

CASE AND WORD ORDER IN ENGLISH AND GERMAN

Allan Ramsay* & Reinhard Schäler⁺

*Department of Language Technology
UMIST, PO Box 88, Manchester M60 1QD, England

⁺Department of Computer Science
University College Dublin, Belfield, Dublin 4, Ireland

Abstract

It is often argued that English is a “fixed word order” language, whereas word order in German is “free”. The current paper shows how the different distributions of verb complements in the two languages can be described using very similar constraints, and discusses the use of these constraints for efficiently parsing a range of constructions in the two languages.

1 Background

The work reported here arises from an attempt to use a single syntactic/semantic framework, and a single parser, to cope with both German and English. The motivation behind this is partly practical — having a uniform treatment of a large part of the two languages should make it easier to develop MT and other systems which are supposed to manipulate texts in both languages; and partly theoretical, just because any shared structural properties of languages with differing surface characteristics are of interest in themselves.

The general framework is as follows:

- Lexical items contain detailed information about the arguments they require and the targets they can modify. This information includes a specification of where a particular argument or target will be found. For arguments, this is done by partially ordering the arguments in terms of which is to be found next and specifying the direction in which it is to be found via a feature that can take one of the values `left` or `right`. This is very similar to the treatment in categorial grammar, save that in the current approach the sequence in which arguments are to be found is given via a partial ordering, rather than the complete linear ordering of standard categorial grammar. Additionally, the feature specifying the direction in which to look for a particular argument may not be instantiated until immediately before that argument is required.
- There is a strictly compositional semantics, expressed using a dynamic version of (?) property theory¹.
- Syntactic, and hence semantic, analysis is performed by a chart parser driven by a “head-corner” strategy, whereby phrases are built up by combining the head with its arguments looking either to the right or the left depending on the direction specified by the next argument.

This system would analyse the sentence *he stole a car* as

$$\begin{aligned} \exists A :: \{ & \text{past}(A) \} \\ & \text{simple}(A, \\ & \quad \lambda B(\iota C :: \{ \text{subset}(C, \lambda D(\text{male}(D))) \\ & \quad \quad |C| = 1 \} \\ & \quad \exists E :: \{ \text{subset}(E, \\ & \quad \quad \quad \lambda F(\text{book}(F))) \\ & \quad \quad |E| = 1 \} \end{aligned}$$

⁰in *Recent Advances in Natural Language Processing*, eds. R. Mitkov & N. Nicolov, 1995

¹Property theory allows you to combine the standard logical truth functional operators with the abstraction operator of the λ -calculus without either running into the paradoxes of self-reference or being restricted by an otherwise unnecessary hierarchy of types. See (?) for phenomena whose analysis is greatly simplified by the absence of types from property theory.

$event(B)$
 $type(B, steal)$
 $object(B, E)$
 $by(B, C))$

This example displays most of the characteristics of our semantic analyses:

- We use an event-based semantics, with aspect interpreted as a relation between event types and temporal objects such as instants. An event type is represented as a λ -abstraction over sentences about events (though remember that we are using property theory rather than typed λ -calculus as the means to interpret such expressions).
- We use anchors to capture dynamic characteristics of referring expressions, so that an expression like $\iota C::\{subset(C, \lambda D(male(D))) \mid |C| = 1\}W$ says that W is true of the contextually unique singleton set of male individuals C if there is one, and is uninterpretable otherwise (in other words, W is true of *he*).
- Thematic relations are named after the prepositions that give them their most obvious syntactic marking, so that $by(A, B)$ means that B is the agent of the event A , since agency is marked by the use of the case-marking preposition *by* when it is marked at all.

This kind of semantic analysis is reasonably orthodox: the use of Davidsonian events has been widely adopted (e.g. see (?)), the treatment of referring expressions via anchors into the context is very similar to the use of anchors in situation semantics (?), the decision to use the names of case-marking prepositions for thematic relations can easily be justified by appeal to (?)'s analysis of the semantics of thematic relations. The most surprising element of the treatment above is the analysis of aspect as a relation between a temporal object and an event type: dealing with aspect this way provides more flexibility than is available in the approach taken by (?), but as far as the present paper is concerned it makes little difference and if you find it unintuitive then the best thing to do is ignore it.

Treatments of a variety of semantic phenomena *in English* have been published elsewhere (?; ?). The purpose of the current paper is to describe the syntactic devices which are used to indicate thematic role in English and to show how these can be adapted with very minor changes to obtain the same information in German.

2 Case and Order in English

English deploys two mechanisms for assigning thematic roles to arguments of a verb. (i) Thematic roles are partially ordered in terms of their affinity for the syntactic role of subject. In particular, if the list of required arguments includes an agent *and this argument is not explicitly case marked* then it must be the subject; and the only time the thematic object of any verb can be the syntactic subject is if there are no other candidates. The subject is always adjacent to the verb, either on the left (simple declarative sentences) or the right (aux-inverted questions). The subject has the surface case marker $+nom$. For passive verbs, the item which would have taken the role of subject for the active form is found and is marked as being optional and obligatorily case marked before the real subject is found. (ii) Any other arguments appear to the right of the verb and are otherwise freely ordered, with the proviso that the argument in what is usually termed direct object position should be required to be marked $+acc$ if possible, while any other arguments should have a case marking which reflects their thematic role. This case marking typically comes in the form of a preposition. Thus in

1 *He gave his mother a picture*

2 *He gave a picture to his mother*

he is the agent of the event, *his mother* is the recipient and *the picture* is the object. In both cases the subject has to be the agent, since agents always take precedence when allocating the role of subject. In (??) the second argument *a picture* has its thematic role assigned by the surface case marking. In this case, that surface case marking is $+acc$, which specifies that this argument is playing the role of object. This leaves the role of recipient to *his mother*. The explicit case marker for the role of recipient is overridden by the assignment of $+acc$ to whatever appears in direct object position, but it doesn't matter because it is already clear that the other two

arguments are the agent and the object, which leaves recipient as the only option. In (??) the second argument *to his mother* has the case marker *to*, and hence is clearly the recipient, leaving object as the only option for *a picture*.

The behaviour of the verb *open* fits the same pattern:

3 *He opened the door with the key*

4 *He opened the door*

5 *The key opened the door*

6 *The door opened*

(??) is just like (??): the role of subject is taken by the agent, the final argument has its thematic role explicitly marked by the case-marking preposition *with*, and the remaining argument gets the role of object because that's all that is left. In the other cases, the role of subject gets allocated to the agent in (??), to the instrument in (??), and the object in (??), in descending order of affinity. The only real problem is that we would expect to get

(??') *He opened the key the door*

as a sort of 'dative shift' variant of (??). We rule this out simply by banning the instrument of the verb *open* from appearing in this position.

This mapping between thematic roles and surface appearance is determined by three sets of rules. (i) Local rules may specify properties of particular arguments, e.g. that the agent of any passive verb must be marked *-nom*, or that the instrument of the verb *open* must be marked *-obj1*. (ii) A set of "subject affinity" rules specifies which thematic role will be realised by an NP playing the surface role of subject. (iii) A set of linear precedence rules of the kind introduced in GPSG (?) specifies the permitted orders in which the arguments of the verb may appear.

Subject affinity rules

The decision as to which item should take the role of subject is determined by a set of rules such as the following:

(S1) $X[+agent, +nom] \ll_{subj} Y$

(S2) $X[+nom] \ll_{subj} Y[+object]$

The first of these says that the agent is a better candidate for the role of subject than anything else is, *provided that it is in fact capable of playing this role at all*. The side-condition that the agent must be capable of playing this role is specified by the requirement that it should satisfy the property of being *+nom* — in certain circumstances, notably in passives, the agent is required to be explicitly case-marked by the preposition *by*, and hence cannot be the subject. In any sensible implementation the explicit case marking of the agent should precede the application of the subject affinity rules, but it is not in fact a logical necessity.

The second rule here says that the thematic object is the worst candidate, among those that are eligible, for this role.

These two rules cover most, if not all, cases in English: the only situations where they fail to determine the subject is if (i) there is no agent or the agent is not eligible, and (ii) there are two other arguments neither of which is the semantic object. Such situations are sufficiently rare to be ignored for the purposes of this paper.

Linear precedence rules

The notion of LP-rule used here is slightly different from the standard GPSG treatment. In particular, because the grammar here is highly lexical our LP-rules deal with the arguments of lexical items, rather than with daughters of ID-rules. We will want to use the LP-rules on the fly, to determine which argument to look for next, and where to look for it. The following are the key rules for the arguments of English verbs:

(LP1) $X \ll_{lp} Y[-nom, mother = X]$

(LP2) $X[+nom, mother = M]$

$\ll_{lp} Y[mother = M]$

(LP3) $X[+nom, mother = Y] \ll_{lp} Y[-inv]$

(LP1) says that any non-subject argument Y of X must follow X ; (LP2) says that the subject of M must precede any non-subject argument; and (LP3) says that if Y is marked as being non-invertible then its subject must precede it.

(S1–2) and (LP1–3) can be utilised within a head-corner parser to determine what argument to look for next and where to look for it, as follows:

- Start by applying the local rules: it’s best to do this before choosing the subject, since the local rules will generally only be compatible with one choice of subject, but it is not strictly necessary to do so.
- Next allocate the role of subject to one of the arguments of the verb by (S1–2). Require this item to be marked $+nom$.
- If there is an argument X of the verb V such that (i) $V \ll_{lp} X$ and (ii) there is no argument Y such that $V \ll_{lp} Y \ll_{lp} X$ then look to the right for X , and delete X from the set of arguments waiting to be found. This step cannot sensibly be performed until the subject has been found, since the LP rules depend on whether some item is $+/-nom$.
- If there is an argument X of the verb V such that (i) $X \ll_{lp} V$ and (ii) there is no argument Y such that $X \ll_{lp} Y \ll_{lp} V$ then look to the left for X , and delete X from the set of arguments waiting to be found.

With one non-trivial extension, these rules cover virtually all the relevant phenomena in English. The key extension concerns the presence or absence of an explicit case marker on the leftmost item after the subject and the verb. If we mark this item as $+obj1$, then we need a default rule of the form

$$\frac{M(X[+acc]) : X[+obj1]}{X[+acc]}$$

This says that if it is possible to require the item in the relevant position to be marked $+acc$ then you should do so. Unlike the previous rules this has to be a default rule and hence cannot be applied until the others have all done their work. The point here is that for a verb like *rely* the item in direct object position must be case-marked by the preposition *on*, as in *He relied on her integrity*: the consistency check in the above rule allows this by noting that the effect of the rule is incompatible with the effect of the lexical properties of *rely*, and hence the rule does not apply. Note that the requirement that the first non-subject argument after the verb has to be $+obj1$ provides the mechanism for ruling out *he opened the door the key* as a “dative” version of *he opened the door with the key*. We simply mark the instrument of *open* as $-obj1$ (though not $-nom$, since the instrument can get promoted to subject position if there is no explicit agent).

The above rather straightforward rules cover virtually all the relevant phenomena in English. In particular they provide appropriate analyses for (??)–(??), and for:

7 *A picture was given to his mother*

8 *His mother was given a picture*

9 *I saw him stealing a car*

10 *He was seen stealing a car*

11 *The ancient Greeks knew that the earth went round the sun*

12 *That the earth went round the sun was known to the ancient Greeks*²

²The case marker for the subject of the active sentence (??) turns out to be the preposition *to*, indicating that *The Greeks* is not the agent of *know*. This reflects the fact that agents typically intend the events that they bring about, which is not the case for the ancient Greeks in (??).

2.1 Extraposition in English

English word order is not, however, as rigidly fixed as these examples suggest. In particular, it is not unusual for one of the arguments which would normally appear to the right of the verb to appear way over to the left in front of the subject. The usual reason for this is that it provides a way of making the semantics of the shifted item available for some discourse operation such as contrast, as with *the book* in

13 *The film was banal, but the book I enjoyed*

A wide variety of more-or-less formal accounts of such discourse operations have been provided (e.g. (?; ?; ?; ?)), and there is no need to discuss their various merits and demerits here. The crucial point for the current paper is that surface word order does frequently get reorganised in this way in English, so that any claim that word order in English is fixed has to be treated very carefully.

It is worth noting at this point that VP modifiers such as PPs and ADVPs seem to be subject to very similar kinds of constraint on where they can appear. Cases such as *He suddenly stopped the car, he ate it in the park, I saw him sleeping by himself, . . .* seem to indicate that there is a general rule in English that says that a VP can be modified by an appropriate modifier, and that the modifier should appear to the left of the VP if it is head-final and to the right if it is head-initial³. This simple rule, however, is violated by examples like

14 *In the park he ate a peach*

15 *She believed with all her heart that he loved her*

In (??) the head-initial PP *in the park* is to the left of the S, rather than to the right of the VP; and in (??) the PP *with all her heart* is *between* the verb *believed* and its sentential complement *that he loved her*. In order to account for (??) we have to argue either that *in the park* can be either a left-modifier of an S or a right-modifier of a VP, in which case it will have to have different semantic types to combine appropriately with the types of its two potential targets; or that it is in fact a right-modifier of the VP which has been shifted to the left, probably in order to reduce ambiguity (since (??) has only one reading, whereas *he ate a peach in the park* has two) rather than to make it the argument of a discourse operator. The easiest way to account for (??) seems to be to argue that the complement *that he loved her* has been *right* shifted, probably again in order to reduce ambiguity (*she believed that he loved her with all her heart* sounds extremely odd, largely because the obvious attachment of *with all her heart* is to the VP *loved her*).

In the system being described here, these “shifts” of some argument or adjunct are dealt with using the standard unification technique of having a category valued feature called `slash` which can be given a value in order to denote the fact that some item is “missing” from its expected position. We extend the standard notion, however, by allowing `slash` to have a stack of items as its value, indicating that more than one thing has gone missing. This is a departure from standard practice — in GPSG, for instance, the foot feature principle specifies that `slash` can be given a non-trivial value by at most one of the daughters of a rule. This extension is required in English to cope with cases like

16 *I was just talking to him when suddenly he collapsed*

where the most obvious analysis assumes that both *when* and *suddenly* have been left-shifted. Much the same also holds for

17 *Quietly, without a word, he turned his face to the wall*

where *quietly* and *without a word* have both been topicalised, and for

18 *where I believed at the time that he had left it*

³Modifiers consisting of a single word, such as *quietly*, are both head-initial and head-final, so that you get both *he ate it quietly* and *he quietly ate it*.

where *where* has been topicalised out of *that he had left it*, which has itself been right-extrapolated in the same way as *that he loved her* in (??). We will assume from now on that there is no pre-determined limit on the number of items that may be shifted either right or left, though there may well be local constraints that prevent extraposition happening in particular cases. The decision to allow multiple extrapositions could easily lead to an explosion in the number of partial parses that might be constructed. We therefore use a mechanism similar to (?)’s notion of “sponsorship” to insist that for each object which you believe has been left-shifted there must indeed be at least one candidate item somewhere to the left. Furthermore, if more than one item has been left-extrapolated then the sponsors must appear in the right order. With this filter on the freedom to hypothesise left-extrapolations, our move to permitting multiple extrapositions does not lead to an unacceptable increase in the number of potential analyses⁴.

3 Case and Order in German

We now turn to German, where surface case marking seems to be rather more important than word order in the allocation of thematic roles to arguments. Very roughly, it seems that in German the following conditions hold:

- General properties of a clause determine whether the verb appears as the first, second or final constituent.
- One argument is marked as being the subject, and undergoes the usual agreement constraints for subjects.
- The arguments of a verb are not subject to a strict set of LP-rules, though quite strong discourse effects can be obtained by putting something other than the subject as the leftmost argument.

To take a simple example,

19 *Er gab seiner Mutter ein Bild*

20 *Er gab ein Bild seiner Mutter*

21 *Seiner Mutter gab er ein Bild*

22 *Ein Bild gab er seiner Mutter*

are all reasonable translations of *he gave a picture to his mother*. In each case, the choice of *er* as the subject indicates that he was the agent, the dative marking of *seiner Mutter* indicates that the mother was the recipient, and the accusative marking of *ein Bild* shows that this is the thematic object. Choosing (??) or (??) would normally presuppose that the speaker wanted to make *seiner Mutter* or *ein Bild* available for some discourse operator, but all four options are certainly permissible.

Similarly,

23 *Gab er seiner Mutter ein Bild?*

24 *Gab er ein Bild seiner Mutter?*

25 *Gab seiner Mutter er ein Bild?*

26 *Gab ein Bild er seiner Mutter?*

are all available as questions about the donation of a book to someone’s mother, with the choice of which argument is to come immediately after the verb indicating whether, as in (??), we don’t know whether the person he gave the book to was his mother, or, as in (??), we don’t know whether what he gave her was a book.

(?) argues that (??)—(??) and (??)—(??) can all be obtained, as for the English cases, from a set of rules which choose the subject, a set of LP-rules, and a mechanism for left-extraposition. The essence of Uszkoreit’s analysis is that there is one basic LP-rule, which in the terms used here would look like

⁴It does not seem possible to extend the notion of sponsorship to deal with right extrapositions, since you can’t anticipate whether sponsors may turn up later on as you proceed. Fortunately the local constraints tend to restrict the number of items that could possibly be right-shifted.

$X \ll_{lp} Y[mother = X]$

and that the simple declarative forms (??)—(??) are obtained by topicalisation.

This looks very straightforward, and the only change that we would argue for at this point is that Uszkoreit deals with cases like

27 *In dem Park aß er ein Apfel*

by treating *in dem Park* as an argument of *aß*, whereas it seems more sensible to treat it as an ordinary post-modifier of the VP and to allow it to be left-shifted just as in (??). It is notable that, in German as in English, cases where a preposition modifier is left-shifted are much less marked than ones where some other non-subject item appears in the leftmost position. The reason is that left-shifting a modifier can be used as a means of reducing ambiguity, and hence is a useful thing to do regardless of any discourse effect you want to produce.

The first point at which this simple rule has to be altered arises when we consider verbs other than the main verbs of major clauses (i.e. non-finite verbs and main verbs of subordinate clauses. Following Uszkoreit we will mark these as $-mc$). In

28 *Ich sah ihn ein Auto stehlen*

29 *Ich habe ein Auto gestohlen*

the NP *ein Auto* is certainly an argument of *(ge)ste(o)hlen*, yet appears to its left. It also seems as though in (??) *ihn* may also be an argument of *stehlen*, as something like a $+acc$ marked subject.

To accommodate these examples, we might adapt our observations about word order by simply saying that the arguments of a minor verb must precede it, and leave it at that. The LP rules would then become

$X[+mc] \ll_{lp} Y[mother = X]$

$Y[mother = X] \ll_{lp} X[-mc]$

These rules have much the same flavour as the ones for English, and could be used in just the same way by a parser which incrementally chose which argument to look for next and which direction to look for it in. Clearly the verb-second examples like (??)—(??) would require you to worry about left-extrapolation, but this is not a major extra burden since you will always have to worry about that anyway.

Unfortunately, you cannot always tell from the appearance of a verb whether it should be marked $+mc$ or $-mc$. Non-finite verbs are always $-mc$, but there are plenty of cases, e.g. *stehlen*, where the appearance of the verb does not determine its form; and even where the form is determined, you cannot know for a tensed verb whether it is $+mc$ or $-mc$ until you know the context in which it appears. This means that any bottom-up parser which depends on the two LP-rules above is frequently going to have to investigate two sets of hypotheses, one looking to the right for all the arguments of the verb and one looking to the left.

At this point it is worth recalling two points: (i) on Uszkoreit's account, the only difference between the polar interrogative form and the simple main clause declarative is that the latter has something (either an argument or a modifier) left-extrapolated. (ii) For entirely independent reasons, it seemed sensible in English to allow multiple items to be extrapolated. We therefore propose the following alternative treatment of $-mc$ verbs in German.

- There is only one LP-rule for verbs, namely $X \ll_{lp} Y[mother = X]$
- Polar interrogatives, simple declaratives and $-mc$ verbs are distinguished entirely by the number of items which have been left-shifted.

With these rules we get all the obvious cases, e.g.

30 *stehle er ein Auto?*

[[*stehle*, *right-er*], *right-[ein, right-Auto]*]

31 *er ein Auto stehle*

[*er*, [[*ein*, *right*–*Auto*],
[[*stehle*, *right*–*trace*], *right*–*trace*]]]

32 *ein Auto stehle er*

[[*ein*, *right*–*Auto*],
[[*stehle*, *right*–*er*], *right*–*trace*]]

The markers *left* and *right* in these indicate where the item in question was found, and *trace* indicates that what was found was a trace of something which has been extraposed. Thus in (??) both arguments were found to the right of the verb, but one of them was a trace which was cancelled by the NP *ein Auto* which itself consisted of a determiner with a noun to its right.

This is exactly as described by Uszkoreit for verb-initial and verb-second clauses. More interestingly, we can cope with embedded clauses without requiring *–mc* verbs to look to the left for their arguments:

33 *ich weiss, er stehle ein Auto,*

[*ich*,
[[*weiss*, *right*–*trace*],
right–[*er*, [[*stehle*, *right*–*trace*],
right–[*ein*, *right*–*Auto*]]]]]

34 *ich weiss, in dem Park stehle er ein Auto,*

[*ich*,
[[*weiss*, *right*–*trace*],
right
–[[*in*, *right*–[*dem*, *right*–*Park*]],
[[[*stehle*, *right*–*er*], *right*–[*ein*, *right*–*Auto*]],
right–*trace*]]]]]

35 *ich weiss, das er ein Auto stehle,*

[*ich*,
[[*weiss*, *right*–*trace*],
right
–[*das*,
right
–[*er*, [[*ein*, *right*–*Auto*],
[[*stehle*, *right*–*trace*], *right*–*trace*]]]]]]]

In (??) *weiss* requires a *+mc* clause as its argument, and hence *er stehle ein Auto*, with one left-shifted argument is fine. Similarly, the presence of the left-shifted PP in (??) means that the embedded clause is *+mc*. In (??), on the other hand, the complementiser *das* requires a *–mc* clause as its argument, and hence both arguments *er* and *ein Auto* of *stehle* have to be left-shifted. The complementiser then returns a *+mc* clause, as required.

Similarly, in

36 *ein Auto, das er stehle,*

[*ein*,
right
–[*Auto*, *right*–[*das*, [*er*, [[*stehle*, *right*–*trace*],
right–*trace*]]]]]]]

the relative clause has to be $-mc$ and hence has all its arguments left-shifted, with the WH-pronoun (!) *das* coming first because of the fact that you can't extrapose anything from a sentence which has already been WH-marked (so you don't get *ein Auto, er das stehle*).

Verbs with non-finite sentential complements work exactly the same way:

37 *ich sah ihn ein Auto stehlen*

[*ich*,
 [[*sah*, *right-trace*],
 right
 - [*ihn*, [[*ein*, *right-Auto*],
 [[*stehlen*, *right-trace*], *right-trace*]]]]]

The embedded clause *ihn ein Auto stehlen* has a non-finite, hence $-mc$, main verb, and therefore both arguments have again been left-shifted.

Auxiliaries are slightly more awkward. In English, auxiliaries and modals take VPs as their arguments, i.e. verbs which have found all their arguments apart from the subject. To deal with that in the current context, we would have to allow slash elimination to occur with VP's as well as with S's, analysing the phrase *ein Auto gestohlen* in

38 *ich habe ein Auto gestohlen*

by taking *gestohlen* as something like $VP[*subcat* = \{NP[+nom]\}, *slash* = \{NP\}]$ and then cancelling the slashed NP against *ein Auto* to obtain a normal VP.

This is a possibility, but the decision to allow slash elimination to occur with items other than S's is a major step. For the moment we prefer to assume that auxiliaries and modals require S's whose subjects have been extraposed, rather than ones whose subjects have not been found, and to retain the principle that slash elimination only occurs with S's. We therefore treat (??) as

[*ich*,
 [[*habe*,
 right
 - [[*ein*, *right-Auto*],
 [[*gestohlen*, *right-trace*], *right-trace*]]],
 right-trace]]]

Here *gestohlen* has had both arguments extraposed, and *habe* has had its subject extraposed. Only one of the arguments for *gestohlen* gets cancelled, namely *ein Auto*, and therefore there is an S with its subject missing immediately to the right of *habe*. This is therefore accepted as one argument, and the other is slashed. When it turns up, namely as *ich*, the whole thing turns out to be a perfectly ordinary declarative main clause.

This may turn out not to be the best solution for auxiliaries. For the moment we will just note that it does at least work, and that it does not require any radical extensions to the analysis developed above for the other cases. We will be looking again at this, but it does at least provide a treatment that works without incurring any substantial extra costs.

4 Implementation

The rules outlined above for computing the properties of the next argument to be found when saturating a verb in English and German have been implemented in a version of the parser and syntax/semantics reported in (?; ?). Within this framework as much information as seems sensible is packed into the descriptions of lexical items, with a very small number of rules being used for saturating and combining structures together. In particular, the description of a lexical item *W* contains the following pieces of information:

- a description of the syntactic properties of the item *W'* that would result from saturating *W*.

- a description of the *set* of arguments which *W* requires. This set may be empty, as in the case of pronouns or simple nouns.
- a description of the items that *W'* might modify (e.g. an adjective like *old* would specify that it could modify an \bar{N} , a preposition like *in* would specify that when saturated it could modify either an \bar{N} or a VP).

The grammar then has four rules:

- An unsaturated item can combine with one of its arguments under appropriate circumstances.
- A modifier can combine with an appropriate target.
- A sentence which has had something extraposed to the left or right can combine with an appropriate item on the left or right.
- If *X'* is a redescription of *X* then any of the first three rules can be applied to *X'*. This rule captures the notion that items can often be viewed from different perspectives — that a generic \bar{N} can be seen as an NP, that certain sorts of WH-clause can be seen as NP's (e.g. *I don't know much about art but I know what I like*), and so on.

These rules are simple enough for it to be reasonable to build the parser around them. The key, of course, is that the first three all talk about “appropriate” items and circumstances, and this notion of appropriate needs to be fleshed out. Part of what is meant here is that feature percolation principles have to be applied in order to complete the descriptions of the required items. These feature percolation principles are essentially dynamic, since they include pre-defaults which say things like “unless you already know that *X* is required to be something else then require it to be *+acc*”; post-defaults, which say things like “unless you know that *X* is capable of functioning as an adjunct then assume it isn't”; and principles like the FFP which depend on properties of the siblings of the item in question. The issue of appropriateness also includes information about which argument from a set of arguments to look for next, and whether to look to the left or the right for it; and about whether a modifier should appear to the left or right of its target, or whether an extraposed item should be found to the left or right of the sentence it has been extracted from.

The question of whether to look to the left or right for an item is also essentially dynamic. Consider, for instance, the following NP's:

39 *a sleeping man*

40 *a quietly sleeping man*

41 *a man sleeping in the park*

In (??) and (??) the modifier has to appear to the left of the target, in (??) it has to appear to the right. The reason seems to be that *sleeping* and *quietly sleeping* are head-final, whereas *sleeping in the park* is not. This is a property of the phrase as a whole, rather than of its individual components, and hence cannot be determined until the whole phrase has been found. Similarly, the discussion of case-marking and argument order in Sections 2 and 3 above suggests that the direction in which the next argument should be found and the details of its syntactic properties depend on what was found last and what properties *it* had.

Given this dynamic view of these otherwise rather skeletal rules, it seems reasonable to embody them directly into the parser. The term “head-corner” reflects the fact that we work outwards from lexical items, trying to saturate them by looking either left or right, as determined dynamically by the LP-rules. This strategy provides a very effective combination of top-down and bottom-up processing. As examples of cases where this pays off, consider the following English sentences:

42 *That she should be so confident says a lot for her education.*

43 *Eating raw eggs can give you salmonella poisoning.*

44 *There is a dead rat in the kitchen.*

In (??), the sentence *that she should be so confident* is the subject of the verb *says*; in (??) the subject of *give* is the VP *eating raw eggs*; and in (??) the subject of *is* is the dummy item *there*. The fact that verbs can require either non-NP's or extremely special NP's as their subjects means that you can't afford to have a simple rule like

$$S \Rightarrow NP, VP[+tensed]$$

since it won't cover (??), it won't cover (??) unless you regard present participle VP's as being a species of NP, and it won't specify the detailed characteristics of the subject NP in (??). You would therefore need a rule more like:

$$S \Rightarrow X, VP[+tensed, subject = X]$$

But any parser which worked generally left to right would produce unacceptable numbers of hypotheses in the presence of a rule like this. By working outwards from the head verb in directions specified by the LP-rules, we can cope with (??)—(??) without drowning in a sea of unwarranted hypotheses. Similarly, by replacing the general rule

$$X \Rightarrow X, conj, X$$

by lexical entries whose subcategorisation frames say that a conjunction can be saturated to an X if you find an X to the left and then one to the right, we can cope with the combinatorial explosion that a rule of this kind would otherwise introduce.

In much the same way, the fact that we determine the direction in which a modifier is to seek its target dynamically means that we can be economical about making hypotheses about where to look for adjunct/target pairs. The main reason for providing distinct mechanisms for combining heads with arguments and adjuncts with targets comes from our desire to treat examples like

45 *In the park there is a playground for pre-school children.*

as involving extraposition of the PP *in the park*. This treatment is motivated on semantic grounds, since otherwise we have to be prepared to treat *in the park* as both a function of type $t \rightarrow t$ when it modifies an S, as would happen in (??); and as a function of type $((e \rightarrow t) \rightarrow (e \rightarrow t))$ when it modifies a VP, as in:

46 *The youths drinking cider in the park looked extremely threatening.*

The key difference is that in head/argument pairs, the argument can be extraposed, whereas in modifier/target pairs the modifier can. We therefore cannot afford to treat a preposition like *in* as being of type $VP \setminus VP/NP$, as would be done in raw categorial grammar, since there is no obvious way of extraposing the partially saturated structure *in the park* from this.

This parser works fine for English. It works even better for German. Consider the verb *gab*. On the analysis outlined above, this generates six possible orders for the arguments, namely *agent-object-recipient*, *agent-recipient-object*, *object-agent-recipient*, *object-recipient-agent*, *recipient-object-agent*, *recipient-agent-object*. Some of these mark strong rhetorical devices, and others may only be possible with particular combinations of +/-heavy NP's, but they are all at least conceivable. Furthermore, if we take it that *gab* can appear in polar questions, +mc declarative sentences, and -mc clauses, then each of these can appear with the verb either at the start, after the first item, or at the end — a total of 18 possible sequences. And then we have to consider the possible presence of adjuncts, which could easily lead to +mc declarative forms in which the verb precedes all three core arguments. And finally, of course, in each case we have to consider the possibility that a given argument may have been extraposed, either for rhetorical reasons or simply to construct a relative clause.

Within the current framework, we initially generate just three hypotheses — that the agent is the leftmost argument, or that the object is, or that the recipient is. We then look to the right for this argument: if we find a concrete instance then the case marking will almost certainly rule out all except one case, and if we decide to hallucinate an extraposed instance then the search for sponsors will ensure that we only do so if there is indeed something of the required kind already lying around. We therefore explore only a very constrained part of the overall search space. The parser we developed initially for English actually works even better for German!

5 English is German

Uszkoreit, rightly, complains that a consequence of the historical concentration on English is that other languages get forced into a framework which really does not fit them at all well. This is particularly unfortunate in view of the fact that English is in fact a rather messy amalgam of other languages, with German being a notable contributor. It is therefore appropriate to finish the current paper by noting a couple of English constructions which do not fit the analysis outlined in Section 2 above, but which do behave very much like the constructions described in Section 3.

The first is a rather archaic form of polar question. It used to be possible to say things like

47 *Know ye not who I am?*

rather than

48 *Don't you know who I am?*

(??) is exactly parallel to the standard form of German polar question, and it is tempting to treat it in exactly the same way. It is also tempting, of course, to treat it using the standard English rules but allowing words other than auxiliaries to be marked *+inv*, and it would be a mistake to make too much of this example, but it is at the very least provocative.

Perhaps more significant is the topicalisation of

49 *An old man was on the bus.*

to

50 *On the bus was an old man.*

The standard rules for topicalisation in English would have produced *On the bus an old man was*, parallel to *On the bus an old man slept*. The German rules, however, would have produced (??). Should we therefore deal with this one as though the English copula was in fact subject to the German LP-rules? Is at least part of English just German?

References

- Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. Bradford Books, Cambridge, MA.
- Dowty, D. R. (1988). Type raising, functional composition and non-constituent conjunction. In R.T. Oehrle, E. B. and Wheeler, D., editors, *Categorial Grammars and Natural Language Structures*, pages 153–198, Dordrecht. Kluwer Academic Press.
- Gazdar, G., Klein, E., Pullum, G. K., and Sag, I. (1985). *Generalised Phrase Structure Grammar*. Basil Blackwell, Oxford.
- Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. Arnold, London.
- Hoffman, B. (1995). Integrating “free”, word order syntax and information structure. In *EACL-95*, pages 245–252, Dublin.
- Johnson, M. and Kay, M. (1994). Parsing and empty nodes. *Computational Linguistics*, 20(2):289–300.
- Krifka, M. (1993). Focus, presupposition and dynamic interpretation. *Journal of Semantics*, 10.
- Moëns, M. and Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

- Ramsay, A. M. (1992). Bare plural nps and habitual vps. In *COLING-92*, pages 226–231, Nantes.
- Ramsay, A. M. (1994). Focus on “only”, and “not”. In *COLING-94*, pages 881–885, Kyoto.
- Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, 13(1).
- Turner, R. (1987). A theory of properties. *Journal of Symbolic Logic*, 52(2):455–472.
- Uszkoreit, H. (1987). *Word Order and Constituent Structure in German*. CSLI, Stanford.
- van Eijck, J. and Alshawi, H. (1992). Logical forms. In Alshawi, H., editor, *The Core Language Engine*, pages 11–40, Cambridge, Mass. Bradford Books/MIT Press.
- Williams, E. (1981). On the notions ‘lexically related’ and ‘head of a word’. *Linguistic Inquiry*, 12:254–274.