Meaning-Text Unification Grammar: A Fully Lexicalizable Dependency Grammar

Sylvain Kahane

Abstract

This paper presents a formal unification grammar based on Meaning-Text theory, focusing on the semantic-syntactic interface. Our grammar associates semantic graphs with sentences via syntactic dependency trees. The same formalism allows us to express a modular grammar in the spirit of MTT or HPSG and a fully lexicalized grammar in the spirit of TAG or CG. We propose a fragment of a grammar for French, including the description of idioms, raising verbs and *tough*-movement.

1 Introduction

The goal of this paper is to present a new formalism called Meaning-Text Unification Grammar. As its name indicates, MTUG is a formal unification grammar based on the Meaning-Text theory (MTT), which has been developed since more than 35 years (Mel'cuk 1988). Following the MTT postulates, we consider that 1) a grammar[1] is a formal system which ensures the bidirectional correspondence between texts and meanings (= semantic representations) and that 2) intermediate levels of representation—a morphological level and a syntactic level—must be considered, a grammar consisting of several modules which establish correspondences between representations of adjacent levels. Our grammar is composed of three modules: The morphological module ensures the correspondence between sentences and morphological representations, the syntactic module ensures the correspondence between morphological and syntactic representations and the semantic module ensures the correspondence between syntactic and semantic representations. Our presentation will focus on the semantic module and on the possible lexicalization of the grammar by combining several rules of different modules. The main goal of our approach is to propose a formalism powerful enough to provide a linguistically well-motivated treatments and to avoid the multiplication of elementary structures as in other fully lexicalized grammars such as TAG or CG (Categorial Grammars). A fragment of a French grammar will exemplify our presentation of MTUG.

In section 2, we present the different levels of representation, in section 3, the semantic module, in section 4, the syntactic module, and in section 5, the morphological module. The combination of the different modules and the lexicalization of the grammar will be studied in section 6.

2 Levels of Representation

2.1 Morphological Representation

A *morphological representation* of a sentence is the sequence of the morphological representations of the words of the sentence; the morphological representation of a word is a *surface lexical unit* accompanied with a list of *surface grammemes*. Consider the sentence:

(1) *Zoé essaye de parler à la belle dame*
Zoé tries to speak to the beautiful woman

The morphological representation of (1) is:[2]

(2) $ZOÉ\ ESSAYER_{ind,présent,3,sg}\ DE\ PARLER_{inf}\ À\ LE_{fém,sg}\ BEAU_{fém,sg}\ DAME_{sg}$

2.2 Syntactic representation

The *syntactic representation* of a sentence is a non ordered *dependency tree* similar to the surface syntactic trees of MTT or the stemmas of Tesnière 1959. The nodes of the structure are labeled with surface lexical units, each being accompanied by a list of *surface grammemes*,[3] and the dependencies are

[1] The term grammar is used in its Chomskian sense of 'linguistic model'.

[2] In French, the adjective agrees in gender with the noun. French nouns have a gender feature, indicated in its lexical entry, but they do not bear gender grammemes like adjective because they are not inflected in terms of gender.

[3] In the syntactic representation, only the surface grammemes which are meaningful appear.

labeled with (surface) *syntactic relations*. The dashed arrow indicates a *quasi-dependency*, which will be described in section 3.

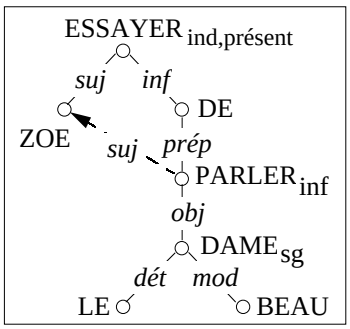

Fig. 1. *Syntactic representation of (1)*

2.3 Semantic Representation

The *semantic representation* of a sentence is a directed graph whose nodes are labeled by *semantemes*. A semanteme acts like a predicate and is linked to its arguments by arrows pointing to them. The different arrows emerging from a semanteme are numbered from 1 to *n* following the increasing syntactic salience of the arguments. Such an arrow, representing a predicate-argument relation, is called a *semantic dependency*.

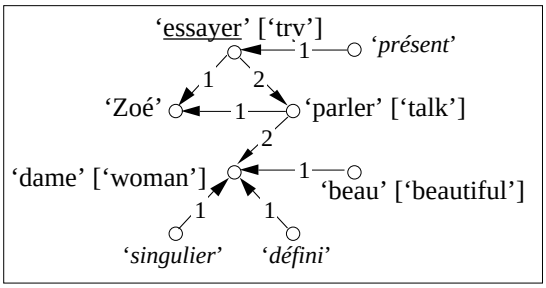

Fig. 2. *The semantic representation of (1)*

Two kinds of semantemes are distinguished: lexical and grammatical semantemes. A *lexical semanteme* or *deep lexical unit* is the meaning of a surface lexical unit or a group of surface lexical units making an idiom. A *grammatical semanteme* or *deep grammeme* is the meaning of a surface grammeme or a complex expression including surface grammemes and surface lexical units, such as the French *passé composé* tense 'perfect' ($AVOIR_{ind,présent}$ + $V_{part\_passé}$).
The underlined node in Fig. 2 indicates the node which corresponds to the root of the syntactic tree. Such an information can be calculated from the communicative structure (Polguère 1990, Kahane & Mel'cuk 1999), which will not be considered here.
A semantic representation can be equivalently described by a logical formula that introduces a variable for each semantic node:[4]

(3) TOP(*e*) *e*:'essayer'(*x*,*e'*) '*présent*'(*e*) *x*:'Zoé' *e'*:'parler'(*x*,*y*) *y*:'dame' '*singulier*'(*y*) '*défini*'*y*) *p*:'beau'(*y*)

3 Semantic Module

3.1 MTT Semantic Rule

In MTT, a *semantic rule* indicates the correspondence between a fragment of a semantic representation and a fragment of a syntactic representation. A *lexical semantic rule* indicates the correspondence between a lexical semanteme with its arguments and the corresponding configuration of surface lexical units (Fig. 3). The grammemes that must be expressed with each surface lexical unit have to be indicated (for a French verb, it is the mood, the tense and the voice). The arrow (→) preceding them indicates that they are deep grammemes and must be translated by *grammatical semantic rules* (see below). These deep grammemes are calculated from the semantic representation: some appears explicitly as semantic node, some are calculated from the communicative structure (such as the voice which depends mainly on the theme-rheme partition) and some can be imposed by the government such as the infinitive mood.

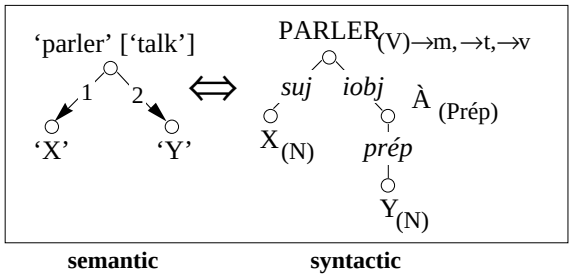

Fig. 3. *A lexical semantic rule (MTT)*

Government and agreement grammemes are introduced by the syntactic module.

[4] On the contrary of most semantic theories, our variables do not refer to objects of the world and our semantic representation does not describe a state of the world but only the linguistic meaning. In MTT, a semantic representation has simply the goal to represent the common meaning of a set of paraphrases. Otherwise some questions such as the scope of quantifier are not considered here (it would need to introduce a second argument for quantifiers pointing on the top node of the scope).

In the spirit of MTT, rules such as the rule of Fig. 3 are used to translate a semantic graph into a syntactic tree (or the converse, according to whether we are considering synthesis or analysis). Therefore, a semantic representation is given and for each lexical semanteme of this representation a rule is triggered, thus giving us a fragment of a syntactic tree which is assembled with the other fragments produced by the other rules. The root of the syntactic tree fragment corresponding to the semanteme at the 'X' position on the left of the rule must be inserted in the X position on the right (see Kahane & Mel'cuk 1999 for a complete presentation).

3.2 MTUG Lexical Semantic Rule

The correspondence rule can be presented in a more generative way: a correspondence rule can be viewed as a generative rule generating two fragments of structures corresponding to each other. Fig. 4 shows a generative MTUG rule equivalent to the MTT rule of Fig. 3. The rule of Fig. 4 combines in one structure the two structures on the left and on the right of the rule of Fig. 3. The label nodes are feature structures. In order to make the figures readable some features are not explicited but indicated by typological means: The word in capitals on the first line is the lexical unit name; on the second line we have the syntactic features between brackets (part of speech, gender for noun…), followed by the grammemes. The semantic information is divided into several features: The feature sém (Fr. *sémantique* 'semantic') gives the semanteme corresponding to the configuration of surface lexical units. The arguments of this semanteme are lintroduced by separate features arg1, arg2 … sharing their value with the sém feature of the argument. Empty words bear a special value ¬sém, which blocks the unification with a feature structure containing the sém feature.

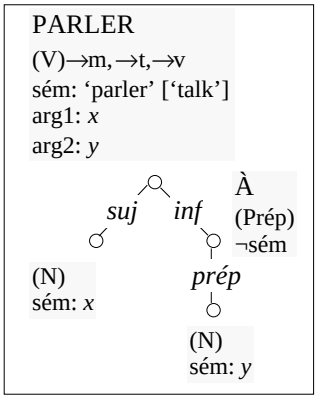

Fig. 4. *A lexical semantic rule (MTUG)*

MTUG lexical semantic rules combine by unification. When two trees combine, the root of one merges with a node of the other one and the labels of the merging nodes are unified (Fig. 5). The resulting structure must be a tree whose all sém features are instantiated. Such a structure describes both a syntactic dependency tree and a semantic graph.

3.3 Idioms

The meaning of an idiom is attached to the root of the corresponding syntactic tree, the other nodes receiving a ¬sém feature (Fig. 6). This avoids modifications of another node than the root, because a modifier is a predicate whose syntactic governor is a semantic argument and must bear a sém feature (see, for instance, Fig. 5 the adjective PETIT 'little' and the adverb ICI 'here'):

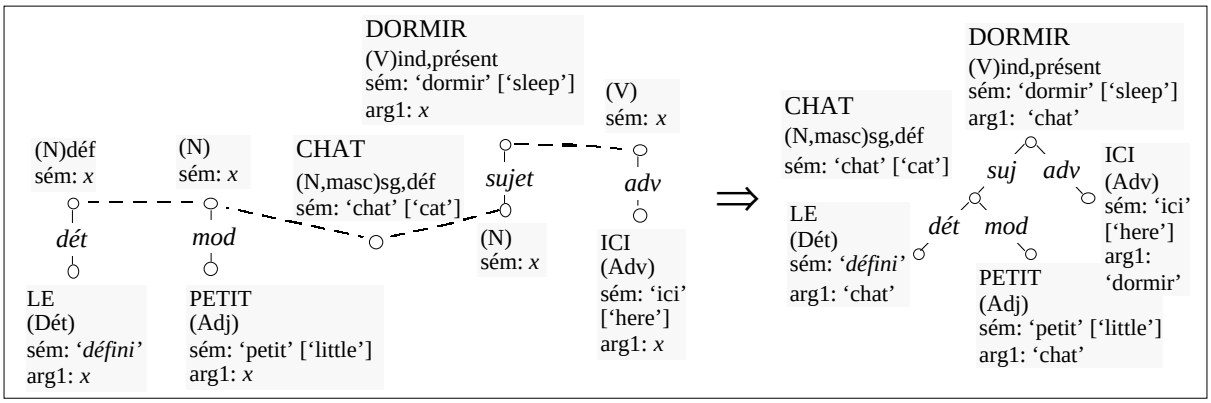

Fig. 5. *Derivation of* le petit chat dort ici *'the little cat is sleeping here'*

(1) a. *La moutarde me monte sérieusement au nez*
The mustard to-me(clit) is-going-up seriously to the nose
'I am greatly flaring up'
b. **La moutarde forte me monte au nez*
The mustard strong to-me(clit) is-going-up to the nose

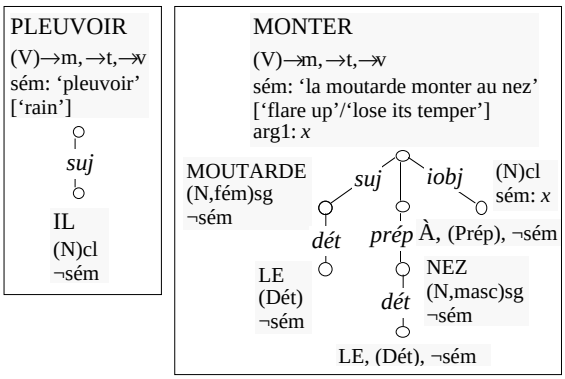

Fig. 6. *Lexical semantic rules for an impersonal verb and an idiom*

3.4 Raising and Control Verbs

The well-known contrast between raising verb and control verb is traditionally encoded, in phrase structure grammar, in the syntactic representation. In our approach, the two kinds of verbs have exactly the same syntactic representation: They govern a subject and an infinitive verb sharing the same subject (we will come back on the subject relation of an infinitive). The contrast comes from the semantic representation: A *control verb* takes its subject as semantic argument but not a *raising verb* (Fig. 7). In both cases the governed infinitive verb controls the subject of its governor and we need a link between them. This link is clearly not a semantic dependency, because the subject can be semantically empty (4a) or belong to the same semantic unit as the verb (4b).

(4) a. *Il commence à pleuvoir*
it is-beginning to rain
b. *La moutarde commence à lui monter au nez*
the mustard is-beginning to-him(clit) go-up to the nose
'he is beginning to flare up'
c. **Il essaye de pleuvoir*
it is-trying to rain
d. **La moutarde essaye de lui monter au nez*
the mustard is-trying to to-him(clit) go-up to the nose

The link between the subject of a raising or control verb and the infinitive verb it governs is clearly a subject syntactic relation. However, it is not a true dependency relation. We will introduce a new type of syntactic relation, *quasi-dependency*, which does not play any role in the tree hierarchy and in the linearization (see Hudson 2000 for a similar proposition). Therefore, infinitive verbs will have a subject but they will be linked to their subject only by a quasi dependency. Quasi-dependencies are represented by dashed arrows.

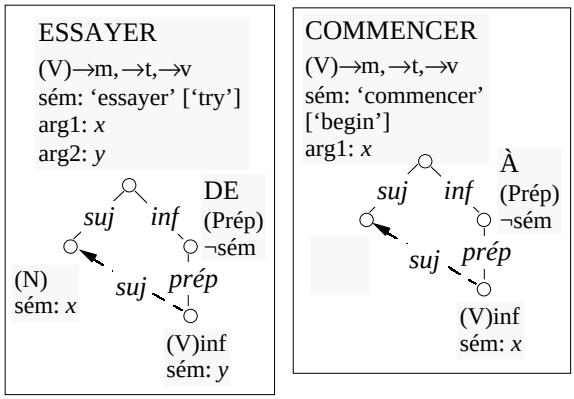

Fig. 7. *Lexical semantic rules for a control verb and a raising verb*

The contrast in (4) between raising and control verbs is simply due to their semantic arity. Contrary to the raising verb COMMENCER 'begin', the control verb ESSAYER 'try' cannot combine with PLEUVOIR 'rain' or the idiom LA MOUTARDE MONTER AU NEZ 'flare up', because it needs a subject with a sém feature.

3.5 Adjectives and Copulative Verbs

Copulative verbs will be treated similarly to raising verb: A *mod(ificative)* quasi-dependency from a noun to the predicative adjective appears in the elementary structure of copulative verbs in order to get a semantic dependency between them (Fig. 8).

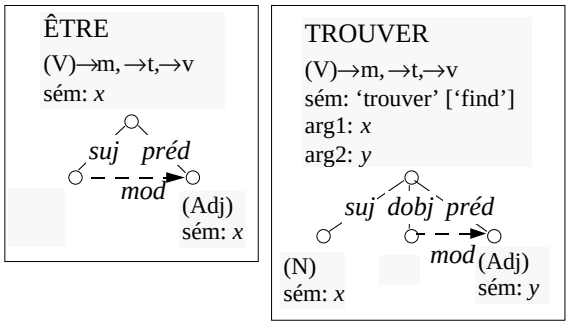

Fig. 8. *Copulative verbs*

This solution allows the same semantic rule (see Fig. 5) to be used for the attributive uses (2a) and predicative uses (2b,c) of an adjective, provided a dependency can merge with a quasi-dependency to give rise to a quasi-dependency.

(5) a. *un petit livre*
a little book
b. *ce livre est petit*
this book is little
c. *Pierre trouve ce livre petit*
Pierre finds this book little

Moreover, in French, in all these cases the adjective must agree with the noun, which could be easily ensured by the syntactic module using the (quasi-)dependency *mod*.

3.6 *Tough*-movement

Tough-movement can be described in the same way as raising. When an adjective such as FACILE 'easy' governs a verb, the noun it modifies is not its semantic argument but this one of the verb (5a). The semantic link between the verb and the modified noun is ensured via a direct object quasi-dependency (Fig. 9). Thus only a transitive verb can combine with FACILE and the "extraction" of its direct object will be realized by the unification of the dobj dependency of the verb and the dobj quasi-dependency. Note that the semantic rule of Fig. 9 can also combine with copulative verbs (Fig. 8) in order to give (5b,c).

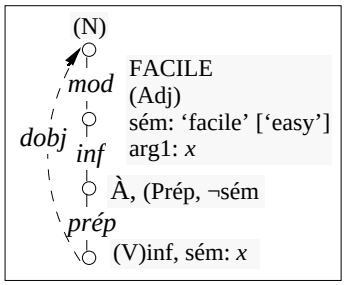

Fig. 9. *Tough-movement*

(5) a. *un livre facile à lire*
a book easy to read
b. *ce livre est facile à lire*
this book is easy to read
c. *Pierre trouve ce livre facile à lire*
Pierre finds this book easy to read

3.7 Grammatical Semantic Rules

A lexical semantic rule cannot be combined with other lexical rules before translating the deep grammemes (preceded by →).

A *grammatical semantic rule* translates a deep grammeme into a surface grammeme or a more complex configuration. In the first case the rule is trivial. Three rules of the second case are proposed Fig. 10. First the inf(initive) mood is expressed by the surface grammeme inf and the transformation of the subject dependency into a quasi-dependency. Second the *passé composé* ('perfect tense') is expressed by the auxiliary AVOIR 'have' with present tense and the past participle. AVOIR is the default auxiliary and another auxiliary can be assigned by the special feature @aux. Third the passive voice is expressed by the auxiliary ÊTRE 'be' and the past participle, while the object node becomes subject and the subject node becomes an optional prepositional complement introduced by the preposition PAR 'by' (in the default case). Note that the voice unlike the mood and the tense is a deep grammeme but not a surface grammeme.

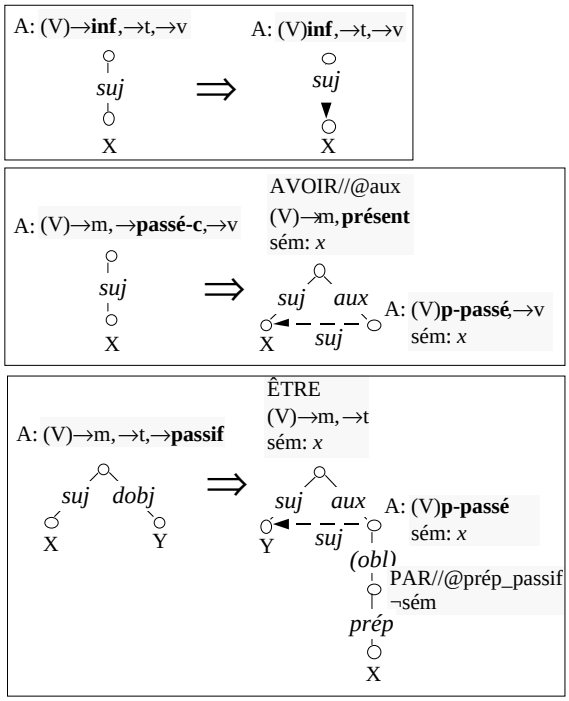

Fig. 10. *Grammatical rules for the infinitive mood,* passé composé *tense and passive voice*

Note that the grammatical semantic rules are not correspondences between a semantic and a syntactic configuration, but rather a function between two syntactic-semantic configurations. A remarkable point is that the order in which the grammatical rules are applied does not matter, that is, mood, tense and voice can be developed in whatever order.

4 Syntactic Module

A syntactic rule indicates the correspondence between a fragment of syntactic representation and a fragment of morphological representation. Some rules ensure the agreement or government requirements: For instance, a verb must agree with its subject in person and number and a pronominal subject receive the nominative case (Fig. 11).

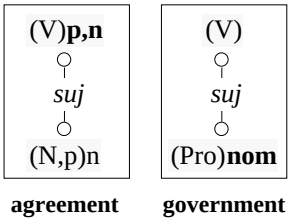

Fig. 11. *Agreement and government rules*

The main family of syntactic rules are the linearization rules. Such a rule indicates how a node is positioned in function of its governor and the type of syntactic relation linking them. For instance, a subject can be before its governor that is, if a node Y is the subject of a verb X, Y can be before X and, vice-versa, if a node Y is before a verb X, Y can be the subject of X (Fig. 12).[5] But the rule must be more precise, because other dependents can be before the verb. Therefore the rule indicates that a non-pronominal subject is positioned at 10 positions before the verb, leaving free 9 positions between the verb and it.[6]

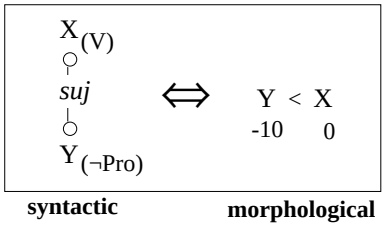

Fig. 12. *A syntactic linearization rule (MTT)*

As the semantic rules, the syntactic rules can be presented in a more generative way. The rule of Fig. 12, equivalent to the rule of Fig. 13, combines a dependency relation and an order relation between two nodes. Such rules can be combined by unification of the node in order to get a linearly ordered dependency tree. Moreover we lay down these ordered tree to be projective.[7] We will not discuss in this paper the very studied question of linearizing non projective tree (see for instance Bröker 1998, Lombardo & Lesmo 1998, Kahane et al. 1998, Hudson 2000).

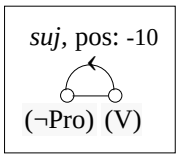

Fig. 13. *A syntactic linearization rule (MTUG)*

5 Morphological Module

We finish our presentation of the grammar by the morphological module. A morphological rule indicates the correspondence between the morphological representation of a word and a word (Fig. 14) and as in the previous cases it can be expressed in a more generative way (Fig. 15).

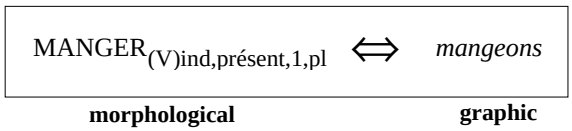

Fig. 14. *A morphological rule (MTT)*

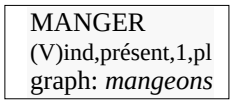

Fig. 15. *A morphological rule (MTUG)*

6 Combination of Modules

6.1 Derivation of a Sentence

Not only the rules of a same module but also rules of different levels can combine by unification. The derivation of a sentence needs the combination of rules of the three modules (Fig. 16). Such a derivation

[5] There is another more complicate rule positioning the subject after the verb: In French the subject can be after the verb if and only if the verb has not a non pronominal object and if another dependent is topicalized.

[6] The values in the linearization rules do not need to be minimal. In French, at least 7 positions must be considered before the verb for verbal clitics (*je* < *ne* < *te* < *le* < *lui* < *en* < *y* <V).

[7] A linearly ordered tree is said *projective* if no arcs cross each other and no arc covers the root. This property is equivalent to require that the projection of each node is a continuous constituent.

associates a sentence to a semantic graph via a morphological string and a syntactic dependency tree.
We cannot describe precisely the process for using the grammar in analysis or synthesis, but we want to focus on two major ways of analysis. The first one consists to process horizontally, that is to analyze module by module: tagging with the morphological module, shallow parsing with the syntactic module and deep analysis with the semantic module. The second one consists to process vertically, that is to fully lexicalize the grammar.

6.2 Full Lexicalization

The lexicalization consists of associating each rule of the derivation with a semantic rule and combinig them to obtain a single rule (Fig. 17). Each deep lexical unit will be associated with a family of rules, called elementary structures, describing all the property of the deep lexical unit from the semantic to the morphology. The result is a fully lexicalized grammar, in the same spirit as TAG or CG. A dependency grammar of this type has been proposed by Nasr 1995.
It is well known that it is hard to develop and maintain wide coverage fully lexicalized grammars and some tools has been proposed for representing them in compact way and generating them from higher level formalisms (Vijay-shanker 1992, Kasper et al. 1995, Candito 1996). Our approach solves this problem in an elegant way proposing a same formalism for the lexicalized grammar and a modular grammar allowing us to generate it.

Note that several choices are possible in the lexicalization. For instance, the linearization rule positioning a clitic complement can be associated to the governor or to the clitic. The second solution is more efficient because the position of the clitic only depends on the nature of the clitic and not of the verb. Thus we avoid the multiplication of the elementary structures associated to a verb. (A third solution consists of adopting a more "morphological" treatment of clitics and putting them in the structure of the verb.)

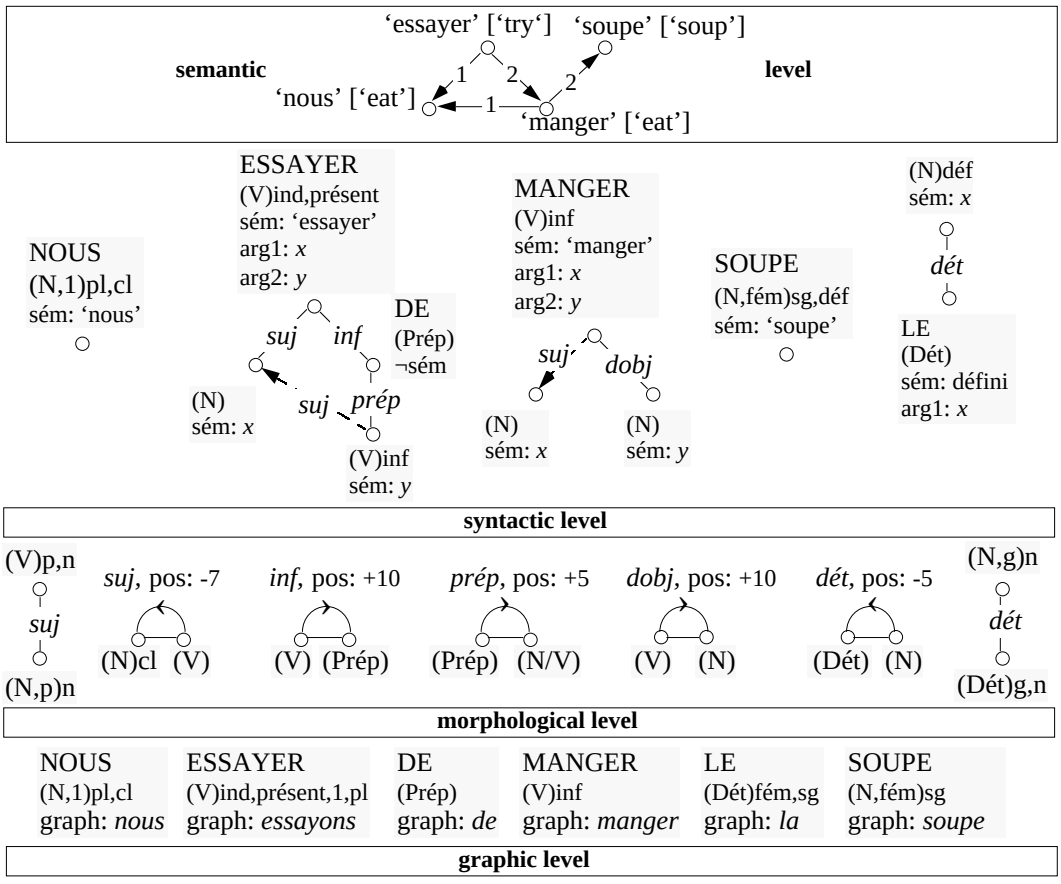

Fig. 16. *Derivation of* Nous essayons de manger la soupe *'We are trying to eat the soup'*

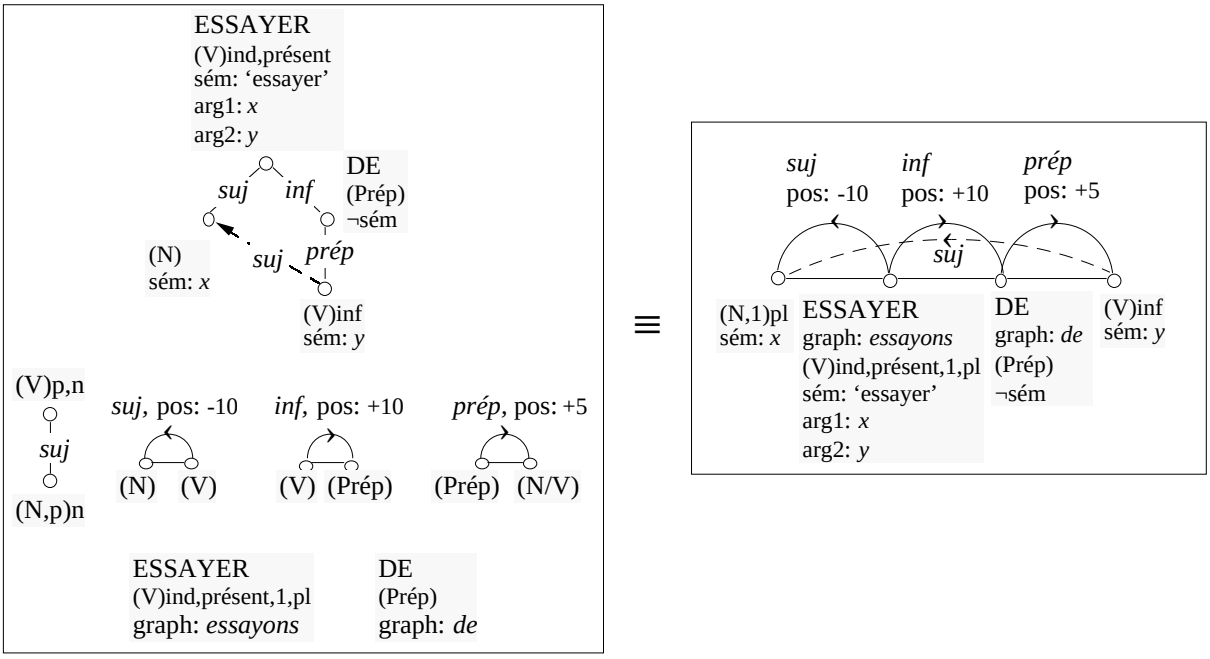

Fig. 17. *Derivation of a fully lexicalized rule*

7 Conclusion

We have presented a formal grammar based on the Meaning-text theory. This grammar provides a clear separation of semantic, syntactic and morphological information. In particular, we adopt a strict separation of the rules of subcategorization, which are described in the semantic module, and the linearization rules, which are described in the syntactic module. We also distinguish deep and surface lexical units and deep and surface grammemes and our analysis of raising and *tough*-movement is based on the clear distinction between semantic and syntactic dependencies.
Resorting to unification allows us to easily express the result of combination of rules of different modules, which is not the case of the traditional MTT formalism.
We hope that our work will help for a better knowledge of the Meaning-text theory, which is one of the most elaborate linguistic theory, and for a better understanding of the various links between MTT and other formalisms and between modular grammars and lexicalized grammars.

References

Candito Marie-Hélène, 1996, "A Principle-based Hierarchical representation of LTAG, *COLING'96*, Copenhagen.

Hudson Richard, 2000, "Discontinuity", in S. Kahane (ed.), *Dependency Grammars*, T.A.L., 41(1): 15-56.

Kahane Sylvain, Alexis Nasr, Owen Rambow, 1998, "Pseudo-Projectivity: a Polynomially Parsable Non-Projective Dependency Grammar", *COLING-ACL'98*, Montreal, 646-52.

Kahane Sylvain, Igor Mel'cuk, 1999, "La synthèse des phrases à extraction en français contemporain. Du réseau sémantique à l'arbre syntaxique", *T.A.L*, 40(2): 25-85, Paris.

Kasper R., B. Kiefer, K. Netter, K. Vijay-Shanker, 1995, "Compilation of HPSG to TAG", *ACL'95*.

Lombardo Vincenzo, Leonardo Lesmo, 1998, "Formal Aspects and Parsing Issues of Dependency Theory", *COLING-ACL'98*, Montreal, 787-93.

Mel'cuk Igor, 1988, *Dependency Syntax: Theory and Practice*, NY : SUNY Press.

Mel'cuk Igor, Nicolas Pertsov, 1987, *Surface syntax of English – A Formal Model within the Meaning-Text Framework*, Amsterdam : Benjamins.

Nasr Alexis, 1995, "A Formalism and a Parser for Lexicalised Dependency Grammars", *4th Int. Workshop on Parsing Technologies*, SUNY Press.

Polguère Alain, 1990, *Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte*, PhD thesis, Univ. de Montréal.

Tesnière Lucien, 1959, *Eléments de syntaxe structurale*, Paris: Kliencksieck.

Vijay-Shanker K., Yves Schabes, 1992, "Structure Sharing in Lexicalized TAG", *COLING'92*.