LACUS FORUM XXIX

Linguistics and the Real World

THE UNIVERSITY OF TOLEDO

REPRINT

TOWARD LINGUISTICALLY PLAUSIBLE LANGUAGE MODELLING IN REAL-WORLD APPLICATIONS

Gerald Penn
Department of Computer Science, University of Toronto

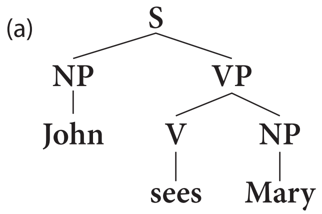
There is a growing awareness among engineers working with speech and text that, in order for the functionality of current natural language applications to progress to the next level, access to thematic roles and grammatical function assignment, i.e., 'who did what to whom', will be just as important as a probabilistic model's ability to predict the next word in a string. The latter is the current litmus test for 'language models', as these probabilistic finite-state or simple context-free grammars are almost mockingly called in engineering circles. In striving to represent meaning and discourse relations, these engineers and the annotated corpora they use are dutifully following the common assumption in the Chomskyan linguistic tradition that they are artifacts of configurational relations—primitives that are evident from the phrase-structure trees licensed by the grammar.

In the case of English, there have been some remarkable successes in the last five years. The most notable is that of Collins (1999) and several successive improvements, who used knowledge about headedness and subcategorisation, a traditional n-gram language model and some information about unbounded dependencies to dramatically improve a context-free parser's ability to predict the most likely phrase-structure tree given a string of words — with the tacit assumption that this tree is sufficient to provide an interpretation. While there have also been more modest successes with purely dependency-based grammars in the realm of freer word-order (fwo) languages (Collins et al. 1999), even agreeing on what the best phrase-structure tree should be in these languages is not easy. Predicting it from data, moreover, seems utterly intractable, given the number of movement operations and empty projections involved.

In the ambitious hunt for such universals of syntactic structure across languages, however, some alternative structures from linguistic theory that maintain a closer relationship to the attested data in fwo languages have been neglected. These challenge the traditional view of constituency, in which word order, phrasal discontinuities, semantic interpretation, and discourse structure all happily agree on the compositional subunits to which their constraints refer.

This paper presents a new discrete language model for parsing that uses parallel phrase-structure-like trees, synchronised by grammatical constraints. It is inspired by these dissenting proposals, beginning with the distinction drawn by Curry (1961) between *tectogrammatical* and *phenogrammatical* structure. There is also a set of interpretation rules with primitives for stating *linear precedence* and *constituent liberation*, roughly in the sense of (Zwicky 1986). A naïve parsing algorithm and some details of an implementation of the model are also discussed.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by
Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.



(b) S ⇒ NP VP
VP ⇒ V NP

Figure 1. *An example phrase structure tree (a) and its corresponding phrase structure rules (b).*

1. PHRASE STRUCTURE GRAMMARS. Phrase structure trees, such as the example shown in Figure 1a, are constrained relative to the rules of a given context-free grammar, such as those shown in Figure 1b. These rules carry with them three implicit assumptions about constituency, namely:

1. constituents are realised as contiguous substrings,
2. constituents are internally ordered (by the right-hand-sides), and
3. constituents guide the assemblage of a semantic interpretation.

Simply put, the goal of the present study has been to characterise languages in which one or more of these three assumptions do not hold, while retaining access to semantic information.

2. PREVIOUS MODELS OF FREER WORD-ORDER. Previous models of FWO languages have typically been either very informal or, if formalised, rather superficial in their consideration of empirical data from attested languages. Many, for example, have taken the still-dubious existence of languages with completely free word-order for granted. They also have generally given attention to the relative linear precedence of constituents with very little regard for the contiguity of constituents. The principal exceptions to this are Zwicky (1986) and Reape (1994). Zwicky (1986) proposed that certain categories be designated as 'liberated,' which removed the requirement of contiguity on every instance of such a category in phrase structure. Reape (1994) uses a feature-structure-based system in which a binary-valued feature controls whether the substring corresponding to each subtree is contiguous and ordered, or merely ordered, in which case other substrings could be inserted among its words in the way that two stacks of playing cards can be shuffled together.

Some (Nash 1980; Barton, Berwick & Ristad 1987) posit no linear precedence whatsoever—phrase structure grammars were simply reduced to specifications of mother-daughter category relationships in trees, later called *immediate dominance* (ID). Many others relax or require explicit statements of linear precedence (LP), most notably the ID/LP (Gazdar et al. 1985). This has enjoyed very wide usage in linguistics, although with some ambiguity as to whether (1) its (relative) linear precedence statements (such as 'NP < VP') are intended to be quantified over all immediate dominance rules

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by
Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

Dem Mann (to) the man	*habe* have	*ich das Buch* I the book	*gegeben* given	
Vorfeld	left Satzklammer	Mittelfeld	right Satzklammer	Nachfeld

Figure 2*. A topological analysis of a simple German sentence.*

and/or over all NPs and VPs in the same rule, if more than one occurs, (2) its truth requires the existence of its argument categories, and (3) its argument categories must be distinct. The last point is important when non-atomic categories that can partially overlap in their denotations, such as feature structures, are used.

3. TOPOLOGICAL FIELDS. There are some empirically attested cases of freer linear precedence constraints for which the < operator is ill-suited. This is primarily because < is a statement of *relative* precedence between two constituents within some predefined region—in the case of ID/LP, for example, within the region spanned by the left-hand-side constituent of the phrase-structure rule.

3.1. EXAMPLE 1: THE GERMAN *MITTELFELD*. Traditional nineteenth century analyses of German sentence structure, for example, distinguish a linear topology that can be applied to every German clause, as shown in Figure 2.

This topology is distinguished by a region, called the *Mittelfeld*, which is bracketed on the left and right (left and right *Satzklammer*) by a closed class of categories, mostly verbal, as is the case in Figure 2. In matrix clauses, a topic position appears before this bracketing, called the *Vorfeld*. Another field at the end (*Nachfeld*) receives clausal arguments and adjuncts of the main verb, as well as prosodically heavy arguments and relative clauses of NPs that have been raised.

A number of linear precedence constraints exist within the *Mittelfeld*. Pronouns, for example, precede prosodically heavier noun phrases in this field (occurring just after the left *Satzklammer*). Pronouns also occur relative to each other in a pre-defined order. Temporal adjunct phrases generally precede locatival adjunct phrases. These constraints only pertain within the *Mittelfeld*, however—nearly every constituent mentioned can alternatively appear in the *Vorfeld*, in which case it must by definition occur before everything in the *Mittelfeld*. Yet '*Mittelfeld*' does not correspond to a node or a subtree in a traditional phrase structure tree. These orderings, furthermore, are purely an issue of intraclausal scrambling among the base-generated constituents of the clause in question—not unbounded dependencies derived through movement.

3.2. EXAMPLE 2: THE GERMAN LEFT *SATZKLAMMER*. The position of the left *Satzklammer* itself is another such example. In many analyses, this position can be filled by a variety of different traditional constituent categories, including finite verbs and auxiliaries, complementisers and subordinating conjunctions. They occur after the *Vorfeld*, and thus form a 'second position,' which, as in the case of many FWO languages, must be treated rather specially since the 'first position' does not always contain a complete

u lepi grad in beautiful city	*je* CL-3S		*Ivan stigao* Ivan arrived
PP	2nd		

***Figure 3**. Clitic placement after a full syntactic phrase (2D).*

u lepi in beautiful city	*je* CL-3S	*grad* city	*Ivan stigao* Ivan arrived
prosodic word	2nd	remainder	

***Figure 4**. Clitic placement after a prosodic word (2W).*

phrasal projection. In German, this mismatch generally arises with instances of partial verb phrase fronting (Kathol 2000). The < operator cannot encode the LP constraint, 'second' without granting some kind of constituency status to whatever occurs first.

3.3. EXAMPLE 3: SERBO-CROATIAN[1] SECOND-POSITION CLITICS. The problem of the constituency status of first and second positions is even more acute in Serbo-Croatian, a language in which the realisation of NPs is generally quite a bit freer than in German. Here, there is a class of pronominal and auxiliary forms that are prosodically enclitic to the topic position and likewise form a 'second position.' In the case of Serbo-Croatian, the first position can either be a full NP or PP projection (2D), as in Figure 3, or a prosodic word (2W), as in Figure 4 (Browne 1974).

Most Serbo-Croatian prepositions, including *u*, are prosodic proclitics to the first word of their objects, so *u lepi* is a prosodic word even though it does not form a traditional syntactic sub-constituent.

There is, in fact, a long and rather embarrassing series of attempts to analyse these linear distributions in purely syntactic or purely prosodic terms, as discussed in Penn 1999[2]. Now, there appears to be a general agreement that both levels of representation are necessary, although there is still no apparent consensus on how to combine them.

When more than one second-position clitic occurs in a single clause, they occur in a cluster with a fixed order among them (Browne 1974).

4. RELATED WORK. The parsing model formulated here gives first-class status to a linear topology for defining word-order and contiguity. This is, of course, not necessarily the five-field topology defined for German clauses, and must be declared in a manner similar to the declaration of phrase structure rules in a context-free grammar. This kind of 'linear constituency' co-exists with the more traditional constituency that guides the assemblage of an interpretation of a sentence through the assignment of thematic roles, grammatical function and (abstract) case, syntactic constraints on the resolution of scope ambiguities, etc.

The dual-constituent nature of this model is inspired by the proposal by Curry (1961) to separate constituency into two kinds: *tectogrammatical* constituency, which guides interpretation, and *phenogrammatical* constituency, in which the role of inflectional morphology and the different behaviour of fixed and free word-order languages is expected to be captured. This initial proposal has also had a very strong influence on work in the later Prague school, as well as that of Dowty (1996) and Kathol (2000), although all of them define this distinction in slightly different ways. The present proposal is again slightly different from these.

The present approach also departs from Kathol (2000) in viewing topological fields as nested within a *region* whose internal word-order is characterised by such a topology. A region can, in turn, be nested inside a field. German embedded clauses, for example, have internally defined topologies, as all German clauses do, but occur themselves within the *Nachfeld* of a higher clause. In principle, the topology within a region need not be the same as the topology within which it is embedded. Noun phrases, for example, do not have the same five fields as clauses in German, but they can occur in the *Vorfeld* or *Mittelfeld* of a clause.

The distinction between constituent structure (C-structure) and functional structure (F-structure) in Lexical-Functional Grammar (Kaplan & Bresnan 1982), in the author's view, has essentially the same motivation. C-structure, the nearest correlate of phenogrammatical constituency, is not intended to guide the assemblage of an interpretation, but still uses tectogrammatical constituents in its rules (and, in fact, still has only one kind of constituent). When dealing with freer word-order languages, C-structure rules must then be numerous and have many daughters in order to capture sufficiently many tectogrammatical categories within a single rule to enumerate all of their possible permutations. As a result, C-structure trees look very flat and rather arbitrary, given that what are essentially phenogrammatical regions are still labelled at their root by tectogrammatical categories.

One could also cite the work of Duchier and Debusmann (2001) in comparison with the present approach. They use topological fields in a dependency grammar, where they are used to resolve the order of dependents within the domain of governing nodes. Although Duchier and Debusmann (2001) generalise topological fields to handle the internal order of noun phrases, rather than just clauses, governors are the only approximation to the regions used here. In a great many FWO languages, however, including German and Serbo-Croatian, regions are not always identical to the arguments of a single syntactic head. A few *ad hoc* devices are introduced for the purpose of handling the exceptions that arise in the authors' implementation of a German grammar. Nevertheless, the result of these repairs yields a dependency tree as the only (tectogrammatical) guide to assembling an interpretation. Other work on dependency grammar within FWO languages, notably that of Hajičova, Sgall and others on Czech, has extensively documented the shortcomings of a single surface-oriented dependency tree in assembling a sufficiently rich semantic interpretation. They typically use transformations on dependency trees to create a *tectogrammatical* dependency tree. Here, tectogrammatical structure is formalised with the retention of phrasal projections.

The present work also bears a resemblance to Autolexical Syntax (Sadock 1991) in that parallel but mutually constrained structural derivations are being posited. In fact, although prosodic words are currently identified as phenogrammatical regions in the present approach, there should be a third parallel structure for prosody, and possibly even a fourth for discourse-linked regions that are structurally constrained in the same way. This has not been fully developed yet, only because of the author's own ignorance of how prosodic and discourse structure should be represented.

5. TOPOLOGICAL PARSING MODEL. We can thus distinguish three kinds of constituent primitives:

PHENOGRAMMATICAL:

1. fields, e.g., the German *Mittelfeld*, a Serbo-Croatian *clitic field* and *pre-clitic field*, etc.;
2. regions, e.g., German clauses and noun phrase regions (whose internal topologies differ from those of clauses), Serbo-Croatian clitic regions (within which the fixed order of clitic clusters is defined), prosodic word regions, which, in 2W position, occupy the pre-clitic field, etc.; and

TECTOGRAMMATICAL:

3. categories, either atomic ones, or more complicated categories such as typed feature structures (Pollard & Sag 1994).

The lexicon assigns one or more category and one or more field or region to every word.

5.1. TOPOLOGICAL RULES. We can then state two kinds of rules, one phenogrammatical and one tectogrammatical, that tell us how to assign these primitives to larger substrings. The phenogrammatical, or topological rules are essentially context-free rules over fields and regions. Each has one of two possible forms:

- $f \rightarrow r$, which indicates that a particular field, f, can contain region, r; and
- $r \rightarrow d_1\ d_2 \ldots d_n$, which indicates that region, r, has a topology defined by the *field descriptors* d_1 through d_n. Each field descriptor is one of:
  - *f*, exactly one occurrence of the topological field, f
  - {*f*}, zero or one occurrence of field, f
  - *f**, zero or more occurrences of field, f, or
  - *f*+, one or more occurrences of field, f.

As examples of the former, we find in German:

nf → *clause*

German embedded clauses occur in the *Nachfeld* in the next higher clause. As an example of the latter, we again find in German:

matrix → vf,cf,mf*,{vc},{nf}

We can assume the existence of a distinguished single field/region, matrix, which corresponds exactly to entire sentences, and is not contained in any other field or region. Here it is assumed that these correspond to matrix clauses, which consist of the five standard fields. Note that the right *Satzklammer* (vc, for verbal complex), and *Nachfeld* (nf) are optional, and any number of regions may occur between the left *Satzklammer* (cf, for complementiser field) and right *Satzklammer*, provided that they can bear the field assignment, mf (*Mittelfeld*).

5.2. INTERPRETATION RULES. The tectogrammatical, or interpretation rules are like traditional phrase structure rules over traditional, tectogrammatical categories, but without any assumptions about linear precedence or contiguity built in. Where they exist, these are explicitly specified as attachments to the rule:

$\text{cat}_0 \Rightarrow \text{cat}_1\ \text{cat}_2 \ldots \text{cat}_n;\phi$

where ϕ consists of:

- *i<j* (linear precedence)
- *i<<j* (immediate precedence)
- *i compacts* (contiguity)

closed under conjunction, for some collection of pairs $1 \leq i,j \leq n$.

As an example, we can consider two interpretation rules that encode a simplified case of PP dislocation in German, in which a PP modifying a noun occurs either immediately after the noun, or in the *Nachfeld* of the clause in which the noun occurs:

NP ⇒ NP PP; 1<<2
NP ⇒ NP PP; 2 matches nf

5.3. SYNCHRONISATION CONSTRAINTS. Topological rules relate the phenogrammatical primitives to each other, and interpretation rules relate the tectogrammatical primitives to each other. Mediation of constraints between phenogrammar and tectogrammar are handled by synchronisation constraints. These constraints specify the circumstances under which an instance of a field or region corresponds to the same string that some category corresponds to:

- (∀) *f/r matched_by* (∃) *cat*
- (∀) *f/r covered_by* (∃) *cat*
- (∀) *cat matches* (∃) *f/r*
- (∀) *cat covers* (∃) *f/r*
- (∀) *cat compacts*

Constraints with ‘covers’ or ‘covered_by’ indicate that the field’s/region’s string is a substring, not necessarily an exact match.

The left-hand-side of each constraint is implicitly universally quantified, and the right-hand-side is implicitly existentially quantified. For example, the first says that *every* string corresponding to an f/r also corresponds to *some* cat. Because of this asymmetry, we also need constraints in which the tectogrammatical category appears on the left—these are the third and fourth forms. The fifth is simply a universally quantified form of the contiguity constraint found in the interpretation rules. It would also be possible to add universally quantified linear precedence statements, although these seem to be adequately handled by topological fields.

As an example, noun phrase regions, the phenogrammatical primitives over which the linear order of noun phrases is stated, are only related to NPs by covering, not matching:

npr covered_by np

The reason for this is that relative clauses or PPs which, tectogrammatically, are part of the NP and its interpretation, may be linearly dislocated to the *Nachfeld* of the containing clause, and thus become subject to the clausal topological ordering.

6. PERFORMANCE. A prototype of a parser based on this model, with grammars for both German and Serbo-Croatian, has been implemented by Mohammad Haji-Abdolhosseini of the University of Toronto Linguistics Department. The implementation is written in SICStus Prolog, and compiles grammars provided by a user in the above form into SICStus Prolog code, which is then further compiled by the SICStus compiler itself. It currently supports only atomic categories, although an extension of this to typed feature structures is planned.

The parser proceeds by first looking up the parts of speech of an input sentence, and then associating these through the structural constraints provided by the grammar, with phenogrammatical fields and/or regions. A phenogrammatical tree is then built in a bottom-up fashion with a standard context-free parsing algorithm. As fields or regions are encountered that are structurally constrained by tectogrammatical categories, those categories are predicted to exist, with their structures generated from the tectogrammatical rules in a top-down fashion. If they can eventually be linked to the lexical parts of speech, then the prediction is certified as having been derived. In this way, both trees are constructed. Other parsing algorithms could be devised.

The performance of the compiled code by this implementation on 10–15 word German sentences is slightly less than 200 ms on average. Its performance on 5–10 word Serbo-Croatian sentences is about 150 ms on average. While this is not quite fast enough for real-world applications, it is promising, particularly in light of the number of compile-time optimisations that could be performed on such grammars to exploit the constraints on linear precedence that they do exhibit. Currently, no optimisations are being applied.

7. FUTURE WORK. A great deal of work remains to be done on this topological parsing model. Of foremost importance for linguistic application is the development of larger experimental grammars, and alternative parsing strategies to improve our understanding of the kinds of constraints that are exhibited in practice in FWO languages. In terms of scaling up to coverage of very large corpora, e.g., newspaper text, the most important issue is discovering the right numerical parametrisation of this model, along with statistical estimators for those parameters, which can be used to select the 'best,' i.e., most probable parse. The current parsing method finds all possible parses, which, although obviously useful for grammar development and testing, can never be as fast on a large-scale because of the sheer number of constructions that must be licensed in such grammars.

As mentioned above, a more mature account of prosody and discourse structure must be devised to be incorporated into this model, and further compiler optimisations must also be implemented. It will also be necessary to allow grammar writers to state their rules in the form of phrase structure rules, and the more traditional idioms of syntactic theory, where enough linear precedence and contiguity does exist to justify them.

It should also be noted that no attempt is made here to deal with unbounded dependencies. It is recognised that these are a very different issue from the intraclausal scrambling that characterises FWO languages, and some account needs to be incorporated into the model. In the author's opinion, a sufficiently rich category system, such as typed feature structures, would be enough to handle these without any modification external to that category system, but there is some evidence to suggest that unbounded dependencies involve 'movement' of phenogrammatical constituents rather than tectogrammatical ones (Penn 1999). This requires further investigation.

[1] Although German and Serbo-Croatian were chosen for independent reasons, it was brought to the present author's attention during the conference that there has been an earlier comparative study of linear precedence in these two languages, with the aim of demonstrating that Slovene is a South Slavic language in transition to a more Germanic-style verb-second word-order (Bennett 1987). It would be very interesting indeed to cast this transition into the terms of the model proposed in this paper by formulating a grammar for Slovene with it.

[2] The notable exceptions to this are Halpern (1995) and Schuetze (1996), who both try to combine syntactic and prosodic influences into a single account.

REFERENCES

BARTON, G. EDWARD, ROBERT C. BERWICK & ERIC S. RISTAD. 1987. *Computational complexity and natural language*. Cambridge MA: MIT Press.

BENNETT, DAVID C. 1987. Word-order change in progress: The case of Slovene and Serbo-Croat and its relevance for Germanic. *Journal of linguistics* 23:269–87.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

Browne, Wayles. 1974. On the problem of enclitic placement in Serbo-Croatian. In *Slavic transformational syntax*, ed. by Richard D. Brecht & Catherine V. Chvany. Ann Arbor: University of Michigan Press.

Collins, Michael. 1999. *Head-driven statistical models for natural language parsing*. University of Pennsylvania doctoral dissertation.

———, Jan Hajic, Lance Ramshaw & Christoph Tillmann. 1999. A statistical parser for Czech. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 505–12. College Park, Maryland.

Curry, Haskell B. 1961. Some logical aspects of grammatical structure. *Structure of language and its mathematical aspects: Proceedings of the twelfth symposium in applied mathematics*, 56–68. American Mathematical Society.

Dowty, David. 1996. Toward a minimalist theory of syntactic structure. In *Discontinuous constituency*, ed. by Harry Bunt & Arthur van Horck, 11–62. Berlin: Mouton de Gruyter.

Duchier, Denys & Ralph Debusmann. 2001. Topological dependency trees: a constraint-based account of linear precedence. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 180–87. Toulouse, France.

Gazdar, Gerald, Geoffrey Pullum, Ewan Klein & Ivan Sag. 1985. *Generalized phrase structure grammar*. Cambridge ma: Harvard University Press.

Halpern, Aaron. 1995. *On the placement and morphology of clitics.* Stanford: csli Publications.

Kaplan, Ronald M. & Joan Bresnan. 1982. Lexical-functional grammar: a formal system for grammatical representation. In *The mental representation of grammatical relations*, ed. by Joan Bresnan, 173–281. Cambridge ma: mit Press.

Kathol, Andreas. 2000. *Linear syntax*. Oxford: Oxford University Press.

Nash, David G. 1980. *Topics in Warlpiri grammar*. mit Doctoral dissertation.

Penn, Gerald. 1999. Linearization and WH-extraction in HPSG: Evidence from Serbo-Croatian. In *Slavic in head-driven phrase structure grammar (Studies in constraint-based lexicalism)*, ed. by Robert D. Borsley & Adam Przepiórkowski, 149–82. Stanford: csli Publications.

Pollard, Carl & Ivan Sag. 1994. *Head-driven phrase structure grammar.* Chicago: University of Chicago Press.

Reape, Michael. 1994. Domain union and word order variation in German. In *German in head-driven phrase structure grammar (Lecture Notes 46)*, ed. by John Nerbonne, Klaus Netter & Carl Pollard, 151–97. Stanford: csli Publications.

Sadock, Jerrold M. 1991. *Autolexical syntax: A theory of parallel grammatical representations (Studies in contemporary linguistics)*. Chicago: University of Chicago Press.

Schuetze, Carson T. 1996. Serbo-Croatian clitic placement: An argument for prosodic movement. *Annual workshop on formal approaches to Slavic linguistics: the College Park meeting 1994 (Michigan Slavic materials 38)*, 225–248. Ann Arbor: Michigan Slavic Publications.

Zwicky, A. 1986. Concatenation and liberation. *Papers from the 22nd regional meeting of the Chicago Linguistic Society*, 65–74. Chicago Linguistic Society.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by
Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.