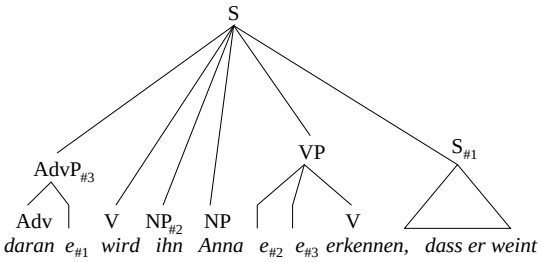
Phenogrammatical annotations require an additional mechanism encoding tectogrammatical struc-

[1] This is what distinguishes corpora used for grammar induction from other collections of language data. For instance, so-called *test suites* (cf. (Lehmann et al., 1996)) consist of typical instances of selected phenomena and thus focus on a subset of real-world language.

ture, e.g., trace-filler dependencies representing discontinuous constituents in a context-free constituent structure (cf. (Marcus et al., 1994), (Sampson, 1995)). A major drawback for annotation is that such a hybrid formalism renders the structure less transparent, as is the phrase-structure representation of sentence (1):

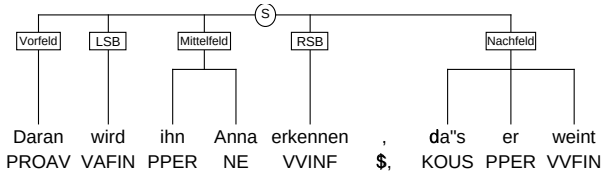
(1) daran wird ihn Anna erkennen, dass er weint
at-it will him Anna recognise that he cries

'Anna will recognise him at his cry'



Furthermore, the descriptivity requirement could be difficult to meet since constituency has been used as an explanatory device for several phenomena (binding, quantifier scope, focus projection).

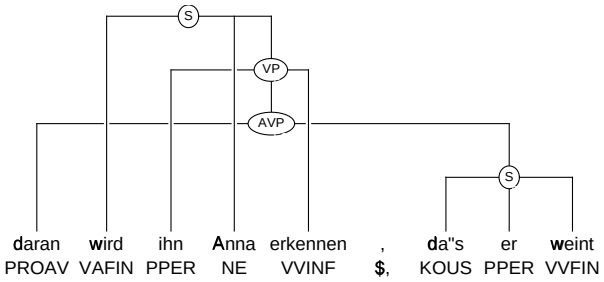
The above remarks carry over to other models of *phenogrammatical structure*, e.g. *topological fields*, cf. (Bech, 1955). A sample structure is given below[2]



Here, as well, topological information is insufficient to express the underlying tectogrammatical structure (e.g., the attachment of the extraposed that-clause)[3]. Thus the *field model* can be viewed as a non-standard phrase-structure grammar which needs additional tectogrammatical annotations.

2.2 Argument Structure Annotations

An alternative to annotating surface structure is to directly specify the tectogrammatical structure, as shown in the following figure:



This encoding has several advantages. Local and non-local dependencies are represented in a uniform way. Discontinuity does not influence the hierarchical structure, so the latter can be determined on the basis of lexical subcategorisation requirements, agreement and some semantic information.

An important advantage of tectogrammatical structure is its proximity to semantics. This kind of representations is also more theory-neutral since most differences between syntactic theories occur at the phenogrammatical level, the tectogrammatical structures being fairly similar.

Furthermore, a constituent tree can be recovered from a tectogrammatical structure (cf. section 4). Thus tectogrammatical representations provide a uniform encoding of information for which otherwise both constituent trees *and* trace-filler annotations are needed.

Apart from the work reported in this paper, tectogrammatical annotations have been successfully used in the TSNLP project to construct a language competence database (cf. (Lehmann et al., 1996)).

2.3 Suitability for German

Further advantages of tectogrammatical annotations have to do with the fairly weak constraints on German word order, resulting in a good deal of discontinuous constituency. This feature makes it difficult to come up with a precise notion of constituent structure. In the effect, different kinds of structures are proposed for German, the criteria being often theory-internal[4].

In addition, phrase-structure annotations augmented with the many trace-filler co-references would lack the transparency desirable for ensuring the consistency of annotation.

3 Methodology

The standard methodology of determining constituent structure (e.g., the *Vorfeld* test) does not carry over to tectogrammatical representations, at least not in all its aspects. The following sections are thus concerned with methodological issues.

[2] LSB, RSB stand for *left* and *right sentence bracket*.

[3] Even annotating grammatical functions is not enough as long as we do not explicitly encode their tectogrammatical attachment of such functions.

[4] Flat or binary right-recursive structures, not to mention the status of the head in verb-initial, verb-second and verb-final clauses, cf. (Netter, 1992), (Kasper, 1994), (Nerbonne, 1994), (Pollard, 1996).

3.1 Structures vs. Labels

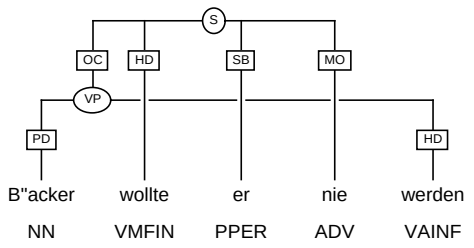
The first question to be answered here is how much information has to be encoded structurally. Rich structures usually introduce high spurious ambiguity potential, while flat representations (e.g., category or function labels) are significantly easier to manipulate (alteration, refinement, etc.).

Thus it is a good strategy to use rather simple structures and express more information by labels.

3.2 Structural Representations

As already mentioned, tectogrammatical structures are often thought of in terms of *dependency grammar* (DG, cf. (Hudson, 1984), (Hellwig, 1988)), which might suggest using conventional *dependency trees* (stemmas) as our representation format. However, this would impose a number of restrictions that follow from the theoretical assumptions of DG. It is mainly the DG notion of heads that creates problems for a flexible and maximally theory-neutral approach. In a conventional dependency tree, heads have to be unique, present and of lexical status, requirements other theories might not agree with.

That is why we prefer a representation format in which heads are distinguished outside the structural component, as in the figure below[5]:



The tree encodes three kinds of information:

tectogrammatical structure: trees with possibly crossing branches (no non-tangling condition);

syntactic category: node labels and part-of-speech tags (Stuttgart-Tübingen Tagset, cf. (Thielen and Schiller, 1995)).

functional annotations: edge labels.

3.3 Classification of Labels

Compared to the fairly simple structures employed by our annotation scheme, the functional annotations encode a great deal of linguistic information. We have already stressed that the notion *head* is distinguished at this level. Accordingly, it seems to be the appropriate stratum to encode the differences between different classes of dependencies.

For instance, most linguistic theories distinguish between complements and adjuncts. Unfortunately, the theories do not agree on the criteria for drawing the line between the two classes of dependents. To this date there is no single combination of criteria such as category, morphological marking, optionality, uniqueness of role filling, thematic role or semantic properties that can be turned into a transparent operational distinction linguists of different schools would subscribe to.

In our scheme, we try to stay away from a theoretical commitment concerning borderline decisions. The distinction between functional labels such as SB and DA – standing for traditional grammatical functions – on the one hand and phrases labelled MO on the other should not be interpreted as a classification into complements and adjuncts. For the time being, functional labels different from MO are assigned only if the grammatical function of the phrase can easily be detected on the basis of the linguistic data. MO is used, e.g., to label adjuncts as well as prepositional objects. Likewise the label OC is used for easily recognisable clausal complements. Other embedded sentences depending on the verb are labelled as MO[6]. This is consistent with our philosophy of stepwise refinement. We are in the process of designing a more fine-grained classification of functional labels together with testable criteria for assigning them. This classification will not contain a distinction between complements and adjuncts. Thus the locative phrase *in Berlin* in the sentence *Peter wohnt in Berlin* will just be marked as a locative MO with the category PP. As linguistic theories disagree on the question, we will not ask the annotators to decide whether this phrase is a complement of the verb.

This strategy differs from the one pursued by the creators of the Penn Treebank. There the difference between complements and adjuncts is encoded in the hierarchical structure. Verbal complements are encoded as siblings of the verb whereas adjuncts are adjoined at a higher level. In a case of doubt, the annotators are asked to select adjunction. We consider this structural encoding less suitable for refinement than a hierarchy of functional labels in which MO can be further specified by sublabels.

3.4 Determining the Structure

It is obvious that word-order based constituency tests are little use for our approach. Instead we have to rely on other criteria such as argument selection, case, agreement and semantic information.

However, these criteria sometimes conflict. For instance, passivisation and argument raising present a discrepancy between case assignment and seman-

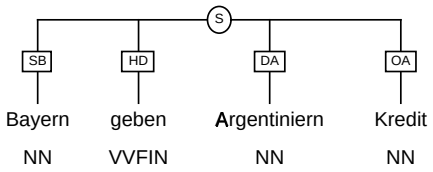
[5] Edge labels: HD head, SB subject, OC clausal complement, PD predicative, MO modifier. Note that crossing edges indicate discontinuous constituency.

[6] MO is inspired by the usage of the term 'modifier' in traditional structuralist linguistics where some authors (Bloomfield, 1933) use it for adjuncts and others also for complements (Trubetzkoy, 1939).

tics. In the following we will show in which cases the above criteria are employed.

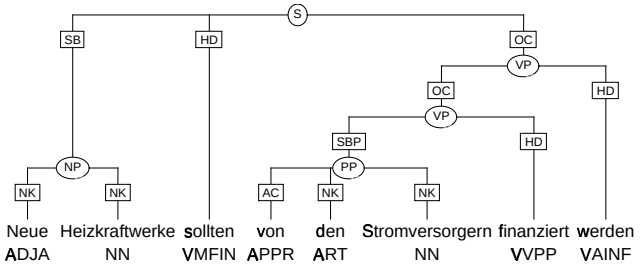
3.4.1 Syntactic Criteria

In spite of the blurred distinction between complementation and adjunction, there is agreement on classifying a number of syntactic dependencies as subcases of complementation (e.g., subjects and objects). Here one can rely on lexical subcategorisation requirements reflected in phenomena such as categorial requirements, case assignment or agreement. In general, the resulting trees comply with semantic argument structure, cf.



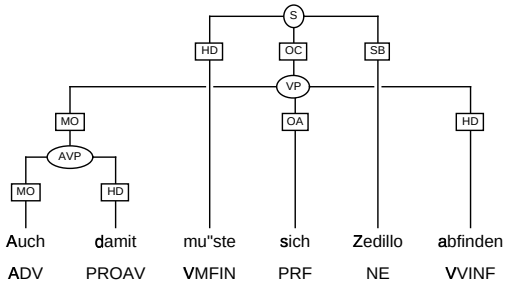
The first rough classification of function labels is based on agreement, case and syntactic category. Thus we distinguish subjects (SB), accusative and clausal objects (OA, OC), datives (DA).

The most important discrepancies between semantics and subcategorisation requirements are phenomena such as passivisation or raising. In such cases we stick to the syntactic criteria and annotate passivised objects and raised subjects as the subject of the finite verb, cf.



(SBP stands for the passivised 'deep' subject).

In linguistic literature, the *argument raising* or *argument composition* technique is also used to account for phenomena like topicalisation, clause union and scrambling, cf. (Hinrichs and Nakazawa, 1994), (Pollard, 1996). Since the main motivation for this analysis is discontinuous constituency, we analyse clause-unioned and topicalised complements *in situ*:



3.4.2 Semantic Criteria

In the case of phenomena involving less syntactic hints for determining dependency structure, we have to resort to semantic criteria.

However, semantic information in general should not be expressed at the level of syntactic annotations (e.g. quantifier scope). Thus we only take into account semantic phenomena having clearly distinguishable syntactic effects, e.g. the scope of focus particles (*nur, auch*), modal and temporal operators, and negation.

Adjunct attachment is often unclear, especially with auxiliaries or modals involved. In such cases a number of empirical tests are employed consisting mainly in paraphrasing the construction in a way making the semantics of the modal/auxiliary explicit and transforming its scope into a that-clause:

(2) hierher mußt du nicht kommen
es ist nicht notwendig, daß du hierher kommst
*es ist notwendig, daß du nicht hierher kommst
*es ist hierher nicht notwendig, daß du kommst

In case multiple readings are acceptable, the highest attachment site is preferred.

3.5 Headedness

In section (3.2) we pointed out the constraints imposed by the DG notion of heads, which are assumed to be unique, present, and of lexical status. However, all the three requirements are problematic.

First, heads do not need to be unique. The DP versus NP discussion has demonstrated that there exist reasons for and against treating the noun as the head of the noun phrase. We circumvent these theory-internal analyses by simply assigning the same functional label (NK) to both determiners and nouns. Theory-specific structures can be recovered from such representations on the basis of part-of-speech information.

Secondly, heads do not have to be present in the annotation. In phrases where a conventional DG approach would be forced to assume an empty head, our annotation simply does not assign a head for the phrase in question.

Thirdly, our format does not require heads to be lexical. In order to avoid stipulations about heads

that are based on a specific linguistic theory and not on language data, we have decided to mark heads in the functional label (e.g., as HD or as NK) and not by assigning them a special status in the hierarchical structure.

4 Extracting Other Representations

Tectogrammatical structures as used in our annotation format can be converted to other (theory or application specific) representations. As an example, we describe an algorithm for converting argument structure representations into standard context-free phrase structure trees. The basic idea is to eliminate crossing edges, which happens in three steps:

1. find the *pivot* of the phrase, i.e. the daughter node that determines the *in situ* position of the phrase and thereby specifies which nodes are 'moved out';
2. for each moved element, determine a position for the corresponding trace;
3. for each moved element, determine a new attachment site such that it no longer yields a crossing edge.

The rest of this section describes these three steps together with default rules. These can be overridden by theory-specific extensions based on the category of the phrase and the moved nodes as well as the direction of movement.

The Pivot: Each phrase is assumed to contain a node determining the *in situ* position of the phrase. It is generally the head or some equivalent function. The *pivot* stays in its original position, and all elements of the phrase that cause crossing edges are assigned traces in a position adjacent to the pivot and its siblings [7]. Given a discontinuous constituent, we refer to its largest contiguous part containing its pivot the *pivot part* of that phrase.

The exact definition of the *pivot* is subject to theory specific parametrisation (e.g., with respect to the analysis of the finite verb in V2-sentences). A sample conversion format is given below:

a) if one of the nodes is explicitly marked as the head (HD), it is the pivot;

b) instead of a unique HD, NPs and PPs contain a number of noun kernel elements (marked NK); the rightmost element of the kernel is the pivot;

c) in a coordination, the first conjunct constitutes the pivot;

d) if none of the above rules apply, the leftmost daughter node marks the *in situ* position.

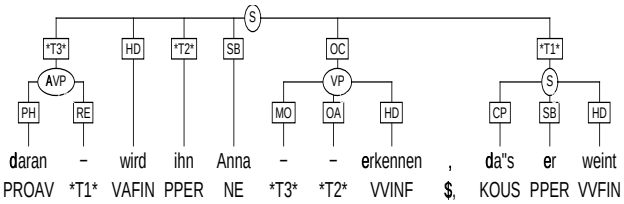
[7] Except in cases of head movement, which has to be handled separately and needs theory-specific rules to determine the position of the phrase.

Position of the Trace The default position of a trace corresponding to a moved constituent is the left or right boundary of the pivot part of the phrase, depending on the direction of the movement. This can be overridden by theory-specific assumptions that, e.g., may place the empty element adjacently to the head.

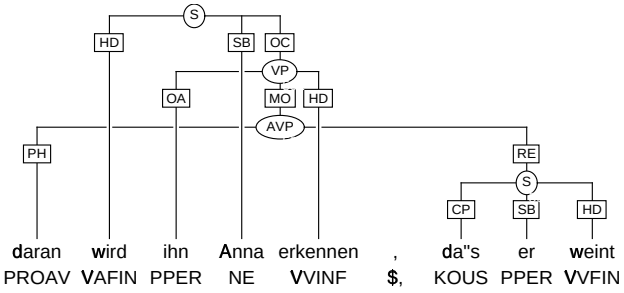
New Attachment Site As the third and last step in eliminating crossing edges we find a new (higher) phrase to attach the moved element (the 'filler') to. The default is to use the lowest phrase that covers the pivot part of the discontinuous constituent and that does not cause crossing edges after attaching the filler. Theory-specific extensions can impose restrictions on the category of the attachment site, e.g., by requiring an S or VP node for particular constellations.

As discontinuity in phrases at lower levels of the structure influences the discontinuity of phrases at higher levels, crossing edges are resolved in a bottom-up way and, in case of conflicts, from left to right.

The procedure described above enables us to convert the tectogrammatical annotations to constituent trees. All information needed is encoded in the tectogrammatical structure and the surface order, e.g.,



is automatically derived from



All discontinuous phrases in our corpus, which currently consists of 4,200 sentences, can be converted using this procedure.

5 Conclusions

The emerging relevance of data-oriented NLP requires the development of a specific methodology, partly different from the generative paradigm which has dominated linguistics for nearly 40 years. The importance of consistent and efficient encoding of linguistic knowledge has absolute priority in this new

approach, and thus we have argued for easing the burden of explanatory claims, which has proved a severe constraint on linguistic formalism.

We have presented a number of linguistic analyses used in our treebank, as well as several examples of the interaction of different syntactic phenomena. The particular representation chosen enables the derivation of other, theory specific representations. Our claims are backed by an annotated corpus of currently about 4,200 sentences, all of which have been annotated twice in order to ensure consistency.

References

Ahrenberg, L. 1990. A grammar combining phrase structure and field structure. In *Proceedings of COLING '90*.

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

Bloomfield, L. 1933. *Language*. New York.

Butt, M., C. Fortmann, and C. Rohrer. 1996. Syntactic analyses for parallel grammars: Auxiliaries and genitive nps. In *COLING 96*.

Hellwig, P. 1988. Chart parsing according to the slot and filler principle. In *COLING 88*, pages 242–244.

Hinrichs, Erhard and Tsuneko Nakazawa. 1994. Linearizing AUXs in German verbal complexes. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in Lecture Notes. CSLI Publications, Stanford University, pages 11–37.

Hudson, Richard. 1984. *Word Grammar*. Basil Blackwell Ltd.

Kasper, R. 1994. Adjuncts in the mittelfeld. In J. Nerbonne, K. Netter, and C. Pollard, editors, *German in HPSG*. CSLI, Stanford.

Kathol, Andreas and Carl Pollard. 1995. Extraposition via Complex Domain Formation. In *Proceedings of the Thirty-Third Annual Meeting of the ACL*, pages 174–180, Cambridge, MA.

Lehmann, Sabine, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Hervé Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP — Test Suites for Natural Language Processing. In *Proceedings of COLING 1996*, Kopenhagen.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*, San Francisco, Morgan Kaufmann.

Nerbonne, John. 1994. Partial verb phrases and spurious ambiguities. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in Lecture Notes. CSLI Publications, Stanford University, pages 109–150.

Netter, Klaus. 1992. On Non-Head Non-Movement. In Günter Görz, editor, *KONVENS '92*, Reihe Informatik aktuell. Springer-Verlag, Berlin, pages 218–227.

Pollard, Carl. 1996. On head nonmovement. In Arthur Horck and Wietske Sijtsma, editors, *Discontinuous Constituency*. Mouton de Gruyter.

Reape, Mike. 1994. Domain union and word order variation in German. In John Nerbonne, Klaus Netter, and Carl Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, number 46 in Lecture Notes. CSLI Publications, Stanford University, pages 151–197.

Sampson, Geoffrey. 1995. *English for the Computer*. Oxford University Press, Oxford.

Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of ANLP-97*, Washington, DC.

Tesnière, L. 1959. *Éléments de Syntaxe Structurale*. Klincksieck, Paris.

Thielen, Christine and Anne Schiller. 1995. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text 17./18. Februar 1994, Schloß Hohentübingen. Lexicographica Series Maior*, Tübingen. Niemeyer.

Trubetzkoy, N. 1939. Le rapport entre le déterminé, le déterminant et ledéfini. In *Mélanges de linguistique, offers á Charles Bally*. Geneva.