Regions and Word Order

Frank Richter Manfred Sailer*

September 20th, 1996

1 Motivation

- Empirical observations: a purely phrase structural account of German seems to be empirically inadequate.
- Theoretical considerations: HPSG provides powerful mechanisms for word order that are partially independent of syntactic structure. However, in existing word order based approaches to grammar (Reape (1992), Richter and Sailer (1995), Kathol (1995), ...), the level on which word order is constrained is still hierarchically organized.
- Experience with implementations: for the implementation of a parser for a linearization-based analysis of German, the recognition of the syntactically relevant subparts of the input string is a major problem.

2 Problems for phrase structure based approaches

In a phrase structure grammar (PSG), the linear order of words can be read off the yield of the structure. In German there are at least two empirical phenomena which are problematic for a PSG-approach:

- scrambling
- extraposition

*{fr,mf}@sfs.nphil.uni-tuebingen.de

2.1 Scrambling

In principle, all complements and all adjuncts of the main verb can be permuted in the German Mittelfeld (X-field). This has led to a flat analysis of German (GPSG, HPSG) or to the postulation of clause internal movement (scrambling, GB).
The order of elements in the Mittelfeld is constrained by (cf. Lenerz (1977), Höhle (1982), Jacobs (1988), ...):

- the complement/adjunct distinction,
- inherent properties of the elements to be ordered,
- inherent properties of the element that syntactically selects the elements to be ordered,
- scope properties,
- the focus structure of the sentence,
- ...?

2.2 Extraposition

Only parts of a syntactic constituent appear in phonologically adjacent positions, PPs and clauses can be 'moved' to the right periphery. This has been treated via meta rules (GPSG), lexical rules (HPSG (Pollard, Levine and Kasper 1994)) or rightward movement (GB, HPSG (Keller 1994)).

(1) a. daß [der alte Mann [den du auch kennst]] gekommen ist
that [the old man [that you too know]] arrived is

b. daß [der alte Mann] gekommen ist [den du auch kennst]

c. der alte Mann [den du auch kennst] ist gekommen

d. der alte Mann ist gekommen [den du auch kennst]

3 Syntactic Structure and Word Order

3.1 Terminology

The following units are relevant for word order:

- *governor*: the smallest unit, consisting just of one or two words, depending on lexical properties.

  [Maria], [die, Frau], ...

- *area*: a part of the phonological string that is limited by the elements of a governor.

  [daß, [Maria], [die, Frau], sah]

  (governor: [daß, sah])

- *region*: a part of the phonological string that consists of areas.

  [sagt],[Peter],[daß, [Maria], [die, Frau], sah]

A word has chunking properties:

- Does the word constitute a governor of its own, or with which word does it form a governor?
- Does the word impose restrictions on the regions in which it may occur?

3.2 Government

Areas and regions have properties which rely on a notion of syntactic (selectional) government:

(2) a. A word s-governs all its arguments and modifiers. (corresponds to some definitions of government in GB, and to the notion of head-projection in some variants of HPSG)

b. A governor G g-governs a governor G' iff some word in G s-governs some word in G'. ($G' \subset G$)

c. g*-government is the reflexive transitive closure under g-government. ($G' \subset^* G$)

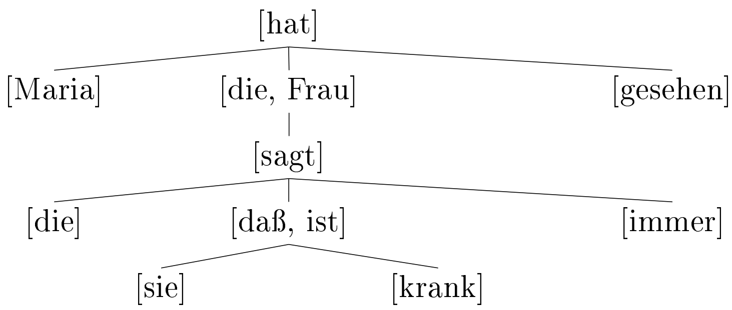
(3) a. Maria hat die Frau gesehen, die immer, sagt, daß sie krank ist.

Maria has seen the woman which always says that she is sick

b. [Maria],[hat],[die,Frau],[gesehen],

[die],[immer],[sagt],[daß,ist],[sie],[krank]

(4) The g*-government relation:



3.3 Areas and regions

(5) **A property of areas:**

For each area $A = [w, A_1, \ldots, A_n, w']$, and for each i with $1 \leq i \leq n$, $A_i \subset^* [w, w']$.

By virtue of this property of areas, the g*-government relation can be generalized from governors to areas.

(6) a. * Maria hat gesehen, die immer sagt, daß sie *die Frau* krank ist

b. complex area: [daß [sie] [*die, Frau*] [krank] ist]

[die, Frau] is not g*-governed by [daß, ist].

(7) **A property of regions:**

For each region $R = A_1 \ldots A_n$, there is an $A_\top$ in R such that

$A_\top$ g*-governs all A_i in R and

for each A_i in R that is not g-governed by $A_\top$, there exist $A'_1, \ldots A'_m$ such that $A_i \subset A'_1 \subset \ldots \subset A'_m \subset A_\top$, and each A'_j either is in R or is included in $A_\top$.

(8) **Definition: strict region**

A region R is a strict region iff all governors that are g*-governed by the top governor $G_\top$ of R occur in some area of R.

In German governors that include a finite verb are the top governor of a strict region.

(9) a. * Maria hat die Frau, daß sie krank ist, gesehen, die immer sagt

b. [Maria] [hat] [die Frau], [daß [sie] [krank] ist], [gesehen], [die] [immer] [sagt]

c. Governors with finite verbs: [hat],[daß,ist], [sagt]

There is no strict region with [sagt] as top governor.

4 The relative order of areas

The relative word order inside an area or a region is restricted. The relevant kind of restrictions may depend on the factors mentioned in section 2.
Therefore: we need entire words as components of governors.
Examples for word order restrictions (WORs). (see (Richter and Sailer 1995) for how these can be expressed in SRL (King 1994))

(10) a. A WOR that forbids preposition stranding: (needed in German)

For each two areas A and A', where A is a preposition and A' is the area governed by the word that is selected by that preposition, A has to precede A' directly.

b. A WOR that orders nominal arguments according to their grammatical function: (proposed for English in (Pollard and Sag 1987))

For each two areas A and A' in an area with top governor G, if the governors of A and A' are nouns N and N', and N has a higher grammatical function in G than N', then A has to precede A' in G.

5 Topological fields

Some parts of the traditional theory of topological fields (Höhle 1986) are directly expressed by the regional requirements of finite verbs. The *Vorfeld* (K) is a region, and the finite verb, the *Mittelfeld* and the verbal complex (FINIT X VK) form a region.

1. **V1-clauses:**

   There is a strict region:

   $$[\text{finite Verb}]\ A_1 \ldots A_n$$

2. ***daß*-clauses:**

   There is a strict region:

   $$[\text{daß}\ A_1 \ldots A_n\ \text{finite verb}]\ A'_1 \ldots A'_{n'}$$

   where
   [daß $A_1 \ldots A_n$ finite verb] is a region

   (The same holds for *ob* (*whether*), *bevor* (*before*), etc.)[1]

[1] In order to account for the whole range of ordering possibilities among verbs, we should really have: [daß $A_1 \ldots A_n$ finite verb] $A''_1 \ldots A''_{n''}$ $A'_1 \ldots A'_{n'}$
where [daß $A_1 \ldots A_n$ finite verb] $A''_1 \ldots A''_{n''}$ is a region.

3. **relative clauses :**

   There is a strict region:

   $$A_1 \ldots A_n \; A'_1 \ldots A'_{n'} \text{ [finite verb] } A''_1 \ldots A''_{n''}$$

   where

   $A_1 \ldots A_n$ is a region (which includes a relative pronoun), and
   $A'_1 \ldots A'_{n'}$ [finite verb] is a region.

   (Similar for embedded w-clauses, in which case the first region has to include a w-word)

4. **V2-clauses:**

   There is a strict region:

   $$A_1 \ldots A_n \text{ [finite verb] } A'_1 \ldots A'_{n'} \; A''_1 \ldots A''_{n''}$$

   where

   $A_1 \ldots A_n$ is a region, and
   [finite verb] $A'_1 \ldots A'_{n'}$ is a region.

The regional requirements of *daß*-clauses explain i.a. the fact that if a PP p is modified by a PP p', then p' must be extraposed if p is extraposed.

(11) a. daß Peter mit dem Auto von Maria gekommen ist
that Peter with the car of Maria come is
that Peter arrived with Maria's car

b. daß Peter gekommen ist mit dem Auto von Maria

c. daß Peter mit dem Auto gekommen ist von Maria

d. * daß Peter von Maria gekommen ist mit dem Auto

The regional requirements of V2-clauses explain i.a. the fact that material that is g*-governed by a noun in the K-field may not occur in FINIT X VK (FXV):

(12) a. *Maria has seen the woman that is sick*

b. $[_K$ die Frau, die krank ist$]$ $[_{FXV}$ hat Maria gesehen$]$
[the woman that sick is] [has Maria seen]

c. * $[_K$ die Frau$]$ $[_{FXV}$ hat Maria, die krank ist, gesehen$]$[2]
[the woman] [has Maria that is sick seen]

d. $[_K$ die Frau$]$ $[_{FXV}$ hat Maria gesehen$]$, die krank ist
[the woman] [has Maria seen] that sick is

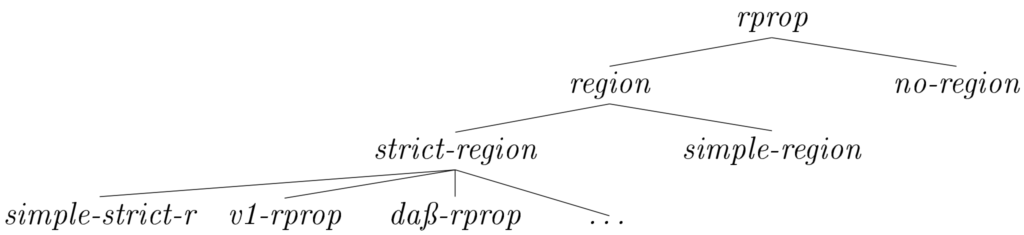
[2]The sentence is grammatical if the relative clause is interpreted as a modifier to *Maria*, but not in the sense of (12a)

6 Word Order in HPSG

Our treatment of word order is fairly independent of a particular theory of grammar or a particular syntactic analysis. However, we intend to integrate it into an HPSG analysis of German as developed and implemented in the B8 project.
Addition to the signature:

- there is a new sort $\left[\begin{array}{ll} \textit{chunking-properties} & \\ \textsc{co-governor} & \textit{word-or-none} \\ \textsc{regional-prop} & \textit{rprop} \\ \ldots & \end{array}\right]$,

- there is a new sort *r(egional-)prop(erties)* which subsorts



- there is a new sort $\left[\begin{array}{ll} \textit{area-bracket} & \\ \textsc{co-bracket} & \textit{area-bracket} \end{array}\right]$

  with subsorts *left-bracket* and *right-bracket*,

- the attribute C(HUNKING-)PROP(ERTIES) is appropriate for *word*, and the value of the appropriateness function $\mathfrak{F}$ on $\langle \textit{word}, \textsc{cprop} \rangle$ is *cprop*.

- the attribute AREAS is appropriate for *phrase*, and the value of the appropriateness function $\mathfrak{F}$ on $\langle \textit{phrase}, \textsc{areas} \rangle$ is *list*.

Our theory contains principles that constrain the AREAS-value of a *phrase*:

- The bracketings are always nested.

- Not more than two *word*-objects may directly follow each other.

- For each pair of corresponding brackets, there exists a sublist of the form

$$\boxed{1}\left[\begin{array}{l} \textit{left-b} \\ \textsc{co-b}\ \boxed{2} \end{array}\right], \left[\begin{array}{l} \textit{word} \\ \textsc{cprop}\ \textsc{co-gov}\ \textit{none} \end{array}\right], \boxed{2}\left[\begin{array}{l} \textit{right-b} \\ \textsc{co-b}\ \boxed{1} \end{array}\right]$$

or

$$\boxed{1}\left[\begin{array}{l} \textit{left-b} \\ \textsc{co-b}\ \boxed{2} \end{array}\right], \boxed{3}\left[\begin{array}{l} \textit{word} \\ \textsc{cprop}\ \textsc{co-gov}\ \boxed{4} \end{array}\right], \ldots, \boxed{4}\left[\begin{array}{l} \textit{word} \\ \textsc{cprop}\ \textsc{co-gov}\ \boxed{3} \end{array}\right] \boxed{2}\left[\begin{array}{l} \textit{right-b} \\ \textsc{co-b}\ \boxed{1} \end{array}\right]$$

and the property of areas holds for '$\boxed{3} \ldots \boxed{4}$'.

- The PHON-value of a phrase is the concatenation of the PHON values of the elements of its AREAS-list.

- If a phrase is fully saturated, its AREAS-list satisfies the word order restrictions and the regional properties.

7 Summary and Future work

- An analysis for scrambling and extraposition in German has been sketched.

⇒ It has to be investigated whether this approach can be applied cross-linguistically.

- The AREAS-list is a flat linear structure. The bracketing can be directly inferred from the governor-restrictions given in the lexical entry of each word on the list.

⇒ From a computational point of view it should be relatively easy to get to the AREA-value in a parsing process. The regional properties can be considered a first approximation to the syntactic structure.

- Adopting the notions defined above, a new theory of extraction can be expressed: (parts of) an area or a region can be extracted instead of a syntactic constituent.

⇒ The German data can be accounted for by assuming syntactic or 'regional' movement. In the light of Serbo-Croatian data presented in (Penn 1996), the second alternative should be explored more seriously.

References

Höhle, T. N.: 1982, Explikationen für "normale Betonung" und "normale Wortstellung", *in* W. Abraham (ed.), *Satzglieder im Deutschen. Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, Tübingen.

Höhle, T. N.: 1986, Der Begriff "Mittelfeld". Anmerkungen über die Theorie der topologischen Felder, *in* A. Schöne (ed.), *Kontroversen alte und neue. Akten des 7. Internationalen Germanistenkongresses Göttingen 1985*, Vol. 3, pp. 329–340.

Jacobs, J.: 1988, Probleme der freien Wortstellung im Deutschen, *Sprache und Pragmatik* **5**, 8–37.

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

Keller, F.: 1994, Extraposition in HPSG, Manuscript.

King, P. J.: 1994, An expanded logical formalism for Head-Driven Phrase Structure Grammar, *Arbeitspapiere des SFB 340 59*, Universität Tübingen.

Lenerz, J.: 1977, *Zur Abfolge nominaler Satzglieder im Deutschen*, Verlag Gunter Narr. Tübingen.

Penn, G.: 1996, Multiple wh-extraction and clitic placement in Serbo-Croatian, Handout of a talk at the International Center Workshop on Computational Linguistics at the University of Tübingen.

Pollard, C. and Sag, I.: 1987, *Information Based Syntax and Semantics. Vol.1: Fundamentals*, CSLI Lecture Notes 13.

Pollard, C., Levine, R. and Kasper, R.: 1994, Studies in Constituent Ordering. Toward a Theory of Linearization in Head-Driven Phrase Structure Grammar, Manuscript. Research Proposal to the National Science Foundation, Ms. Ohio State University.

Reape, M.: 1992, *A Formal Theory of Word Order: A Case Study of West Germanic*, PhD thesis, University of Edinburgh.

Richter, F. and Sailer, M.: 1995, *Remarks on Linearization. Reflections on the Treatment of LP-Rules in HPSG in a Typed Feature Logic*, Master's thesis, Universität Tübingen.