Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language Technology
P.O. Box 4, FIN-00014 University of Helsinki, Finland

Abstract

Syntactic models should be descriptively adequate and parsable. A syntactic description is autonomous in the sense that it has certain explicit formal properties. Such a description relates to the semantic interpretation of the sentences, and to the surface text. As the formalism is implemented in a broad-coverage syntactic parser, we concentrate on issues that must be resolved by any practical system that uses such models. The correspondence between the structure and linear order is discussed.

1 Introduction

The aim of this paper is to define a dependency grammar framework which is both linguistically motivated and computationally parsable.

A linguistically adequate grammar is the primary target because if we fail to define a descriptive grammar, its application is less useful for any linguistically motivated purposes. In fact, our understanding of the potential benefits of the linguistic means can increase only if our practical solutions stand on an adequate descriptive basis.

Traditionally, grammatical models have been constructed by linguists without any consideration for computational application, and later, by computationally oriented scientists who have first taken a parsable mathematical model and then forced the linguistic description into the model which has usually been too weak to describe what a linguist would desire.

Our approach is somewhere between these two extremes. While we define the grammar strictly in linguistic terms, we simultaneously test it in the parsing framework. What is exceptional here is that the parsing framework is not restricted by an arbitrary mathematical model such as a context-free phrase structure grammar. This leads us to a situation where the parsing problem is extremely hard in the general, theoretical case, but fortunately parsable in practise. Our result shows that, while in general we have an NP-hard parsing problem, there is a specific solution for the given grammar that can be run quickly. Currently, the speed of the parsing system[1] is several hundred words per second.

In short, the grammar should be empirically motivated. We have all the reason to believe that if a linguistic analysis rests on a solid descriptive basis, analysis tools based on the theory would be more useful for practical purposes. We are studying the possibilities of using computational implementation as a developing and testing environment for a grammatical formalism. We refer to the computational implementation of a grammar as a *parsing grammar*.

1.1 Adequacy

A primary requirement for a parsing grammar is that it is descriptively adequate. Extreme distortion results if the mathematical properties of the chosen model restrict the data. However, this concern is not often voiced in the discussion. For example, McCawley (1982, p. 92) notes that such a basic assumption concerning linguistic structures that "*strings are more basic than trees and that trees are available only as a side product of derivations that operate in terms of strings*" was attributable to *the historical accident that early transformational grammarians knew some automata theory but no graph theory.*"

One reason for computationally oriented syntacticians to favour restricted formalisms is that they are easier to implement. Those who began

[1] Demo: *http://www.conexor.fi/analysers.html*

to use dependency models in the 1960's largely ignored descriptive adequacy in order to develop models which were mathematically simple and, as a consequence, for which effective parsing algorithms could be presented. These inadequacies had to be remedied from the beginning, which resulted in ad hoc theories or engineering solutions[2] without any motivation in the theory.

There have been some serious efforts to resolve these problems. Hudson (1989), for example, has attempted to construct a parser that would reflecs the claims of the theory (Word Grammar) as closely as possible. However, it seems that even linguistically ambitious dependency theories, such as Hudson's Word Grammar, contain some assumptions which are attributable to certain mathematical properties of an established formalism rather than imposed by the linguistic data[3]. These kinds of unwarranted assumptions tend to focus the discussion on phenomena which are rather marginal, if a complete description of a language is concerned. No wonder that comprehensive descriptions, such as Quirk et al. (1985), have usually been non-formal.

1.2 The European structuralist tradition

We argue for a syntactic description that is based on dependency rather than constituency, and we fully agree with Hajičová (1993, p. 1) that *"making use of the presystemic insights of classical European linguistics, it is then possible that constituents may be dispensed with as basic elements of (the characterization of) the sentence structure."* However, we disagree with the notion of "presystemic" if it is used to imply that earlier work is obsolete. From a descriptive point of view, it is crucial to look at the data that was covered by earlier non-formal grammarians.

As far as syntactic theory is concerned, there is no need to reinvent the wheel. Our description has its basis in the so-called "classical model" based on the work of the French linguist Lucien Tesnière. His structural model should be capable of describing any occurring natural language. His main work, (1959) addresses a large amount of material from typologically different languages. It is indicative of Tesnière's empirical orientation that there are examples from some 60 languages, though his method was not empirical in the sense that he would have used external data inductively. As Heringer (1996) points out, Tesnière used data merely as an expository device. However, in order to achieve formal rigour he developed a model of syntactic description, which obviously stems from the non-formal tradition developed since antiquity but without compromising the descriptive needs. We give a brief historical overview of the formal properties inherent in Tesnière's theory in Section 5 before we proceed to the implementational issues in Section 6.

1.3 The surface syntactic approach

We aim at a theoretical framework where we have a dependency theory that is both descriptively adequate and formally explicit. The latter is required by the broad-coverage parsing grammar for English that we have implemented. We maintain the parallelism between the syntactic structure and the semantic structure in our design of the syntactic description: when a choice between alternative syntactic constructions in a specific context should be made, the semantically motivated alternative is selected[4].

Although semantics determines what kind of structure a certain sentence should have, from the practical point of view, we have a completely different problem: how to resolve the syntactic structure in a given context. Sometimes, the latter problem leads us back to redefine the syntactic structure so that it can be detected in the sentence[5]. Note, however, that this redefinition is now made on a linguistic basis. In order to achieve parsability, the surface descrip-

[2] See discussion of an earlier engineering art in applying a dependency grammar in Kettunen (1994).

[3] For instance, the notion of adjacency was redefined in WG, but was still unsuitable for "free" word order languages.

[4] In such sentence as *"I asked John to go home"*, the noun before the infinitive clause is analysed as the (semantic) subject of the infinitive rather than as a complement of the governing verb.

[5] For instance, detecting the distinct roles of the *to*-infinitive clause in the functional roles of the purpose or reason is usually difficult (e.g. Quirk et al. (1985, p. 564): *"Why did he do it?*; purpose: *"To relieve his anger"* and reason: *"Because he was angry"*). In such sentence as *"A man came to the party to have a good time"*, the interpretation of the infinitive clause depends on the interaction of the contextual and lexical semantics rather than a structural distinction.

tion should not contain elements which can not be selected by using contextual information. It is important that the redefinition should not be made because an arbitrary mathematical model denies e.g. crossing dependencies between the syntactic elements.

2 Constituency vs. dependency

A central idea in American structuralism was to develop rigorous mechanical procedures, i.e. "discovery procedures", which were assumed to decrease the grammarians' own, subjective assessment in the induction of the grammars. This practice was culminated in Harris (1960, p. 5), who claimed that *"the main research of descriptive linguistics, and the only relation which will be accepted as relevant in the present survey, is the distribution or arrangement within the flow of speech of some parts or features relative to others."*

The crucial descriptive problem for a distributional grammar (i.e. phrase-structure grammar) is the existence of non-contiguous elements. The descriptive praxis of some earlier IC theoricians allows discontiguous constituents. For example, already Wells (1947) discussed the problem at length and defined a restriction for discontiguous constituents[6]. Wells' restriction implies that a discontiguous sequence can be a constituent only if it appears as a contiguous sequence in another context. This means that Wells' characterisation of a constituent defines an element which is broadly equivalent to the notion of bunch in Tesnière's (1959) theory. Consequently, these two types of grammars are capable of describing the equivalent syntactic phenomena and share the assumption that a syntactic structure is compatible with its semantic interpretation. However, the extended constituent grammar thus no longer provides a rigorous distributional basis for a description, and its formal properties are unknown.

We can conclude our argument by stating that the reason to reject constitutional grammars is that the formal properties for descriptively adequate constitutional grammars are not known. In the remaining sections, we show that a descriptively adequate dependency model can be constructed so that it is formally explicit and parsable.

3 Parallelism between the syntactic and semantic structures

Obviously, distributional descriptions that do not contribute to their semantic analysis can be given to linguistic strings. Nevertheless, the minimal descriptive requirement should be that a syntactic description is compatible with the semantic structure. The question which arises is that if the correspondence between syntactic and semantic structures exists, why should these linguistic levels be separated. For example, Sgall (1992, p. 278) has questioned the necessity of the syntactic level altogether. His main argument for dispensing with the whole surface syntactic level is that there are no strictly synonymous syntactic constructions, and he therefore suggests that the surface word order belongs more properly to the level of morphemics. This issue is rather complicated. We agree that surface word order does not belong to syntactic structure, but for different reasons.

In contradistinction to Sgall's claim, Mel'čuk (1987, p. 33) has provided some evidence where the morphological marker appears either in the head or the dependent element in different languages, as in the Russian *"kniga professor+a"* (professor's book) and its Hungarian equivalent *"professzor könyv+e"*. Consequently, Mel'čuk (1987, p. 108) distinguishes the morphological dependency as a distinct type of dependency. Thus morphology does not determine the syntactic dependency, as Tesnière (1959, Ch. 15) also argues.

For Tesnière (1959, Ch. 20:17) meaning (Fr. *sens*) and structure are, in principle, independent. This is backed by the intuition that one recognises the existence of the linguistic structures which are semantically absurd, as illustrated by the structural similarity between the nonsensical sentence *"Le silence vertebral indispose la voie licite"* and the meaningful sentence *"Le signal vert indique la voie libre"*.

The independence of syntactic and semantic levels is crucial for understanding Tesnière's thesis that the syntactic structure follows from

[6] Wells (1947): *"A discontinuous sequence is a constituent if in some environment the corresponding continuous sequence occurs as a constituent in a construction semantically harmonious with the constructions in which the given discontinuous sequence occurs."* Further, Wells notes that "The phrase *semantically harmonious* is left undefined, and will merely be elucidated by examples."

the semantic structure, but not vice versa. This means that whenever there is a syntactic relation, there is a semantic relation (e.g. complementation or determination) going in the opposite direction. In this view, the syntactic head requires semantic complementation from its dependents. Only because the syntactic and semantic structures belong to different levels is there no interdependency or mutual dependency, though the issue is sometimes raised in the literature.

There is no full correspondence between the syntactic and semantic structures because some semantic relations are not marked in the functional structure. In Tesnière (1959, p. 85), for example, there are anaphoric relations, semantic relations without correspondent syntactic relations.

4 Surface representation and syntactic structure

4.1 The nucleus as a syntactic primitive

The dependency syntactic models are inherently more "word oriented" than constituent-structure models, which use abstract phrase categories. The notion of word, understood as an orthographic unit in languages similar to English, is not the correct choice as a syntactic primitive. However, many dependency theories assume that the orthographic words directly correspond[7] to syntactic primitives (nodes in the trees). Although the correspondence could be very close in languages like English, there are languages where the word-like units are much longer (i.e. incorporating languages).

Tesnière observed that because the syntactic connexion implies a parallel semantic connexion, each node has to contain a syntactic and a semantic centre. The node element, or *nucleus*, is the genuine syntactic primitive. There is no one-to-one correspondence between nuclei and orthographic words, but the nucleus consists of one or more, possibly discontiguous, words or parts of words. The segmentation belongs to the linearisation, which obeys language-specific rules. Tesnière (1959, Ch 23:17) argued that the notion *word*, a linear unit in a speech-chain, does not belong to syntactic description at all. A *word* is nothing but a *segment* in the speech chain (1959, Ch 10:3).

[7] See Kunze (1975, p. 491) and Hudson (1991).

The basic element in syntactic description is the nucleus. It corresponds to a node in a dependency tree. When the sentence is represented as a dependency tree, the main node contains the whole verb chain.

There are at least two reasons why the concept of the nucleus is needed. In the first place, there are no cross-linguistically valid criteria to determine the head in, say, a prepositional phrase. One may decide, arbitrarily, that either the preposition or the noun is the head of the construction. Second, because the nucleus is also the basic semantic unit, it is the minimal unit in a lexicographical description.

4.2 Linearisation

Tesnière makes a distinction between the *linear order*, which is a one-dimensional property of the physical manifestations of the language, and the *structural order*, which is two-dimensional. According to his conception, constructing the structural description is converting the linear order into the structural order. Restricting himself to syntactic description, Tesnière does not formalise this conversion though he gives two main principles: (1) usually dependents either immediately follow or precede their heads (projectivity) and when they do not, (2) additional devices such as morphological agreement can indicate the connexion.

Although Tesnière's distinction between the linear and structural order corresponds to some extent with the distinction between the linear precedence (LP) and the immediate dominance, there is a crucial difference in emphasis with respect to those modern syntactic theories, such as GPSG, that have distinct ID and LP components. Tesnière excludes word order phenomena from his structural syntax and therefore does not formalise the LP component at all. Tesnière's solution is adequate, considering that in many languages word order is considerably free. This kind of "free" word order means that the alternations in the word order do not necessarily change the meaning of the sentence, and therefore the structural description implies several linear sequences of the words. This does not mean that there are no restrictions in the linear word order but these restrictions do not emerge in the structural analysis.

In fact, Tesnière assumes that a restriction that is later formalised as an *adjacency princi-*

ple characterizes the neutral word order when he says that there are no syntactic reasons for violating adjacency in any language, but the principle can be violated, as he says, for *stylistic reasons* or to save the metric structure in poetics. If we replace the stylistic reasons with the more broader notion which comprises the discourse functions, his analysis seems quite consistent with our view. Rather than seeing that there are syntactic restrictions concerning word order, one should think that some languages due to their rich morphology have more freedom in using word order to express different discourse functions. Thus, linearisation rules are not formal restrictions, but language-specific and functional.

There is no need for constituents. Tesnière's theory has two mechanisms to refer to linearisation. First, there are static functional categories with dynamic potential to change the initial category. Thus, it is plausible to separately define the combinatorial and linearisation properties of each category. Second, the categories are hierarchical so that, for instance, a verb in a sentence governs a noun, an adverb or an adjective. The lexical properties, inherent to each lexical element, determine what the governing elements are and what elements are governed.

There are no simple rules or principles for linearisation. Consider, for example, the treatment of adjectives in English. The basic rule is that attributive adjectives precede their heads. However, there are notable exceptions, including the postmodified adjectives[8], which follow their heads, and some lexical exceptions[9], which usually or always are postmodifying.

5 Historical formulations

In this section, the early formalisations of the dependency grammar and their relation to Tesnière's theory are discussed. The dependency notion was a target of extensive formal studies already in the first half of the 1960's[10].

[8]Example: *"It is a phenomenon consistent with ..."*

[9]Example: *"president elect"*

[10]A considerable number of the earlier studies were listed by Marcus (1967, p. 263), who also claimed that *"Tesnière was one of the first who used (dependency) graphs in syntax. His ideas were repeated, developed and precised by Y. Lecerf & P. Ihm (1960), L. Hirschberg and I. Lynch, particularly by studying syntactic projectivity and linguistic subtrees."*

5.1 Gaifman's formulation

The classical studies of the formal properties of dependency grammar are Gaifman (1965) and Hays (1964)[11], which demonstrate that dependency grammar of the given type is weakly equivalent to the class of context-free phrase-structure grammars. The formalisation of dependency grammars is given in Gaifman (1965, p. 305): For each category X, there will be a finite number of rules of the type $X(Y_1, Y_2 \cdots Y_l * Y_{l+1} \cdots Y_n)$, which means that $Y_1 \cdots Y_n$ can depend on X in this given order, where X is to occupy the position of $*$.

Hays, referring to Gaifman's formulation above, too strongly claims that *"[d]ependency theory is weakly omnipotent to IC theory. The proof is due to Gaifman, and is too lengthy to present here. The consequence of Gaifman's theorem is that the class of sets of utterances [...] is Chomsky's class of context-free languages."* This claim was later taken as granted to apply to any dependency grammar, and the first, often cited, attestation of this apparently false claim appeared in Robinson (1970). She presented four axioms of the theory and claimed they were advocated by Tesnière and formalised by Hays and Gaifman.

Thus, the over-all result of the Gaifman-Hays proof was that there is a weak equivalence of dependency theory and context-free phrase-structure grammars. This weak equivalence means only that both grammars charac-

[11]Tesnière is not mentioned in these papers. Gaifman's paper describes the results "... obtained while the author was a consultant for the RAND Corporation in the summer of 1960." Whereas phrase-structure systems were defined by referring to Chomsky's Syntactic Structures, the corresponding definition for the dependency systems reads as follows: "By dependency system we mean a system, containing a finite number of rules, by which dependency analysis for certain language is done, as described in certain RAND publications (Hays, February 1960; Hays and Ziehe, April 1960)." Speaking of the dependency theory, Hays (1960) refers to the Soviet work on machine translation using the dependency theory of Kulagina et al. In Hays (1964), the only linguistic reference is to the 1961 edition of Hjelmslev's Prolegomena: *"Some of Hjelmslev's empirical principles are closely related to the insight behind dependency theory, but empirical dependency in his sense cannot be identified with abstract dependency in the sense of the present paper, since he explicitly differentiates dependencies from other kinds of relations, whereas the present theory intends to be complete, i.e. to account for all relations among units of utterances."*

terize the same sets of strings. Unfortunately, this formulation had little to do with Tesnière's dependency theory, but as this result met the requirements of a characterisation theory, interest in the formal properties of dependency grammar diminished considerably.

5.2 Linguistic hypotheses

Tesnière's Hypothesis, as Marcus (1967) calls it, assumes that each element has exactly one head. Marcus also formulates a stronger hypothesis, the *Projectivity hypothesis*, which connects the linear order of the elements of a sentence to the structural order of the sentence. The hypothesis is applied in the following formulation: let $x = a_1a_2 \ldots a_i \ldots a_n$ be a sentence, where a_i and a_j are terms in the sentence. If the term a_i is subordinate to the term a_j, and there is an index k which holds $min(i,j) < k < max(i,j)$, then the term a_k is subordinate to the term a_j.

This is the formal definition of projectivity, also known as *adjacency* or *planarity*. The intuitive content of adjacency is that modifiers are placed adjacent to their heads. The intuitive content behind this comes from Behaghel's First Law[12] (Siewierska, 1988, p. 143).

The adjacency principle is applicable only if the linear order of strings is concerned. However, the target of Tesnière's syntax is structural description and, in fact, Tesnière discusses linear order, a property attributable to strings, only to exclude linearisation from his conception of syntax. This means that a formalisation which characterises sets of strings can not even be a partial formalisation of Tesnière's theory because his syntax is not concerned with strings, but structures. Recently, Neuhaus and Bröker (1997) have studied some formal properties of dependency grammar, observing that Gaifman's conception is not compatible either with Tesnière's original formulation or with the "current" variants of DG.

There are several equivalent formalisations for this intuition. In effect they say that in a syntactic tree, where words are printed in linear order, the arcs between the words must not cross. For example, in our work, as the arc between the node "what" and the node "do" in

[12] *"The most important law is that what belongs together mentally (semantically) is placed close together syntactically."*

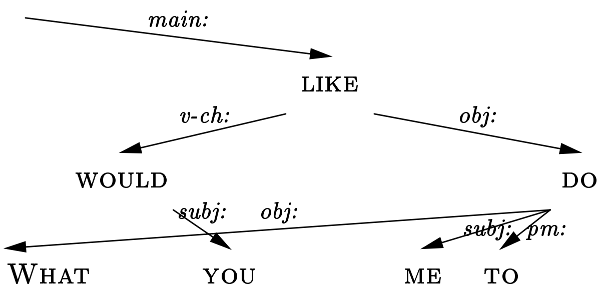

Figure 1: Non-projective dependency tree

Figure 1 violates the principle, the construction is non-projective.

5.3 Formal properties of a Tesnière-type DG

Our current work argues for a dependency grammar that is conformant with the original formulation in Tesnière (1959) and contains the following axioms:

- The primitive element of syntactic description is a nucleus.
- Syntactic structure consists of connexions between nuclei.
- Connexion (Tesnière, 1959, Ch. 1:11) is a binary functional relation between a superior term (regent) and inferior term (dependent).
- Each nucleus is a node in the syntactic tree and it has exactly one regent (Tesnière, 1959, Ch. 3:1).
- A regent, which has zero or more dependents, represents the whole subtree.
- The uppermost regent is the central node of the sentence.

These axioms define a structure graph which is acyclic and directed, i.e. the result is a tree. These strong empirical claims restrict the theory. For example, multiple dependencies and all kinds of cyclic dependencies, including mutual dependency, are excluded. In addition, there can be no isolated nodes.

However, it is not required that the structure be projective, a property usually required in many formalised dependency theories that do not take into account the empirical fact that non-projective constructions occur in natural languages.

6 The Functional Dependency Grammar

Our parsing system, called the Functional Dependency Grammar (FDG), contains the following parts:

- the lexicon,
- the CG-2 morphological disambiguation (Voutilainen, 1995; Tapanainen, 1996), and
- the Functional Dependency Grammar (Tapanainen and Järvinen, 1997; Järvinen and Tapanainen, 1997).

6.1 On the formalism and output

It has been necessary to develop an expressive formalism to represent the linguistic rules that build up the dependency structure. The descriptive formalism developed by Tapanainen can be used to write effective recognition grammars and has been used to write a comprehensive parsing grammar of English.

When doing fully automatic parsing it is necessary to address word-order phenomena. Therefore, it is necessary that the grammar formalism be capable of referring simultaneously both to syntactic order and linear order. Obviously, this feature is an extension of Tesnière's theory, which does not formalise linearisation. Our solution, to preserve the linear order while presenting the structural order requires that functional information is no longer coded to the canonical order of the dependents[13].

In the FDG output, the functional information is represented explicitly using arcs with labels of syntactic functions. Currently, some 30 syntactic functions are applied.

To obtain a closer correspondence with the semantic structure, the *nucleus format* corresponding to Tesnière's stemmas is applied. It

[13]Compare this solution with the Prague approach, which uses horizontal ordering as a formal device to express the topic-focus articulation at their tectogrammatical level. The mapping from the tectogrammatical level to the linear order requires separate rules, called *shallow rules* (Petkevič, 1987). Before such a description exists, one can not make predictions concerning the complexity of the grammar.

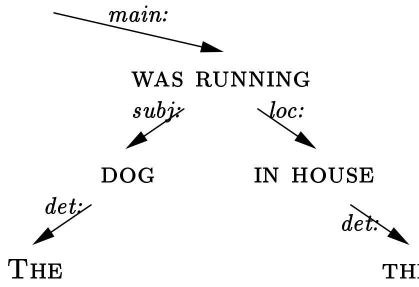

Figure 2: *"The dog was running in the house"*

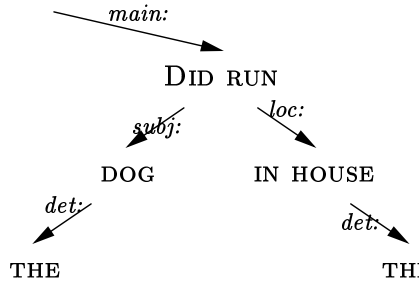

Figure 3: *"Did the dog run in the house"*

is useful for many practical purposes. Consider, for example, collecting arguments for a given verb "RUN". Having the analysis such as those illustrated in Figure 2, it is easy to excerpt all sentences where the governing node is verbal having a main element that has "run" as the base form, e.g. *ran*, *"was running"* (Figure 2), *"did run"* (Figure 3). The contraction form *"won't run"* obtains the same analysis (the same tree although the word nuclei can contain extra information which makes the distinction) as a contraction of the words *"will not run"*. As the example shows, orthographic words were segmented whenever required by the syntactic analysis.

This solution did not exist prior the FDG and generally is not possible in a monostratal dependency description, which takes the (orthographic) words as primitives. The problem is that the non-contiguous elements in a verb-chain are assigned into a single node while the subject in between belongs to its own node.

For historical reasons, the representation contains a lexico-functional level closely similar to the syntactic analysis of the earlier English Constraint Grammar (ENGCG) (Karlsson et al. , 1995) parsing system. The current FDG formalism overcomes several shortcomings[14] of the earlier approaches: (1) the FDG does not rely on the detection of clause boundaries, (2) parsing is no longer sequential, (3) ambiguity is represented at the clause level rather than word level, (4) due to explicit representation of dependency structure, there is no need to refer to phrase-like units. Because the FDG rule formalism is more expressive, linguistic generalisation can be formalised in a more transparent way, which makes the rules more readable.

7 Descriptive solutions

7.1 Coordination

We now tackle the problem of how coordination can be represented in the framework of dependency model. For example, Hudson (1991) has argued that coordination is a phenomenon that requires resorting to a phrase-structure model.

Coordination should not be seen as a directed functional relation, but instead as a special connexion between two functionally equal elements. The coordination connexions are called junctions in Tesnière (1959, Chs. 134-150). Tesnière considered junctions primarily as a mechanism to pack multiple sentences economically into one. Unfortunately, his solution, which represents all coordinative connexions in stemmas, is not adequate, because due to cyclic arcs the result is no longer a tree.

Our solution is to pay due respect to the formal properties of the dependency model, which requires that each element should have one and only one head.[15] This means that coordinated elements are chained (Figure 4) using a specific arc for coordination (labeled as *cc*). The coordinators are mostly redundant markers (Tesnière, 1959, Ch. 39:5)[16], especially, they do not have any (governing) role in the syntactic structure as they do in many word-based forms of dependency theory (e.g. Kunze (1975) and Mel'čuk (1987)).

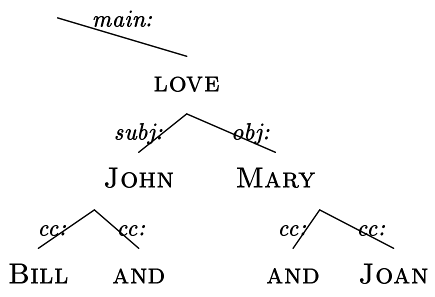

Figure 4: Coordinated elements

Unlike the other arcs in the tree, the arc marking coordination does not imply a dependency relation but rather a functional equivalence. If we assume that the coordinated elements have exactly the same syntactic functions, the information available is similar to that provided in Tesnière's representation. If needed, we can simply print all the possible combinations of the coordinated elements: *"Bill loves Mary"*, *"John loves Mary"*, etc.

7.2 Gapping

It is claimed that gapping is even a more serious problem for dependency theories, a phenomenon which requires the presence of non-terminal nodes. The treatment of gapping, where the main verb of a clause is missing, follows from the treatment of simple coordination.

In simple coordination, the coordinator has an auxiliary role without any specific function in the syntactic tree. In gapping, only the coordinator is present while the verb is missing. One can think that as the coordinator represents all missing elements in the clause, it inherits all properties of the missing (verbal) elements (Figure 6). This solution is also computationally effective because we do not need to postulate empty nodes in the actual parsing system.

From a descriptive point of view there is no problem if we think that the coordinator obtains syntactic properties from the nucleus that

[14] Listed in Voutilainen (1994).

[15] The treatment of coordination and gapping in Kahane (1997) resembles ours in simple cases. However, this model maintains projectivity, and consequently, both multiple heads and extended nuclei, which are essentially phrase-level units, are used in complex cases, making the model broadly similar to Hudson (1991).

[16] The redundancy is shown in the existence of asyndetic coordination. As syntactic markers, coordinators are not completely void of semantic content, which is demonstrated by the existence of contrasting set of coordinators; 'and', 'or', 'but' etc.

```
<John>
    "John" N SG @SUBJ subj:>2
<gave>
    "give" V PAST @+FV #2 main:>0
<the>
    "the" DET ART SG/PL @DN> det:>4
<lecture>
    "lecture" N SG @OBJ #4 obj:>2
<on>
    "on" PREP @ADVL #5 tmp:>2
<Tuesday>
    "Tuesday" N SG @<P pcomp:>5
<and>
    "and" CC @CC #7 cc:>2
<Bill>
    "Bill" N SG @SUBJ subj:>7
<on>
    "on" PREP @ADVL #9 tmp:>7
<Wednesday>
    "Wednesday" N SG @<P pcomp:>9
<.>
```

Figure 5: Text-based representation

it is connected to. Thus, in a sentence with verbal ellipsis, e.g. in the sentence *"Jack painted the kitchen white and the living room blue"*, the coordinator obtains the subcategorisation properties of a verb. A corresponding graph is seen in Figure 6.

Due to 'flatness' of dependency model, there is no problem to describe gapping where a subject rather than complements are involved, as the Figure 5 shows. Note that gapping provides clear evidence that the syntactic element is a *nucleus* rather than a word. For example, in the sentence *"Jack has been lazy and Jill angry"*, the elliptic element is the verbal nucleus *has been.*

8 Conclusion

This paper argues for a descriptively adequate syntactic theory that is based on dependency rather than constituency. Tesnière's theory seems to provide a useful descriptive framework for syntactic phenomena occurring in various natural languages. We apply the theory and develop the representation to meet the requirements of computerised parsing description. Simultaneously, we explicate the formal properties of Tesnière's theory that are used in constructing a practical parsing system.

A solution to the main obstacle to the utilisation of the theory, the linearisation of the syntactic structure, is presented. As a case study, we reformulate the theory for the description of coordination and gapping, which are difficult problems for any comprehensive syntactic theory.

Acknowledgments

We thank Fred Karlsson, Atro Voutilainen and three Coling-ACL '98 workshop referees for useful comments on earlier draft of this paper.

References

Haim Gaifman. 1965. Dependency systems and phrase-structure systems. *Information and Control*, 8:304–337.

Eva Hajičová. 1993. *Issues of Sentence Structure and Discourse Patterns*, volume 2 of *Theoretical and Computational Linguistics.* Institute of Theoretical and Computational Linguistics, Charles University, Prague.

Zellig S. Harris. 1960. *Structural Linguistics.* Phoenix Books. The University of Chicago Press, Chicago & London, first Phoenix edition. Formerly entitled: Methods in Structural Linguistics, 1951.

David G. Hays. 1960. Grouping and dependency theories. Technical Report RM-2646, The RAND Corporation, September.

David G. Hays. 1964. Dependency theory: A formalism and some observations. *Language*, 40:511–525.

Hans Jürgen Heringer. 1996. Empirie und Intuition bei Tesnière. In Gertrud Greciano and Helmut Schumacher, editors, *Lucien Tesnière – Syntaxe structurale et operations mentales*, volume 330 of *Linguistische Arbeiten*, pages 15–31. Niemeyer.

Richard Hudson. 1989. Towards a computer-testable word grammar of English. *UCL working papers in Linguistics*, 1:321–338.

Richard Hudson. 1991. *English Word Grammar.* Basil Blackwell, Cambridge, MA.

Timo Järvinen and Pasi Tapanainen. 1997. A dependency parser for English. Technical Report TR-1, Department of General Linguistics, University of Helsinki, Finland, March.

Sylvain Kahane. 1997. Bubble trees and syn-

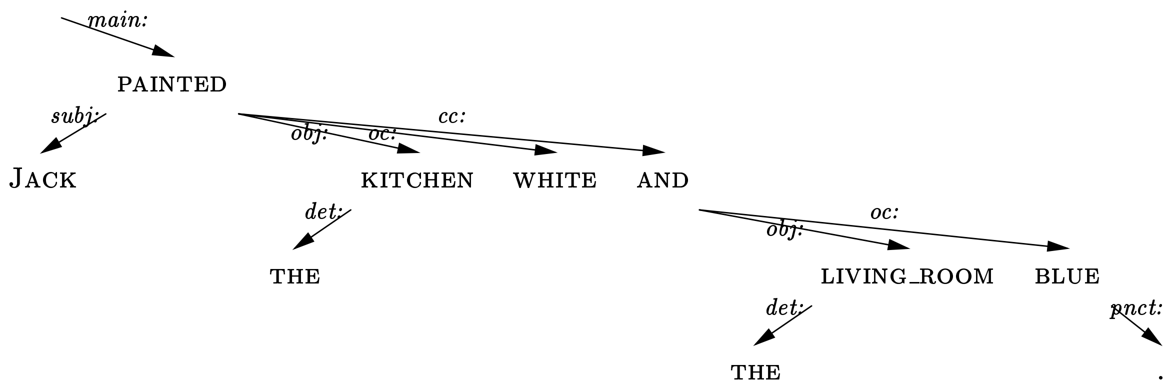

Figure 6: *Jack painted the kitchen white and the living room blue.*

tactic representations. In Becker and Krieger, editors, *Proceedings 5th Meeting of Mathematics of language*, Saarbrücken, DFKI.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4 of *Natural Language Processing.* Mouton de Gruyter, Berlin and New York.

Kimmo Kettunen. 1994. Evaluating FUNDPL, a dependency parsing formalism for Finnish. In *Research in Humanities Computing*, volume 2, pages 47–63. Clarendon Press, Oxford.

Jürgen Kunze. 1975. *Abhängigkeitsgrammatik.* Akademie-Verlag, Berlin.

Solomon Marcus. 1967. *Introduction mathématique à la linguistique structurale.* Dunod, Paris.

James D. McCawley. 1982. Parentheticals and discontinuous constituent structure. *Linguistic Inquiry*, 13(1):91–106.

Igor A. Mel'čuk. 1987. *Dependency Syntax: Theory and Practice.* State University of New York Press, Albany.

Peter Neuhaus and Norbert Bröker. 1997. The complexity of recognition of linguistically adequate dependency grammars. In *ACL-EACL'97 Proceedings*, pages pp. 337–343, Madrid, Spain, July. Association for Computational Linguistics.

Vladimir Petkevič. 1987. A new dependency based specification. *Theoretical Linguistics*, 14:143–172.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* Longman, Harcourt.

Jane J. Robinson. 1970. Dependency structures and transformational rules. *Language*, 46:259–285.

Petr Sgall. 1992. Underlying structure of sentences and its relations to semantics. *Wiener Slawistischer Almanach*, Sonderband 30:349–368.

Anna Siewierska. 1988. *Word Order Rules.* Croom Helm, London.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D.C*, pages 64–71, Washington, D.C., April. Association for Computational Linguistics.

Pasi Tapanainen. 1996. The constraint grammar parser CG-2. Publications 27, Department of General Linguistics, University of Helsinki, Finland.

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

Atro Voutilainen. 1994. *Designing a Parsing Grammar.* Publications of Department of General Linguistics, University of Helsinki, No. 22, Helsinki.

Atro Voutilainen. 1995. Morphological disambiguation. In Karlsson et al. (1995), chapter 6, pages 165–284.

Rulon S. Wells. 1947. Immediate constituents. *Language.* Reprinted in Martin Joos: Readings In Linguistics I, 1957, pp. 186–207.