The a3Treebank:
Its History, Development, and Prospects

Wolfgang Sternefeld

Potsdam, 29. November 2003

The a3Treebank

Staff:
Sam Featherston (Mädchen für alles)
Tanja Kiziak (data acquisition, web-interface, cgi)
Ilona Steiner (web-interface, applets)
Wolfgang Sternefeld (annotation schemes, grammar)
Monika Toth (tree annotation)

1 History: Background

As some of you might know the SFB was originally planned as a much smaller research group which centered around data bases. In this context the idea came up that it would be nice to have a treebank of German data, in particular we wanted to develop a data base with ungrammatical or suboptimal data that could become representative for the Generative Literature on the syntax of German. With this as a starting point the A3-project of the current SFB developed into something much more ambitious, and the data base only became a side aspect of the work done in A3.

However, for various reasons I will explain in a minute, the data base became more demanding than intended; more and more effort was spent for its development and it seems to be time now to give a report of the intermediate state of our work.

At the beginning we were quite confident to be able to rely on previous work and on the work of other projects. First of all there was a large treebank for German already existing within the SFB; this was an inheritance of the **Verbmobil treebank**. This Treebank was developed in a project led by Erhard Hinrichs, and I heard him boast several times that it is the world's largest treebank with German data. In Sept. 2000 it had approximately 38 000 entries. We believed that the experience gathered there, and in particular the annotation tool that created the structures, would help us to build our own structures.

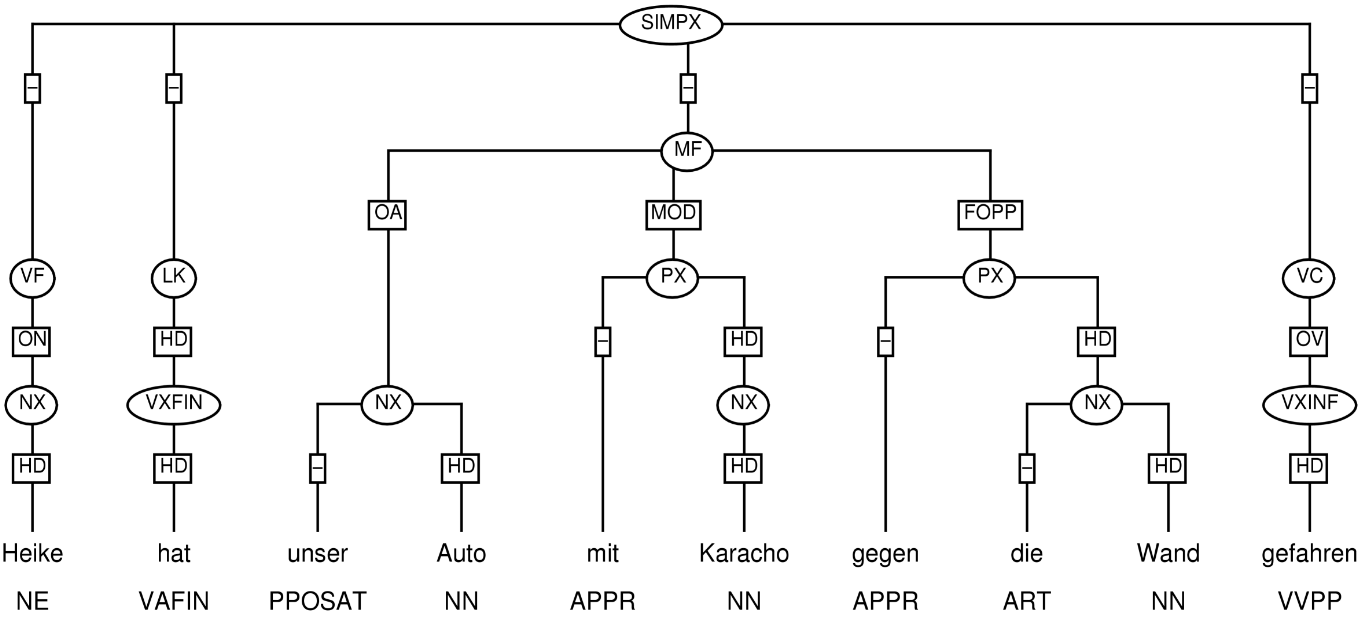
Secondly, we were not only optimistic with respect to being able to build on already exising work; we also heavily counted on a search tool for tree structures that was being newly developed within the SFB.

We then spent a hard time with customizing the annotation tool, a program called **Annotate** that was developed by our dear colleagues from Saarbrücken. It now works fine, except that part of its functionality, namely its ability to partially predict structures is still poor. This not being too disturbing, we had to experience, however, that both the newly developed search tool, namely the Viqtoria program, and the existing VM-treebank turned out to be completely unusable for the intent and purpose of the project.

2 History: The VM-Treebank

To see why this is so, let us first look at a typical tree within the VM treebank. Here is one:

(1)



The first thing to note is that the sentential coarse structure is based on the model of topological fields. This decision has been defended in the stylebook for German treebank from which I quote in (2):

(2) “To ensure the reusability of the data, a theory neutral and surface-oriented annotation scheme has been adopted that is inspired by the notion of topological fields... the linguistic inventory used in the treebank is based on a minimal set of assumptions that are uncontroversial among major syntactic theories.”

It might in fact be true that — as a **descriptive** tool — the topological model is uncontroversially successful. On the other hand, it is not uncontroversial as a hypothesis about constituent structure. There might thus be two different levels of syntactic description, and the **one** level we want to model in our treebank **is** constituent structure. And I do not believe that this level can be modelled in a theory neutral way. Even the decision to adopt the notion of middlefield as a basic descriptive category should not uncontroversial. For example when looking into the original work of Bech (1955/57), who contributed most to present day terminology, you will find that there is no such category like *middlefield*. The closest we can come is the so called *restfield*. As the terminology already indicates, this notion is not a basic descriptive concept but is defined only negatively, and as Bech makes it clear it even need not correspond to a contiguous string of words.

Given that there is no really theory neutral description of constituent structure a natural decision would of course be to take the most up-to-date theory as a guideline for tree structures The problem here is that such a theory would not be usable for practicle purposes. As I have shown elsewhere, when taking minimalist assumptions to the extreme we could easily analyse strings of three words as requiring more than a dozen empty categories. Given that such analyses are not feasible, the only way to proceed is to find a compromise between what we consider to be an almost correct description, which to my mind need not be minimalist at all, and what the average linguist coming from a language like English would expect to find.

Before coming back to the issue of feasability, let me reconsider the VM tree (1) in some detail. The guiding theoretical assumptions that are codified in the stylebook of VM are the longest match principle and the high attachment principle:

(3) "**High Attachment Principle:**
The high attachment principle prescribes that in case of syntactic and semantic ambiguity in the attachment of modifiers such ambiguous modifiers are attached to the highest possible level in a tree structure."

Honestly, I have only a vague idea of what this should amount to. In practice, high attachment seems to be the opposite of low attachment and I will comment on the latter in the final section of my talk. The next principle is the

(4) "**Longest Match Principle:**
The *longest match priciple* demands that as many daughter nodes as possible are combined into a single mother node, provided that the resulting construction is syntactically as well as semantically well-formed."

In other word, structure should be as flat as possible. Here I again do not quite understand what the proviso about syntactic and semantic well-formedness should mean. Alluding to syntactic well-formedness in this context is circular; semantic well-formedness, however, practically does not play a role either, since otherwise the middlefield would not be a constituent: no semantic theory I know of is such that it could treat the middlefield as a constituent.

Flat structures of course are quite the opposite of what we have got used to in Generative Grammar. Flatness prevents one defining a useful notion of c-command, and thereby prevents one from developing useful search strategies. In that respect, another practical decision of the VM treebank turns out contra-productive as well: It's the

(5) No empty category principle

(5) is no doubt reasonable when the ultimate aim is, as it were, parsing of spoken discourse, but it frustrates elementary expectations of the ordinary generative syntactician. To some extent, the VM annotation tried to compensate for this deficiency, but the strategy that was employed is not only ad-hoc but also inapplicable to any of the more interesting phenomena, like the investigation of long distance phenomena, or extraction from NP. This means that a central topic of theoretical research cannot be dealt with in the tree structures of VM.

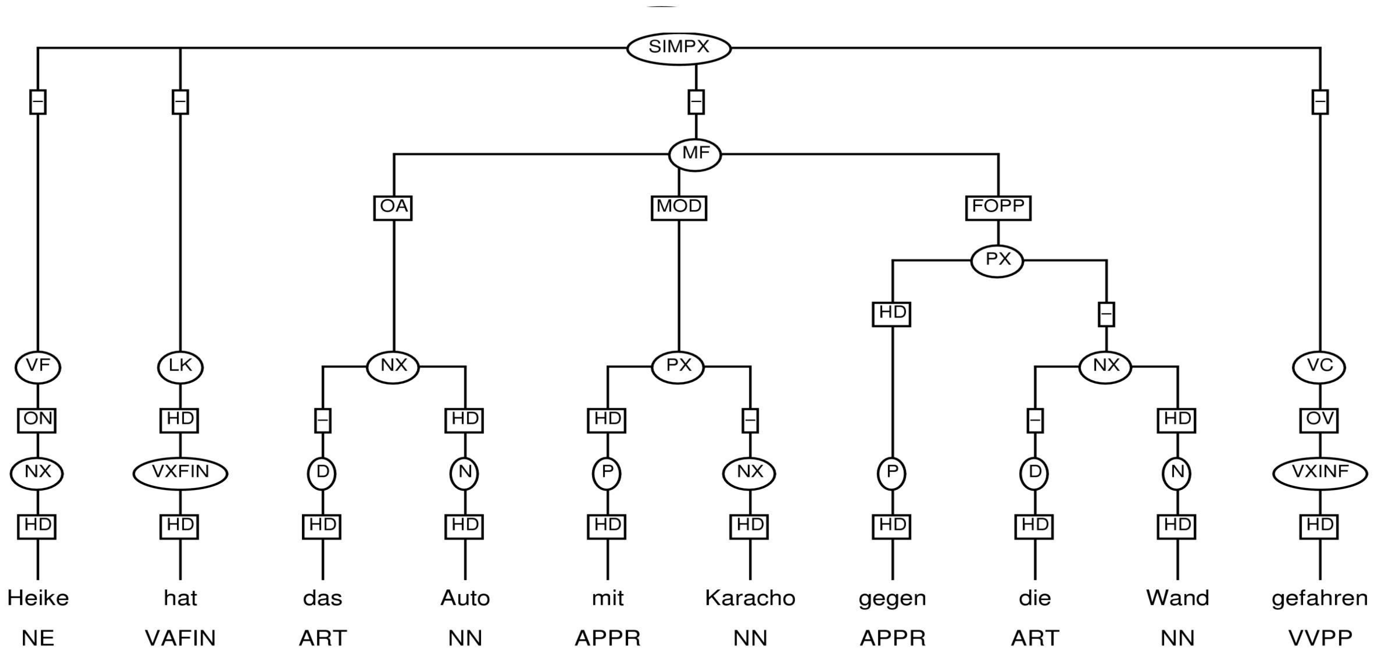
There are also some decisions of minor importance, whose motivation seems rather weak. Consider again (1). Somewhat surprisingly, the head of an PX, that is of PP in the usual notation, is an NP. This is justified as follows:

(6) "In order to facilitate the identification of dependencies between verbs and their nominal complements and adjuncts and in keeping with basic assumptions in Dependency Grammar, not the preposition itself but the complement in prepositional phrases is annotated as the head of the phrase."

Now, why should Dependency Grammar be more adequate and theory neutral, if the distinction between adjuncts and complements is also encoded in the edge label? Why not simply follow ordinary X-bar theory? Much the same applies to the decision not to include node labels for determiners. Why not simply employ a category like Det as usual?

An alternative structure would be this:

(7)



(7) shows that the edge label immediately above a lexical item would always be labelled as head. This is a redundancy which becomes clear immediately when considering the conceptual oddness of imaginating a non-branching structure without a head.

It seems to me, then, that some essentials of X-bar theory are misrepresented in the trees.

3 Present: The “Beispieldatenbank”

Given this conclusion it transpired that we needed an essential redesign of the format of our trees and that this was not a small task. As mentioned already it was also not clear to us, how to customize *Annotate*, and how it could be made to interact as a research tool. So in order not to lose too much time we began to collect some data from the literature and with the help of **Dirk Wiebel** we were able to make the data accessable on the internet. Each entry contained the following information:

(8) *Entries in the* Beispieldatenbank:

- a reference source
- the judgment we found in the literature
- a labelled bracketing as suggested in the literature
- a set of related sentences and their judgments, so that the context of the evaluation becomes clear (a so-called **group**)
- a set of key-words that can be used for the research, e.g.:
  — moved category . . . (list of choices)
  — extraction from . . . (list of choices)
  — landing site . . . (list of choices)
  — remnant movement
  — . . .

It is obvious that the usability of the data base heavily depends on the **key-words** one has at disposal. The idea was to keep key-words quite general, and that a query combines different key words with different parameters, like syntactic categories, so that the query could nontheless characterize highly specific structures. In some cases, however, we had to use rather construction specific key words, such as **partial movement** or **remnant movement construction**. An interactive beta version is accessable under the address in (9):

(9) www.sfb441.uni-tuebingen.de/a3/db/

It was clear to me from the outset that the limitation to key-words was somewhat arbitrary and unsatisfying, and that this method should ultimately be replaced by a search tool that was able to analyse genuine structure. So, after a while, when the outlines of the intended tree structures became clearer to us, we endevoured on translating labelled bracketings into trees. **Monika Toth** will illustrate some of the guidelines we were observing in a minute.

4 History: The Viqtoria Tool

Before going into the design of our trees I would like to mention that during this ongoing annotation we were still not sure about how the search tool could handle our structures. This was frustrating and remained being so for a long period of time. Here is why:

(10) No Viqtory: Why Viqtoria got us nowhere…

1. The present version is by far not powerful enough to express elementary syntactic relations. Universal quantification and negaton were beyond its expressive power; important notions like headedness could not be expressed…
2. After a while we came to believe that the relevant projects within the SFB were no more interested in extending or developing the program in a way that could meet our demands.
3. Developing the programm on our own was obstructed, access to the source code was denied to us.
4. We were informed that Viqtoria was not intended to be used on the internet.

This situation lead to much uncertainty concerning the question of how morphological and syntactic information should ultimately be presented in the tree. I sometimes simply ignored these difficulties, and occasionally I decided in favor of redundancies, hoping that they would help to overcome anticipated weaknesses of any future search tool. Only recently one of our collegues, namely **Stefan Kepser**, kindly provided us with an ingenious way out of the mess. He developed a really powerful search tool that helped to solve all our problems. **Ilona Steiner** will present it to you in the next talk.

5 Present: The Grammar of a3Trees (Monika Toth)

Now, coming back to the specific design of the trees, I pass over to Monika…

6 Linguistic Motivation: One or Two Case Studies

6.1 Binding and Extraposition

One of the problems with binding is that the naive linguist understands binding as involving c-command. This expectation, however, might be difficult to satisfy in extraposed clauses. Let us first look a sentence like

(11) *weil er$_i$ glaubt [$_{\text{CP}}$ dass Fritz$_i$ doof ist]

Here it would indeed be possible to derive a condition C violation in a model that uses high attachment. This is because everything is flat. In a more articulated structure, however, we get into trouble. This would certainly be the case if we were to adopt the IP hypothesis:

(12) *weil [$_{\text{IP}}$ er$_i$ glaubt][$_{\text{CP}}$ dass Fritz$_i$ doof ist]

Here, the finite verb is in I and the subject in SpecI, so that it never has a chance to c-command the CP. However, if we assume that there is no IP, we can get the naively expected result, namely that a condition C violation is triggered by attaching the CP to the verb:

(13) *weil [$_{\text{VP}}$ er$_i$ [$_{\text{V}}$ glaubt [$_{\text{CP}}$ dass Fritz$_i$ doof ist]]]

Going into some more detail, however, it seems that the structure cannot be as simple as that. This is because in our treebank we adhere to the traditional assumption that material in the *Nachfeld* has generally been extraposed from the middle-field, so that the correct structure would involve a trace, t$_j$ in (14):

(14) *weil [$_{\text{VP}}$ er$_i$ [$_{\text{V}}$ [$_{\text{V}}$ t$_j$ glaubt [$_{\text{CP}}$ dass Fritz$_i$ doof ist]$_j$]]]

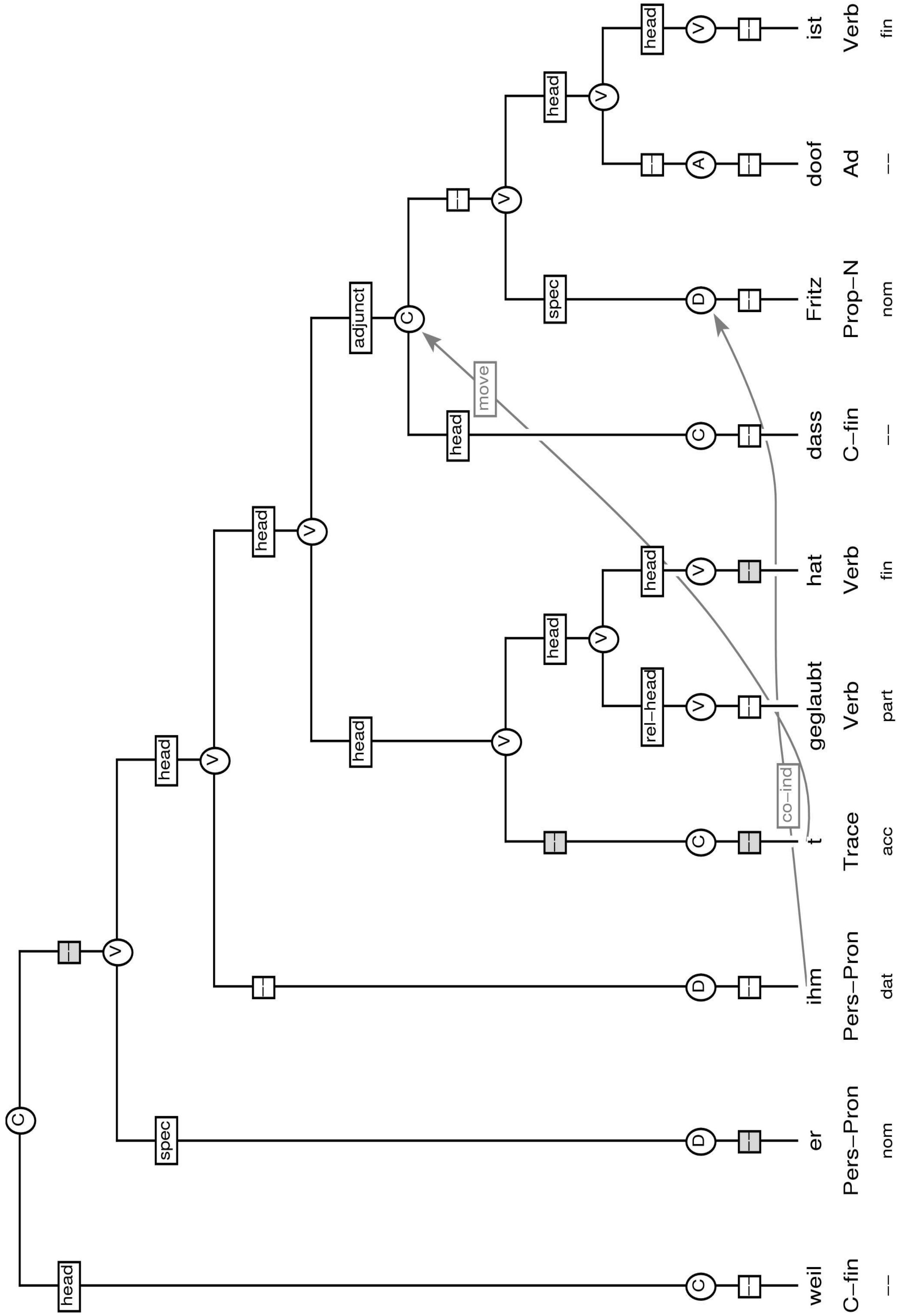
This structure, however, is only possible if we depart from the standard assumption in supposing that extraposition does not necessarily target a maximal projection. Another assumption needed to derive the desired result is that of low attachment: We assume that the path between the trace and its antecedent is as short as possible.

Turning next to more complicated structures, we will see that the c-command requirement leads to another assumption that deviates from the default. Consider

(15) *weil [$_{\text{VP}}$ er [$_{\text{V}}$ ihm$_i$ t$_j$ geglaubt hat [$_{\text{CP}}$ dass Fritz$_i$ doof ist]$_i$]]

For this configuration to confirm to the naive expectation it must be the case that the CP c-commands its trace and the pronoun c-commands the antecedent. This in turn is possible only if *geglaubt hat* forms a verbal complex to which the CP is adjoined, as shown in (16):

(16) *weil [$_{\mathrm{VP}}$ er [$_{\mathrm{V}}$ ihm$_i$ [$_{\mathrm{V}}$ [$_{\mathrm{V}}$ t$_j$ geglaubt hat][$_{\mathrm{CP}}$ dass Fritz$_i$ doof ist]$_i$]]



We thus have to depart from our initial default assumption that we do not build verbal complexes.

The principle which is actually more important here than avoidance of V-clustering is the minimize chain condition for extraposition.

It should be noted here that at the time these decisions were made we did not know how to encode the binding relation itself. We finally decided to use secondary edges. Given this, it is clear that the search for a condition C violation would involve at least the following:

(17) a. a secondary edge from A to B;
b. a c-command relation from A to B.

One might also want to add that A is a pronoun.

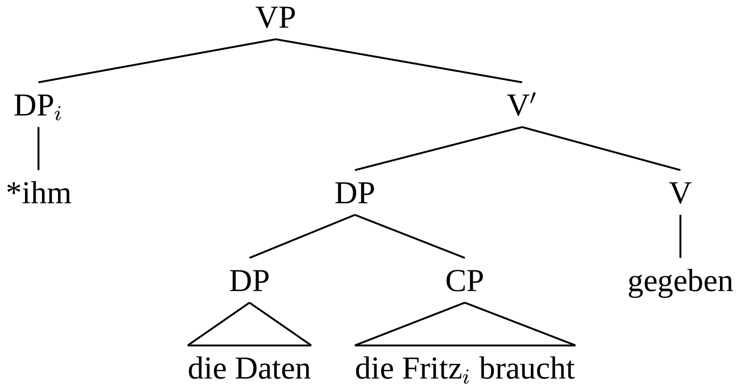
What we did here is to satisfy the naive expectations as closely as possible. I would like to stress, however, that the naive strategy is not necessarily guided by the correct theory.

As an alternative theory I would like to discuss the claim pronounced by Büring and Hartmann, namely that binding must operate before extraposition, or equivalently, that extraposition has to be reconstructed for the purpose of binding. They first consider cases like

(18) a. weil wir dem Fritz$_i$ die Daten, die er$_i$ braucht, gegeben haben
b. *weil wir ihm$_i$ die Daten, die Fritz$_i$ braucht, gegeben haben

Here it is still easy to explain the contrast, because the antecent c-commands the anaphor, as shown in (19):

(19)

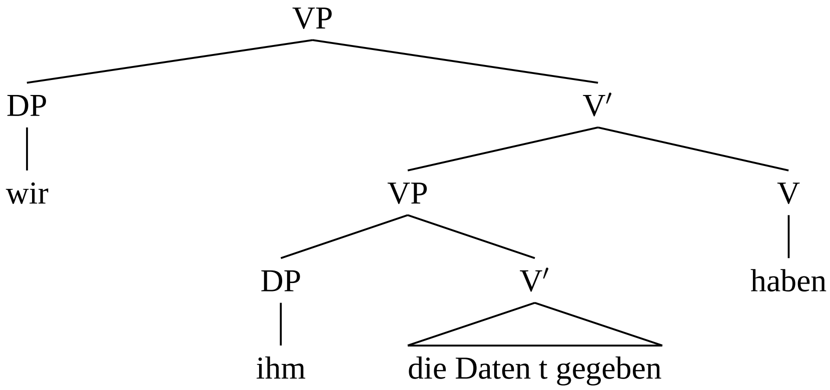


The problem comes in with the extraposition in (20):

(20) a. weil wir dem Fritz$_i$ die Daten gegeben haben, die er$_i$ braucht
b. *weil wir ihm$_i$ die Daten gegeben haben, die Fritz$_i$ braucht

The structure presupposed by B/H is roughly that in (21):

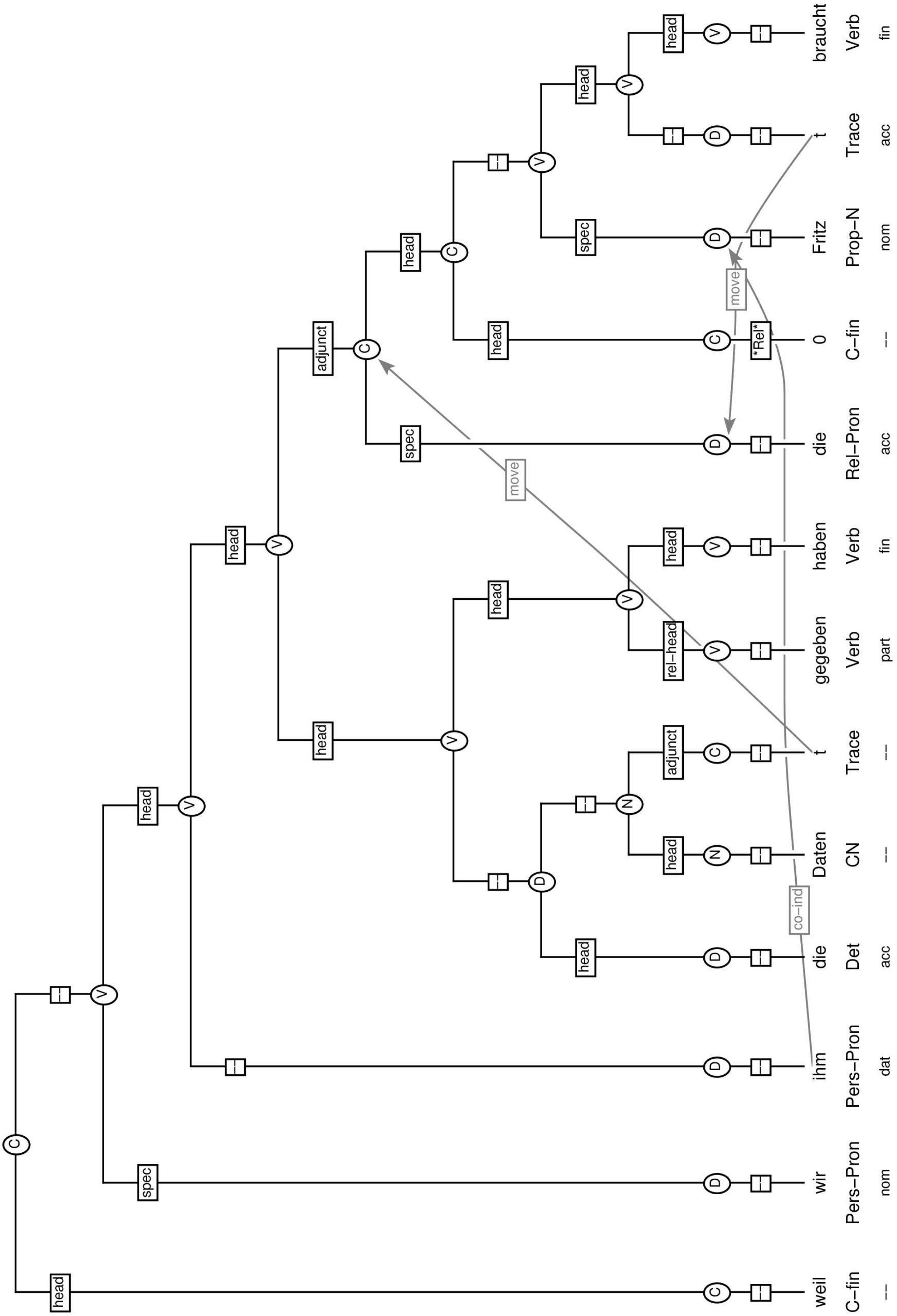
(21)



The problem here is where to attach the extraposed clause in such a way that it follows the finite verb *haben* and is c-commanded by *ihm*. Since this is impossible, they claim to have evidence for reconstruction.

From the present perspective, however, this argument does not hold because it is easy to see that the alternative analyses based on a verbal complex in (22) does the job.

(22)



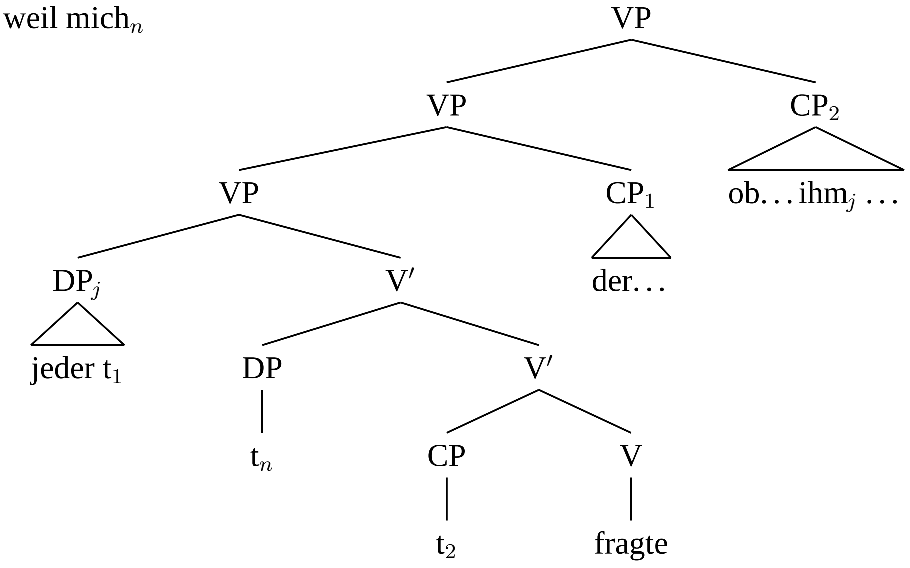
This possibility is not far fetched, indeed it is a standard assumption in the topological model, so that one might wonder why it has not been mentioned in the articles cited above.

Now, it is not the case, however, that the theory of B&H (the bra theory) is necessarily wrong. For one thing, one might wonder, why we are forced here to build a verbal complex, whereas in other cases the default is not to do it. I have no answer to this question, so we might conclude that the reconstruction theory is still true, although the argument given by B&H was incomplete. Another question is whether we can find data that could proof the point in a more conclusive way. A relevant construction is (23), taken form Haider (1994):

(23) weil mich jeder$_j$ fragte, der mich kannte, ob ich ihm$_j$ helfen könne.

The relevant partial structure is displayed in (24):

(24) weil mich$_n$



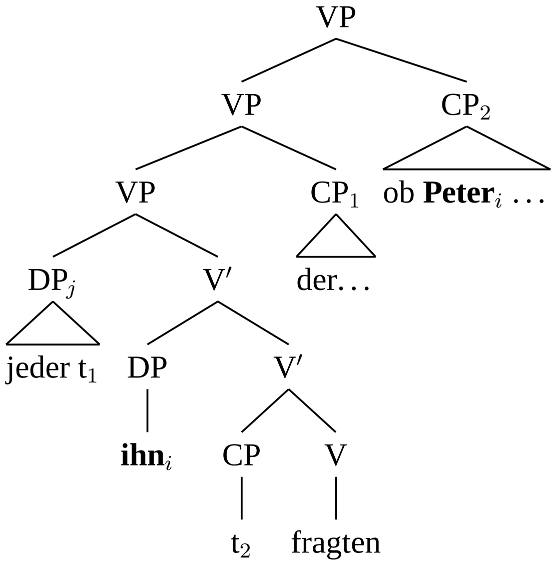
The problem here is that the quantified DO *jeder* should bind the pronoun, but cannot do so because a relative clause intervenes; the assumption here is again, that the extraposed relative clause has to c-command its trace. It follows that the surface structure is not interpretable without reconstruction.

Now it seems to me that the very same point can be made with condition C effects in (25):

(25) *weil jeder ihn$_i$ fragte, der ihn$_i$ kannte, ob Peter$_i$ kommen würde

The structure is analogous to that in (24):

(26)



This is exactly the case that corroborates the BH-theory: a violation of condition C without surface structural c-command.

What can we learn from this? There are some obvious consequences, namely that we need sophisticated structure and sophisticated search tools. But we also have to put them to use in an intelligent way. In order to detect the very existence of cases like the one shown above, it is not sufficient to look at what is naively considered a condition C-*violation*. We must always also look at the complementary cases, namely those where we would expect grammaticality because there is no offending c-command. It is only here that we can detect the relevant data, namely cases of ungrammaticality despite lack of c-command. This is precisely the context that requires reconstruction.

On the other hand one might reconsider one of the premisses of the above argument. What is at stake here is that the relative clause need not be moved from the position adjacent to its head. Instead, it could be base generated in a position that is low enough to permit c-command of the complement clause. Such an analysis was proposed by Haider, although he does not mention the interpretative problem of providing his structures with a compositional semantics. In this respect the most important principle of the syntactic annotation is semantic interpretability. Haider's structures clearly must be rejected, as long as nothing is said about the semantics.

Now, contrary to standard assumptions, it might indeed be possible to develop an alternative semantics for such cases, namely one that is based on the idea that each quantifier presupposes its own domain of dicourse. The restriction of the quantifier expresses that the domain of discourse is included in the denotation of restriction; this peace of information is presupposed rather than asserted. We might than say that the relative clause also expresses the presupposition that the discourse is part of its denotation. It might then become possible to interpret Haider's structures, since the information expressed by the restriction of the quantifier and that expressed by the relative clause need not be interpreted as a unit. The question, however, remains, whether this should motivate a drastic departure from generally accepted principles.

My personal reaction is that in this particular case it should not. In general, however, there might be cases where the decision is not guided by more or less naive expectations. The only feasable solution I can think of is that the treebank could provide for different alternative structures for the same sentence. This may well be the price that has to be paid as a consequence of there not being a theory neutral way of describing structure. The problem here is to decide between alternative theories, but this is of course the problem we were starting off with right at the beginning. Some decisions are more difficult than others, and in some cases no decision will lead to multiple structures for the same sentence.

6.2 Subjects within SpecC

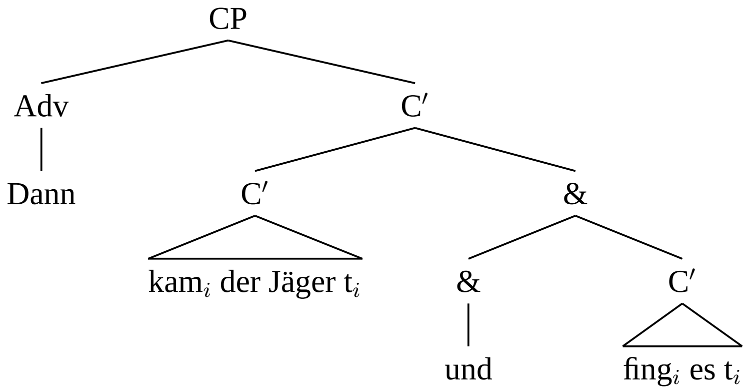
Above Monika already mentioned that subjects may be base generated in SpecC. In the literature there were some attemps to justify something similar, namely that SVO-sentences are IPs, whereas all other types of sentences are CPs. Looking at the details of this argument I found that they are not particularly convincing because they presuppose an IP. On the other hand, they make perfectly sense if we interpret the IP in such sentences as a CP whose subject is base generated.

Still the arguments are not very strong; what I would like to present here is the strongest argument I presently know of. It is a rather indirect one, it can be derived from Höhle's analysis of a particular kind of construction illustrated in (27):

(27) a. Dann kam der Jäger und fing es
b. Lässig betrat Ede den Raum und brachte auch ein Bier mit
c. „Plötzlich habe sich die Brezel auf den Verdauungstrakt des Präsidenten gestürzt und sei auf dem Weg dorthin im Hals steckengeblieben." (*taz*)

If this were a coordination of V/1-sentences, as suggested in (28),

(28)

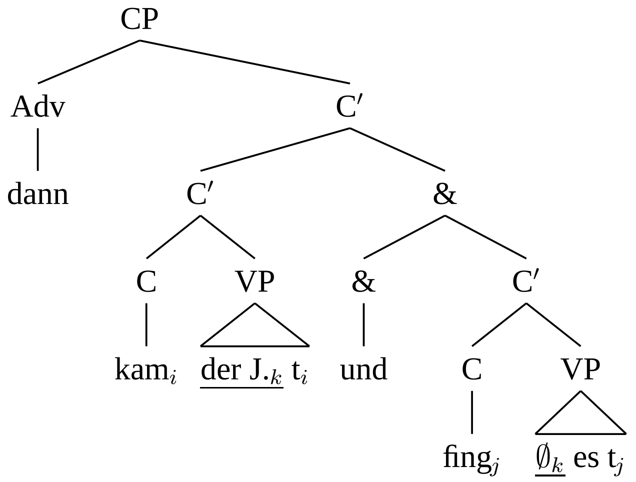


we would not know where the subject is in the second clause and how to describe agreement.

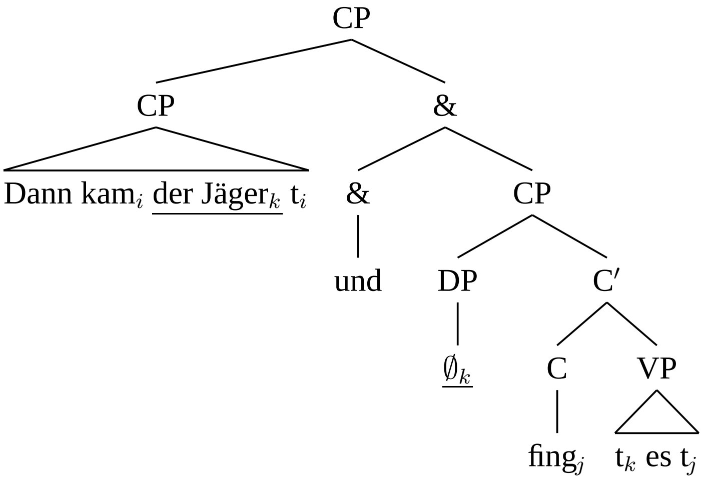
(29) Dann kamen die Jäger und fingen/*fing es

Now given the traditional assumption that only identical material can be coordinated, one would have to conclude that the subjects is an empty category, giving rise to two possible analyses as in (30):

(30) a.



b.



Problem 1: Warum ist die leere Kategorie $\emptyset_k$ gerade in diesem speziellen Kontext (in einem Zweitglied einer Koordination) möglich, in normalen Kontexten jedoch nicht? („The exact nature of this EC is unclear"; cf. Valin (1986))

Problem 2: Die leere Kategorie kann zwar in (30) wie ein koreferentes Pronomen interpretiert werden:

(31) Dann kam der Jäger und er (=der Jäger, der gekommen ist) fing es.

Wie jedoch schon Höhle (1983) bemerkte, versagt diese Methode bei Beispielen wie (32-a-c); ein eigener Hörbeleg ist (32-d):

(32) a. Dennoch kam niemand und machte die Tür auf
b. Dann kommt wieder jeder und beschwert sich
c. Hoffentlich kommt keiner nach Hause und sieht da den Gerichtsvollzieher
d. Nicht umsonst kamen viele Bomber zurück und hatten ihre Ladung noch bei sich

Lesart von (32-d) mit $\emptyset$ als koreferentiellem leeren Pronomen ergibt das falsche Resultat (33-a); Lesart von $\emptyset$ als gebundene Variable in (33-b) ergibt das Richtige:

(33) a. Viele Bomber kamen zurück und sie (=die Bomber, die zurückkamen) hatten ihre Ladung noch bei sich
Implikatur: *Nicht alle Bomber kamen zurück
b. Es gibt viele Bomber, für die gilt: sie kamen zurück und hatten ihre Ladung noch bei sich
Implikatur: Nicht alle Bomber hatten ihre Ladung noch bei sich, als sie zurückkamen

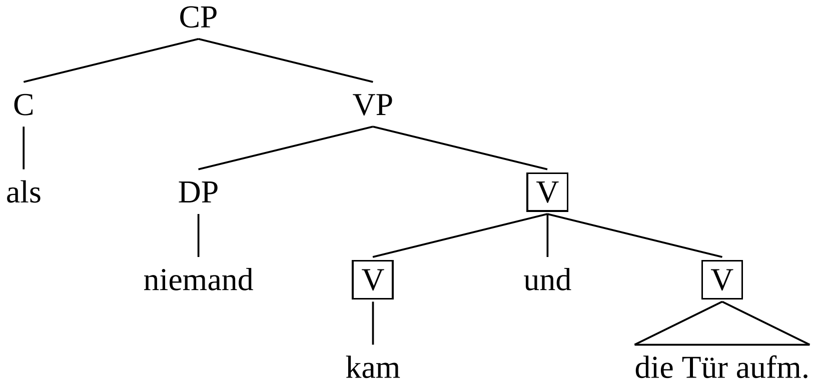
Folgerung: das Subjekt muss weiten Skopus über die Konjunktion haben; gebundene Variablen müssen von ihrem Antezedens c-kommandiert werden. Man sieht dies am Kontrast in (34):

(34) a. [$_{\text{CP}}$ Dann kam wieder so ein Idiot$_i$] und [$_{\text{CP}}$ wieder beschwerte er$_i$ sich]
b. *[$_{\text{CP}}$ Dann kam wieder jeder/niemand$_i$] und [$_{\text{CP}}$ wieder beschwerte er$_i$ sich]

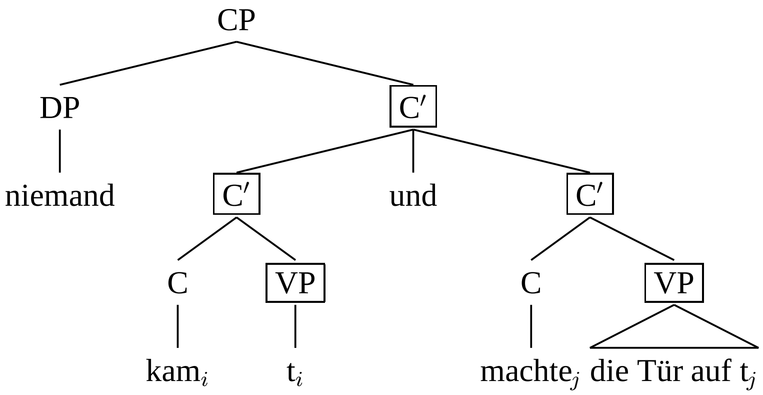
Lösung in Anlehnung an Höhle (1990) und Rambow and Santorini (1995):

(35) a. First assumption: coordination does not require strict categorial identity, but functional identity.
b. The relevant property here is the sharing of an external (subject) theta role

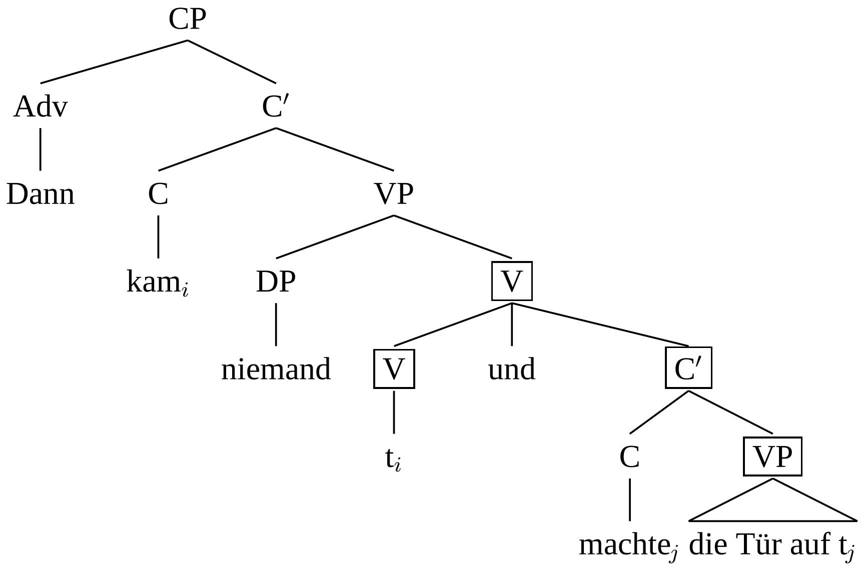
(36) a.



b.



c.

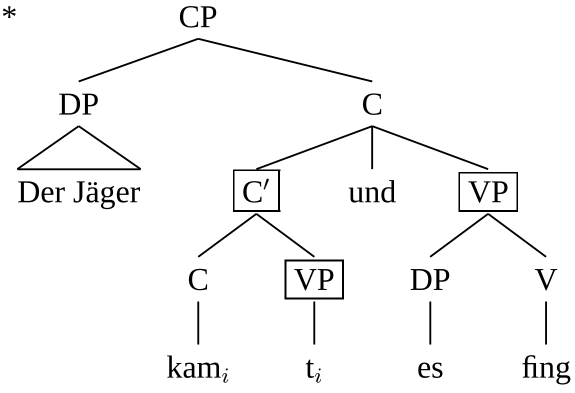


(36) a. In accordance with the analysis of Höhle, I assume that the first conjunct is totally transparent for movement, whereas the second one is totally opaque.

b. These properties imply that C′ in (35-c) does have a base generated external argument: both V and C′ share an external theta role.

c. Syntactically, it is assumed that the *and*-clause is adjoined to the V-projection, which might explain why it is opaque for extraction. This asymmetry prevents (37) from being generated:

(37) *



Similar asymmetries also exist in other coordinations that do not involve Subjekt-*sharing*, cf. (38):

(38) a. Wenn jemand nach Hause kommt (V/E) und da steht der Gerichtsvollzieher vor der Tür (V/2)…

b. Kommst du nach Hause (V/1) und da steht der Gerichtsvollzieher vor der Tür (V/2)...

Therefore I assume that some construction specific assumptions are needed anyway.

Apart from the evidence one can find in Höhle (1990), Fortmann (2003) draws attention to the following data:

(39) a. auf der Lichtung entkam der Hase dem Förster und
begegnete dem begegnete dem Fuchs
b. ??auf der Lichtung entkam dem Förster der Hase und begegnete dem Fuchs
c. auf der Lichtung entkam dem Förster der Hase und begegnete der Fuchs

(39-b) can be explained under current assumptions, if is assumed that argument sharing is not only possible for nominative subjects, but also for those that may appear highest in the argument hierarchy. This implies that verbs like *entkommen* and *begegnen* may have alternative theta grids, one with a nominative and one with a dative as highest argument. If identity of highest non-subject arguements is a sufficient condition, we would also expect the following to be grammatical:

(40) a. in der Regel gebührt dem Besitzer nicht nur die Nutzung der Sache sondern obliegt auch die Pflicht zu deren Erhaltung
b. heute gefiel dem Publikum das Stück und misfiel die Inszenierung
c. beim Spiel reitet ihn der Teufel und sticht der Hafer (Fortmann, unpublished)

The ungrammaticality of (39-b) is more difficult to explain. Fortmann assumes a general filter to the effect that the order of grammatical functions is identical in both conjuncts. Given the possibility of scrambling in German, this implies that scrambling out of the first conjunct is blocked, but as in examples like (41) topicalization is fine.

(41) das Ventil hat der Klempner gestern gebracht und hat dann heute die Heizung repariert

An interesting unsolved question then would be whether the ungrammaticality of scrambling could be explained on independent grounds, or whether we have to take Formanns condition as a plain surface filter.

Literatur

Bech, Gunnar (1955/57): *Studien über das deutsche verbum infinitum*. Munksgaard, Kopenhagen. Reprint bei Niemeyer, Tübingen 1983.

Fortmann, Christian (2003): Die Lücke im Bild von der Subjektlücken-Konstruktion. IMS, unpublished.

Haider, Hubert (1994): Detached Clauses — The Later the Deeper. Bericht Nr. 41 des Sonderforschungsbereich 340, Universität Stuttgart/Tübingen.

Höhle, Tilman (1983): Subjektlücken in Koordinationen. Unveröffentlichtes Manuskript, Universität Tübingen.

Höhle, Tilman (1990): Assumptions about Asymmetric Coordination in German. *In:* J. Mascaro and M. Nespor, eds, *Grammar in Progress*. Foris, Dordrecht, pp. 221–236.

Rambow, Owen and Beatrice Santorini (1995): ‘Incremental Phrase Structure Generation and a Universal Theory of V2’, *NELS* **25**, 373–387.

Stegmann, Rosemary, Heike Telljohann and Erhard W. Hinrichs (2000): Stylebook for the German Treebank in VERBMOBIL. Technical report, VM-Report 239, Seminar für Sprachwissenschaft, Tübingen.

Sternefeld, Wolfgang (2003): *Syntax. Eine merkmalbasierte Analyse des Deutschen. Band 1*. Stauffenburg Verlag, Tübingen.

Valin, Robert D. Van jr. (1986): ‘An Empty Category as the Subject of a Tensed S in English’, *Linguistic Inquiry* **17**, 581–586.