Discontinuous constituency: Introduction[1]

Harry Bunt

1 Discontinuity and linguistic theory

This book is concerned with the phenomenon of discontinuous constituency, studied from different theoretical perspectives.

Is discontinuous constituency a theory-dependent notion? It might certainly appear that way, since the notion of constituency and the role that constituent structure plays vary from theory to theory. The notion of phrase structure plays a central role in the formulation of many of the best developed grammatical theories, and the assignment of structural descriptions to sentences in terms of constituents is a major goal of the theory, the main differences between theories in this respect being the variety of constituent types they employ and the operations for buiding representations of constituent structure.

Head-driven Phrase Structure Grammar (HPSG) and Categorial Grammar (CG) are somewhat different from more traditional phrase-structure based approaches. Although obviously phrase-structure based, HPSG approaches the relations between constituents in a special way, in that

- the theory focuses on the immediate-dominance aspect of constituent structure and does not use phrase markers representing the dominance- and precedence relations among constituents, but instead encodes dominance relations as part of a sign, the unit of all linguistic information in HPSG;
- linear precedence relations are not represented in signs, and are only considered in separately expressed LP-constraints.

In Categorial Grammar the situation is sometimes claimed to be entirely different from phrase-structure based approaches, since CG does not pursue

[1] *H. Bunt and A. van Horck (eds.), Discontinuous Constituency, 1–10.*

the derivation of constituent structures of any kind. Still, any categorial analysis of a sentence corresponds to an *implicit* constituent structure being assigned to the sentence, since the categorial derivation makes decisions on which words and word groups to combine, based on their categories and their linear precedence relations. This is brought out clearly in the contributions by Moortgat and Bunt in this volume, where CG rules are considered that allow non-adjacent elements to be combined by introducing discontinuity in categories (functors and arguments).

The case of CG illustrates that, more generally, in those linguistic theories that do not aim at assigning explicit descriptions of constituent structures to sentences, the construction of such descriptions does nonetheless occur implicitly and is in fact unavoidable, since semantic interpretation requires decisions on which elements in the sentence are combined. Whenever non-adjacent elements are combined, this requires the description of discontinuity in some way or other.

It therefore seems that the phenomenon of discontinuous constituency is essentially theory-independent, and does not go away when we choose one theory rather than another. There is some variation in the linguistic data that are recognized as manifestations of the phenomenon, because different theories have somewhat different views on what counts as a constituent; still, there is a great deal of consensus about this matter across linguistic theories. As a result, the phenomenon of discontinuous constituency provides a potentially fruitful domain for discussion across the borders between linguistic theories. The present book aims at contributing to this discussion.

2 Empirical facts about discontinuity

When we describe the syntactic organization of a sentence, we group the words at various levels and divide the sentence into noun phrases, verb phrases, relative clauses, adjective phrases, prepositional phrases, amount expressions, and the like. The use of such word groups may be syntactically motivated by the possibility to achieve a higher level of generalization by describing syntactic structure in terms of such groups than in terms of individual words. Semantically, there is the perhaps even stronger motivation formed by the Principle of Compositionality, which requires the decomposition of sentences into meaningful parts in order to systematically derive sentence meanings.

More often then not, the syntactically and semantically meaningful word groups, or 'constituents', are contiguous: they consist of words standing next to one another. But sometimes this is not the case, as in example sentence (1):

(1) *Mary* woke *me* up *at seven-thirty.*

The verb form 'woke up' is split into two non-adjacent parts here, the particle 'up' being separated from 'woke' by 'me', which causes a discontinuity in the verb phrase.

Exactly which discontinuities a particular grammar formalism has to deal with, depends on the general views on constituency that the grammatical theory in question takes. Theories which recognize verb phrases, for instance, have to deal with VP discontinuities. As most theories do assume verb phrases, it is not surprising that VP discontinuities form an important topic of study (see e.g. the contributions to this volume by Hepple, Hoekstra, and Reape).

Noun phrases are not so often discontinuous. But of course, relative clause extraction leads to discontinuous NPs, and so do the VP discontinuities in NP-embedded VPs, as in premodifying relative clauses in Dutch or German. A genuine NP discontinuity is formed by Dutch determiner structures of the form *Wat voor*, as in sentence (2):

(2) Wat *heb jij* voor auto *gekocht?*
(What did you car buy?, meaning *What car did you buy?)*

Such discontinuities are studied by Corver (this volume).

PPs and APs are mostly contiguous, but again, it is certainly not impossible for them to be discontinuous. Examples of discontinuous adjective- and adverbial phrases are presented in (3):

(3) a. *This is a* better *movie* than I expected.
b. *Leo is* harder *gegaan* dan ooit tevoren.
(*Leo has been going faster than ever before.*)

Moreover, virtually any phrase can become discontinuous by the insertion of *metacomments*, such as parentheticals. Some examples:

(4) a. *John* talked *of course* about politics.
    b. *Peter bought a house* on, *or almost on,* Angels Beach.
    c. *Leo is going* faster, *I would say,* than ever before.
    d. *He invited* the vice-chairman, *I think it was*, of the committee.
    e. An undergraduate student, *supposedly*, who had witnessed the event *reported it to him.*
    f. *This designer is* equally cool, *you know*, as Armani.

(Examples d and e taken from Dowty, this volume.)

Even idiomatic expressions, which one may be inclined to think of as fixed, inseparable, sequences of words *par excellence*, can be discontinuous:

(5) a. Will *John soon* kick the *proverbial* bucket?
    b. Le marchand de sable va *bientôt* passer.

(See Abeillé and Schabes, this volume.)

It is thus not an exaggeration to say that discontinuities may arise almost everywhere; discontinuous constituents are by no means rare or exceptional. Still, most grammar formalisms are explicitly or implicitly based on the idea of a constituent as a continuous sequence of words, being primarily designed to describe continuous constituents and having to take recourse to special operations for dealing with discontinuities. This is because concatenation, and therefore the notion of *adjacency*, plays a central role in most grammar formalisms. The description of structures made up of non-adjacent elements therefore in general presents difficulties.

3 Approaches and techniques for handling discontinuity

In this book a wide variety of approaches to the study and treatment of discontinuous constituents is presented, varying from attempts to reduce the problems they pose by re-examining the notion of constituency and minimizing the articulation of syntactic description in constituent structures, to mathematical and computational techniques for dealing effectively with discontinuous constituents in linguistic descriptions and computer programs for interpreting and generating sentences with discontinuities.

Dowty in his contribution *Towards a Minimalist Theory of Syntactic Structure* takes a fundamental approach to discontinuous constituency by addressing the more general question what might count as a constituent and for what reasons. Dowty notes that

> "No assumption is more fundamental in the theory (and practice) of syntax than that natural languages should always be described in terms of constituent structure, at least whereever possible."

and goes on to note that, in spite of this,

> "..syntacticians don't usually ask questions .. as to what the nature of evidence for a constituent structure in a particular sentence in a particular language is: we just take whatever structure our favorite syntactic theory would predict as the expected one for the string of words in question - by current X-bar theory, etc. - unless and until that assumption is contradicted by some particular fact. .. I suspect syntacticians today have almost come to think of 'the primary empirical data' of syntactic research as phrase structure trees, so firm are our convictions as to what the right S-structure for most any given sentence is."

Dowty advocates a more skeptical, 'minimalist' approach to syntactic structure which takes the default assumption that a clause or group of words is only a linear structure: a string, and that a hierarchical structure should be assigned only when necessary for getting the data right. By taking linear structure to be the norm, Dowty argues that the description of various discontinuous syntactic phenomena can be simpler. This minimal, string-like syntactic structure Dowty calls the *phenogrammatical* structure, which he contrasts with the *tectogrammatical* structure, thereby making a distinction originally due to Curry (1963). The tectogrammatical structure describes the steps by which a sentence is built up from its parts, but without regard to the form that these combinations of parts take, i.e. without information about the order of words and phrases, whether inflectional morphology marks the syntactic organization or not, and so on. The representation of the latter information is the phenogrammatical structure.

Abeillé and Schabes in their contribution use Tree Adjoining Grammar as their theoretical framework, and argue that the formalism of Lexicalized TAGs, where sets of elementary trees are associated to lexical items, defining linguistic units of *extended domain of locality* that have both syntactic and semantic relevance, offers a natural representation for constituents that may exhibit internal discontinuities. Such constituents follow regular syntactic TAG composition rules, but lack semantic compositionality. This representation considers noncompositional constituents as one semantic unit without having to stipulate any element to be semantically empty, and is thus kept strictly increasing.

Corver's contribution approaches discontinuous constituency within the Barriers framework of Chomsky (1986). Specific attention is given to the discontinuity in Dutch determiner phrases that may occur in structures of the form *wat voor*, as in (2). It is argued that such noun phrases consist of an interrogative quasi-argument *wat*, that functions as the head of the NP, and a predicative PP headed by *voor*, which is conjoined to *wat*. The presence or absence of discontinuity in *wat voor*-phrases in various syntactic environments can be accounted for in terms of the Subjacency Condition and the Empty Category Principle on the basis of the proposed internal structure of the *wat voor*-phrases.

Hoekstra's contribution, that falls within the same theoretical tradition, is concerned not just with discontinuity but more generally with phenomena of movement and binding. Two types of binding relationships are distinguished, the stronger of which is subject to a strict locality requirement. The proposed approach to binding results in analyses that, among other things, account for a recently discovered problem concerning the effect of move-alpha on the binding of negative polarity items.

The contribution by Bunt is concerned with the development of tools for formal description and effective processing of bounded discontinuities. Bunt argues first that articulate syntactic descriptions of hierarchical constituent structure do not necessarily take the form of *trees*. He argues, with McCawley (1982), that syntactic descriptions in general should be allowed to have crossing branches, and that there is nothing objectionable about such structures, which he calls *discontinuous trees* (or *disco-trees*, for short). A precise mathematical definition of these structures is provided. A fairly intriguing matter is the definition of a linear precedence relation among discontinuous trees in a way, useful in the formulation of grammatical rules for combining such trees. The key notion used for this purpose is the *adjacency sequence*, which is defined in such a way that the concatenation of two or more discontinuous trees, forming an adjacency sequence, has the effect that the elements of those trees are interwoven. Complexity-theoretical issues are briefly addressed, as are computational issues concerning their implementation in chart parsers.

The notions introduced by Bunt are theory-neutral in that they can be used in many different grammatical frameworks. In his contribution, Bunt explains their use in phrase-structure grammars in some detail, developing a formalism called *Discontinuous Phrase Structure Grammar* (DPSG), and very briefly considers their possible use in Categorial Grammar. In a categorial framework, their use could be in allowing *discontinuous categories*. A simple example would be that of a discontinuous verb phrase like *'woke Mary up'*, here we could use a category discontinuous category `(NP§)[/NP]/PART` for the verb (*'woke'*), which concatenates to the right in an adjacency sequence `(NP§)[/NP]/PART` + NP + PART as follows:

(6) `(NP§)[/NP]/PART` + NP + PART $\Longrightarrow$ `(NP§)/NP`

The part [/NP] in the discontinuous category indicates that a noun phrase (*'Mary'*) may separate a VP-final particle (*'up'*) from the verbal stem (*'woke'*); the verb and the particle form a discontinuous functor which applies to the separating NP as its right-hand side argument.

Moortgat, in his contribution, goes at great length in studying the consequences and possibilities of discontinuous categories. He considers two *discontinuous type constructors*, intended to capture extraction and infixation operations at a level suitable for linguistic generalization. In the first case, the operator constructs a discontinous category of which the two constituent parts together form a discontinuous functor, applicable to the intervening material; in the other case (infixation) the constructor constructs a discontinuous argument for the functor corresponding to the intervening material. These operators had been introduced before in Moortgat (1988) in an attempt to accomodate Bach's (1981; 1984) work on discontinuity within a categorial type logic approach. In his contribution, Moortgat improves on his earlier proposal by presenting a *complete* logic for extraction and infixation type constructors in terms of the sign-based approach to categorial deduction.

Sanfilippo in his contribution develops the idea that the relevant constraints on discontinuous verb-object dependencies, arising in Italian sentences like *'A Maria, Carlo non gli ha ancora scritto'* ('To Mary, Carlo has not written to-him yet'), can be expressed in terms of access to the thematic domain of a sentence. It is shown how a strong generalization expressed in these terms can be given a precise syntactic, semantic and computational interpretation within the framework of Unification Categorial Grammar augmented with a neo-Davidsonian verb semantics and predicate-argument association. The resulting approach can easily be extended to also provide an empirically motivated account of dislocated subjects, and is thus interesting for the treatment of discontinuous subject-VP dependencies as well.

Hepple's contribution is also in the framework of Categorial Grammar.

He presents a novel view of the relation between word order and subcategorization, according to which the canonical word order in any language arises through the interaction of three factors: (i) the 'underlying' order of subcategorization by the verb for its complements corresponding to obliqueness; (ii) the direction (left or right) of subcategorization by the verb of each complement; (iii) a lexical process of 'verb displacement', causing the verb to be displaced from its position relative to its complements according to the underlying subcategorization.

Pollard's contribution is concerned with a detailed study of verb-second and raising phenomena in the framework of Head-driven Phrase Structure Grammar. In Government-Binding theory and associated frameworks it has become standard to regard to occurrence in second position of the finite verb in declarative sentences (in English, Dutch, German, and many other languages) as a case of head movement. The finite verb or auxiliary, on this account, undergoes movement into the COMP position, and some phrasal constituent moves to the left of it. The apparent discontinuity between the finite auxiliary and its VP sister in such sentences might be thought to be problematic for nontransformational theories. Pollard discusses this issue in detail for verb clusters in German, and gives an account of the possibilities and constraints to front parts of such clusters. A satisfactory analysis of these phenoma is proposed within the HPSG framework, with one issue not entirely resolved: the resulting set of combination schemata allow different parses for certain sentences, with no demonstrable linguistic (e.g., pragmatic) differences between the alternative analyses. Exactly how this poblem of spurious ambiguity can be resolved is not quite clear.

Reape's contribution is also in the framework of HPSG. He argues that surface syntax should be rejected as the basis of word order. Instead, he claims that word order is determined within locally definable 'word order domains', where nonlexical word order domains are composed from smaller word order domains, and where the word order domain of a daughter may be the same as that of the mother. This last stipulation is the basis for allowing discontinuities, as it means, informally, that the elements of the daughters's domain can be interleaved with those of the mother's domain. Technically, Reape introduces an operation called *sequence union* to be able to describe an 'interleaving concatenation' of discontinuous constituents. The sequence union operator applied to two strings A and B can produce any sequence of elements from A and B which conserves the relative order among the A-elements and that among the B-elements. This concept is similar to Bunt's notion of concatenating *adjacency sequences*, the latter being slightly more constraining.

Aarts in his contribution introduces a variant of Bunt's DPSG framework, which he calls *Discontinuous Rewrite Grammar*. He shows how the formalism can describe both bounded and unbounded dependencies in terms of discontinuous trees with crossing branches, and pays particular attention to the design of an elegant and efficient chart parser for this formalism.

De Smedt and Kempen describe a grammar formalism designed for incremental generation of sentences, called *Segment Grammar*. This formalism is somewhat reminiscent of Tree-Adjoining Grammar, but differs in that sister nodes can be incrementally added to an existing (partial) structure. The formalism distinguishes two levels of syntactic description: *f-structures*, which are unordered functional structures, and *c-structures*, which represent left-to-right order of constituents. True discontinuities are viewed as differences between immediate dominance relations in c-structures and those in corresponding f-structures. Constructions that are treated in this way include clause union (including Dutch cross-serial dependencies), right dislocation, and fronting.

Their contribution is concerned with discontinuous constituency not only from a structural viewpoint but also from a processing point of view. They note that several kinds of discontinuities seem to offer advantages for an incremental generation strategy; this is especially true of optional dislocations. Fronting of focused constituents is natural if we assume prominent concepts to be passed on to the grammatical encoder earlier than other parts of the semantic input. Similarly, right dislocation allows the generator to utter constituents which are ready and postpone uttering more complex ('heavy') ones, which are still being processed semantically. In addition, right dislocation allows the incorporation of new semantic input as afterthoughts.

Saint-Dizier's contribution is devoted to new techniques for expressing constraints and relations on discontinuous elements in a structure in a computationally attractive way. To this end, an extension to the logic programming language Prolog is presented, called *Dislog*, especially designed for logic programming with discontinuities. A Dislog clause is a set of definite clauses, possibly sharing common variables, which must co-occur in a parse (or proof) tree. The notion of discontinuity is expressed by the fact that there are no *a priori* restrictions on the locations of these definite clauses in the parse (proof) tree. Saint-Dizier explains the concepts of Dislog and shows how Dislog can be applied to the parsing of discontinuous constituents in natural language.

References

Aarts, E. (1994) Discontinuous Rewrite Grammar and Unbounded Dependencies. *This volume.*

Abeille, A. & Schabes, Y. (1994) Non-compositional discontinuous constituents in Tree Adjoining Grammar. *This volume.*

Bach, E. (1981) Discontinuous constituents in generalized cetegorial grammar. *NELS XI*, Amherst (MA), 1-12.

Bach, E. (1984) Some generalizations of Categorial Grammars, in F. Landman & F. Veltman (eds.) *Varieties of Formal Semantics*, 1-23.

Bunt, H.C. (1994) Formal tools for describing and processing discontinuous constituency structure. *This volume.*

Chomsky, N. (1986) *Barriers*, MIT Press, Cambridge (MA).

Corver, N. (1994) Discontinuity and the 'wat voor'-construction in Dutch. *This volume.*

Dowty, D.R. (1994) Toward a Minimalist Theory of Syntactic Structure. *This volume.*

Emonds, J.E. (1976) *A transformational approach to English syntax.* Academic Press, New York etc.

Emonds, J.E. (1979) Appositive relatives have no properties. *Linguistic Inquiry*, vol. 10, 211-243.

Hepple, M. (1994) Verb Movement in Categorial Grammar and the Dowty-Bach Analysis of Grammatical Relations. *This volume.*

Hoekstra, E. (1994) Discontinuity and the Binding Theory. *This volume.*

McCawley, J.D. (1982) Parentheticals and Discontinuous Constituent Structure. *Linguistic Inquiry*, vol. 13, no. 1, 91-106

Moortgat, M. (1994) The Logic of Discontinuous Type Constructors. *This volume.*

Pollard, C.H. (1994) On Head Non-Movement. *This volume.*

Reape. M. (1994) Getting Things in Order. *This volume.*

Ross, J.R. (1973) Slifting. In M. Gross, M. Halle & M.P. Schützenberger (eds.) *The formal analysis of natural language.* Mouton, The Hague.

Saint-Dizier, P. (1994) Expressing Discontinuous Constituency in DISLOG. *This volume.*

Sanfilippo, A. (1994) Thematic Accessability in Discontinuous Dependencies. *This volume.*

Smedt, K. De & Kempen, G. (1994) Discontinuous Constituency in Segment Grammar. *This volume.*