Gerdes K., Kahane S. (2013) Defining dependency (and constituency), in K. Gerdes, E. Hajičová, L. Wanner (éds.), *Computational Dependency Linguistics*, IOS Press.

Defining dependencies (and constituents)

Kim GERDES[a] and Sylvain KAHANE[b]
[a]*LPP, Sorbonne Nouvelle, Paris*
[b]*Modyco, Université Paris Ouest Nanterre*

Abstract. The paper proposes a mathematical method of defining dependency and constituency provided linguistic criteria to characterize the acceptable fragments of an utterance have been put forward. The method can be used to define syntactic structures of sentences, as well as discourse structures for texts, or even morphematic structures for words. Our methodology leads us to put forward the notion of *connection*, simpler than dependency, and to propose various representations which are not necessarily trees. We also study criteria to refine and hierarchize a connection structure in order to obtain a tree.

Keywords. Connection graph, dependency tree, phrase structure, syntactic representation, syntactic unit, catena

Introduction

Syntacticians generally agree on the hierarchical structure of syntactic representations. Two types of structures are commonly considered: Constituent structures and dependency structures (or mixed forms of both, like headed constituent structures, sometimes even with functional labeling, see for example Tesnière's nucleus [34], Kahane's bubble trees [21], or the Negra corpus' trees [6]). However, these structures are often rather purely intuition-based than well-defined and linguistically motivated, a point that we will illustrate with some examples. Even the basic assumptions concerning the underlying mathematical structure of the considered objects (ordered constituent tree, unordered dependency tree) are rarely motivated (why should syntactic structures be trees to begin with?).

In this paper, we propose a definition of syntactic structures that supersedes constituency and dependency, based on a minimal axiom: *If an utterance can be separated into two fragments, we suppose the existence of a connection between these two parts.* We will show that this assumption is sufficient for the construction of rich syntactic structures.

The notion of *connection* stems from Tesnière who says in the very beginning of his *Éléments de syntaxe structurale* that

> "Any word that is part of a sentence ceases to be isolated as in the dictionary. Between it and its neighbors, the mind perceives **connections**, which together form the structure of the sentence."

Our axiom is less strong than Tesnière's here, because we do not presuppose that the connections are formed between words only. In the rest of the paper, we use the term *connection* to designate a non-directed link (*the* — *dog*; *the* and *dog* are connected) and the term *dependency* to designate a directed link (*dogs* ← *slept*; *dogs* depends on *slept* or *slept* governs *dogs*).

We will investigate the linguistic characteristics defining the notion of "fragment" and how this notion leads us to a well-defined graph-based structure, to which we can apply further conditions leading to dependency or constituency trees. We will start with a critical analysis of some definitions in the field of phrase structure and dependency-based approaches (Section 1). Connection structures are defined in Section 2, and are applied to discourse, morphology, and deep syntax in Section 3. The case of surface syntax is explored in Section 4. Dependency structures are defined in Section 5 and refined in Section 6. In Section 7, we show how to derive constituent structures from dependency structures.

1. Previous definitions

1.1. Defining dependency

Tesnière 1959 [34] does not go any further in his definition of dependency and remains on a mentalist level ("the mind perceives connections"). The first formal definition of dependency stems from Lecerf 1960 [24] and Gladkij 1966 [13] (see also [21]) who showed that it is possible to infer a dependency tree from a constituent tree with heads (what is commonly called *phrase structure*). Further authors have tried to overcome these first definitions of constituency. Mel'čuk 1988 ([25]) states that two words are connected as soon as they can form a fragment, and he gives criteria for characterizing acceptable two-word fragments. But it is not always possible to restrict the definition to two-word fragments. Consider:

(1) *The dog slept.*

Neither *the slept* nor *dog slept* are acceptable syntactic fragments. Mel'čuk resolves the problem by connecting *slept* with the head of *the dog*, which means that his definitions of fragments and heads are mingled. Moreover Mel'čuk's definition of the head is slightly circular: "In a sentence, wordform w1 directly depends syntactically on wordform w2 if the passive [surface] valency of the phrase w1+w2 is (at least largely) determined by the passive [surface] valency of wordform w2." However, the concept of passive valency presupposes the recognition of a hierarchy, because the passive valency of a word or a fragment designates the valency towards its governor (see Section 5).[1]

Garde 1977 [12] does not restrict his definition of dependency to two-word fragments but considers more generally "significant elements" which allows him to construct the dependency between *slept* and *the dog*. However, he does not show how to reduce such a dependency between arbitrary "significant elements" to links between words. The goal of this article is to formalize and complete Garde's and Mel'čuk's definitions.

Schubert 1987 ([29]) attempts to define dependency as "directed co-occurrence" while explicitly including co-occurrence relations between "distant words". He explains the directedness of the co-occurrence by stating that the "occurrence of certain words [the dependent] is made possible by the presence of other words," the governor. However, "form determination should not be the criterion for establishing co-occurrence lines." This

[1] The valency of an element is the set of all the connections it awaits commonly. The idea is that a connection is generally controlled by the governor, which is then active, while the dependent is passive. In the case of modifiers, however, the contrary holds as they rather control the connection with their syntactic governor. The passive valency of an element can also be seen as its distribution. This supposes that when looking at the distribution of an element we exclude all elements depending on it, which supposes that we already know which slot of the valency is the governor's slot (see Section 5).

adds up to lexical co-occurrences which can describe relationships between words on a semantic or on a discourse level. Consider the relation between *radio* and *music* in:

(2) *The radio is playing my favorite music.*
(3) *I heard a great piece of music on the radio this morning.*

It is clear that *radio* and *music* co-occur frequently in a statistical sense of the word "co-occur", *i.e.* the occurrence of one word is highly correlated with the occurrence of the other word within the same sentence. However, in both sentences, *music* and *radio* do not form an acceptable text fragment.

Moving closer to syntax, consider the relationship between *radio* and *play* in sentences (2). This relation describes something we would name a "semantic dependency", a type of dependency that Hudson [18] precisely proposes to show in his dependency structures. For our part, we want to restrict a connection and a dependency to couples of elements that can form an acceptable text fragment in isolation (which is not the case of *the radio playing* and even less so for *music* and *radio*). We do not disagree that some sort of dependency exists between these words, but we consider this link as a lexical or semantic dependency (see Mel'čuk [25], [27]) rather than as a surface syntactic one.

1.2. Defining constituency

In order to evaluate the cogency of a definition of dependency based on a pre-existing definition of constituency, we have to explore how constituents are defined. Bloomfield 1933 [4] does not give a complete definition of syntactic constituents. His definition of the notion of *constituent* is first given in the chapter *Morphology* where he defines the morpheme. In the chapter on syntax, he writes:

> "Syntactic constructions are constructions in which none of the immediate constituents is a bound form. […] The actor-action construction appears in phrases like: *John ran, John fell, Bill ran, Bill fell, Our horses ran away*. […] The one constituent (*John, Bill, our horses*) is a form of a large class, which we call *nominative expressions*; a form like *ran* or *very good* could not be used in this way. The other constituent (*ran, fell, ran away*) is a form of another large class, which we call *finite verb expressions.*"

Bloomfield does not give a general definition of constituents: They are only defined by the previous examples as instances of distributional classes. The largest part of the chapter is dedicated to the definition of the head of a construction. We think that in some sense Bloomfield should rather be seen as a precursor of the notions of connection (called *construction*) and dependency than as the father of constituency.

For Chomsky, a constituent exists only inside the syntactic structure of a sentence, and he never gives precise criteria of what should be considered as a constituent. In Chomsky 1986 [9], quarreling with the behaviorist claims of Quine [31], he refutes it as equally absurd to consider the fragmentation of *John contemplated the problem* into *John contemplated – the problem* or into *John contemp – lated the problem* instead of the "correct" *John – contemplated the problem.* No further justification for this choice is provided.

Gleason 1961 ([14]) proposes criteria for defining constituents (like substitution by one word, possibility to be a prosodic unit) and to build a constituent structure bottom up:

> "We may, as a first hypothesis, consider that each of [the words of the considered utterance] has some statable relationships to each other word. If we can describe these interrelationships completely, we will have described the syntax of the utterance in its entirety. […] We might start by marking those pairs of words which are felt to have the closest relationship. "

But he makes the following assumption without any justification: "We will also lay down the rule that each word can be marked as a member of only one such pair." Gleason then declares the method of finding the best among all the possible pairings to be "the basic problem of syntax" and he notes himself that his method is "haphazard" as his "methodology has not as yet been completely worked out" and lacks precise criteria. We are not far from agreeing with Gleason but we do not think that one has to choose between various satisfactory pairings. For instance, he proposes the following analysis for the NP *the old man who lives there*:

<table>
<tr><td>the</td><td>old</td><td>man</td><td>who</td><td>lives</td><td>there</td></tr>
<tr><td>the</td><td colspan="2">graybeard</td><td>who</td><td colspan="2">survives</td></tr>
<tr><td>the</td><td colspan="2">graybeard</td><td colspan="3">surviving</td></tr>
<tr><td>the</td><td colspan="5">survivor</td></tr>
<tr><td colspan="6">he</td></tr>
</table>

We think that other analyses are possible, such as

<table>
<tr><td>the</td><td>old</td><td>man</td><td>who</td><td>lives</td><td>there</td></tr>
<tr><td>the</td><td colspan="2">graybeard</td><td colspan="2">living</td><td>there</td></tr>
<tr><td colspan="3">someone</td><td colspan="3">surviving</td></tr>
<tr><td colspan="6">he</td></tr>
</table>

and these analyses are not in competition, but complementary; both (and others) can be exploited to find the structure of this NP.

Today, the definition of 'constituent' seems no longer to be a significant subject in contemporary literature in syntax. Even pedagogical books in phrase structure based frameworks tend to skip the definition of constituency, for example Haegeman 1991 [17] who simply states that "the words of the sentence are organized hierarchically into bigger units called phrases.", and take constituency for granted.

Commonly proposed tests for constituency include proform substitution tests (including interrogation and clefting), the "stand-alone test", meaning that the segment can function as an "answer" to a question, "movement tests" (including insertion and suppression), and coordinability, the latter being fraught with confounding factors of multiple constituents, gapping, and right node raising (RNR). However, the application of these criteria to our previous example (*the old man who lives there*) does not clearly favor the first decomposition over the second one.

In phrase structure frameworks, constituents are nothing but a global approach for the extraction of regularities, the only goal being the description of possible constructions with as few rules as possible. However, it is never actually shown that the proposed phrase structure really is the most efficient way of representing the observed utterances.

We see that the notion of constituency is either not defined at all or in an unsatisfactory way, often based on the notion of one element, the *head*, being linked to another, its *dependent*, modifying it. It is clear that the notion of dependency cannot be defined as a derived notion of constituency, as the definition of the latter presupposes head-daughter relations, making such a definition of dependency circular. Conversely, we will see that, as soon as the dependency relations are constructed, it is possible to select some fragmentations between all those that are possible, these fragmentations being the ones that are aimed at in phrase structure based approaches.

1.3. Intersecting analyses

An interesting result of the vagueness of the definitions of constituency is the fact that different scholars invent different criteria that allow choosing among the possible constituent structures. For example, Jespersen's lexically driven criteria select particle verbs as well as idiomatic expressions. For instance, the sentence (4) is analyzed as “S W O” where W is called a “composite verbal expression” (Jespersen 1937[16])

(4) *She* [*waits on*] *us*.

As a point of comparison, Van Valin & Lapolla 1997 [35] oppose *core* and *periphery* of every sentence and obtain another unconventional segmentation of sentences as in example (5).

(5) [*John ate the sandwich*] [*in the library*]

Assuming one of these various fragmentations necessitates that one put forward additional statements (all legitimate) based on different types of information like head-daughter relations (for X-bar approaches), idiomaticity (for Jespersen) or argument structure or information packaging (for VanValin & Lapolla) and serve merely for the elimination of unwanted fragments.

For us, multiple decompositions of an utterance are not a problem. There is no reason to restrict ourselves to one particular fragmentation, as it is done in phrase-structure based approaches. On the contrary, we think that the best way to compute the syntactic structure of an utterance is to consider all its possible fragmentations and this is the idea we want to explore now. Steedman [33] may have been one of the first linguists to develop a formal grammar that allows various groupings of words. Steedman’s articles corroborate the multi-fragment approach to syntactic structure which is pursued here.

2. Fragmentation and connection

2.1. Fragments

We will relax the notion of syntactic constituent and define a new syntactic notion: We call a part of an utterance a *fragment* if it is a linguistically acceptable phrase with the same semantic contribution as in the initial utterance. Let us take an example:

(6) *Peter wants to read the book.*

We consider the acceptable fragments of (6) to be: *Peter, wants, to, read, the, book, Peter wants, wants to, to read, the book, Peter wants to, wants to read, read the book, Peter wants to read, to read the book, wants to read the book*.

We will not justify this list of fragments at this point, but rather we point for the moment just to the fact that *wants to read,* just like *waits on* in (4), fulfills all the commonly considered criteria of a constituent: It is a "significant element", "functions as a unit" and can be replaced by a single word (*reads*).[2] In the same way, *Peter wants* could be a perfect utterance. Probably the most unnatural fragment of (6) is the VP *wants to read the book,* that, together with the corresponding subject, is traditionally considered as the main constituent of a clause in a phrase structure analysis.

Our fragments correspond more or less to the catenae of Osborne et al. [28] (see Section 2.8 for details). We both think that fragments and catenae are very relevant units (more than constituents in particular), but we consider that the relationship between fragments and dependencies goes the other way around: Osborne et al. [28] define the catenae of from the dependency tree as the connected subparts of this tree, but do not say how they define dependency. We think that the best way to define dependency is to define fragments first.[3]

2.2. Fragmentations

A *fragmentation* (*tree*) of an utterance U is a recursive partition of U into acceptable fragments, that is, a tree-like decomposition of the utterance into fragments. Figure 1 shows two of the various possible fragmentations of (6).[4]

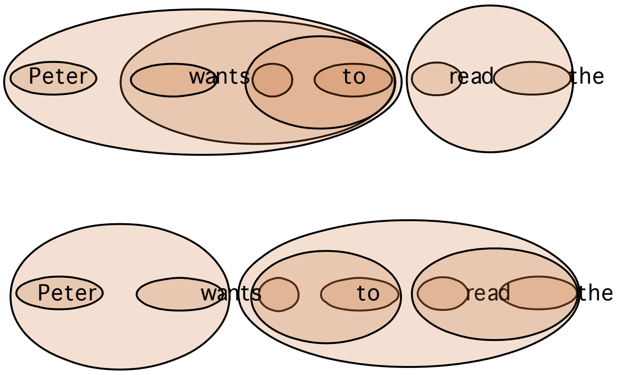


Figure 1. Two fragmentation trees of (6)

[2] The substitution test is generally limited to proforms. But this test does not work for adjectival or adverbial phrases for instance. And although English has the verbal proform DO, such a proform does not exist in many other languages, including close languages such as French (for instance, *Pierre a parlé à Marie* 'Peter talked to Mary' gives *Pierre* ***l'****a fait* 'Peter did (it)', where an accusative clitic is obligatory even if the verb PARLER has only an indirect object). The possible substitution by a single lexical item can be a useful relaxing of the substitution test.

[3] It may seem paradoxical that we do not think that fragments are syntactic primitives. In some sense we agree with Tesnière when he says that "the mind perceives connections". A fragment is a witness to a possible connection and it allows us to postulate one connection or another.

[4] We represent a constituency tree as a bracketing. See Figure 15 for the equivalence to the traditional representation Another equivalent representation, introduced by Hockett at the end of the 50s, was used above when we presented Gleason 1961's decomposition of *the old man who lives there*.

More formally, if X is set of minimal units (for instance the words of (6)), *fragments* are subsets of X and a *fragmentation* F is a subset of the powerset[5] of X ($F \subseteq P(X)$), which is *well-partitioned*, that is, which verifies the two following properties:

1. For every f_1, $f_2 \in F$, either $f_1 \subseteq f_2$, $f_2 \subseteq f_1$, or $f_1 \cap f_2 = \varnothing$;
2. Each fragment is partitioned by its immediate sub-fragments.

Written out less formally this means that a fragmentation is just a selection of subsets composed of minimal units (the fragments) such that:

1. The fragments cannot overlap strangely: If two fragments overlap, then one must be completely contained in the other.
2. Each (non-minimal) fragment can be decomposed into fragments.

A fragmentation whose fragments are constituents is nothing else than a constituent tree.[6] A fragmentation is *binary* if every fragment is partitioned into 0 or 2 fragments.

2.3. Connection structure and fragmentation hypergraph

We consider that each partition of a fragment in two pieces induces a *connection* between these two pieces.[7] This allows us to define the graph of connections between the fragments of a set X. Such a graph, defined on a set of sets (that is, on a subset of the powerset of X) is called a *hypergraph*. More formally, a *hypergraph* H on X is a triplet (X,F,φ) where $F \subset P(X)$ (F is the set of fragments) and φ is a graph on F. If F is only composed of singletons, H corresponds to an ordinary graph on X. For each binary fragmentation F on X, we will define a *fragmentation hypergraph* H = (X,F,φ) by introducing a connection between every couple of fragments which partitions another fragment.

Let us illustrate this with an example:

(7) *Little dogs slept.*

There are two natural fragmentations of (7) whose corresponding hypergraphs are given Figure 2.[8]

[5] Every set *S* has a *powerset*, noted *P(S)*. It is the set of all subsets of *S*. For example, if *S={a,b,c}* then *P(S)* contains the following elements: the whole set *S* itself, all subsets of two elements *{a,b}*, *{a,c}*, *{b,c}*, all subsets of one element *{a}*, *{b}*, *{c}*, and ⍰, the empty set. This gives the identity *P(S)={ {a,b,c}, {a,b}, {a,c}, {b,c}, {a}, {b}, {c}, ⍰}*. A powerset is thus a set of sets, as is a fragmentation, which is a set of fragments, which are sets of words (or whatever minimal units we have chosen). And if such a set of sets has good properties (i.e. if it is well-partitioned), we can represent it by a tree or a bracketing.

[6] There are two views on constituents. From a purely formal point of view, every fragmentation tree is a constituent tree, that is, a tree where each node *C* represents a subpart of X and the daughters of *C* are subparts of *C* partitioning *C*. From a linguistic point of view, only some subparts of X are (syntactic, prosodic, …) constituents, and this is why some linguists would consider that only some of our fragmentation trees are constituent trees.

[7] The restriction of the connections to binary partitions can be traced back all the way to Becker (1827:469 [3]), who claims that "every organic combination within language consists of no more than two members." (*Jede organische Zusammensetzung in der Sprache besteht aus nicht mehr als zwei Gliedern).* Although we have not encountered irreducible fragments of three or more elements in any linguistic phenomena we looked into, this cannot be *a priori* excluded. It would mean that we encountered a fragment XYZ where no combination of any two elements forms a fragment, i.e. is autonomizable in any without the third element. Our formal definition does not exclude this possibility at any point and a connection can in theory be ternary.

[8] It is possible that, for most readers, H_1 seems to be more natural than H_2. From our point of view, that is not the case: *dogs slept* is a fragment just as valid as *little dogs*. Nevertheless, see footnote 12.

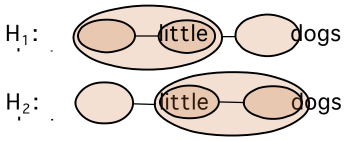


Figure 2. The two fragmentation hypergraphs of (7)

These two hypergraphs show both constituents and connections (i.e. non-hierarchized dependency). This is redundant and we will now see how to keep the connections only. We remark that *little* is connected to *dogs* in H_1 and *dogs* to *slept* in H_2. H_2 also shows a connection between *little* and *dogs slept*, but in some sense, this is just a rough version of the connection between *little* and *dogs* in H_1. The same observation holds for the connection between *little dogs* and *slept* in H_1, which corresponds to the connection between *dogs* and *slept* in H_2. In other words, the two hypergraphs contain the same connections (in more or less precise versions). We can thus construct a finer-grained hypergraph H with the finest version of each connection (Figure 3). We will call this hypergraph (which is equivalent to a graph on the words in this case) the *connection structure* of the utterance. We will now see how to define the connection structure in the general case.

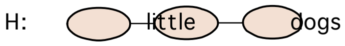


Figure 3. Connection structure of (7)

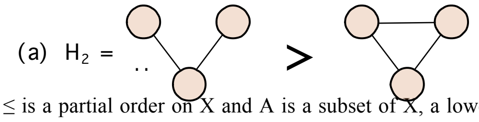
2.4. A complete partial order on hypergraphs

We saw with our example that the connection structure is a finer-grained version of the different fragmentation hypergraphs of the utterance. So we propose to define the connection structure as the *infimum*[9] of the fragmentation hypergraphs for a natural order of fineness. The definition we expose in this subsection requires good skills in mathematics and can be skipped without loss of continuity.

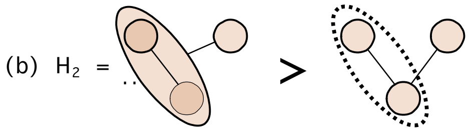
A connection f — g is *finer* than a connection f' — g' if f ⊆ f' and g ⊆ g'. For instance the connection [*dogs*]–[*slept*] is finer than the connection [*little dogs*]–[*slept*]. A connection is *minimal* when it cannot refine.

Intuitively, the *fineness order*, henceforth noted ≤, represents the precision of the hypergraph, ie. $H_1 \leq H_2$ if H_1 is a finer-grained analysis than H_2. A hypergraph H_1 is *finer* than a hypergraph H_2 (that is $H_1 \leq H_2$) if every connection in H_2 has a finer connection in H_1.

In other words, H_1 must have more connections than H_2, but H_1 can have some connections pointing to a smaller fragment than in H_2, and in this case the bigger fragment can be suppressed in H_1 (if it carries no other connections) and H_1 can have less fragments than H_2. This is illustrated with the following schemata:



[9] If ≤ is a partial order on X and A is a subset of X, a lower bound of A is an element b in X such that b ≤ x for each x in A. The infimum of A, noted ΛA, is the greatest lower bound of A. A partial order for which every subset has an infimum is said to be complete. (As a classical example, consider the infimum for the divisibility on natural integers, which is the greatest common divisor: 9 ⍰ 12 = 3).



In case (a), H_1 is finer because it has one connection more. In case (b), H_1 is finer because it has a finer-grained connection and the dotted fragment can be suppressed. It is suppressed when it carries no further connection.

We think that this partial order on hypergraphs is *complete* (see note 9). We have not proven this claim but it appears to be true on all the configurations we have investigated.

If we have an utterance U and linguistic criteria characterizing the acceptable fragments of U, we define the *connection structure* of U as the infimum of all its fragmentation hypergraphs.

2.5. Constructing the connection structure

Our definitions are complicated, perhaps. In practice however, it is easy to build the connection graph of an utterance as soon as you have decided what the acceptable fragments of an utterance are. Indeed, because the fineness order on hypergraphs is complete, one can begin with any fragmentation and refine its connections until there are no further refinements to be made. The connection structure is obtained when all the connections are minimal. The completeness ensures, due to the uniqueness of the greatest lower bound, that one always obtains the same structure.[10]

Let us see what happens with example (6). Suppose the first step of the fragmentation is:

f_1 = *Peter wants to*
f_2 = *read the book*

One has a connection here between f_1 and f_2 that will correspond to a link between two minimal fragments in the final connection, possibly words. Now, one wants to discover these minimal fragments. To accomplish that, one seeks the minimal fragment g overlapping both f_1 and f_2: g = *to read*. The fragment g can be decomposed into *to* and *read*. Therefore the connection between f_1 and f_2 is finally a connection between *to* and *read*. It now remains to calculate the connection structures of f_1 and f_2 in order to obtain the complete connection structure of the whole sentence (Figure 4).

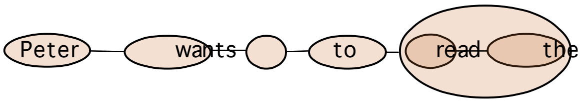


Figure 4. Connection structure of (6)

[10] This is more complicated if the connection structure contains cycles (Section 2.7). The previous process, starting with a (fragmentation) tree, gives us acyclic structures and we must verify that no connection can be added to obtain the whole connection structure.

2.6. Irreducible fragment

The connection structure of (6) is not equivalent to a graph on its words because some fragments are irreducible. An *irreducible fragment* is a fragment bearing connections which cannot be attributed to one of its parts. For instance, *the book* in (6) is irreducible because there is no fragment overlapping *the book* and including only *the* or only *book* (neither *read the* nor *read book* are acceptable).

(8) *The little dog slept.*

Example (8) poses the same problem, because *little* can be connected to *dog* (*little dog* is acceptable), but *slept* must be connected to *the dog* and cannot be refined (neither *dog slept* or *the slept* is acceptable). One easily verifies that (8) has the fragmentation hypergraphs F_1, F_2, and F_3 of Figure 5 and the connection graph H (which is their infimum). Note that the fragment *the dog* persists in the final connection graph H because it carries the link with *slept* but *little* is connected directly to *dog* and not to the whole fragment *the dog*.

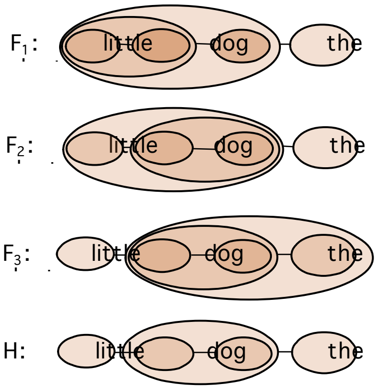


Figure 5. The fragmentations hypergraphs of (8) and its connection structure $H = F_1 \wedge F_2 \wedge F_3$

Irreducible fragments are quite common with grammatical words. We have seen the case of determiners, but conjunctions, prepositions, or relative pronouns can also cause irreducible fragments:

(9) *I think [that [Peter slept]]*
(10) *Pierre parle [à Marie]*[11]
Peter speaks [to Mary]
(11) *[the (old) man] [who lives] there*

[11] Preposition stranding (**Pierre parle à* 'Pierre speaks to') is impossible in French.

2.7. Cycles

Usually the connection graph is acyclic (and could be transformed into a tree by choosing a node as the root, as we have shown for example (7)). But we can have a *cycle* when a fragment XYZ can be fragmented into XY+Z, YZ+X, and XZ+Y. This can happen in examples like:

(12) *Mary gave advice to Peter.*
(13) *I saw him yesterday at school.*
(14) *the rise of nationalism in Catalonia*

In (12), *gave advice, gave to Peter*, and *advice to Peter* are all acceptable. We encounter a similar configuration in (12) with *saw yesterday, saw at school,* and *yesterday at school* (*It was* yesterday at school *that I saw him*). In (13), *in Catalonia* can be connected both with *nationalism* and *the rise* and there is no perceptible change of meaning. We can suppose that the hearer of these sentences constructs one connection or the other (or even both) and does not need to favor one.[12]

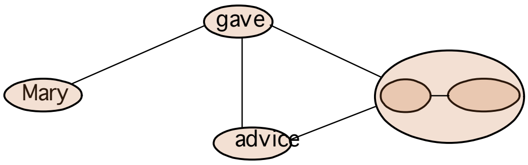


Figure 6. The cyclic connection structure of (12)[13]

Personal names are another interesting example.[14]

(15) *I met Fernando Manuel Rodriguez Pérez*

For such a Spanish name the possible fragments are *Fernando, Rodriguez, Fernando Manuel, Fernando Manuel Rodriguez, Fernando Rodriguez*, and *Fernando Rodriguez Pérez*, giving us the following connection graph with a cycle, because both *Fernando* and *Rodriguez* can be connected to the verb (Figure 7).

[12] The fact that we cannot always obtain a tree structure due to irreducible fragment and cycle suggests that we could add weights on fragments indicating that a fragment (or a fragmentation) is more likely than another. We do not pursue this idea here, but we think that *weighted connection graphs* are certainly cognitively motivated linguistic representations.
Note also that the preferred fragmentation is not necessary to the constituent structure. For instance, the most natural division in (i) occurs right before the relative clause, which functions as a second assertion in this example and can be preceded by a major prosodic break (Deulofeu *et al.* [11]).
(i) *He ran into a girl, who just after that entered the shop.*

[13] The irreducibility of *to Peter* is conditioned by the given definition of fragments. If we considerrelativization as a criterion for fragments, the possibilities of preposition stranding in English may induce the possibility to affirm that *gave* and *advice* are directly linked to the preposition *to*.

[14] We would like to thank Orsolya Vincze and Margarita Alonso Ramos for presenting us with this data.

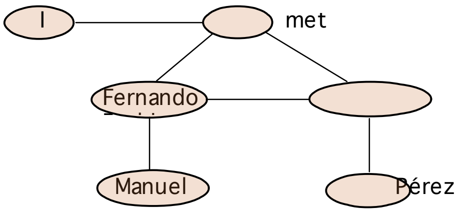


Figure 7. The cyclic connection structure of (15)

2.8. Connection structures and fragments

We have seen that the connection structure is entirely defined from the set of fragments. Conversely, the set of fragments can be reconstructed from the connection graph.

We extend the notion of *catena* that Osborne et al. [28] define for dependency trees. For a dependency tree T on X, a catena of T is a subset of X composed of elements that are connected together by dependencies in T. In other words, catenae of T are supports of connected subparts of T (the support of a structure is just the set of its vertices). For instance, if we take the previous example (Figure 7), *Manuel, Fernando* and *met* are connected and form a connected subpart of the complete connection structure: the support of this subgraph, *met Fernando Manuel*, is a catena.

For hypergraphs, we need a slightly more complex definition, in particular because the notion of "connectedness" is not immediate in this case. In a (well-partitioned) hypergraph, we said that two fragments are *weakly connected* if there is a connection between them or if one is included in the other (they cannot intersect if the hypergraph is well-partitioned). A hypergraph is *weakly connected* if its fragments are weakly connected. For instance, the connection structure of Figure 5 is weakly connected because *little* is connected to *dog*, which is included in [*the dog*], which is connected to *slept*. But it is not connected in a strong sense, because there is no chain of connections from *little* to *slept* without considering the inclusion between *dog* and [*the dog*].

A *catena* of a well-partitioned hypergraph H is the support of a weakly connected sub-hypergraph of H. A weakly connected sub-hypergraph of H is a subpart of H, the vertices of which are weakly connected together. If H is a hypergraph on X, the support of a subgraph of H is a subset of X.; therefore, the support of a sub-hypergraph is not exactly the set of its vertices, because its vertices are subsets of X, but the union of its vertices. For instance, *read* — [*the book*] is a weakly connected sub-hypergraph of the connection structure of Figure 4; this sub-hypergraph has two vertices, *read* and *the book*, and its support, *read the book* is their union. But *read book* is not a catena because *read* and *book* are not weakly connected.

Every catena can be obtained by cutting connections in the structure and keeping the segment of the utterance corresponding to continuous pieces of the connection structure. For instance in the connection structure of (6), cutting the connections between *to* and *read,* gives the segment *read the book*. But the segment *read book* cannot be obtained because even when cutting the connection between *the* and *book*, *read* remains connected to the entire group *the book*.

We have the following result: If F is a set of fragments and C is the connection structure associated to F, then the set of catenae of C is F. It means that the connection structure contains the memory of all the fragments and all the fragmentation trees that this set of fragments allows us to construct. In this sense, the connection structure is a very powerful structure and the most economical way we can imagine to encode a set of fragments. Considering that our fragments are representative syntactic units, the connection structure they define is a representative syntactic structure (and maybe the most motivated syntactic structure we can imagine).

We will now see that different criteria allow us to define different sets of fragments, which in turn define connection structures for different domains.

3. Discourse, morphology, semantics

Dependency structures are usually known to describe the syntactic structures of sentences, i.e the organization of the sentence's words. In the next sections, we will give a precise definition of fragments for surface syntax in order to obtain a linguistically motivated connection structure and to transform it into a dependency tree. Let us now at first apply our methodology to construct connection structures for discourse, morphology, and the syntax-semantics interface.

3.1. Discourse

Nothing in our definition of connection graphs is specific to syntax. We obtain syntactic structures if we limit our maximal fragment to be sentences and our minimal fragments to be words. But if we change these constraints and begin with a whole text and take “discourse units” as minimal fragments, we obtain a discourse connection graph. This strategy can be applied to define discourse relations and discourse structures such as RST or SDRT. Of course, to obtain linguistically motivated structures, we need to define what an acceptable sub-text of a text is (generally it means to preserve coherency and cohesion).

(16) (π_1) *A man walked in.* (π_2) *He sported a hat.* (π_3) *Then a woman walked in.* (π_4) *She wore a coat.* (Asher & Pogodalla [2])

We have the fragments $\pi_1\pi_2$, $\pi_1\pi_3$, $\pi_3\pi_4$ but we do not have $\pi_2\pi_3$ (no coherency) nor $\pi_1\pi_4$ (no cohesion). This gives us the connection graph of Figure 8.

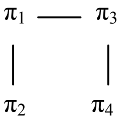


Figure 8. Connection structure of discourse (16)

3.2. Morphology

On the other side, we can fragment words into morphemes. To define the acceptable fragmentations of a word, we need linguistic criteria like the commutation test. As an example for constructional morphology, consider the word “*unconstitutionally*”, which has two possible fragmentations presented in Figure 9. These two possible decompositions can be summarized in a unique structure, i.e. the connection structure induced by the two possible fragmentations.

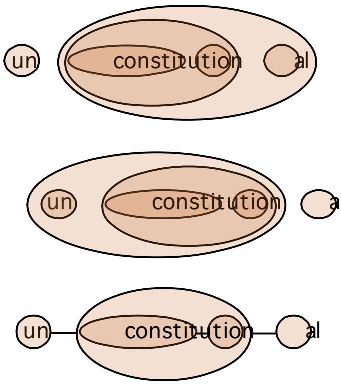


Figure 9. Two fragmentation trees and the resulting connection structure for a word decomposition

3.3. Deep Syntax

The *deep syntactic representation* is the central structure of the semantics-syntax interface (Mel'čuk [25], Kahane [22]). If we take compositionality as a condition for fragmentation, we obtain a structure that resembles Mel'čuk's deep syntactic structure. In other words, the deep syntactic structure is obtained by the same method as the surface syntactic structure except that idioms are not fragmented and semantically empty grammatical words are not considered as fragments (Figure 10).

(17) *Pierre donne du fil à retordre à ses parents*.
lit. Peter gives thread to twist to his parents
'Peter has become a major irritant to his parents'

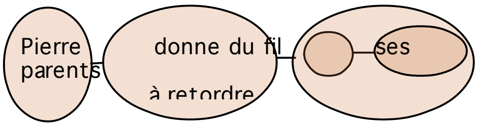


Figure 10. Deep-syntactic connection structure

4. Fragmentations for surface syntax

4.1. Criteria for syntactic fragments

The connection structure we obtain depends completely on the definition of acceptable fragments. We are now interested in the linguistic criteria we need in order to obtain a connection structure corresponding to a usual surface syntactic structure. As a matter of fact, these criteria are more or less the criteria usually proposed for defining constituents. A *surface syntactic fragment* of an utterance U

- is a subpart of U (generally in its original order),[15]

[15] For instance, the NP *a wall 3 meters high* has the subfragment *a high wall* and not **a wall high*.

- is a linguistic sign and its meaning is the same when it is taken in isolation and when it is part of U,[16]
- can stand alone (for example as an answer to a question),[17]
- belongs to a distributional class (and can for instance be replaced by a single word).

Mel'čuk [26] proposes, in his definition of wordforms, to weaken the stand-alone property (or autonomizability). For instance in (8), *the* or *slept* are not autonomizable, but they can be captured by subtraction of two autonomizable fragments: *slept* = *Peter slept* \ *Peter*, *the* = *the dog* \ *dog*.[18] We call such fragments *weakly autonomizable*.[19]

Of course, even if our approach resolves most of the problems arising when trying to directly define constituents, some problems remain. For instance, if you consider the French noun phrase *le petit chien* 'the little dog', the three fragments *le chien*, *petit chien*, and *le petit* 'the little one' are acceptable. Eliminating the last fragment *le petit* necessitates that one assume nontrivial arguments: *le petit,* when it stands alone, is an NP (it commutes with NPs) but it cannot commute with NPs like for example *la fille* 'the girl' in *le petit chien* as **la fille chien* 'the girl dog' is ungrammatical. Many exciting questions posed by other phenomena like coordination or extraction cannot be investigated here for lack of space.

4.2. Granularity of the fragmentation

Syntactic structures can differ in the minimal units. Most authors consider that the wordforms are the basic units of dependency structure, but some authors propose to consider dependencies only between chunks and others between lexemes and grammatical morphemes. The following figure shows representations of various granularities for the same sentence (18).

(18) *A guy has talked to him.*

Tree A is depicting an analysis in chunks (Vergne [36]), Tree B in words, Tree D in lexemes and inflectional morphemes (and can be compared to an X-bar structure with an IP, governed by agreement and tense). The tree C (corresponding to the surface syntactic structure of Mel'čuk [25]) can be understood as an underspecified representation of D.

These various representations can be captured by our methods. The only problem is to impose appropriate criteria to define what we accept as minimal fragments. For instance,

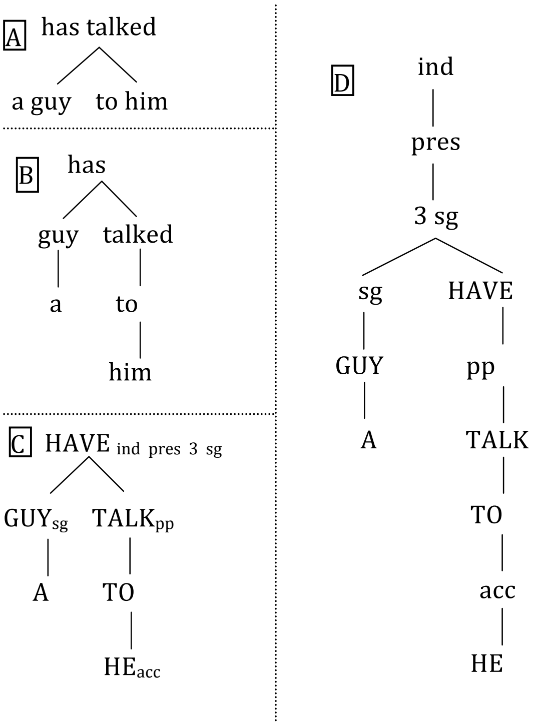
[16] This condition has to be relaxed for the analysis of idiomatic expressions as they are precisely characterized by their semantic non-compositionality. The fragments are in this case the elements that appear autonomizable in the paradigm of parallel non-idiomatic sentences.

[17] Mel'čuk (1988 [25], 2011[27]) proposes a definition of two-word fragments. Rather than the stand alone criterion, he proposes that a fragment must be a prosodic unit. This is a less restrictive criterion, because the possibility to stand alone supposes to be a speech turn and therefore to be a prosodic unit. For instance *little dog* can never be a prosodic unit in *the little dog* but it is a prosodic unit when it stands alone. We think that this criterion is interesting, but not easy to use because the delimitation of prosodic units can be very controversial and seems to be a graded notion. Note also that clitics can form prosodic units which are unacceptable fragments in our sense, like in:
(i) *the king* | ***of England's*** | *grandmother*
(ii) *Je crois* | ***qu'hier*** | *il n'est pas venu*
'I think | that yesterday | he didn't come'

[18] Note that singular bare noun like *dog* are not easily autonomizable in English, but they can for instance appear in titles.

[19] Some complications arise with examples like Fr. *il dormait* 'he slept'. Neither *il* (a clitic whose strong form *lui* must be used in isolation), nor *dormait* are autonomizable. But if we consider the whole distributional class of the element which can commute with *il* in this position, containing for example *Peter,* we can consider *il* to be autonomizable *by generalization over the distributional class*.

trees C and D are obtained if we accept parts of words which commute freely to be "syntactic" fragments (Kahane [22]). Conversely, we obtain tree A if we only accept strongly autonomizable fragments.



Syntactic trees of various granularities for (18)

5. Heads and dependencies

Most syntactic theories suppose that the syntactic structure is hierarchized.[20] This means that connections are directed. A directed connection is called a *dependency*. For a dependency from A to B, A is called the *governor* of B, B, the *dependent* of A, and A, the

[20] The only dependency-based grammar we know that uses non-hierarchized connections (and even cycles) is Link Grammar [32], which has developed one of the most efficient parsers of its time.

head of the fragment AB.[21] The introduction of the term "head" into syntax is commonly attributed to Henry Sweet (1891-96, I:16, Sections 40 and 41 [30]):

> "The most general relation between words in sentences from a logical point of view is that of **adjunct-word** and **head-word**, or, as we may also express it, of **modifier** and **modified.** […] The distinction between adjunct-word and head-word is only a relative one: the same word may be a head-word in one sentence or context, and an adjunct-word in another, and the same word may even be a head-word and an adjunct-word at the same time. Thus in *he is very strong*, *strong* is an adjunct-word to *he,* and at the same time head-word to the adjunct-word *very*, which, again, may itself be a head-word, as in *he is not very strong*."

Criteria for the recognition of the direction of relations between words have been proposed by Bloomfield [4], Zwicky [37], Garde [12], or Mel'čuk [25]. The most common criterion is that the head of a constituent is the word controlling its distribution, which is the word that is the most sensitive to a change in its context. But for any fragment, its distribution does not depend only on its head (and, as we have said in the introduction, constituents cannot easily be defined without using the notion of head). As an example, consider the fragment *little dogs* in (19):

(19) *Very little dogs slept.*

As *little* is connected to *very* and *dogs* to *slept, little dogs* does not have the distribution of *dogs* nor of *little* in 17 as *very dogs slept* and *very little slept* are both unacceptable. Determining the head of the fragment *little dogs* (i.e. the direction of the relation between *little* and *dogs*) is equivalent to the identification of the governor of this fragment (between *very* and *slept*). But, as soon as we have identified the governor of the fragment, the head of the fragment is simply the word of the fragment which is connected to the fragment's governor – the main word outside the fragment. For example, in (19), the identification of *slept* as the governor of the fragment *little dogs* also chooses *dogs* as the head of *little dogs*.

Problems occur only if we are dealing with an irreducible fragment like the determiner-noun connection.[22] To sum up: In order to identify the directedness of the connections and to define a dependency structure for a sentence, it is central to define the head of the whole sentence (and to resolve the case of irreducible fragments if we want a dependency tree). We consider that the head of the sentence is the main finite verb, because it bears most of the illocutionary marks: Interrogation, negation, and mood morphemes are linked to the main finite verb. In English, interrogation changes the verbal form (20), and in French, interrogation (20), negation (20), or mood (20) can be marked by adding clitics or inflectional morphemes on the finite verb even if it is an auxiliary verb.

[21] Dependency relations are sometimes called head-daughter relations in phrase structure frameworks. Note the distinction between *head* and *governor*. For a fragment *f*, the governor of *f* is necessary outside *f*, while the head of *f* is inside *f*. The two notion are linked by the fact that the governor *x* of *f* is the head of the upper fragment composed of the union of *f* and *x*.

[22] Various criteria have been proposed in favor of considering either the noun or the determiner as the head of this connection, in particular in the generative framework (Principles and Parameters, Chomsky (1981 [8]), remains with NP, and, starting with Abney (1986 [1]), DP is preferred). It seems that the question is triggered by the assumption that there has to be one correct directionality of this relation, in other words that the syntactic analysis is a (phrase structure) tree. This overly simple assumption leads to a debate whose theoretical implications do not reach far as any DP analysis has an isomorphic NP analysis. The NP/DP debate was triggered by the observation of a parallelism in the relation between the lexical part of a verb and its inflection (reflected by the opposition between IP and VP in the generative framework). This carries over to dependency syntax: The analysis D of sentence (18) captures the intuition that the inflection steers the passive valency of a verb form.

(20) **a.** *Did very little dogs sleep?*
b. *Pierre a**-t-il** dormi?*
lit. Peter has-he slept?
'Did Peter sleep?'

c. *Pierre **n'a pas** dormi.*
lit. Peter neg has neg slept
'Peter didn't sleep'
d. *Pierre **aurait** dormi.*
lit. Peter have-COND slept?
'Peter would have slept'

Once the head of the sentence has been determined, most of the connections can be directed by a top down strategy. Consequently the main criterion to determine the head of a fragment *f* is to search if one of the words of *f* can form a fragment with the possible governors of *f*, that is, if one of the words of *f* can be connected with the possible governors of *f*. If not, we are confronted with an irreducible fragment, and other criteria must be used, which will be discussed in the next section (see also Mel'čuk [25], [27]).[23] Nevertheless, it is well known that in many cases, the head is difficult to find (Bloomfield [4] called such configurations *exocentric*). It could be advocated not to attempt to direct such connections and thus settle with an only *partially directed connection structure*.[24]

6. Refining the dependency structure

Even when the connection structure is completely directed, the resulting dependency structure is not necessarily a tree due to irreducible fragments and cycles. We can use new principles to refine the dependency structure and to get closer to a dependency tree.

The situation is as follows: C is connected to AB and neither AC nor BC is an acceptable fragment. We thus have the following configuration: [A — B] — C.

A first case can be solved only by structural considerations: it is the case where C governs AB and B governs A, which means [A ← B] ← C. In this case there is only one solution for obtaining a tree. It is not possible to connect C and A, and we necessary have A ← B ← C. This can be illustrated by the sentence *Peter thinks Mary snored*, with A = *Mary*, B = *snored*, and C = *thinks*.

In any other case of A — B being irreducible, C can be connected either to A or to B and two solutions are structurally possible. We need an additional linguistic principle to decide. To train our intuition, let us consider an example: *the most famous of the world* can be analyzed in [[*the most*] ← [*famous*]] → [*of the world*] and neither *famous of the world* nor *the most of the world* are acceptable.[25] But we think, rather, that [*of the world*] is rather

[23] Conversely, whenever the fragmentation tests do not give clear results on whether or not a connection must be established, criteria used to determine the head can be helpful to confirm the validity of the connection.

[24] Equally, the problem of PP attachment in parsing is certainly partially based on true ambiguities, but in many cases, it is an artificial problem of finding a tree structure where the human mind sees multiple connections, like for instance in *He reads a book about syntax* or in the examples (12) to (13). We can assume that a statistical parser will give better results when trained on a corpus that uses the (circular) graph structure, reserving the simple tree structures for the semantically relevant PP attachments.

[25] Most English speakers prefer *the most famous in the world*, which may have another structure because *famous in the world* is an acceptable fragment. French has only one construction, *le plus célèbre du monde*, lit. 'the most famous of the world', for which the analysis discussed here also holds true.

selected by the superlative marker *the most* rather than by the adjective *famous*, because for any adjective *X* we have *the most X of the world*, while *the most* cannot commute with other adjective modifiers (**very famous of the word*).

Generalizing this idea, we propose the following principle.

> **Principle of selection:** If in the configuration [A — B] — C, B commutes less freely than A, then C can be connected directly to B, which gives us the configuration A — B — C.[26]

In our previous example, *famous* commutes more freely than *the most* in *the most famous of the world* because *famous* can commute with every adjective while *the most* cannot commute with most other adjective modifiers. This means that it is *the most* and not *famous* that selects *of the world* and if we refine the connection between *the most famous* and *of the world*, this connection must be attributed to *the most*.

We can give another example: in the sentence *Mary snored*, if we segment *snored* into a verbal lexeme SNORE and an inflection *-ed*, one questions to what element the subject *Mary* must be connected (Figure 11). As SNORE can commute with most verbal lexeme but *-ed* cannot commute with non-finite inflections (non-finite forms of the verb cannot fill the subject slot of the verbal valency), it follows that it is the inflection and not the lexical part of the verb that selects the subject.[27]

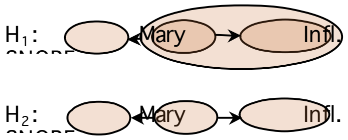


Figure 11. A dependency structure H_1 of and its refinement H_2

Selection must not be confused with subcategorization. Following Tesnière [34], we take the subject to be a verbal actant just like the direct and the indirect object. The connection of the subject with the inflection can be compared with other cases of raising. For instance, in *Peter seems to snore*, *Peter* is clearly the subject of *seems* even if it is subcategorized for by SNORE and not SEEM.

The connection between *parle* and *à Marie* in (10) can also be refined using the principle of selection. As *Marie* can commute with any noun while the preposition *à* cannot commute with other prepositions, the connection is attributed to the preposition (Figure 12).

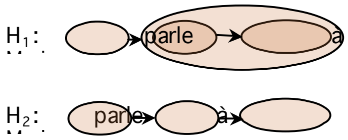


Figure 12. Dependency structure of (10) and its refinement H_2

Sometimes the principle of selection is not discriminating. This is the case for the choice between determiner and noun in *The dog slept*. Indeed *the* as well as *dog* can

[26] The term *selection* is often used in linguistics for selectional restrictions. This is the same idea here: C selects B rather than A because C restricts the distributional paradigm of B more than the one of A.

[27] X-bar syntax makes the same assumption: the subject is a daughter of InflP while other actants are daughters of VP. Note also that, as discussed in Section 5, the inflection is clearly the head of the fragment composed of the verbal lexeme and its inflection.

commute freely: *the/a/each/this… dog/girl/man/idea… slept*. In such a case another principle must be introduced in order to obtain a tree, but we will not discuss this point further.

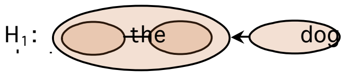


Figure 13. Dependency structure of *The dog slept*

Other principles can be proposed in order to decide which of the two categories determiner and noun is the head of this fragment. As discussed before (Note 16), we are not convinced that such a decision is linguistically relevant. Nevertheless, from a computational point of view it can be motivated to want to manipulate trees rather than hypergraphs and an arbitrary decision can be taken. The most common decision in dependency-based theory (Tesnière [34], Mel'cuk [25], but see Hudson [19] for the reverse choice) is to choose the noun as the head. This choice privileges the semantic dependency between the noun and the governor of the NP (*I bought two books* means 'I bought books, the number of which is two').

We will conclude this section by discussing the consequences of the refinement of the structure. Let us recall that the set of possible fragments can be recovered from the connection structure (Section 2.8). As soon as we refine the structure we increase the number of catenae: [A — B] — C has only two catenae (AB and ABC) while A — B — C has three catenae (AB, BC and ABC). It is possible to label the connections in order to indicate which connection has been refined by the principle of selection and thus does not correspond to a fragment: Each time connections external to a fragment are attributed to internal nodes of this fragment, we label *e* the external connections that has been refined and *i* the internal connections of the fragment. For instance, if the connection between C and the fragment f = AB is refined and attributed to B, we obtain the labeled graph A –*i*– B –*e*– C. The initial hypergraph can be reconstructed by attributing each *e*-connection to the fragments obtained by aggregating all the nodes connected by adjacent *i*-connections (Figure 14). It is sometimes necessary to coindex *e*-links and *i*-links obtained from the refinement. Let us, for instance, consider the configuration [A — B –]– [C — D], where we have two irreducible fragments AB and CD and a connection between B and CD. If we reduce both AB and CD, the connection between B and CD will become an *e*-link, but we must indicate that this *e*-link corresponds to the *i*-link between C and D but not to the *i*-link between A and B. See also the second refinement of Figure 14.

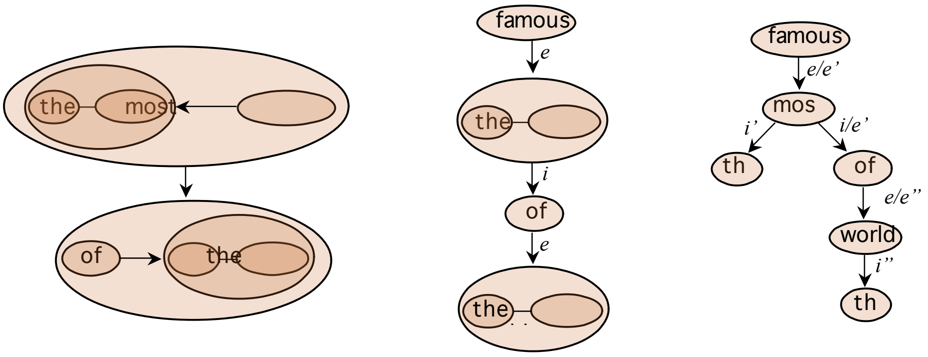


Figure 14. A dependency structure and two successive refinements (with encoding of fragments by labels)

7. Constituency

We saw in Section 2.8 that any fragmentation can be recovered from the connection structure. As soon as the connections are directed, some fragmentations can be favored and constituent structures can be defined.

Let us consider nodes A and B in a dependency structure. A *dominates* B if A = B or if there is a path from A to B starting with a dependency whose governor is A. The fragment of elements dominated by A is called the *maximal projection* of A (following [24] and its definition of projectivity). Maximal projections are major constituents (XPs in X-bar syntax). The maximal projection of A can be fragmented into {A} and the maximal projections of its dependents. This fragmentation gives us a flat constituent structure (with possibly discontinuous constituents). Discontinuous constituents are often postulated in phrase structure grammars and are particularly difficult to characterize because they require a relaxation of the defining criteria. The difficulty is to authorize discontinuous constituents without obtaining an explosion of the units that can possibly acquire the status of constituent. In our approach we accept at the beginning discontinuous fragments and the constituents are selected among the fragments using the hierarchization provided by the heads.

Partial projections of A are obtained by considering only a part of the dependencies governed by A. By defining an order on the dependency of each node (for instance by deciding that the subject is more "external" than the object), we can privilege some partial projections and obtain our favorite binary fragmentation equivalent to the phrase structure trees we prefer. In other words, a phrase structure for a given utterance is just one of the possible fragmentations and this fragmentation can only be identified if the notion of *head* is considered.

We can thus say that phrase structure contains a definition of dependency at its very base, a fact that already appears in Bloomfield's work, who spends much more time on defining head-daughter relations than on the notion of constituency. Jackendoff [20]'s X-bar theory is based on a head-centered definition of constituency, as each XP contains an X being the (direct or indirect) governor of the other elements of XP.

If we accept a mix of criteria for identifying fragments and heads, it is possible to directly define a constituent structure without considering all the fragmentations. The strategy is recursive and top-down (beginning with the whole sentence at first constituent); each step consists of first identifying the head of the constituent we want to analyze and then looking at the biggest fragments of the utterance without its head: These biggest fragments are constituents.[28] Let us exemplify this with sentence (4): *wants* is the head of the sentence and all the biggest (i.e. in a sense of inclusion: fragments not contained in other fragments) remaining fragments are *Peter* and *to read the book*. At each step we have first to take the head off and to go on with the subfragments we obtain, which give us successively *to* and *read the book*, *read* and *the book*, *the* and *book*. The resulting constituent structure is given in Figure 15.[29]

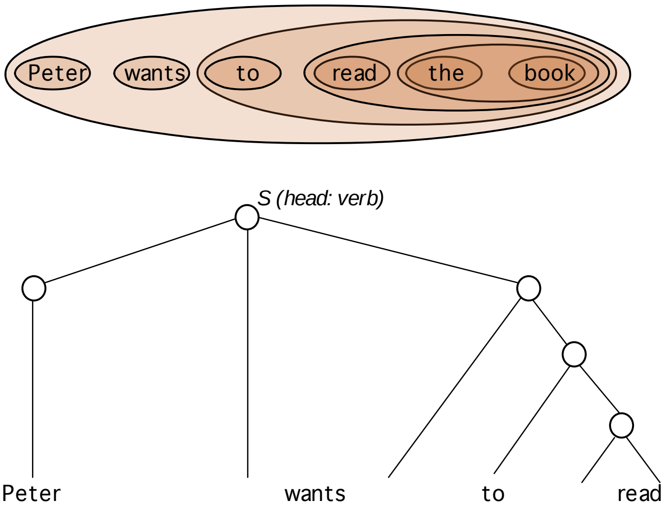


Figure 15. Two equivalent representation of the constituent structure of (6)

Conclusion

We have shown that it is possible to formally define a syntactic structure solely on the basis of fragmentations of an utterance. The definition of fragments does not have to keep the resulting constituent structure in mind, but can be based on simple observable criteria like different forms of autonomizability. Even (and especially) if we obtain intersecting fragmentations, we can obtain a connection graph. This operation can be applied to any type of utterance, yielding connections from the morphological to the discourse level. This delegates the search for the head of a fragment to a secondary optional operation. It is again possible to apply the known criteria for heads only when they provide clear-cut answers,

[28] If the head of the constituent is a finite verb, clefting can be a useful test for characterizing sub-constituents. But clefting can only capture some constituents and only if the head of the constituent has been identified and is a finite verb. As noted by Croft [10] considering the typological point of view, such constructions can only be used to characterize the constituents once we have defined them. We know that constructions like clefting select constituents because we were able to independently define constituents with other techniques. We cannot inversely define constituents by use of such language-specific constructions.

[29] Our constituent tree does not contain a VP. Indeed the maximal projection of the main verb is the whole sentence. The VP can be obtained as a maximal projection only if we separate the verbal lexeme from its inflection (see Figure 11).

leaving us with partially unresolved connections, and thus with a hypergraph, and not necessarily a tree structure. It is possible, and even frequent, that the syntactic structure is a tree, but our definition does not presuppose that it must be one. This two-step definition (connection and directionality) allows for a more coherent definition of dependency as well as constituency avoiding the commonly encountered circularities. It takes *connection* as a primary notion, preliminary to constituency and dependency.

Another interesting feature of our approach is not to presuppose a segmentation of a sentence into words and even not suppose the existence of words as an indispensable notion.

In this paper, we could explore neither the concrete applicability of our approach to other languages nor the interesting interaction of this new definition of dependency with recent advances in the analysis of coordination in a dependency based approach, like the notion of pile put forward in Gerdes & Kahane [15]. It also remains to be shown that the order on hypergraphs is really complete, i.e. that we can actually always compute a greatest connection graph refining any set of fragmentation hypergraphs. We also leave it to further research to explore the inclusion of weights on the connection which could replace the binary choice of presence or absence of a connection.

Acknowledgments

We would like to thank Igor Mel'čuk, Timothy Osborne, Federico Sangati, and our three anonymous reviewers.

References

[1] S. Abney, *The English Noun Phrase in its Sentential Aspect*, Unpublished Ph.D., MIT, 1986.

[2] N. Asher, S. Pogodalla, SDRT and Continuation Semantics, *Logic and Engineering of Natural Language Semantics 7* (LENLS VII), 2010.

[3] K. F. Becker, *Organismus der Sprache*, 2nd edition. Verlag von G.F. Kettembeil, Frankfurt am Main, 1841 [1827].

[4] L. Bloomfield, *Language*. Allen & Unwin, New York, 1933.

[5] R. Bod. *Beyond grammar: an experience-based theory of language*. Stanford, CA: CSLI Publications, 1998.

[6] Th. Brants, W. Skut, and H. Uszkoreit, Syntactic annotation of a German newspaper corpus. *Treebanks*, pp. 73-87. Springer Netherlands, 2003.

[7] A. Carnie, *Modern Syntax: A Coursebook,* Cambridge University Press, 2011.

[8] N. Chomsky, *Lectures On Government and Binding*. Foris, Dordrecht, 1981.

[9] N. Chomsky, *New horizons in the study of language and mind*, Cambridge University Press, 1986.

[10] W. Croft, *Radical construction grammar: syntactic theory in typological perspective* Oxford University Press, 2001

[11] J. Deulofeu, L. Dufort, K. Gerdes, S. Kahane, P. Pietrandrea, Depends on what the French say, *The Fourth Linguistic Annotation Workshop (LAW IV)*, 2010.

[12] P. Garde, Ordre linéaire et dépendance syntaxique : contribution à une typologie, *Bull. Soc. Ling. Paris*, 72:1, 1-26, 1977.

[13] A. V. Gladkij, *Leckii po matematiceskoj linguistike dlja studentov NGU*, Novosibirsk, 1966 (French translation: *Leçons de linguistique mathématique*, fasc. 1, 1970, Paris, Dunod).

[14] H. A. Gleason, *An Introduction to Descriptive Linguistics.* New York: Holt, Rinehart & Winston, 503 p., 1955, Revised edition 1961.

[15] K. Gerdes, S. Kahane, Speaking in piles: Paradigmatic annotation of a French spoken corpus, *Corpus Linguistics 2009*, Liverpool, 2009

[16] O. Jespersen, *Analytic syntax*. Copenhagen, 1937.

[17] L. M. V. Haegeman, *Introduction to Government and Binding Theory*. Blackwell Publishers, 1991.

[18] R. Hudson, Discontinuous phrases in dependency grammars, *UCL Working Papers in Linguistics*, 6, 1994.

[19] R. Hudson, *Language Networks: The new Word Grammar,* Oxford University Press, 2007.

[20] R. Jackendoff, *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press, 1977.

[21] S. Kahane, "Bubble trees and syntactic representations", *Proceedings of MOL5*, Saarbrücken, 70-76, 1997

[22] S. Kahane, “Defining the Deep Syntactic Structure: How the signifying units combine”, *Proceedings of MTT* 2009, Montreal.

[23] S. Kahane, Why to choose dependency rather than constituency for syntax: a formal point of view, in J. Apresjan, M.-C. L’Homme, L. Iomdin, J. Milićević, A. Polguère, L. Wanner, eds., *Meanings, Texts, and other exciting things: A Festschrift to Commemorate the 80th Anniversary of Professor Igor A. Mel’čuk*, Languages of Slavic Culture, Moscow, 257-272.

[24] Y. Lecerf, Programme des conflits, module des conflits, *Bulletin bimestriel de I'ATALA*, 4,5, 1960.

[25] I. Mel’čuk, *Dependency Syntax*: Theory and Practice. The SUNY Press, Albany, N.Y., 1988

[26] I. Mel’čuk *Aspects of the Theory of Morphology*. de Gruyter, Berlin, New York, 2006.

[27] I. Mel’čuk, Dependency in language, *Proceedings of Dependency Linguistics 2011*, Barcelona, 2011.

[28] T. Osborne, M. Putnam, Th. Groß, Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15:4, 354-396, 2012.

[29] K. Schubert, *Metataxis: Contrastive dependency syntax for machine translation.* http://www.mt-archive.info/Schubert-1987.pdf, 1987.

[30] H. Sweet, *A New English Grammar*, 2 vols. Clarendon Press. Oxford, 1891-1896.

[31] W. Quine. 1986. Reply to Gilbert H. Harman, in E. Hahn and P.A. Schilpp, eds., *The Philosophy of W.V. Quine*. La Salle, Open Court.

[32] D. Sleator, D. Temperley. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report*, CMU-CS-91-196,1991.

[33] M. Steedman. 1985. Dependency and coordination in the grammar of Dutch and English, *Language*, 61:3, 525-568.

[34] L. Tesnière, Éléments de syntaxe structurale. Klincksieck, Paris, 1959.

[35] R. D. Van Valin, R. J. LaPolla, *Syntax: Structure, meaning, and function*, Cambridge University Press, 1997.

[36] J. Vergne, A parser without a dictionary as a tool for research into French syntax, in: *Proceedings of the 13th conference on Computational linguistics*-Volume 1, pp. 70-72. Association for Computational Linguistics, 1990.

[37] A. M. Zwicky, Heads, *Journal of Linguistics*, 21: 1-29, 1985.