Microsyntactic annotation

Sylvain Kahane, Kim Gerdes, Rachel Bawden

Abstract. This chapter describes the microsyntactic analysis of the Rhapsodie corpus in terms of dependency syntax. Microsyntax studies the relations between words that are characterized by a strong syntactic cohesion, traditionally called "government". The different steps in the annotation are presented: segmentation into words and labeling in parts of speech, dependency structure, and basic syntactic functions. We justify our decision to use a small set of tags without redundancies, but to introduce a predicative relation for elements that form a complex predicate including copula, auxiliaries, and some modal verbs. Complex cases of annotations such as extraction (relative and interrogative clauses, cleft sentences) and negation are also presented. In addition, we show how a constituent structure can be computed from the dependency structure.

1. Introduction

Syntax describes the way in which linguistic units combine. Microsyntax describes the relations between words that are characterised by a strong syntactic cohesion, traditionally called government (Fr. *rection*). Microsyntax is the syntax proper, which is encoded in all syntactic treebanks, for written as well as spoken languages, from the Penn Tree Bank (PTB, Marcus et al. 1993) and the Prague Dependency Treebank (PDT, Hajič et al. 1999) to more recently developed treebanks.

Our microsyntactic annotation follows the dependency syntax tradition (Tesnière 1959, Mel'čuk & Pertsov 1987, Mel'čuk 1988, Kahane 2001), but differs, as mentioned in Chapter 3, from other treebanks on three points:

- the *whole text*, including disfluent segments, is syntactically analyzed;
- paradigmatic phenomena, called "lists" (see Chapter 5), including coordination, apposition, reformulation, and also disfluency, are richly annotated;
- microsyntactic relations are considered beyond the boundaries of the illocutionary units (Chapter 6) or even beyond speech turns. More precisely, we do not have an aprioristic notion of the sentence inside which all elements must be syntactically connected. On the contrary, we adopt a bottom-up strategy consisting of building government islands (our government units), whose autonomy is studied at the macrosyntactic level.

Section 2 describes the segmentation into words and our set of parts of speech. The dependency structure, which is the backbone of our microsyntactic analysis, is defined in Section 3, while our set of syntactic functions is described in Section 4. Complex cases of annotations, such as extraction and negation, are studied in Section 5 (coordination and other cases of listing are the topic of Chapter 5). In Section 6 we show how a constituent structure has been computed from our dependency structure.

Our choices will be compared with the PDT, two treebanks for spoken language, that is, the oral part of the P7T, the Paris 7 Treebank (Abeillé & Crabbé 2013) and the CGN, the Corpus of Spoken Dutch (Van der Wouden 2002), and the output of the Stanford dependency parser (De Marneffe & Manning 2008).

2. Morpho-syntactic Analysis

Morpho-syntactic analysis consists of the segmentation of the text into words (hereafter referred to as lexemes to avoid confusion with orthographic words), lemmatization, and morpho-syntactic tagging.

The first issue of syntactic annotation is to determine which units are covered by the analysis. In this section we are concerned with the minimal units of microsyntax, the word-forms.

2.1. Word-forms and lexemes

Definition. Broadly defined, word-forms are the smallest units that cannot be split into two freely combining, separable, or dissociable segments. Two elements are dissociable if one element can be used without any element of the paradigm of the other. Two segments X and Y freely combine if they belong to two regular paradigms such that for any X' and Y' belonging to these paradigms, X'Y' is acceptable and has comparable properties to XY. Syntax is the study of this type of free combination (Kahane 2008).

Inflections freely combine with verbal lexemes, but they are neither separable (nothing can be inserted between the verbal root and its ending), nor dissociable (neither the inflection, nor the lexeme can be used independently of the other). Therefore, together they form a word-form, although they combine freely. The first step of the morpho-syntactic analysis of such word-forms must therefore be the identification of the combined elements, that is to say the lexeme and any grammatical morphemes. The last step of the morpho-syntactic analysis is the attribution of a part-of-speech and a lemma to each lexeme. The lemma is the citation form of the lexeme (e.g. the infinitive form for verbal lexemes in French).

Word-forms and multi-word expressions. The theoretical limits of our segmentation are illustrated by the analysis of complex conjunctions of subordination: *dès que* 'as soon as' is analysed as two lexemes (*dès* + *que*), while *alors que* is analysed as one (*alors_que*). Our criterion is the commutation of *que*:

(1) dès **que je suis arrivé** 'from the moment I arrived' ~ dès **mon arrivée** 'from my arrival'

(2) alors que je partais 'whilst I was leaving' ≠ *alors mon départ '*whilst my departure'

Another example is *à peu près* 'about/almost'. Although it is possible to analyse it as a single, cohesive unit, we have decided to analyse its internal structure, given that *peu* can commute with other nouns (*à une semaine près* 'give or take a week'*, à deux centimètres près* 'give or take two centimetres'…).

A significant complication is introduced by frozen expressions: if in a combination XY a meaning cannot be attributed to X and Y, the previous criteria cannot be applied, which does not imply that XY should be divided into two segments from a syntactic point of view. For example in *pomme de terre* 'potato' (lit. 'apple of ground'), *pomme* and *terre* do not commute freely (since *pomme* and *terre* do not have their own semantic contribution), but it is clear that *pomme de terre* is a fixing of the free expression *pomme de terre* which is constructed on the same syntactic schema as *corpus de français* 'French corpus' (N *de* N), which is itself a free combination: *corpus/texte/livre… de français/ chinois/syntaxe...* 'French/Chinese/syntax… corpus/text/book…'. We consider *pomme de terre* to be analogous to *corpus de français* and that it must therefore be segmented in the same way. The segment XY is said to be analogous to X'Y' if meanings of X and Y exist such that X and Y behave in the same way as X' and Y' and such that XY behaves in the same way as X'Y'. A segment XY which does not freely combine but is analogous to a segment which does is called an idiom, a phraseme or a set phase.

We have chosen to keep our analysis to a surface syntactic level and to therefore decompose set phrases and analyse them in the same way as free combinations to which they are analogous.

Word-forms and orthographic words. We have avoided segmenting orthographic words into lexemes. For example, *afin* (*de faire ça*) 'in order (to do that)' is considered a single lexeme even if we could potentially recognise the combination *à* + *fin* 'at + end' and that the two parts are separable as in *à seule fin (de faire ça*) 'for the sole purpose (of doing that)'. The only exceptions are the amalgams, such as *au* /o/ merging two lexemes: *au* = à + *le* 'to the'. For *des*, we have distinguished the cases where *des* commutes with *ces* 'these' (see (3)) from those where it commutes only with *de ces* 'of these' (see (4)). In the latter case only is *des* treated as a combination of *de* + *les*.

(3) ensuite c'est des escaliers [Rhap-M0010, Avanzi]

‘then there are **some** stairs’

(4) dans le vingtième c'est le problème **des** écoles maternelles et primaires dans lequel... [Rhap-D0002, CFPP2000]
‘in the twentieth (arrondissement) it’s the problem **of the** nursery and primary schools in which…’

We have made the same decisions for *du* (and *de la, de l’*) depending on whether it is a partitive determiner that commutes with *ce* or whether it introduces a prepositional group with the preposition *de*.

Orthographic words containing a hyphen are considered a single lexematic word, except in the case of the combination of a verb form and a clitic. In (5), *a-t-il* is decomposed as *a* + *-t-il* and receives the lemmas *avoir* + *il* ‘have + it’.

(5) qui il y a dans qui y **a -t-il** dans la voiture noire [Rhap-D2010, Rhapsodie]
‘who it’s in who is it in the black car’

Conclusion. The criteria we applied in the lexeme segmentation are based on solely syntactic considerations, which gives them a high level of coherence. The result is a quite fine-grained tokenization compared to most other treebanks, which attempt to encode semantically bound “multi-word expressions” into the segmentation, like for instance the P7T, which makes the division dependent on the presence or absence of the multi-word forms in a dictionary.

2.2. Parts of speech

We take into account 13 parts of speech:1

- **V** for verbs;

- **N** for nouns;
- **Adj** for adjectives;
- **Adv** for adverbs;
- **Pre** for prepositions;
- **CS** for subordinating conjunctions;
- **J** for junctors: coordinating conjunctions and other elements that connect layers of listing relations, such as *c'est-à-dire* 'that is to say' or *y compris* 'including'; general extenders such as *et caetera* are also categorised as junctors;
- **D** for determiners (including partitive *du*);
- **I** for interjections, including discourse markers such as *bon, ben, euh, hein,* (the imperative forms such as *allons* 'let's go, come on'*, écoute* 'listen'*, tiens* 'well' are treated as verbs in the imperative rather than as interjections);
- **Qu** for qu-words that are relatives and interrogatives (~ wh-words);
- **Cl** for clitics, including subject clitics (*je, tu il, on, ce, ils, etc.*) and the adverb of negation *ne*;
- **Pro** for the other pronouns;
- **X** for the elements for which the part of speech cannot be determined: inaudible words (noted *XXX*), certain false starts (when the lexeme and part of speech cannot be recovered), as well as unpronounced positions marked by **&**, exceptionally when the part of speech is ambiguous.

Remarks:

- Numerals are categorised as **D** or **Adj** depending on their position. Therefore *deux* ‘two’ is **D** in *deux chaises* ‘two chairs’ and **Adj** in *les deux chaises* ‘the two chairs’. The same goes for *quelques* ‘some’ in *quelques chaises* ‘some chairs’ et *ces quelques chaises* ‘these few chairs’. This choice is notably justified by the fact that we assume no functional annotation of determiners (since all dependents of an **N** have the same function) and that the label **D** is therefore as functional as categorical.

- The modifier *tout* ‘all’ is categorised in the same way as the numerals (above) when they qualify a noun, except that *tout* preposes a determiner. Therefore *toute* is **D** in *de toute façon* ‘in any case’ and *tout* is **Adj** in *tout le monde* ‘everybody’.
- Deictics such as *demain* ‘tomorrow’ are categorised amongst the **Adv** according to tradition, even if there are good arguments to class them under nouns, for the same reasons as *lundi* ‘Monday’*: il vient demain/lundi/lundi prochain/ce lundi* ‘he is coming tomorrow/Monday/next Monday/this Monday’.
- On the other hand, *grâce* in *grâce à lui* ‘thanks to him’ is categorised amongst the **N** even if here it is the head of an adverbial phrase.

Pro is one of the most difficult categories to comprehend. We define as **Pro**:

- stressed personal pronouns (*moi* ‘me’, *toi* ‘you’, *soi* ‘oneself’, *elle* ‘her’, *lui* ‘him’, *eux* ‘them’ …), as well as possessive pronouns (*mien* ‘mine’, *tien* ‘yours’, *leur* ‘theirs’ …);

- the demonstrative pronouns *ça* ‘it/that’, *cela* ‘that’ and *ceci* ‘this’ and the form *ce* ‘it’ only when it appears in an indefinite relative construction of the form *ce que* *j'aime* 'what I like' or *ce qui fait ça* 'what it does' (Note that *ce* as the subject of a verb is considered a **Cl**.);

- indefinite pronouns (*rien* 'nothing', *chacun* 'each person/one', *tout* 'everything/ everyone'…) except for certain exceptions such as *un petit rien* 'a small nothing', where *rien* takes on the role of an **N**;
- numbers when they represent quantifiable pronouns:

(1) (6) il faut faut compter autour de **soixante soixante-dix** [Rhap-D2009, Mertens]
'you should should expect around **sixty seventy** (euros)'

(2) (7) parce que c' était une frange de **vingt trente** [Rhap-D2009, Mertens]
'because it was a fringe of **twenty thirty** (centimeters)'

(3) (8) corner à **deux** 'corner kick with **two** (players)' [Rhap-D2003, Rhapsodie]

It is important not to confuse numbers representing nominal concepts and pronouns. For example the following uses are not considered pronouns but nouns: bus numbers, floor numbers, arrondissements (*dans le* ***vingtième*** 'in the twentieth'), football scores (***zéro*** *à* ***un*** '**zero** to **one**'), page numbers (*page* ***cent douze*** *de votre ouvrage* 'page **one hundred and twelve** of your work'), dates, etc.

2.3. Morpho-syntactic features

Verbs are assigned a mode feature which can take 6 values: *indicative, subjunctive, imperative, infinitive, past_participle, present_participle.* Only indicative verbs vary in tense; the feature *tense* possesses 5 values: *present, imperfect, future, conditional* and *perfect* (corresponding to the simple past and present only once in our corpus). The compound tenses are annotated on the syntactic level rather than the morphological level: a compound past is therefore a **V** *être* or *avoir* (e.g. of mode="indicative, tense="present") whose dependent *pred* is a **V** of mode="past_participle". There is no specific marking for the distinction between the compound past *il est venu* ‘he has come’ and the passive *il est compris* ‘it is included’.

Verbs are also assigned agreement features: the feature *number* has two values *sg* and *pl*, the feature *genre* (for participles) two values *fem* and *masc* and the feature *person* three values *1,2* and *3*.

N, **Adj** and **D** have the features *number* and *genre*. **Cl** and **Pro**, in addition, have a feature *person.*

Certain features can be under-specified. For example, the names of towns and the nominal use of numbers have a feature number=“masc/fem”.

3. Microsyntactic structure

3.1. Government

Microsyntax is limited to *government* relations. We adopt a rather restrictive notion of government. We speak of government when an element imposes on another element its nature, its markers and/or its position. Government relations include subcategorised elements as well as adjuncts. For example, the object of a verb is *governed* by this verb. In *Pierre admire* ***le paysage*** ‘Pierre admires **the scenery**’, *le paysage* ‘the scenery’ is governed by the verb form *admire* ‘admires’. We can see that:

- the form is imposed: the paradigm of elements that can commute with *le paysage* is limited to nominal phrases;

- the markers are imposed: in the case of the direct object in French, there is no explicit marker, but if the complement is pronominalised (*Pierre **l'**admire* 'Pierre admires **it**'), a particular form of the pronoun must be used;

- the position is imposed: the direct object must follow the verb, except particular pronominalised forms or rare cases of anteposition (***deux euros** ça coûte*, '**two euros** it costs').

We use the possibility of clefting as one of the major tests to characterise elements governed by a verb (***c'est** le paysage **que** Pierre admire* '**it's** the scenery **that** Pierre admires' / ***c'est** ici **que** Pierre a dormi* 'i**t's** here **that** Pierre slept').

In (9) and (10), the phrases *simplement* 'in short' and *ces années d'école* 'those school years' are not dependent as they are considered to be non-governed, since they are not cleftable:

(1) (9) a. **simplement** vous êtes un peu plus jeune que moi [Rhap-D0001, CFPP2000]

'**in short** you are a little younger than me'

(2) b. *c'est simplement que vous êtes un peu plus jeune que moi

'*it's in short that you are a bit younger than me'

(3) (10) a. et euh donc **ces ces années d'école** ça a été des bonnes années [Rhap-

D0001, CFPP2000]

‘and uh so **those those school years** they were good years’

(4) b. *c’est ces années d'école que ça a été des bonnes années

‘*it’s those school years that they were good years’

The syntactic relation between *simplement* 'in short' or *ces années d'école* 'those school years' and the rest of utterances (9) and (10) is taken into account at the macrosyntactic level, where these groups are categorised as pre-nuclei (see Chapter 6).

3.2. Dependency

We choose to encode the microsyntactic structure by a dependency graph. Formally, a dependency is a directional relation between two words, which we represent with an arrow: The origin of the arrow is called the *governor* and the target the *dependent*. Each dependency represents a government relation. In (9), *vous* ‘you’ is the subject of *êtes* ‘are’: we represent this with a dependency from *êtes* (the governor) towards *vous* (the dependent) (see Figure 1).

Figure 1. *Microsyntactic analysis of (9)*

In the same way, *un peu plus jeune que moi* ‘a bit younger than me’ is the predicative complement of the verb form *êtes*: we represent this with a dependency from *êtes* towards the head of the adjectival phrase *un peu plus jeune que moi*, that is, the adjective *jeune*.

The verb *êtes*, which is the main verb of this utterance, is therefore not governed. This is noted by a vertical dependency. The adverb *simplement* ‘in short’, which is not governed either, also receives a vertical dependency.

Note that making *que moi* ‘than me’ dependent on *plus* ‘more’ is justified by the fact that the presence of *que moi* is validated by the presence of *plus* (**vous êtes jeune que moi* ‘*you are young than me’) and that *plus que moi* ‘more than me’ can form an autonomous group (*vous êtes jeune //+ plus que moi > en tout cas //* ‘you are young //+ more so than me > in any case //’; see Chapter 6 for macrosyntactic marking).

3.3. Government unit

A *government unit* (GU) is a maximal unit for government. Put simply, it is the collection of all the lexemes of a dependency graph. A GU has a head, which is not governed, and all the elements of the GU are dominated by this head, that is, they are governed by a lexeme, which is governed by a lexeme, …, which is governed by the head of the GU. In other words, a GU is the *maximal projection* of a non-governed lexeme.

We distinguish the GU from the *illocutionary unit* (IU) (cf. Chapter 6). As mentioned in Chapter 3, an IU can be composed of several GUs and the other way around a GU can be composed of several IUs. The IU in (9) is composed of two GUs; the adverbial phrase *simplement* and the clause *vous êtes un peu plus jeune que moi*, as can be seen by the two unconnected parts of the graph in Figure 1.

Note that government does not necessarily stop at the end of a speech turn.

(1) (11) $L1 donc < moi < "ben" je vais je je prends le mét~ je prends le métro le matin "bon" jusqu’au Palais Royal //+

$L2 à quelle heure "excusez-moi" //

$L1 "oui oui" je prends le métro le matin à huit heures et demie // [D0001, CFPP2000]

‘$L1 so < me < "well" I go I I take the met~ I take the metro in the morning "well" to Palais Royal //+’

$L2 at what time "sorry" //’

$L1 "yes yes" I t**ake the metro in the morning at eight thirty** //’

In (11), $L2’s question (*à quelle heure* ‘at what time’) continues the preceding microsyntactic construction (*je prends le métro le matin jusqu’à Palais Royal* ‘I take the metro in the morning "well" to Palais Royal’) and $L1’s reply (*je prends le métro le matin à huit heures et demie* ‘I take the metro in the morning at eight thirty’) has exactly the same structure as the concatenation of the two preceding turns (*je prends le métro le matin à quelle heure* ‘I take the metro in the morning at what time’) (see Figure 2).

Figure *2. Microsyntactic analysis of (11): a GU on two IUs*

Although we consider government beyond IUs and speech turns, our dependency structure is more restrictive than the dependency structure used in other theoretical frameworks (cf. for example Mel'čuk 1988 or the CGN that considers "satellite" dependency relations for so-called "dislocations"). Indeed, we only represent microsyntactic information in our structure, that is, information relevant to government. Dependency is a formal means of representing different syntactic relations. We could also have decided to represent macrosyntactic information, such as the dependency of post or pre-nuclei to the nucleus (see Chapter 6). We have decided to encode micro and macrosyntax separately (see Chapter 3) and to only use dependency for microsyntactic information, because we consider that macrosyntax is a rather topological phenomenon related to the linear position of the elements, whereas microsyntax describes relations that are foremost defined on other grounds than linear order.

3.4. Choice of the head

The *head* of a microsyntactic unit (a phrase for instance) is intuitively its most important element. On the one hand it is the element that controls the distribution of the unit (the external head) and on the other hand it is the element that validates the presence of the other elements of the unit (the internal head). We shall now look at some configurations that could pose a problem.

Auxiliary: The head of a clause is what is traditionally known as the main verb. In the case of a complex verb form, we treat the auxiliary as the head (sse Figure 3).

(12) et en fait euh j'ai pas été acceptée parce qu'il y avait un entretien oral [M1001, Rhapsodie]

‘and well uh I wasn’t accepted because there was an oral interview’

Figure 3. *Microsyntactic analysis of (12): the auxiliary*

In (12), *ai* governs *été*, which governs *acceptée* (Figure 3). This choice is justified by the fact that the auxiliary is a finite form, which therefore carries the mode (*elle* ***a*** *été acceptée* ‘she was accepted’ vs *il est étonnant qu’elle* ***ait*** *été acceptée* ‘it’s surprising that she was accepted’) and illocutionary markers such as enclitics (***a-t-elle*** *été acceptée ?* ‘was she accepted?’). Moreover, the auxiliary imposes the mode of the content verb: *elle a* ***accepté*** ‘she was accepted’ (past participle) vs. *elle va* ***accepter*** ‘she will accept’ (infinitive). While the CGN annotation follows this classical syntax-based choice, the Prague Dependency Treebank and the Stanford Dependency Parser make a more semantic choice and choose the content verb as the head, while P7T has a flat structure.2

Determiner: We consider that in noun phrases, the noun is the head and governor of the determiner. This choice is partly arbitrary (the determiner also has the role of marker of the noun phrase, which would justify it being the head; Hudson 2004, Gerdes & Kahane 2013), but is the most common in dependency syntax (Tesnière 1959, Mel’čuk 1988, CGN).

Note that, in noun phrases of the form “**Adv** de **N**” (*peu de gens* ‘few people’*, trop de gras* ‘too much fat’ ...), we consider the **Adv** to be the head (Figure 4), because “de N” can be pronominalised (see Figure 4).

(13) a. mh mh donc c’est vrai que ça fait quand même **beaucoup de**

changements [Rhap-D0004, CFPP2000]

‘mm mm so it’s true that that makes actually a lot of changes’

b. ça **en** fait quand même **beaucoup**

'that makes **a lot of them**'

Figure 4. *Microsyntactic analysis of (13): a complex determiner*

Completive subordinate clauses and adjectives: We would also like to point out that in (13), constructions “*c’est **Adj** que **V***” ‘it’s **Adj** that **V**’ are analysed such that the completive clause “que **V**” depends on the **Adj**. Yet again it is a surface analysis which does not take account of the fact that the completive is a deep subject (que **V** est **Adj**), since on the surface the subject of *être* is the pronoun *ce*. The analysis is based on the possibility of making the **Adj** and the completive an autonomous phrase in certain cases (*impossible qu’il vienne* ‘impossible that he will come’).

Incompletion: When a GU is clearly incomplete, that is, an obligatory position is not filled, we note it by an & (see Figure 5). This symbol indicates the non-represented position, and is not a representation of a position by an empty element (cf. traces in generative grammar).

‘at a given moment it became &’	‘but it c~’
[Rhap-D0005, PFC]	[Rhap-D0005, PFC]

Figure *5. Incomplete GUs*

The sequences *First name Surname* (e.g. *Françoise Giroud*) are considered the combination of two words (if only because each name can be deleted in favour of the other) and the first is treated as the head.

Even if they have an internal syntax, all numbers have been analysed as a single word-form: *deux mille quatorze* ‘two thousand and fourteen’.

4. Syntactic functions

We have decided to keep the number of syntactic functions to a minimum. Syntactic functions can be introduced for two quite different goals:

- to regroup dependents that behave in the same way and to distinguish those that behave differently. Therefore, the complement *à Marie* ‘to Marie’ in *parler à Marie* ‘speak to Marie’ behaves in the same way as that in *donner quelque chose à Marie* ‘give something to Marie’ (*lui parler* ‘speak to her’*, lui donner quelque* *chose* ‘give her something’) but differently from that of *penser à Marie* ‘think of Marie’ (*penser à elle* ‘think of her’*, y penser* ‘think of her’).

- to distinguish the different dependents of a same word. For example, in *Pierre a nommé Louis général* ‘Pierre named Louis general’, *Louis* and *général* are two dependents of the same verb and only the first can be cliticised (*Louis, Pierre l’a nommé général* ‘Louis, Pierre named him general’; **Général, Pierre l’a nommé Louis* ‘*General, Pierre named Louis it’).

Our approach is clearly along the lines of the second goal. Contrarily to the 30-odd relations used in some of the dependency annotations of the CoNLL shared task treebanks (Nilsson et al. 2007) or the 16 relations for verb dependents of the Dutch CGN treebank, some of which are redundant and can be retrieved in combination with the POS annotation and other lexical features of the annotation, we refer to only seven plain dependency relations (similar to P7T):

- *root:* elements that are not governed by another element;
- *sub:* grammatical subjects of verbs;
- *obj:* direct objects of verbs;
- *obl:* oblique complements of verbs, including indirect objects;
- *ad:* adjuncts to the verb;
- *pred:* all elements that form a complex predicate with a verb;
- *dep:* all the dependents of non-verbal forms.

Two other relations are considered for paradigmatic lists (see Chapter 5).

- *para:* paradigmatic links between phrases occupying the same position (in case of coordination, disfluencies, reformulation, etc.)
- *junc:* to link conjuncts to junctors (= coordinating conjunctions), giving the asymmetrical annotation of coordination put forward by Mel'čuk (1988).

The *pred* relation, which is the only originality of our functional annotation, deserves some clarification. It is used for:

- constructions known as predicative complements of the subject (*il est **gentil*** ‘he is **nice**’) or of the object (*il trouve Marie **gentille*** ‘he finds Mary **nice**’) (see Figure 6);

‘mh mh and I find it nice to see them in Paris’ [Rhap-D0006, CFPP2000]

Figure *6. Predicative adjective*

- complex verb forms (*avait mangé* ‘had eaten’*, est parti* ‘has left’);
- constructions with a modal verb (*peut venir* ‘can come’*, doit manger* ‘has to eat’), where the infinitive does not easily commute with a noun phrase (see Figure 7).

‘now on the other hand it must be more expensive’ [Rhap-D0009, PFC]

Figure *7. Modal verb*

- except in light verb constructions (*avoir envie* ‘want’, *avoir l’intention* ‘have the intention’*, avoir besoin* ‘need’*, faire peur* ‘scare’ ...), where the predicative noun is treated as a direct object (see Figure 8). This is coherent with our syntax-centered approach where idiomatic structures are broken up if the parts construct in a common way to form the idiom (see Section 2.1).

We use the *obl(ique)* relation for all indirect objects, that is, subcategorised complements of the verb which are not *obj*. Prepositional complements which are part of a set phrase (*mettre* ***en doute*** ‘question’, lit. ‘put in doubt’) and locative constructions with *être* ‘be’ are also treated as *obl*.

Complements of verbal expressions are treated as dependents of the predicative element (**Adj** or **N**), and therefore receive a *dep* relation as dependents of a non-verbal element: see the *dep* relation between *envie* ‘desire’ and *d’entendre de la musique* ‘to hear music’ in Figure 8.

‘so they wanted to hear (some) music’ [Rhap-D0003, PFC]

Figure *8. Light verb construction*

5. Some complex cases

5.1. Extraction and qu-words

It is possible to consider, following Tesnière (1959) for example, that in relative phrases, the relative pronoun’s role is a pronoun within the relative clause and at the same time a complementiser allowing the relative clause to modify a noun. This analysis implies that the relative pronoun should occupy a double syntactic position, as head of the clause (as complementiser) and dependent within the clause (as pronoun).3 Some analyses even go so far as to say that certain *qu*-words, notably *que* in relative clauses, are above all complementisers (Kayne 1975).

In our analysis we only encode the position of the pronoun within the relative clause, the main verb is therefore the head of the relative clause and a direct dependent of the modified noun (see Figure 9).

‘the forces that it is capable of taking on’ [Rhap-M2001, C-PROM]

Figure *9. Relative clause*

We do not deny the fact that relative pronouns and *qu-* words in general have this complementiser role, but we decide, in the interest of simplicity to not encode this position, which can be recovered. On the other hand, the pronominal position of the *qu-*word and its function cannot be easily recovered, notably because of long distance dependencies, that is, cases where the relative pronoun occupies a deep position in the relative clause, which results in a non-projective structure, since the relative pronoun is not found next to the governor of the extracted position (Figure 9). The same choice is made by the PDT and the Stanford dependency parser, but the CGN and the PTB choose to encode both positions.

For cleft constructions of the form “c’est X qui/que V” (*c’est l’utilisateur qui a fait une fausse manœuvre*), we treat the subordinated clause as a relative clause but dependent on the cleft marker, that is, the verb *être* ‘be’ rather than the phrase extracted by clefting. Moreover, given the particular nature of the relation between the cleft construction and the subordinate clause, we assign it the function *dep*, whereas the extracted element has the function *pred* (the idea is that the underlying construction is *ce qui/que X est V: ce qui nous revient est un virus qui a fait le tour du monde* 'what comes back to us is a virus that travelled round the world', see Figure 10).

'and each year it's a virus that has travelled round the world that comes back to us uh' [Rhap-D2008, Rhapsodie]

Figure *10. Cleft sentence*

The analysis is extended to interrogative cleft constructions of the form *qu'est-ce que*, bearing in mind that it is the interrogative form of a cleft construction (*c'est **quoi** que tu fais là* 'it is what that you are doing there' ***qu'****est-ce que tu fais là*, see Figure 11):

'he said to me but what are you doing there very surprised' [Rhap-D2001, Mertens]

Figure *11. Interrogative form*

5.2. Negation

Although there are good reasons to consider *pas de X* 'no X/not X' as being a phrase (see *beaucoup de X* 'lots of X'*, peu de X* 'few X'), *pas de X* does not always function as a cohesive unit, since it can be separated by adverbs or verb forms (Figure 12, left) and, as for *beaucoup de* (see (13b)), *de X* can be pronominalised. Therefore *pas* is always considered the adjunct to the verb and the noun the object, *de* being treated as a partitive determiner. In the case of a verb in the compound past, the *ne* and *pas* are always considered as adjuncts of the auxiliary. It is only when there is no governing verb that *pas de X* is treated as a noun phrase with *pas* as head (see Figure 12, right).

‘you are not going to add a fringe’	‘no problem’
[Rhap-D0009, Mertens]	[Rhap-D0008, Avanzi]

Figure *12. Syntactic position of negation* pas

6. Phrase structure

Although Rhapsodie is above all a treebank conceived with dependency syntactic considerations, it remains interesting to provide a phrase structure version of the annotation, one reason being the comparison with data from phrase structure based treebanks and prosodic units, the other that many syntacticians have been trained to read and understand phrase structure rather than dependency structure. A phrase structure can be computed automatically from the dependency structure, albeit some complications in case of non-projectivity to be explained below.

Each lexeme in the corpus gives two constituents: a phrasal constituent, which is its maximal projection and a lexical constituent.4 A lexeme of part-of-speech X gives a projection of the category XP and a lexical constituent of the category X. More precisely:

- Nouns (N) and tonic pronouns (Pro) project an NP (noun phrase)
- Inflected verbs (V) projects an S (sentence)
- Uninflected verbs (V) (infinitives and participles) project a VP (verb phrase)
- Conjunctions of subordination (CS) project a CP (complementiser phrase)
- Prepositions (P) project a PP (prepositional phrase)
- Junctors (J) project a JP (junctor phrase)

So as not to unnecessarily weigh down the structures, adjectives (Adj), adverbs (Adv), determinants (D) and light pronouns (Cl, Qu) only project an AdjP, AdvP etc. when they have dependents.

Overhanging relations between microsyntactic constituents give a tree structure, which we call the *microsyntactic constituent tree*. The existence of non-projective dependencies adds a complication illustrated by the example provided in Figure 13 (D0003).

‘and there were people in there who were prisoners’ [Rhap-D0003, PFC]

Figure 13. *A non-projective dependency tree*

The projection of *gens* 'people' is the constituent *des gens qui étaient prisonniers* 'people who were prisoners'. The non-projection of the dependency structure (the fact that the adverb *là-dedans* interrupts the noun phrase) means that:

- either the order of the words is only partially conserved in the microsyntactic constituent structure as in Figure 14.
- or it is necessary to consider phrase structure trees with overlapping branches as in Figure 15, where phrases are no longer linearly contiguous units.
-

Figure 14. *A derived phrase structure not preserving word-order*

Figure 15. *A derived phrase structure with crossing branches*

This situation, a common problem related to non-projective sentence structure where both phrase structures seem suboptimal, has actually been at the basis of the development of "traces" in the Chomskyan approaches to syntax (Gerdes 2006), because putting a trace directly under S (e.g. instead of the adverb's position in Figure 14) linked to the adverb's actual linear position (e.g. by attaching it inside the NP) constitutes a compromise that allows to express the government relation between the verb (*avait*) and the adverb (*là-dedans*) without resorting to crossing branches. All this of course to the detriment of the simplicity of the syntactic structure as a whole, leave alone the arbitrary placement of the (invisible) trace.

Under these considerations and in order to remain as close as possible to the dependency structure, we distribute Rhapsodie with an automatically derived phrase structure that does not preserve word order in case of non-projective sentence structure (as in Figure 14).

7. Conclusion

Rhapsodie's microsyntactic annotation has the ambition to be theory-driven in the sense that our goal is to provide a theoretically sound syntactic annotation, based solely on well-defined syntactic criteria, and non-redundant with other levels of annotations (such as the existing macrosyntactic annotation or an envisioned semantic annotation). Most other treebanks have been developed under different considerations, e.g. they are a result of transcoding a phrase structure treebank into dependency (PTB) or their raison d'être is to provide machine-training material for NLP tasks where most often closeness to semantics gives better results, as in Universal Dependencies (Nivre et al. 2016), where relations between "content words" are favoured.

The Rhapsodie annotation scheme has been developed from scratch, in a tradition of multi-layered dependency analysis, adapted to the needs of spoken language, and with the goal to deepen its understanding. Rather than eliminating disfluencies from the syntactic analysis (as has been done for example in the CGN) on the basis of an underlying supposedly "correct" structure, we take the stance that each word, even a repetition or a correction, has its role to play in the meaning construction of the utterance, and should therefore appear in syntax. The identification itself of syntactic units has been reconsidered as a purely syntactic task, and categories, dependency relations and macrosyntactic positions have been thought as complementary and non-redundant syntactic information.

This approach has proven its capacities also in the analysis of renowned complex syntactic structures such as extraction, negation, and coordination, therefore providing a solid base for further work on dependency linguistics.