Defining dependencies (and constituents)

Kim Gerdes
LPP
Sorbonne Nouvelle, Paris
kim@gerdes.fr

Sylvain Kahane
Modyco
University Paris Ouest
sylvain@kahane.fr

Abstract

The paper proposes a mathematical way of defining dependency and constituency as soon as 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 morphematic structures for words. We particularly investigate the linguistic criteria for defining syntactic structures.

Keywords: connection graph, dependency tree, phrase structure.

1 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). However, these structures are rarely clearly defined and often purely intuition-based as 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 syntactic structures should be trees?). The chosen structures lead to important well-known problems in the analysis of various syntactic phenomena like, among others, extraction and coordination.

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."[1] Our axiom is less strong than Tesnière's, because we do not presuppose that the connections are formed between words only.

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 constituent trees.

The coherence of the definition of the considered syntactic structures is essential not only for theoretical reasons but also from an NLP perspective because statistical parsers reproduce annotations and these parsers are naturally limited by any incoherence in the annotation used for training. This incoherence can be due to the insufficiency of the mathematical structures used for the representation. In other words, the measures of quality of parsers are also measures of the adequacy of the syntactic structures used for the annotation with the actual complexity of the syntactic phenomena in the corpus.

[1] "Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins l'esprit aperçoit des **connexions**, dont l'ensemble forme la charpente de la phrase." (Tesnière 1959:11)

If a syntactic theory, like for example X-bar based approaches, considers an overly simple underlying model (eg. ordered constituent trees), this brings about an arsenal of add-ons like "movement" and "traces" to make up for this shortcoming. These add-ons are not reflected in the geometry of the structure, which makes an oxymoron from *parsing in Government and Binding*.

We will start with a critical analysis of some definitions in the field of phrase structure and dependency based approaches (Section 2). Connection structures are defined in Section 3. They are applied to discourse, morphology, and deep syntax in Section 4. The case of surface syntax is explored in Section 5. Dependency structures are defined in Section 6 and constituent structures in Section 7.

2 Previous definitions of dependency

2.1 Defining dependency

Tesnière (1959) 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 constituency stems from Lecerf (1961) and Gladkij (1965) (see also Kahane 1997) 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 this prior definition of constituency. Mel'čuk (1988:113) shows how to define a dependency on a segment consisting of two words: "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."

This definition is slightly circular as the concept of valency presupposes the recognition of a dependency. Moreover, this definition does not show how to decide which word couples have to be considered. In other words, he explains how to direct a connection but he does not define the connection itself. Consider:

(1) *The dog slept*.

It is impossible to test the "passive valency" of *the slept* or *dog slept* as these two word couples are not phrases and therefore have no distribution. Mel'čuk thus should only obtain the dependency *the* ← *dog* and no further analysis.[2]

Garde (1977) does not restrict his definition of dependency to wordforms but consider 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 wordforms. The goal of this article is to formalize and complete Garde's definition attempt.

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 saying 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."[3] This adds up to lexical co-occurrences rather than syntactic dependencies. If, for example, we applied his criteria to sentence (2), the co-occurrence will first yield the dependency relation *it* ← *raining* as the lexical co-occurrence relation *it* ← *keeps* is less strong.

(2) *It keeps raining*.

Hudson (1994) precisely proposes to keep this dependency. For our part, we want to restrict connection and dependency to couples of element which can form an acceptable fragment of text in isolation (which is not the case of *it raining*). We do not disagree that some sort of dependency exists between *it* and *raining*, but we consider this link as a lexical or semantic dependency (Mel'čuk 1988) rather than a surface syntactic one.

2.2 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) 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 Syntax it is said that "Syntactic constructions are constructions in which none of the immediate constituents is a bound form. […] The actor-action construction

[2] This omission in his definition could be attributed to the fact that his first definition was done with examples in Russian, where there is no obligatory determiner.

[3] He gives the following reasons for this restriction:
1. Form determination can be done by two governors and we would not construct a tree structure.
2. Inflection does not exist in all languages, which reduces the universality of an inflection based criterion.

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*; a form like *John* or *very good* could not be used in this way." 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 of the syntactic structure of a sentence, and he never gives precise criteria of what should be considered as a constituent. In Chomsky (1986), quarreling with the behaviorist claims of Quine (1986), he refutes it as equally absurd to consider the fragmentation of *John contemplated the problem* into *John contemplated – the problem* as *John contemp – lated the problem* instead of the "correct" *John – contemplated the problem.* No further justification for this choice is provided.

Today, the definition of 'constituent' seems no longer be a significant subject in contemporary literature in Syntax. Even pedagogical books in this framework tend to skip the definition of constituency, for example Haegeman (1991) who simply states that "the words of the sentence are organized hierarchically into bigger units called phrases." More recently, Carnie (2011:111) requires a constituent to be a subpart of a sentence that "functions as a unit" and notes that two words are part of the same constituent "if one word modifies another". He then notices that this also holds for *quickly scratched* in *A black cat quickly scratched the rather large couch* and pursues "while we know that *quickly* modifies *scratches*, the constituent that contains them is actually *quickly scratches the rather large couch*, not *quickly scratched*." He then refers to other tests to be defined later, among them the "replacement test" by a single word which applies perfectly to *quickly scratched*. Other commonly proposed tests include the "stand-alone test", meaning that the segment can function as an "answer" to a question, the "movement test" including clefting and topicalization, and coordinability, the latter causing the "problems" of coordination of multiple constituents, gapping, and right-node raising.

In phrase structure frameworks, constituents are nothing but a global approach for the extraction of regularities, the only goal being the description of possible orders with few rules. However, it is never actually shown that the proposed phrase structure really is the most efficient way of representing the observed utterances. An exception to this observation is the work of Bod (1998) who carries out statistical tests to compute a minimal phrase structure grammar for a given corpus. The results, however, are not at all the commonly known rewriting rules that one might expect.

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.

2.3 Intersecting analyses

An interesting result of the vagueness of the definitions of constituency is the fact that different scholars invent different criteria that allow to choose among the possible constituent structures. For example, Jespersen's lexically driven criteria select particle verbs as well as idiomatic expressions. For instance, the sentence (3) is analyzed as "S W O" where W is called a "composite verbal expression" (Jespersen 1937:16)

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

Inversely, Van Valin & Lapolla 1997:26) oppose *core* and *periphery* of every sentence and obtain another unconventional segmentation of (4).

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

Their criteria seem to be *ad hoc* choice for the elimination of the unwanted fragments passing common tests.

We consider that the fact that we find multiple decomposition of an utterance is 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.

3 Fragmentation and connection

3.1 Fragments

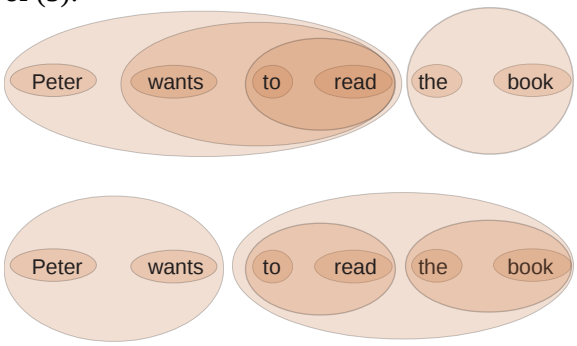
We will relax the notion of syntactic constituent. We call *fragment* of an utterance any of its subparts which is a linguistically acceptable phrase with the same semantic contribution as in the initial utterance. Let us take an example :
(5) *Peter wants to read the book*.

We consider that the acceptable fragments of (5) are: *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 give a justification of that at this point (see section 5). We just say for the moment that *wants to read*, just like *waits on*, 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*). In the same way, *Peter wants* could be a perfect utterance. Probably the most unnatural fragment of (5) is the VP *wants to read the book*, traditionally considered as a major constituent in a phrase structure analysis.

3.2 Fragmentations

A *fragmentation* (*tree*) of an utterance U is a recursive partition of U into acceptable fragments. The following figure shows two fragmentations of (5):



More formally, if X is set of units (for instance the words of (5)), *fragments* are subsets of X and a *fragmentation* F is a subset of the powerset of X ($F \subset P(X)$) such that:

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

A fragmentation whose fragments are constituents is nothing else than a constituency tree.

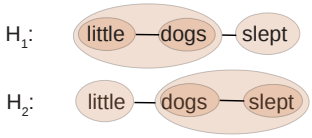
A fragmentation is *binary* if every fragment is partitioned into 0 or 2 fragments.

3.3 Connection structure and fragmentation hypergraph

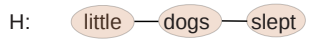
We consider that each segmentation of a fragment in two pieces induces a *connection* between these two pieces. This allows us to define graphs on the fragments of a set X.An *hypergraph* H on X is a triple (X,F,φ) where $F \subset P(X)$ and φ is a graph on F. If F is only composed of singletons, H corresponds 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.[4]

Let us illustrate this with an example:
(6) *Little dogs slept*.
There are two natural fragmentations of (6) whose corresponding hypergraphs are:



As you can see, these two hypergraphs tell us that *little* is connected to *dogs* and *dogs* to *slept*. H_2 also show 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 correspond to the connection between *dogs* and *slept* in H_2. In other words, the two hypergraphs contains 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:



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.

[4] The restriction of the connections to binary partitions can be traced back all the way to Becker (1827:469), quoted after Graffi (2001:137), who claims that "every organic combination within language consists of no more than two members." 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, for example, ternary.

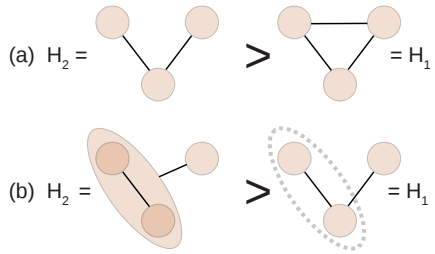
3.4 A complete partial order on hypergraphs

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*[5] of the fragmentation hypergraphs for a natural order of fineness.

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. Formally, the hypergraph $H_1 = (X,F_1,\varphi_1)$ is finer than the hypergraph $H_2 = (X,F_2,\varphi_2)$ (that is $H_1 \leq H_2$):

- if globally $H_2 \subseteq H_1$ (that is $F_2 \subseteq F_1$ and $\varphi_2 \subseteq \varphi_1$).
- but the following exception can occur: If H_2 has moreover a connection between f and g and H_1 has instead a connection between f and g' with $g' \subset g$, g can be absent from H_1.

In other words, H_1 must *a priori* have more fragments and more connections than H_2, but H_1 can have some connections that are "more precise" than corresponding connections of H_2. "More precise" means to point to a smaller fragment, and in this case the bigger fragment can be suppressed (if it carries not other connections). This can be resumed by the following schemata:



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. Overall, we have the order H_2 > "H_1 without the dotted fragment" > "H_1 with the dotted fragment", because any additional fragment refines a fragmentation.

We think that this partial order on hypergraphs is *complete* (see note 5). We have not proven it but it appears to be true on all the configurations we have investigated.

[5] 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 $b \leq x$ for each *x* in A. The *infimum* of A, noted ⋀A, 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: $9 \wedge 12 = 3$).

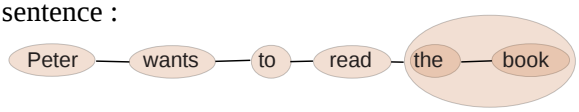
If we have an utterance U and linguistic criteria characterizing the acceptable fragments of U, we define the *connection structure* of U as the infimum of its all fragmentation hypergraphs.

3.5 Constructing the connection structure

Our definition could appear as being a bit complicated. In practice, it is very easy to build the connection graph of a utterance as soon as you have decided what the acceptable fragments of an utterance are. Indeed, because the fineness order on hypergraphs is complete, you can begin with any fragmentation and refine it until you cannot refine it any further. Let us see what happens with example (5). Suppose the first step of your fragmentation is :

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

This means that you have a connection between f_1 and f_2 that will correspond in the final connection structure to a link between two minimal fragments, possibly words. Now, you want to discover these minimal fragments. For that you are looking for the minimal fragment g overlapping both f_1 and f_2: g = *to read*. It is fragmentable 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 :



Connection structure of (5)

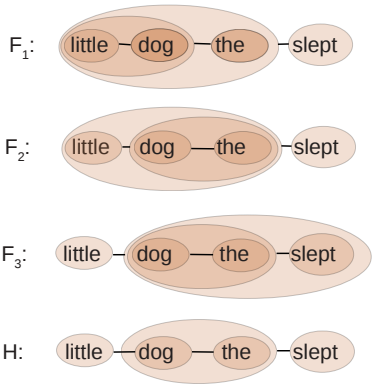
3.6 Irreducible fragment

The connection structure of (5) 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 (5) 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).

(7) *The little dog slept*.

Example (7) 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 (7) has the fragmentation hypergraphs F_1, F_2, and F_3 and the connection graph H (which is their infimum). Note that the fragmentation *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 fragmentation *the dog*.



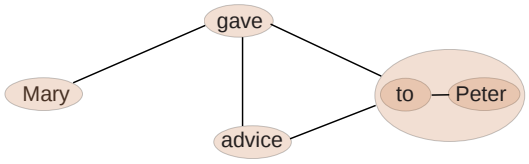
Connection structure of (7): $H = F_1 \wedge F_2 \wedge F_3$

3.7 Cycles

Usually the connection graph is acyclic (and could be transformed into a tree by choosing a node as the root). 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 :

(8) *Mary gave advice to Peter*.
(9) *I saw him yesterday at school*.
(10) *the rise of nationalism in Catalonia*

In (8), *gave advice*, *gave to Peter*, and *advice to Peter* are all acceptable. We encounter a similar configuration in (9) with *saw yesterday*, *saw at school*, and *yesterday at school* (*It was* yesterday at school *that I saw him*). In (10), *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 both connections and does not need to favor one.[6]



Cyclic connection graph for (8)[7]

[6] 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 is more likely than another. We do not pursue this idea here, but we think that *weighted connection graph* are certainly cognitively motivated linguistic representations.

[7] The irreducibility of *to Peter* is conditioned by the given definition of fragments. If we considered relativization as a criteria 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.

3.8 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. Every initial fragment can be obtained by cutting connections in the structure and keeping the segment of the utterance corresponding to continuous piece of the connection structure.

For instance in the connection structure of (5) if you cut the connections between *to* and *read*, you obtain the segment *read the book*. But the segment *read the* cannot be obtained because even if you cut the connection between *the* and *book*, *read* remains connected to the entire group *the book*.

4 Discourse, morphology, semantics

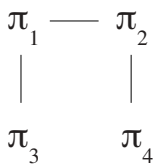
Dependency structures are usually known to describe the syntactic structures of sentences, that is the organization of words into the sentence. We will see in the next sections how to give a precise definition of fragments for surface syntax in order to obtain a linguistically motivated connection structure and to transform it into a dependency tree. Let us see now other application of our methodology to construct connection structures for discourse, morphology or syntax-semantics interface.

4.1 Discourse

Nothing in our definition of connection graphs (and it will be the same for the dependency graphs after) 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 relax these constraints and begin with a whole text, we obtain a discourse connection graph. These strategy can be applied to define discourse relations and discourses structures as RST or SDRT. Of course, to obtain linguistically motivated structures, we need to define what is an acceptable sub-text of a text (generally it means to preserve coherency and cohesion).

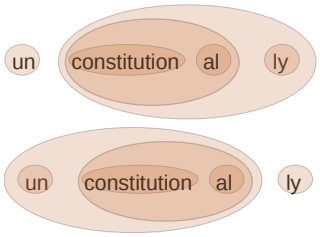
(11) (π_1) *A man walked in.* (π_2) *He sported a hat.* (π_3) *Then a woman walked in.* (π_4) *She wore a coat.*(Asher & Pogodalla 2010)

We have the fragments $\pi_1\pi_2$, $\pi_1\pi_3$, $\pi_3\pi_4$ but we don't have $\pi_2\pi_3$ nor $\pi_1\pi_4$. This gives us the following connection graph:

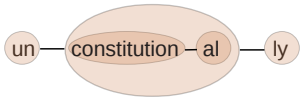


4.2 Morphology

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*". The two possible fragmentations are:



giving us the connection structure:



4.3 Deep Syntax

The *deep syntactic representation* is the central structure of the semantics-syntax interface (Mel'čuk 1988, Kahane 2009). If we take compositionality as a condition for fragmentation, we obtain a structure that resembles Mel'čuk's deep syntactic structure. In other words, idioms must not be fragmented and semantically empty grammatical words are not considered as fragments.

(12) *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'

Interestingly, it is sometimes not very easy to analyze an idiom synchronically. In (11), two surface syntactic structures are possible (*donner à retordre* or *fil à retordre*) and the choice can not be done without making assumptions about the underlying compositional meaning.

5 Fragmentations for surface syntax

5.1 Criteria for syntactic fragments

The connection structure we obtain completely depends 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 subparts of U (in its original order),
- is a linguistic sign and its meaning is the same when it is taken in isolation and when it is part of U,[8]
- can stand alone (for example as an answer of a question),
- belongs to a distributional class (and can for instance be replaced by a single word).

Mel'čuk (2006) proposes, in its definition of wordform to weaken the stand alone property, which he called automizability, to capture some fragments. For instance in (7), *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*.[9] We call such fragments *weakly autonomizable*.[10]

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* supposes to put forward non trivial semantic arguments.

Many exciting questions posed by other phenomena like coordination or extraction cannot be investigated here for lack of space.

5.2 Granularity of the fragmentation

Syntactic structures can differ on the minimal units. Most of the author consider that the wordforms are the basic units of the dependency, but some authors propose to consider dependencies only between chunks and others between lexemes and grammatical morphemes. The following figure shows representations of various granularity for the same sentence (13).

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

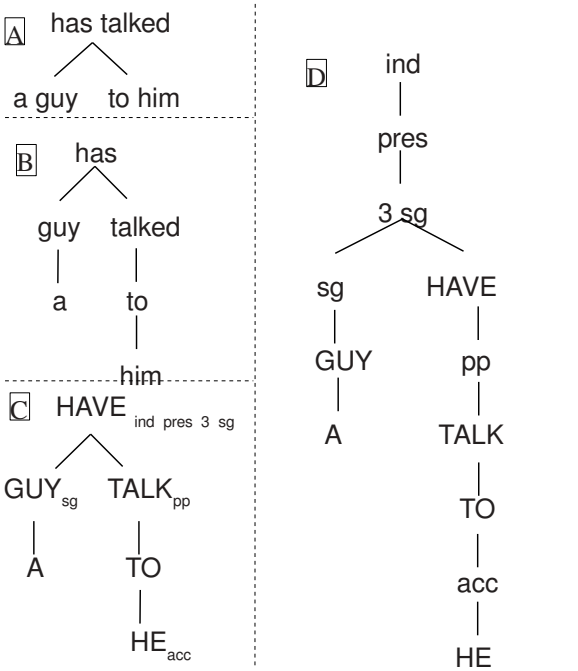
Tree A is depicting an analysis in chunks (Vergne 1990), Tree B in words, Tree D in lexemes and inflectional morphemes (and can be

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

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

[10] Some complications arise with examples like *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*.

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 1988) 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 want as minimal fragments. For instance, trees C and D are obtained if we accept parts of words which commute freely to be "syntactic" fragments (Kahane 2009). Conversely, we obtain tree A if we only accept strongly autonomizable fragments.

6 Heads and dependencies

6.1 Defining head and dependency

A *dependency* is a directed connection. A connection between A and B can be directed by choosing the head of the fragment AB. Criteria have been proposed by Bloomfield (1933), Zwicky (1985), Garde (1977), or Mel'čuk (1988). In short, B is the *head* of AB if most of the distributional property of AB come from B. In other words, A can be more easily erased than B, A can be commuted more easily than B, B is more sensitive than A to a change in the context of AB (Garde 1977). If B is a *head* of AB, B is called the *governor* of A and A the *dependent* of B. The connection between A and B is *directed* from B to A.

A dependency structure can be obtained by choosing a head for each connection. Nevertheless, it is well known that in many cases the head is difficult to find (Bloomfield called such configurations *exocentric*). In such cases, it could be advocated not to attempt to direct the connections and to have an only *partially directed connection structure*.

The most famous of these cases is the determiner-noun connection. 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 framwork (Principles and Parameters, Chomsky (1981), remains with NP, and, starting with Abney (1986), 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 (13) captures the intuition that the inflection steers the distribution (or passive valency) of a verb form.

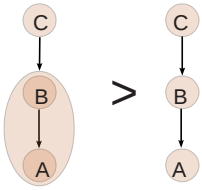
Equally, the problem of PP attachment in parsing is certainly partially based on true ambiguities, but in many cases, it is an artificial problem of finding a tree structure where the human mind sees multiple connections, like for instance in *He reads a book about syntax* or in the examples (8) to (10). 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.

6.2 Refining the dependency structure

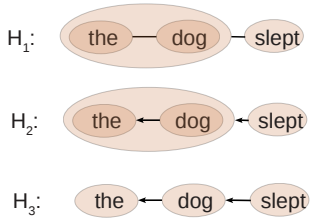
We have seen that, even when the connection structure is completely directed, the resulting dependency structure is not necessary a tree due to irreducible fragments and cycles. We can use two principles to refine the dependency structure and to get closer to a dependency tree. The fineness order on hypergraphs will be prolonged for directed hypergraph in accordance with these principles.

The first principle consists of avoiding double government: if C governs AB and B is the head of AB, then the dependency from C to AB can be replaced by a dependency from C to B (if $[A \leftarrow B] \leftarrow C$, then $A \leftarrow B \leftarrow C$). In other words, the directed hypergraph with the connec-

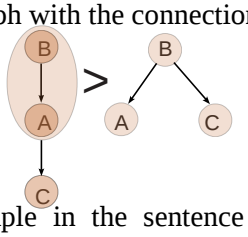
tion B ← C is finer than the hypergraph with the connection [AB] ← C.



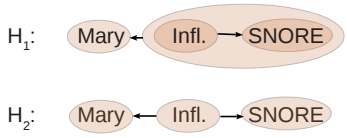
If, for instance, for the sentence (1) *The dog slept*, we obtained the connection graph H_1 below. We can then add directions: The head principle easily gives the link from *slept* to the rest of the sentence, and some additional criteria may direct the connection between *dog* and *the* to give us H_2. We can now carry over this directionality to a complete dependency graph H_3.



Inversely, the second principle consist of avoiding the creation of unacceptable projections: if C depends on AB and B is the head of AB, then the dependency from AB to C can be replaced by a dependency from B to C (if [A ← B] → C, then A ← B → C). In other words, the directed hypergraph with the connection B → C is finer than the hypergraph with the connection [AB] → C.[11]



If, for example in the sentence *Mary snored*, based on the distribution of the sentence depends on the inflection of the verb, we decide to direct the relation between the inflection and the lexical part of the verb *snored* as *inflection → SNORE*, this implies, following Principle 2, that the subject depends on the inflection, and not on the lexical part of the verb. This corresponds to the observation that other, non-finite forms of the verb cannot fill the subject slot of the verbal valency.



7 Constituency

We saw in section 3.8 that any fragmentation can be recovered from the connection structure. As soon as the connections have been 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. Maximal projections are major constituents (we mean XPs in X-bar syntax). The maximal projection of A can be fragmented into {A} and maximal projections of its dependents. This fragmentation gives us a flat constituency structure (with possibly discontinuous constituants).

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 shows in Bloomfield's work, who spends much more time on defining head-daughter relations than on the notion of constituency. Jackendoff's X-bar theory is based on a head-centered definition of constituency, as each XP contains an X being the (direct of indirect) governor of the other elements of XP.

If we accept to mix criteria for identifying fragments and heads, it seems 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); at each step it consists to first find the head of the constituent we want

[11] The two principles could be generalized by only one: if C is connected to AB and B is the head of AB, then the connection between AB and C can be replaced by a connection between B and C (if [A ← B] – C, then A ← B – C). Nevertheless we think that the two principles are different and that the second one is less motivated. For instance, *the most famous of the world* can be analyzed in [[*the most*] ← [*famous*]] → [*of the world*] and neither *famous of the world* or *the most of the world* are acceptable, but we think that [*of the world*] is rather 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*). The problem can be also solved by declaring *the most of the world* acceptable based on previous more general arguments.

to analyze and then to look at the biggest fragments of the utterance without its head: these biggest fragments are constituents.[12]

8 Conclusion

We have shown that it is possible to formally define dependency 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 done on 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, 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 finds *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 (2009).

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.

[12] If the head of the constituent is a finite verb, clefting can be a useful test for characterizing sub-constituents. But clefting can be used only to capture some constituents and only if the head of the constituent has been identified and is a finite verb.

References

Steven Abney. 1986. *The English Noun Phrase in its Sentential Aspect*. Unpublished Ph.D., MIT.

Nicholas Asher, Pogodalla S., 2010. "SDRT and Continuation Semantics" in *Logic and Engineering of Natural Language Semantics 7* (LENLS VII)

Leonard Bloomfield. 1933. *Language*. Allen & Unwin, New York.

Rens Bod. 1998, *Beyond grammar: an experience-based theory of language*. Stanford, CA: CSLI Publications.

Andrew Carnie. 2011. *Modern Syntax: A Coursebook*. Cambridge University Press.

Noam Chomsky. 1981. *Lectures On Government and Binding*. Foris, Dordrecht.

Noam Chomsky. 1986. *New horizons in the study of language and mind*, Cambridge University Press.

Giorgio Graffi. 2001. *200 years of syntax: a critical survey*, John Benjamins Publishing Company.

Kim Gerdes, Kahane S.. 2009. "Speaking in piles: Paradigmatic annotation of a French spoken corpus", *Proceedings of Corpus Linguistics 2009*, Liverpool.

Otto Jespersen. 1937. *Analytic syntax*. Copenhagen.

Liliane M. V. Haegeman. 1991. *Introduction to Government and Binding Theory*. Blackwell Publishers

Richard Hudson. 1994. "Discontinuous phrases in dependency grammars", *UCL Working Papers in Linguistics*, 6.

Sylvain Kahane. 1997. "Bubble trees and syntactic representations", *MOL5*, Saarbrücken, 70-76.

Sylvain Kahane. 2009. "Defining the Deep Syntactic Structure: How the signifying units combine", *MTT* 2009, Montreal.

Igor Mel'čuk. 1988. Dependency Syntax: Theory and Practice. The SUNY Press, Albany, N.Y.

Igor Mel'čuk. 2006. Aspects of the Theory of Morphology. Berlin - New York: de Gruyter.

Klaus Schubert 1987. *Metataxis: Contrastive dependency syntax for machine translation*. http://www.mt-archive.info/Schubert-1987.pdf

Willard 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.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

Arnold M. Zwicky. 1985. "Heads", *Journal of Linguistics*, 21: 1-29.