GIDLP: A Grammar Format For Linearization-based HPSG

Michael W. Daniels and W. Detmar Meurers
The Ohio State University

Proceedings of the HPSG04 Conference
Center for Computational Linguistics
Katholieke Universiteit Leuven
Stefan Müller (Editor)
2004
CSLI Publications
http://csli-publications.stanford.edu/

Abstract

Linearization-based HPSG theories are widely used for analyzing languages with relatively free constituent order. This paper introduces the Generalized ID/LP (GIDLP) grammar format, which supports a direct encoding of such theories, and discusses key aspects of a parser that makes use of the dominance, precedence, and linearization domain information explicitly encoded in this grammar format. We show that GIDLP grammars avoid the explosion in the number of rules required under a traditional phrase structure analysis of free constituent order. As a result, GIDLP grammars support more modular and compact grammar encodings and require fewer edges in parsing.

1 Introduction

Within the framework of Head-Driven Phrase Structure Grammar (HPSG), the so-called linearization-based approaches have argued that constraints on word order are best captured within domains that extend beyond the local tree. A range of analyses for languages with relatively free constituent order have been developed on this basis (see, for example, Reape 1993; Kathol 1995; Müller 1999a; Donohue and Sag 1999; Bonami et al. 1999) so that it is attractive to exploit these approaches for processing languages with relatively free constituent order.

This paper introduces a grammar format that supports a direct encoding of linearization-based HPSG theories. The Generalized ID/LP (GIDLP) format explicitly encodes the dominance, precedence, and linearization domain information and thereby supports the development of efficient parsing algorithm making use of this information. We make this concrete by discussing key aspects of a parser for GIDLP grammars that integrates the word order domains and constraints into the parsing process.

2 Linearization-based HPSG

The idea of discontinuous constituency was first introduced into HPSG in a series of papers by Mike Reape (see Reape 1993, and references therein).[1] The core idea is that word order is determined not at the level of the local tree, but at the newly introduced level of an *order domain*, which can include elements from several local trees. We interpret this in the following way: Each terminal has a corresponding

*This paper includes material from (Daniels and Meurers 2004). The authors would like to thank Stefan Müller and the anonymous reviewers for HPSG04 and COLING04, as well as the HPSG04 and COLING04 audiences for advice and helpful comments.

[1] Apart from Reape's approach, there have been proposals for a more complete separation of word order and syntactic structure in HPSG (see, for example, Richter and Sailer 2001; Penn 1999), an option outside the scope of this paper.

order domain, and just as constituents combine to form larger constituents, so do their order domains combine to form larger order domains.

Following Reape, a daughter's order domain enters its mother's order domain in one of two ways. The first possibility, *domain union*, forms the mother's order domain by shuffling together its daughters' domains. The second option, *domain compaction*, inserts a daughter's order domain into its mother's. Compaction has two effects:

- **Contiguity**: The terminal yield of a compacted category contains all and only the terminal yield of the nodes it dominates; there are no holes or additional strings.
- **LP Locality**: Precedence statements only constrain the order among elements within the same compacted domain. In other words, precedence constraints cannot look into a compacted domain.

Note that these are two distinct functions of domain compaction: defining a domain as covering a contiguous stretch of terminals is in principle independent of defining a domain of elements for LP constraints to apply to. In linearization-based HPSG, domain compaction encodes both aspects.

Later work (Kathol and Pollard 1995; Kathol 1995; Yatabe 1996) introduced the notion of *partial compaction*, in which only a portion of the daughter's order domain is compacted; the remaining elements are domain unioned.

3 Processing linearization-based HPSG

Formally, a theory in the HPSG architecture consists of a set of constraints on the data structures introduced in the signature. As such, word order domains are just additional structures, and the constraints on word order domains are no different from constraints on any other structure, and so the incorporation of linearization into a linguistic theory creates no formal difficulties. On the computational side, however, most systems employ parsers to efficiently process HPSG-based grammars organized around a phrase structure backbone. Phrase structure rules encode immediate dominance (ID) and linear precedence (LP) information in local trees, so they cannot directly encode linearization-based HPSG, which posits word order domains that can extend the local trees.

The ID/LP grammar format (Gazdar et al. 1985) was introduced to separate immediate dominance from linear precedence, and several proposals have been made for direct parsing of ID/LP grammars (see, for example, Shieber 1984). However, the domain in which word order is determined still is the local tree licensed by an ID rule, which is insufficient for a direct encoding of linearization-based HPSG.

The LSL grammar format as defined by Suhre (1999) (based on (Götz and Penn 1997)) allows elements to be ordered in domains that are larger than a local tree; as a result, categories are not required to cover contiguous strings. Linear precedence

constraints, however, remain restricted to local trees: elements that are linearized in a word order domain larger than their local tree cannot be constrained. The approach thus provides valuable worst-case complexity results, but it is inadequate for encoding linearization-based HPSG theories, which crucially rely on the possibility to express linear precedence constraints on the elements within a word order domain.

In sum, no grammar format is currently available that adequately supports the encoding of a processing backbone for linearization-based HPSG grammars. As a result, implementations of linearization-based HPSG grammars have taken one of two options. Some simply do not use a parser, such as the work based on ConTroll (Götz and Meurers 1997); as a consequence, the efficiency and termination properties of parsers do not (automatically) transfer to such approaches.

The other approaches use a minimal parser that can only take advantage of a small subset of the requisite constraints. Such parsers are typically limited to the general concept of resource sensitivity – every element in the input needs to be found exactly once – and the ability to require certain categories to dominate a contiguous segment of the input. Some of these approaches (Johnson 1985; Reape 1991) lack word order constraints altogether. Others (van Noord 1991; Ramsay 1999) have the grammar writer provide a combinatory predicate (such as concatenate, shuffle, or head-wrap) for each rule specifying how the string coverage of the mother is determined from the string coverages of the daughter. In either case, the task of constructing a word order domain and enforcing word order constraints in that domain is left out of the parsing algorithm; as a result, constraints on word order domains either cannot be stated or are tested in a separate clean-up phase following the generate-and-test paradigm.

4 Defining GIDLP Grammars

To develop a grammar format for linearization-based HPSG, we take the syntax of ID/LP rules and augment it with a means for specifying which daughters form compacted domains. A Generalized ID/LP (GIDLP) grammar consists of four parts: a start declaration, a set of lexical entries, a set of grammar rules, and a set of global order constraints. We begin by describing the first three parts, which are reminiscent of context-free grammars (CFGs), and then address order constraints in section 4.1.[2]

- The **start declaration** has the form $start(S, L)$ and states the start symbol S of the grammar and any linear precedence constraints L constraining the start domain.

[2]We base the discussion in this paper on simple term categories; nothing hinges on this, and when using the formalism to encode linearization-based HPSG grammars, one will naturally use the feature descriptions known from HPSG as categories.

- **Lexical entries** have the form $A \rightharpoonup t$ and link the pre-terminal A to the terminal t, just as in CFGs.
- **Grammar rules** have the form $A \rightharpoonup \alpha; C$. They specify that a non-terminal A immediately dominates a list of non-terminals α in a domain where a set of order constraints C holds.
- **Global LP constraints**, as described in section 4.1.1.
- **Global compaction statements**, as described in section 4.1.2.

Note that in contrast to CFG rules, the order of the elements in α does not encode immediate precedence or otherwise contribute to the denotational meaning of the rule. Instead, the order can be used to generalize the head marking used in grammars for head-driven parsing (Kay 1990; van Noord 1991) by additionally ordering the non-head daughters; this is discussed further in section 6.

If the set of order constraints is empty, we obtain the simplest type of rule, exemplified in (1).

(1) S $\rightharpoonup$ NP, VP

This rule says that an *S* may immediately dominate an *NP* and a *VP*, with no constraints on the relative ordering of *NP* and *VP*. If no other rule in the grammar imposes additional constraints, the lexical material dominated by *NP* may appear before, after, or intermingled with the material dominated by *VP*; material from other constituents not dominated by *S* may also intervene.

4.1 Order Constraints

GIDLP grammars include two types of order constraints: LP constraints and compaction statements.

4.1.1 Linear Precedence Constraints

All LP constraints enforce the intuitive idea that any instance of the LHS of the constraint must precede any instance of the RHS within the same context: individual rules (as *rule-level* constraints) or word order domains (as *domain-level* constraints). Domain-level constraints can also be specified as *global* order constraints, which has the effect that they are specified for each single domain.

The formal definition of precedence is as follows: consider all pairs of elements in a context where the first completely precedes the second. If any of these pairs jointly matches[3] the pair description $\langle B, A \rangle$, the constraint is violated.[4]

[3]The precise definition of 'match' will depend on the nature of A and B. For instance, matching involves an identity test when categories are atomic and a subsumption test when categories are feature structures.

[4]This definition is due to (Kasper et al. 1995) and is intended to deal with cases where the nature of the match between the first element and A will influence whether or not the second element matches B, and vice versa.

LP constraints may optionally require that there be no intervening material between the two elements: this is referred to as immediate precedence. LP constraints are notated as follows:

- **Weak precedence:** $A < B$.
- **Immediate precedence:** $A \ll B$.

The symbols *A* and *B* may be *descriptions* or *tokens*. A constraint involving descriptions applies to any pair of elements in any domain in which the described categories occur; it thus can also apply more than once within a given rule or domain. Tokens, on the other hand, can only occur in rule-level constraints and refer to particular RHS members of a rule. In this paper, tokens are represented by numbers referring to the subscripted indices on the RHS categories.

In (2) we see an example of a rule-level linear precedence constraint.

(2) $A \rightharpoonup NP_1, V_2, NP_3;\ 3 < V$

This constraint specifies that the token 3 in the rule's RHS (the second *NP*) must precede any constituents described as *V* occurring in the same domain (this includes, but is not limited to, the *V* introduced by the rule).

4.1.2 Compaction Statements

As with LP constraints, compaction statements exist as rule-level and as global order constraints; they cannot, however, occur within other compaction statements. A rule-level compaction statement has the form $\langle \alpha, A, L \rangle$, where α is a list of tokens, *A* is the category representing the compacted domain, and *L* is a list of domain-level precedence constraints. Such a statement specifies that the constituents referenced in α form a compacted domain with category *A*, inside of which the order constraints in *L* hold. As specified in section 2, a compacted domain must be contiguous (contain all and only the terminal yield of the elements in that domain), and it constitutes a local domain for LP statements.

It is because of partial compaction that the second component *A* in a compaction statement is needed. If only one constituent is compacted, the resulting domain will be of the same category; but when multiple categories are fused in partial compaction, the category of the resulting domain needs to be determined so that LP constraints can refer to it.

The rule in (3) illustrates compaction: each of the *S* categories forms its own domain. In (4) partial compaction is illustrated: the *V* and the first *NP* form a domain named *VP* to the exclusion of the second *NP*.

(3) $S \rightharpoonup S_1, Conj_2, S_3;\ 1 \ll 2, 2 \ll 3, \langle [1], S, \langle [] \rangle \rangle, \langle [3], S, \langle [] \rangle \rangle$

(4) $VP \rightharpoonup V_1, NP_2, NP_3;\ \langle [1, 2], VP, \langle [] \rangle \rangle$

One will often compact only a single category without adding domain-specific LP constraints, so we introduce the abbreviatory notation of writing such a compacted category in square brackets. In this way (3) can be written as (5).

(5) $S \rightharpoonup [S_1], Conj_2, [S_3];\ 1 \ll 2, 2 \ll 3$

A final abbreviatory device is useful when the entire RHS of a rule forms a single domain, which Suhre (1999) refers to as "left isolation". This is denoted by using the token 0 in the compaction statement if linear precedence constraints are attached, or by enclosing the LHS category in square brackets, otherwise. (See rules (23d) and (23j) in section 8 for an example of this notation.)

The formalism also supports *global compaction statements*. A global compaction statement has the form $\langle A, L \rangle$, where A is a description specifying a category that always forms a compacted domain, and L is a list of domain-level precedence constraints applying to the compacted domain.

4.2 Examples

We start with an example illustrating how a CFG rule is encoded in GIDLP format. A CFG rule encodes the fact that each element of the RHS immediately precedes the next, and that the mother category dominates a contiguous string. The context-free rule in (6) is therefore equivalent to the GIDLP rule shown in (7).

(6) $S \rightarrow Nom\ V\ Acc$

(7) $[S] \rightharpoonup V_1, Nom_2, Acc_3;\ 2 \ll 1, 1 \ll 3$

In (8) we see a more interesting example of a GIDLP grammar.

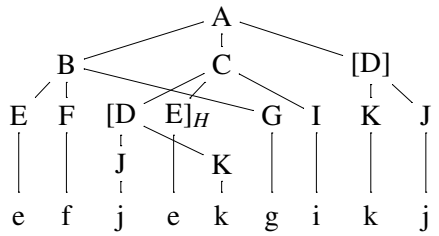
(8)
a) start(A, [])
b) $A \rightharpoonup B_1, C_2, [D_3];\ 2 < 3$
c) $B \rightharpoonup F_1, G_2, E_3$
d) $C \rightharpoonup E_1, D_2, I_3;\ \langle [1,2], H, \langle [] \rangle \rangle$
e) $D \rightharpoonup J_1, K_2$
f) Lexical entries: $E \rightharpoonup e, \ldots$
g) $E < F$

(8a) is the start declaration, stating that an input string must parse as an A; the empty list shows that no LP constraints are specifically declared for this domain. (8b) is a grammar rule stating that an A may immediately dominate a B, a C, and a D; it further states that the second constituent must precede the third and that the third is a compacted domain. (8c) gives a rule for B: it dominates an F, a G, and an E, in no particular order. (8d) is the rule for C, illustrating partial compaction: its first two constituents jointly form a compacted domain, which is given the name H. (8e) gives the rule for D and (8f) specifies the lexical entries (here, the preterminals

just rewrite to the respective lowercase terminal). Finally, (8g) introduces a global LP constraint requiring an E to precede an F whenever both elements occur in the same domain.

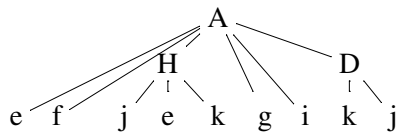
Now consider licensing the string **efjekgikj** with the above grammar. The parse tree, recording which rules are applied, is shown in (9). Given that the domains in which word order is determined can be larger than the local trees, we see crossing branches where discontinuous constituents are licensed.

(9)



To obtain a representation in which the order domains are represented as local trees again, we can draw a tree with the compacted domains forming the nodes, as shown in (10).

(10)



There are three non-lexical compacted domains in the tree in (9): the start A, the compacted D, and the partial compaction of D and E forming the domain H within C. In each domain, the global LP constraint $E < F$ must be obeyed. Note that the string is licensed by this grammar even though the second occurrence of E does not precede the F. This E is inside a compacted domain and therefore is not in the same domain as the F, so that the LP constraint does not apply to those two elements. This illustrates the property of LP locality: domain compaction acts as a 'barrier' to LP application.

The second aspect of domain compaction, contiguity, is also illustrated by the example, in connection with the difference between total and partial compaction. The compaction of D specified in (8b) requires that the material it dominates be a contiguous segment of the input. In contrast, the partial compaction of the first two RHS categories in rule (8d) requires that the material dominated by D and E, taken together, be a continuous segment. This allows the second e to occur between the two categories dominated by D.

Finally, the two tree representations above illustrate the separation of the combinatorial potential of rules (9) from the flatter word order domains (10) that the GIDLP format achieves. It would, of course, be possible to write phrase structure rules that license the word order domain tree in (10) directly, but this would amount to replacing a set of general rules with a much greater number of flatter

rules corresponding to the set of all possible ways in which the original rules could be combined without introducing domain compaction. Müller (2004) discusses the combinatorial explosion of rules that results for an analysis of German if one wants to flatten the trees in this way. If recursive rules such as adjunction are included – which is necessary since adjuncts and complements can be freely intermixed in the German Mittelfeld – such flattening will not even lead to a finite number of rules. We will return to this issue in section 8.

4.3 The Formal Definition of a GIDLP Grammar

The formal definition of a GIDLP grammar arises from the intuition behind the formal definition of a context-free grammar:

> A *context-free grammar* G is a quadruple (V, Σ, R, S), where V is an alphabet, Σ (the set of *terminals*) is a subset of V, R (the set of *rules*) is a finite subset of $(V - \Sigma) \times V^*$, and S (the *start symbol*) is an element of $V - \Sigma$. The members of $V - \Sigma$ are called *nonterminals*. For any $A \in V - \Sigma$ and $u \in V^*$, we write $A \rightarrow_G u$ whenever $(A, u) \in R$. For any strings $u, v \in V^*$, we write $u \Rightarrow_G v$ if and only if there are strings $x, y \in V^*$ and $A \in V - \Sigma$ such that $u = xAy$, $v = xv'y$, and $A \rightarrow_G v'$. The relation $\Rightarrow_G^*$ is the reflexive, transitive closure of $\Rightarrow_G$. Finally, $L(G)$, the *language generated by* G, is $\{w \in \Sigma^* : S \Rightarrow_G^* w\}$; we also say that G *generates* each string in $L(G)$ (Lewis and Papadimitriou 1998).

In particular, just as a context-free derivation is expressed as a series of strings over the corresponding alphabet, a GIDLP derivation will need to be expressed in terms of a series of domain objects that may contain categories and other domain objects.

A **GIDLP grammar** is a quintuple $\langle T, N, R, L, G \rangle$ where T is a set of terminals, N is a set of nonterminal categories,[5] R is the start domain object (defined below), L a relation from $T \rightarrow N$ (the lexicon), and G a set of grammar rules (the grammar).

A **grammar rule** is a triple $\langle A, \alpha, C \rangle$ where $A \in N$ is the left-hand side category of the rule, α is a string of category-token pairs $\langle a, b \rangle$, where $a \in N$ and $b \in \mathbb{N}$, and/or domain objects (the right-hand side of the rule), and C is a set of LP constraints.

A **domain object** is a triple $\langle A, \alpha, C \rangle$ where $A \in N$ is the result category of the domain, α is a string of nonterminals and/or domain objects, and C is a set of LP constraints.

An **LP constraint** is a triple $\langle a, b, t \rangle$ where $a, b \in (N \cup \mathbb{N})$ and $t \in \{w, i\}$ (representing weak and immediate precedence, respectively). Such an LP constraint is **satisfied** by a domain object when, for all pairs of distinct domain elements x, y such that x precedes y, it must be the case that x, y does not match the pair description b, a. In addition, if $t = i$, then for all pairs of distinct domain elements x, y

[5] The definitions in this section are intended to be independent of the nature of the elements of N; the only requirement is that the operation of pairwise-matching is defined. This allows the definitions to work with both atomic categories and feature structure categories.

such that x does not immediately precede y, it must be the case that x, y does not match the pair description a, b.

Now let A be the domain object $\langle a, \alpha A \beta, C_d \rangle$ and A' the domain object $A' = \langle a, \gamma, C_d \rangle$. If there is a rule $r \in G$ such that $r = \langle A, \delta, C_r \rangle$ and γ is a permutation of $\alpha \cdot \delta \cdot \beta$ such that, for all $c \in (C_d \cup C_r)$, γ satisfies c, then we say that $A \Rightarrow A'$ (read A **derives** A' **in one step**). In effect, γ represents a valid insertion of δ into $\alpha\beta$.

The transitive closure of $\Rightarrow$ is denoted $\Rightarrow^*$; when $A \Rightarrow^* A'$, we say A **derives** A', and A' is derived from A. Finally, let a **preterminal string** s of a terminal string t with length n be a string of length n such that for all $0 \leq i < n$, $\langle t_i, s_i \rangle \in L$. Then a string of terminals is **recognized** by a grammar if there exists a corresponding preterminal string that can be derived from the start domain object of the grammar.

As an example, the grammar in (8) is formally described in (11) (for clarity, rule RHSs are given in terms of categories only instead of category-token pairs).

(11)

$$
\begin{aligned}
T &= \{e, f, g, i, j, k\} \\
N &= \{A, B, C, D, E, F, G, H, I, J, K\} \\
R &= \langle A, [A], \{\langle E, F, w \rangle\} \rangle \\
L &= \{\langle e, E \rangle, \langle f, F \rangle, \ldots\} \\
G &= \{ \langle A, [B, C, \langle D, [D], \{\langle E, F, w \rangle\} \rangle], \{\langle 2, 3, w \rangle\} \rangle, \\
&\qquad \langle B, [F, G, E], \emptyset \rangle, \\
&\qquad \langle C, [\langle H, [D, E], \{\langle E, F, w \rangle\} \rangle, I], \emptyset \rangle, \\
&\qquad \langle D, [J, K], \emptyset \rangle\}
\end{aligned}
$$

Note that, aside from the fact that the compaction statements appear 'inside' the rules, (8) and (11) only differ in the absence of global order statements; these are merely an abbreviatory device for grammar writers. The derivation of the string *efjekgikj* is given in (12).

(12)

$$
\begin{aligned}
&\langle A, [A] \rangle \\
&\Rightarrow \langle A, [B, C, \langle D, [D] \rangle] \rangle \\
&\Rightarrow \langle A, [B, C, \langle D, [D] \rangle] \rangle \\
&\Rightarrow \langle A, [E, F, G, C, \langle D, [D] \rangle] \rangle \\
&\Rightarrow \langle A, [E, F, \langle H, [D, E] \rangle, G, I, \langle D, [D] \rangle] \rangle \\
&\Rightarrow \langle A, [E, F, \langle H, [J, E, K] \rangle, G, I, \langle D, [D] \rangle] \rangle \\
&\Rightarrow \langle A, [E, F, \langle H, [J, E, K] \rangle, G, I, \langle D, [K, J] \rangle] \rangle
\end{aligned}
$$

5 A Parsing Algorithm for GIDLP

We have developed a GIDLP parser based on Earley's algorithm for context-free parsing (Earley 1970). In Earley's original algorithm, each edge encodes the interval of the input string it covers. With discontinuous constituents, however, that is no longer an option. In the spirit of Johnson (1985) and Reape (1991), and following Ramsay (1999), we represent edge coverage with bitvectors, stored as integers. For instance, 00101 represents an edge covering words one and three of a five-word sentence.[6]

Our parsing algorithm begins by seeding the chart with passive edges corresponding to each word in the input and then predicting a compacted instance of the start symbol covering the entire input; each final completion of this edge will correspond to a successful parse.

As with Earley's algorithm, the bulk of the work performed by the algorithm is borne by two steps, *prediction* and *completion*. Unlike the context-free case, however, it is not possible to anchor these steps to string positions, proceeding from left to right. In order for the parser to operate as efficiently as possible, it must be possible for the prediction step to intelligently take word order constraints into account. Once a daughter of an active edge has been found, the other daughters should only be predicted to occur in string positions which are compatible with the word order constraints of the active edge. For example, consider the edge in (13).

(13) $A \rightharpoonup B_1 \bullet C_2 \; ; \; 1 < 2$

This notation represents the point in the parse during which the application of this rule has been predicted, and a B has already been located. Assuming that B has been found to cover the third position of a five-word string, two facts are known. From the LP constraint, C cannot precede B, and from the general principle that the RHS of a rule forms a partition of its LHS, C cannot overlap B. Thus C cannot cover positions one, two, or three.

5.1 Compiling LP Constraints into Bitmasks

We can now discuss the integration of GIDLP word order constraints into the parsing process. A central insight of our algorithm is that the same data structure used to describe the coverage of an edge can also encode restrictions on the parser's search space. This is accomplished with two classes of bitvectors: *negative masks* (n-masks) and *positive masks* (p-masks). Efficient bitvector operations (Daniels and Meurers 2002) can then be used to compute, manipulate, and test the encoded constraints.

Negative Masks The n-mask constrains the set of possible coverage vectors that could complete the edge. The 1-positions in a masking vector represent the positions that are masked out: the positions that cannot be filled when completing this

[6] Note that the first word is the rightmost bit.

edge. The 0-positions in the negative mask represent positions that may potentially be part of the edge's coverage. For the example above, the coverage vector for the edge is 00100 since only the third word *B* has been found so far. Assuming no restrictions from a higher rule in the same domain, the n-mask for *C* is 00111, encoding the fact that the final coverage vector of the edge for *A* must be either 01000, 10000, or 11000 (that is, *C* must occupy position four, position five, or both of these positions). The negative mask in essence encodes information on where the active category cannot be found.

Positive Masks The p-mask encodes information about the positions the active category **must** occupy. This knowledge arises from immediate precedence constraints. For example, consider the edge in (14).

(14) $\mathrm{D} \rightharpoonup \mathrm{E}_1 \bullet \mathrm{F}_2 \; ; \; 1 \ll 2$

If *E* occupies position one, then *F* must at least occupy position two; this would be represented by a p-mask of 00010.

Thus in the prediction step, the parser considers each rule in the grammar that provides the symbol being predicted, and for each rule, it generates bitmasks for the new edge, taking both rule-level and domain-level order constraints into account. The resulting masks are checked to ensure that there is enough space in the resulting mask for the minimum number of categories required by the rule.[7]

Then, as part of each completion step, the parser must update the LP constraints of the active edge with the new information provided by the passive edge. As edges are initially constructed from grammar rules, all order constraints are initially expressed in terms of either descriptions or tokens. As the parse proceeds, these constraints are updated in terms of the actual locations where matching constituents have been found. For example, a constraint like $1 < 2$ (where 1 and 2 are tokens) can be updated with the information that the constituent corresponding to token 1 has been found as the first word, i.e. as position 00001.

In summary, compiling LP constraints into bitmasks in this way allows the LP constraints to be integrated directly into the parser at a fundamental level. Instead of weeding out inappropriate parses in a cleanup phase, LP constraints in this parser can immediately block an edge from being added to the chart.

6 Beyond Head-driven Parsing

As described in the GIDLP grammar format defined above, the order of the RHS of a grammar rule does not encode the terminal order of the daughters. Instead, it expresses the order in which the parser will search for these elements.

[7] This optimization only applies to epsilon-free grammars. Further work in this regard can involve determining the minumum and maximum yields of each category; some optimizations involving this information can be found in Haji-Abdolhosseini and Penn (2003).

For a simple example of a construction where ordering the non-head daughters is useful, consider a grammar covering raising verbs in Icelandic as discussed in (Sag et al. 1992). Many verbs in Icelandic assign "quirky case" (i.e. non-nominative) to their subjects; these case assignments persist when the subject is raised to be the subject or object of a matrix verb. This is illustrated by the data in (15) – (20).

(15) Hana virðist vanta peninga
her.ACC seems to-lack money
'She seems to lack money.'

(16) Barninu virðist hafa batnað veikin
the-child.DAT seems to-have recovered-from the-disease
'The child seems to have recovered from the disease.'

(17) Verkjanna virðist ekki gæta
the-pains.GEN seem not to-be-noticeable
'The pains don't seem to be noticeable.'

(18) Hann telur mig vanta peninga
he.NOM believes me.ACC to-lack money
'He believes that I lack money.'

(19) Hann telur barninu hafa batnað veikin
he believes the-child.DAT to-have recovered-from the-disease
'He believes the child to have recovered from the disease.'

(20) Hann telur verkjanna ekki gæta
he believes the-pains.GEN not to-be-noticeable
'He believes the pains to be not noticeable.'

In other words, the fact that the subject in (15) and (18) is accusative is a reflection of the embedded verb 'lack' rather than the matrix verbs 'seem' or 'believe'; the same situation holds for the dative [(16), (19)] and genitive [(17), (20)] examples. In all other respects, however, the matrix verb is still the head of its clause (it must agree in number with the subject, for example). Thus from a parsing perspective, the embedded verb must be known before it can be determined whether a given noun phrase is an acceptable subject for the matrix verb.

Consider a head-driven parser (van Noord 1997): a variant of a phrase-structure parser in which a designated element (the head) is parsed before any other complement); the non-head daughters that occur to the right of the head are then parsed in the usual left-to-right order. With such a parser, the grammar writer would write a rule like (21) to license the matrix clause.

(21) $S \rightharpoonup NP_{subj}\ V^{head}\ VP_{inf}$

With such a rule, the parser will first locate the head (here, the *V*), then the *NP*, and finally the *VP*. As a consequence, the constraints in the *VP* on the case of the subject will not be known until after the subject has been found. The parser will therefore try all possible *NP*s as subjects, and then see which the embedded verb phrase rejects.

With the GIDLP formalism, in contrast, the grammar writer could specify the rule as (22) to avoid this generate-and-test pattern.

(22) $\mathrm{S} \rightharpoonup \mathrm{V}^{1}, \mathrm{VP}^{2}_{\mathrm{inf}}, \mathrm{NP}^{3}_{\mathrm{subj}}$

Now the parser will not look for the subject of the clause until the embedded verb phrase has been located, and so only *NP*s with the appropriate case will even be considered.

7 Computational Complexity

Suhre (1999) shows that the membership problem for his LSL grammar formalism (a subset of the GIDLP formalism; thus comparable results for GIDLP grammars could be no better) is NP-complete, both when considering the grammar plus the string as input (general membership problem) as well as when only the string is considered as input (fixed membership problem). It has been known since Huynh (1983) that the general membership problem for unordered context-free grammars (ID/LP grammars without LP statements) is also NP-complete, so Suhre's first result is not surprising. That the fixed membership problem for LSL grammars is also NP-complete is less straightforward; fortunately, Suhre (1999, 61ff) demonstrates that it stems from the potential for recursive growth of discontinuities. As a result, when the parser can assume an upper bound on the number of discontinuities in any given constituent, the fixed membership problem becomes polynomial. Formally, this can be achieved by requiring that the number of discontinuities introduced by a recursive non-terminal is bounded by some constant.

Interestingly, a related practical proposal based on linguistic argumentation is discussed by Müller (1999b). He proposes a continuity constraint for linearization-based HPSG which requires saturated phrasal elements (that is, maximal projections) to be continuous.[8] Müller shows that adding his continuity constraint results in a significant reduction in the number of passive edges and thereby significant improvements in parsing performance. This continuity constraint is weaker than Suhre's condition in that recursion on the level of adjunction is not restricted. It is, however, interesting to note in this context that a grammar incorporating the $\overline{\mathrm{X}}$-schema (Jackendoff 1977) will require all non-head constituents to be maximal projections. In sum, Müller's result strongly suggests that further research on linguistically-motivated continuity constraints can result in efficient parsing of those GIDLP grammars which include such constraints.

[8] If extraposition is handled via discontinuous constituents, a more complex constraint is required.

8 Evaluation

As discussed at the end of section 4.2, it is possible to take a GIDLP grammar and write out the discontinuity. All non-domain introducing rules must be folded into the domain-introducing rules, and then each permitted permutation of a RHS must become a context-free rule on its own – generally, at the cost of a factorial increase in the number of rules.

This construction indicates the basis for a preliminary assessment of the GIDLP formalism and its parser. The grammar in (23) recognizes a very small fragment of German, focusing on the free word order of arguments and adjuncts in the so-called *Mittelfeld* that occurs to the right of either the finite verb in yes-no questions or the complementizer in complementized sentences.[9]

(23) a) start(s, [])

b) $s \rightharpoonup s(cmp)_1$

c) $s \rightharpoonup s(que)_1$

d) $s(cmp) \rightharpoonup cmp_1, clause_2;\ \langle[0], s(cmp), \langle cmp < \_, \_ < v(\_)\rangle\rangle$

e) $s(que) \rightharpoonup clause_1;\ \langle[0], s(que), \langle v(\_) < \_\rangle\rangle$

f) $clause \rightharpoonup np(n)_1, vp_2$

g) $vp \rightharpoonup v(ditr)_1, np(a)_2, np(d)_3$

h) $vp \rightharpoonup adv_1, vp_2$

i) $vp \rightharpoonup v(cmp)_1, s(cmp)_2$

j) $[np(Case)] \rightharpoonup det(Case)_1, n(Case)_2;\ 1 \ll 2$

k)	v(ditr) ⇀ *gab*	o)	n(acc) ⇀ *Buch*	s)	det(acc) ⇀ *das*
l)	comp ⇀ *dass*	p)	adv ⇀ dort	t)	n(dat) ⇀ *Frau*
m)	det(dat) ⇀ *der*	q)	v(cmp) ⇀ *denkt*	u)	adv ⇀ *gestern*
n)	n(nom) ⇀ *Mann*	r)	det(nom) ⇀ *der*		

The basic idea of this grammar is that domain compaction only occurs at the top of the head path, after all complements and adjuncts have been found. When the grammar is converted into a CFG, the effect of the larger domain can only be mimicked by eliminating the clause and vp constituents altogether.

As a result, while this GIDLP grammar has 10 syntactic rules, the corresponding flattened CFG has 201 rules (with the number of adverbs artificially limited to two). In an experiment, the four sample sentences in (24)[10] were parsed with both our prototype GIDLP parser (using the GIDLP grammar) as well as a vanilla Earley CFG parser (using the CFG); the results are shown in (25).

[9] The symbol _ is used to denote the set of all categories.

[10] The grammar and example sentences are intended as a formal illustration, not a linguistic theory; because of this, we have not provided glosses.

(24) a) *Gab der Mann der Frau das Buch?*

b) *dass das Buch der Mann der Frau gab.*

c) *dass das Buch gestern der Mann dort der Frau gab.*

d) *Denkt der Mann dass das Buch gestern der Mann dort der Frau gab?*

(25)

	Active Edges		Passive Edges	
Sentence	GIDLP	CFG	GIDLP	CFG
a)	18	327	16	15
b)	27	338	18	16
c)	46	345	27	27
d)	75	456	36	24

Averaging over the four sentences, the GIDLP grammar requires 89% fewer active edges.[11] It is important to keep in mind that the GIDLP grammar is more general than the CFG: in order to obtain a finite number of CFG rules, we had to limit the number of adverbs. When using a grammar capable of handling longer sentences with more adverbs, the number of CFG rules (and active edges, as a consequence) increases factorially.

Timings have not been included in (25); it is generally the case that the GIDLP parser/grammar combination was slower than the CFG/Earley parser. This is an artifact of the use of atomic categories, however. For the large feature structures used as categories in HPSG, we expect the larger numbers of edges encountered while parsing with the CFG to have a greater impact on parsing time, to the point where the GIDLP grammar/parser is faster.

9 Summary

In this paper, we have introduced a grammar format that can be used as a processing backbone for linearization-based HPSG grammars that supports the specification of discontinuous constituents and word order constraints on domains that extend beyond the local tree. We have presented a prototype parser for this format illustrating the use of order constraint compilation techniques to improve efficiency. Future work will concentrate on additional techniques for optimized parsing as well as the application of the parser to feature-based grammars. We hope that the GIDLP grammar format will encourage research on such optimizations in general, in support of efficient processing of relatively free constituent order languages using linearization-based HPSG.

[11] It also generates additional passive edges corresponding to the extra non-terminals vp and clause.

References

Bonami, Olivier, Godard, Danièle and Marandin, Jean-Marie. 1999. Constituency and word order in French subject inversion. In Gosse Bouma, Erhard W. Hinrichs, Geert-Jan M. Kruijff and Richard T. Oehrle (eds.), *Constraints and Resources in Natural Language Syntax and Semantics*, Studies in Constraint-Based Lexicalism, pages 21–40, Stanford, CA: CSLI.

Daniels, Michael W. and Meurers, W. Detmar. 2002. Improving the efficiency of parsing with discontinuous constituents. In Shuly Wintner (ed.), *Proceedings of NLULP-02: The Seventh International Workshop on Natural Language Understanding and Logic Programming*, pages 49–68, Roskilde University, Computer Science Department, Copenhagen, Denmark.

Daniels, Michael W. and Meurers, W. Detmar. 2004. A grammar formalism and parser for linearization-based HPSG. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 169–175, Geneva, Switzerland.

Donohue, Cathryn and Sag, Ivan A. 1999. Domains in Warlpiri. In *Abstracts of the Sixth Int. Conference on HPSG*, pages 101–106, Edinburgh: University of Edinburgh.

Earley, Jay. 1970. An Efficient Context-Free Parsing Algorithm. *Communications of the ACM* 13(2), 94–102, also in Grosz et al. (1986).

Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey K. and Sag, Ivan A. 1985. *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard UP.

Götz, Thilo and Meurers, Walt Detmar. 1997. The ConTroll System as Large Grammar Development Platform. In *Proceedings of the Workshop "Computational Environments for Grammar Development and Linguistic Engineering (ENVGRAM)" held at ACL/EACL*, pages 38–45, Association for Computational Linguistics, Madrid: Universidad Nacional de Educación a Distancia.

Götz, Thilo and Penn, Gerald. 1997. A Proposed Linear Specification Language. Volume 134 in Arbeitspapiere des SFB 340, Universität Tübingen.

Grosz, Barbara, Jones, Karen Sparck and Webber, Bonnie Lynn (eds.). 1986. *Readings in Natural Language Processing*. Los Altos, CA: Morgan Kaufmann.

Haji-Abdolhosseini, Mohammad and Penn, Gerald. 2003. ALE Reference Manual. Manuscript, University of Toronto. `http://www.ale.cs.toronto.edu/docs/ref/ale_ref.pdf`.

Huynh, Dung T. 1983. Commutative Grammars: The Complexity of Uniform Word Problems. *Information and Control* 57(1), 21–39.

Jackendoff, Ray. 1977. *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, Mass.: MIT Press.

Johnson, Mark. 1985. Parsing with discontinuous constituents. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 127–132, Chicago.

Kasper, Robert, Kathol, Andreas and Pollard, Carl. 1995. A relational interpretation of linear precedence constraints, the Ohio State University.

Kathol, Andreas. 1995. *Linearization-Based German Syntax*. Ph. D.thesis, The Ohio State University.

Kathol, Andreas and Pollard, Carl. 1995. Extraposition via Complex Domain Formation. In *Proceedings of ACL*, pages 174–180.

Kay, Martin. 1990. Head-Driven Parsing. In Masaru Tomita (ed.), *Current Issues in Parsing Technology*, Dordrecht: Kluwer, previously published in the proceedings of the International Workshop on Parsing Technologies, 1989.

Lewis, Harry R. and Papadimitriou, Christos H. 1998. *Elements of the Theory of Computation*. Upper Saddle River, NJ: Prentice-Hall, second edition.

Müller, Stefan. 1999a. *Deutsche Syntax deklarativ. Head-Driven Phrase Structure Grammar für das Deutsche*. Linguistische Arbeiten, No. 394, Tübingen: Niemeyer.

Müller, Stefan. 1999b. Restricting Discontinuity. Verbmobil Report 237, DFKI, Saarbrücken, also published in the Proceedings of GLDV 99 (Frankfurt/Main).

Müller, Stefan. 2004. Continuous or discontinuous constituents? A comparison between syntactic analyses for constituent order and their processing systems. *Research on Language and Computation, Special Issue on Linguistic Theory and Grammar Implementation* 2(2), 209–257.

Penn, Gerald. 1999. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croatian. In Robert D. Borsley and Adam Przepiórkowski (eds.), *Slavic in HPSG*, pages 149–182, Stanford, CA: CSLI.

Ramsay, Allan M. 1999. Direct parsing with discontinuous phrases. *Natural Language Engineering* 5(3), 271–300.

Reape, Mike. 1991. Parsing bounded discontinuous constituents: Generalisations of some common algorithms. In Mike Reape (ed.), *Word Order in Germanic and Parsing*, pages 41–70, DYANA R1.1.C, ESPRIT BR 3175.

Reape, Mike. 1993. *A Formal Theory of Word Order: A Case Study in West Germanic*. PhD thesis., Univ. of Edinburgh.

Richter, Frank and Sailer, Manfred. 2001. On the left periphery of German finite sentences. In Detmar Meurers and Tibor Kiss (eds.), *Constraint-Based Approaches to Germanic Syntax*, Studies in Constraint-Based Lexicalism, pages 257–300, Stanford, CA: CSLI.

Sag, Ivan, Karttunen, Lauri and Goldberg, Jeffrey. 1992. A Lexical Analysis of Icelandic Case. In Ivan Sag and Anna Szabolcsi (eds.), *Lexical Matters*, pages 302–318, Stanford, CA: CSLI.

Shieber, Stuart M. 1984. Direct Parsing of ID/LP Grammars. *Linguistics and Philosophy* 7, 135–154.

Suhre, Oliver. 1999. *Computational Aspects of a Grammar Formalism for Languages with Freer Word Order*. Diplomarbeit, Department of Computer Science, University of Tübingen, published 2000 as Volume 154 in Arbeitspapiere des SFB 340.

van Noord, Gertjan. 1991. Head corner parsing for discontinuous constituency. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 114–121.

van Noord, Gertjan. 1997. An Efficient Implementation of the Head-Corner Parser. *Computational Linguistics* 23(3).

Yatabe, Shuichi. 1996. Long-distance scrambling via Partial Compaction. In Masatoshi Koizumi, Masayuki Oishi and Uli Sauerland (eds.), *Formal Approaches to Japanese Linguistics 2*, pages 303–317, Cambridge, MA: MITWPL.