Protocol for micro-syntactic coding

Version: 1 November 2013

Authors: Sylvain Kahane (in collaboration with the Rhapsodie syntactic team and in particular Kim Gerdes, Paola Pietrandrea and Christophe Benzitoun)

Revised version by Rachel Bawden and the other annotators (Marie-Amélie Botalla and Adèle Désoyer)

Translated by Rachel Bawden

This document is divided into four sub-sections:

- morphosyntactic analysis
- microsyntactic dependency analysis
- pile constructions
- microsyntactic constituency analysis

Morpho-syntactic Analysis

(Sylvain Kahane, Kim Gerdes)

Morphosyntactic analysis consists of the segmentation of the text into words (hereafter referred to as lexemes to avoid confusion with orthographic words), lemmatisation and morphosyntactic tagging.

A text is segmented into words. With certain exceptions (amalgamations such as *du* and *au*), a word is a lexeme or an inflected form of a lexeme, i.e. a lexeme combined with its grammatical morphemes. (N.B. words are used in their linguistic sense. We distinguish them from orthographic words, which nevertheless often share the same form).

Lemmatisation is the attribution of a lemma (the lemma is the name conventionally used for the lexeme; in the case of verbs it is the infinitive) to each word/lexeme. Morphosyntactic tagging is the attribution of the lemma's part of speech, and where relevant categorical and inflectional features, to each word.

Segmentation into lexemes

Word: By *lexematic word* (also referred to as *word form* in the structuralist tradition), we refer to a particular linguistic unit, generally considered to be the minimal syntactic unit, which we will define broadly. In the Rhapsodie project, a second segmentation into words was carried out by the prosodists for the calculation of rhythmic groups, with a view to eventually aligning the two segmentations.

Hereafter the term *word* will always refer to a lexematic word.

The notion of the word is directly linked to that of the lexeme - the minimal lexical unit. A word is an invariable lexeme or an inflected form of a lexeme, or (very marginally) the amalgamation of two lexemes.

Token: We define *token* (or *orthographic word*) as segments of the orthographic transcription which appear between two whitespaces or a whitespace and a punctuation mark. The apostrophe is also considered to be the right boundary of a token and *l'enfant* 'the child' is therefore the combination of two tokens (*l'* + *enfant* 'the + child'), as are consequently *aujourd'hui* 'today' or *quelqu'un* 'someone'. The hyphen is not considered to be the boundary of a token and *dit-on* is a token which we decompose into two lexemes.

Orthographic conventions are to a large extent motivated by linguistic considerations, and tokens (i.e. orthographic words) correspond in the vast majority of cases to lexematic words and vice versa.

We shall now set out the criteria used to define the (lexematic) word and then indicate the cases for which we will consider words not to be tokens.

Definition of the (lexematic) word

A segment XY is divided into two segments X and Y if X and Y have independent distributions, i.e. if X and Y can be used in other words with other elements whilst retaining the same meaning. Moreover, if X' and Y' commute freely with X and Y in other contexts, they must also commute with X and Y in the context of the combination XY, i.e. X'Y, XY' and X'Y' must also be acceptable and have comparable properties to XY.

Words are not the smallest unit of this type of segmentation. For example, in the verb form *chantons* 'we sing', *chant-* (stem of the verb 'sing') and *–ons* (3rd person plural present tense) commute freely (*chant-* with other radical verbs and *–ons* with other verbal inflections). But *chant-* and *–ons* are highly cohesive: it is not possible to dissociate them (use a verbal radical without inflection and vice versa), nor separate them or modify them independently of each other.

Words are the smallest units that cannot be split into two freely combining, dissociable and separable segments.

If a word can be decomposed, only one of the parts really belongs to an open paradigm and is a lexeme. The other parts are grammatical morphemes associated with this lexeme. The word is therefore the inflected form of a lexeme.

Idiom: 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 mean that we do not want to divide XY into two segments from a syntactic point of view. For example in *pommes de terre* 'potatoes' (lit. 'apples 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 syntactic level and to therefore decompose set phrases and analyse them in the same way as free combinations to which they are analogous.

We will provide a long list of examples to clarify this definition. It is important to understand that the notions of free commutation and analogy are gradual notions and that there exist units whose word status is unclear and for which our choices may seem arbitrary. Nevertheless the decision to treat a unit XY as a whole or as a combination of X and Y (so as to describe the relation between X and Y) does not impact on the rest of the analysis of the utterance. Even if we had wanted to perform the most rigorous segmentation possible, the decisions we have made have local scopes which only effect problematic units, regardless of the chosen analysis.

Lexemes within a token

We have avoided segmenting tokens into words. For example, *afin* 'in order (to)' is considered a single word 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)'.

Amalgam: We have separated the amalgams *au* and *aux* into two lexemes: *au* = à + *le*. For *des*, we have distinguished the cases where *des* commutes with *ces* from those where it commutes only with *de ces*. In the latter case only is *des* treated as a combination of *de* + *les*:

- ensuite c'est **des** escaliers (M0010:2)
  'then there are some stairs'
- **...**dans le vingtième c'est le problème des (**de les)** écoles maternelles et primaires dans lequel... (D0002:21)
  'in the twentieth (arrondissement), it's the problem of the nursery and primary schools in which…'

We have made the same decisions for *du* depending on whether it is a partitive determiner that commutes with *ce* or whether it introduces a prepositional phrase in *de*:

- ça j'avoue qu'on a **du** mal quand on voit que Paul Valéry passe... (D001:112)
  'ok, I admit that it's a little difficult when you see that Paul Valéry passes…'
- **...**le sherpa du (**de le)** président le porteur de valises le conseiller influent du (**de le**) prince... (D2005:6)
  '…the Sherpa of the president, the porter, the influential adviser of the prince…'

In the interest of homogeneity, *de la* and *de l'* also undergo two different analyses, treated either as one or two words.

- et il y a aussi de la (**de_la)** très bonne culture (D1001:26)
  'and there is very good culture too'
- sa femme est originaire **de la** région (D009:182)
  'his wife comes from the region'

Hyphens: Tokens containing a hyphen are considered a single lexematic word, except in the case of the combination of a verb form and a clitic:

- dit-on = dit + -on 'one say'
- a-t-il = a + -t-il : *qui il y a dans qui y* ***a -t-il*** *dans la voiture noire* 'who it's in who is it in the black car' (D2010:186)

Finally, *là* in combinations such as *ce N-là* 'that N' is also considered a lexeme in its own right:

- ...très difficile d' d'apprendre le français à des petits enfants de cet âge –**là** (D002:52)
  '…very difficult to to teach French to small children of that age'

The tokens *là-bas* 'down there', *là-dedans* 'in there' and *là-dessus* 'on there' are considered single lexematic words. We could consider isolating *là*, but its syntax would be analogous to no other lexical element and the paradigm of elements which combine with it remain quite restricted compared to postposed *–là* which combines with all N.

Words composed of several tokens

Here is the list of words composed of several tokens that we have taken into account:

Grammatical words

à nouveau 'anew'
à part 'separate'
à peine 'hardly'
a priori 'a priori'
à savoir 'namely'
à travers 'across/through'
alors que 'while'
au moins 'at least'
autre chose 'something else'
bien sûr 'of course'
c'est-à-dire 'that is to say'
d'abord 'first'
d'ailleurs 'moreover'
de nouveau 'again'
de plus 'more', 'furthermore'
de plus en plus 'more and more'
du tout 'at all'
eh ben 'well'
eh bien 'well'
encore que 'even though'
en fait 'actually'
en tant que 'as'
en tout cas 'in any case'
en quelque sorte 'sort of'
et caetera 'et caetera'
et puis 'and then' (but not *ou bien* 'or else', *ou encore* 'or even', …)

jusqu'à 'until/up to' (when an Adv (jusqu'à chez moi 'to mine'), but not when Adv + Pre (jusqu'à Paris 'to Paris'))
l'un 'one' (but not *l'autre* 'the other', since we have *les deux autres* 'the other two')
lors de 'at the time of' (but not *faute de* 'for lack of')
n'importe quel 'whatever'
n'importe quand 'whenever'
n'importe qui 'whoever'
parce que 'because'
petit à petit 'little by little'
peut-être 'maybe'
quand meme 'all the same'
quelqu'un 'someone'
quelque chose 'something'
quelque part 'somewhere'
sauf que 'except that'
sur ce 'whereupon/with this'
surtout que 'especially that'
tout à fait 'exactly'
tout de suite 'straight away'
vis-à-vis (de) 'towards/by comparison with'
y compris 'including'

All numbers

deux mille neuf 'two thousand and nine'
dix-neuvième 'nineteenth'
dix-huit cent 'eighteen hundred'
dix-huit cent quatre-vingt 'eighteen eighty'
neuf cent cinquante 'nine hundred and fifty'
quatre-vingt-douze 'ninety two'
trois cents 'three hundred'
vingt-deux 'twenty two'
…

Compound nouns

All compound nouns spelt with a hyphen have been considered words: après-midi 'afternoon', arrière-grand-mère 'great grandmother', aujourd'hui 'today', baby-sitter, belle-mère 'step-mother', centre-ville 'town centre', chef-d'œuvre, contre-attaques 'counter-attacks', contre-littérature 'counter-literature', enseignant-chercheur 'teacher and researcher', fauteuil-crapaud 'easy chair', grands-parents 'grandparents', mathématicien-écrivain 'mathematician and writer', mi-temps 'part-time', outre-mer 'overseas', pâtissier-boulanger 'baker and pastry chef', rendez-vous 'meeting', rond-point 'roundabout', week-end …

The noun *face à face* 'confrontation, one-on-one' (*un face_à_face très attendu* 'a much anticipated one-on-one') is treated as a word but the adverbial expression (*ils sont face à face* 'they are face to face') is treated as three distinct words.

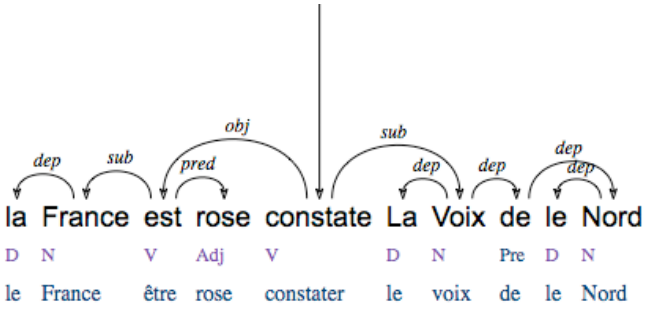
Proper nouns

General Motors
Hauts-de-Seine

Jean-Paul
Pointe-à-Pitre
Proche-Orient 'Near East'
Royaume-Uni 'United Kingdom'
Saint-Jean
Saint-Jean-de-Maurienne
(avenue) Alsace-Lorraine

The sequences First name Surname (*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.

Compound nouns of the form *N de N* are analysed as a combination of words, including proper nouns such as *La Voix du Nord* 'The Voice of the North':



[la France est rose //] >+ constate La Voix du Nord //

'[France is pink //] >+ reported The Voice of the North //'[1]

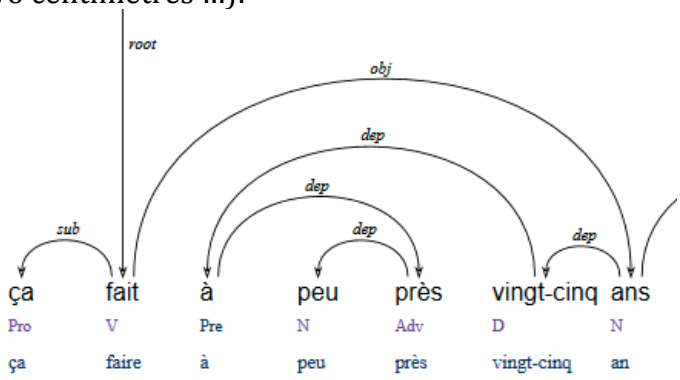
Even so, a certain number of very cohesive compounds have also been considered combinations of words because they possess a particular syntactic configuration:
à côté 'close by' (à droite 'to the right', à gauche 'to the left', …), but not à_travers 'across/through' which is formed differently : à côté **de** la maison 'next **to** the house' vs à travers la maison 'through the house'
à la fois 'at the same time'
à le fond de 'deep down within', à le milieu de 'in the middle of', à le dessus de 'above', (tout) à le long de 'throughout', à le moment de '(just) when', à le sein de 'at the heart of', à le travers de 'through/by means of', à l'égard de 'with regard to'
à partir de 'from/starting from'
à raison de 'at a rate of, à cause de 'because of', à propos de 'concerning'
de ce que 'from what'
de côté 'to one side/sideways' (de face 'full-face', de près 'closely', de loin 'far away')
en gros 'roughly', en cours 'ongoing', en général 'in general', en bas 'down(stairs)'
en matière de 'as regards', en face de opposite', en raison de 'because of'
faute de 'for lack of'
grâce à 'thanks to' (face à 'facing')

[1] See the protocol for macro-syntactic coding for the syntax and the semantics of our tags. Note that // indicates the end of Illocutionary Unit, < and > the limits between the nucleus and a pre- or post-nucleus. And + indicates that the macrosyntactic break is not a microsyntactic break. In other words, *la France est rose* is an embedded IU ([…//]) and the nucleus of the main IU, *constate La Voix du Nord* is a post-nucleus (>), and there is a microsyntactic dependency between the nucleus and the post-nucleus of the main IU (+).

l'autre 'the other' (un autre 'another', deux autres 'two others', les autres 'the others' …)
but not l'un 'one', les uns '(*deux uns)
même si 'even if'
par hasard 'by chance' (par chance 'by chance')
par rapport à 'in relation to'
petit à petit 'little by little' (pas à pas 'step by step', face à face 'face to face')
plus jamais 'never again'
qui que ce soit 'whosoever'
sans doute 'without a doubt'
un peu 'a bit' (un petit peu 'a little bit', à peu près 'about/nearly', pour le peu que j'en sais 'however little I know about it'…)

Conjunctions of subordiantion are considered two words when there is commutation of *que:*
dès que je suis arrivé 'from the moment I arrived' ~ dès mon arrivée 'from my arrival' -> dès + que
alors que je partais 'whilst I was leaving' ≠ *alors mon départ '*whilst my departure' -> alors_que

A last example: Although it is possible to analyse *à peu près* 'about/almost' 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'…):



ça fait à peu près vingt-cinq ans que je suis dans ce laboratoire //

'it's about twenty five years that I'm in this laboratory'

Parts of speech

We take into account 13 parts of speech:

- V for verbs
- N for nouns
- Adj for adjectives
- Adv for adverbs
- Pre for prepositions
- CS for conjunctions of subordination

- J for junctors; traditional coordinating conjunctions and other elements that connect layers of pile relations, such as *c'est-à-dire* 'that is to say' or *y compris* 'including'. General extenders such as *et caetera* are also classed as junctors.
- D for determiners (including partitive *de*)
- 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
- Cl for clitics, including subject clitics (*je, tu il, on, ce*) 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, (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 unambiguous.

Remarks:

- Numerals are classed 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 classed 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 classed 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 classed amongst the N even if here it functions like an Adv.

Pro: 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 posessive 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, e.g. ***l'un*** 'one' (but not *l'autre* 'the other'), *il y en a* ***dix*** 'there are ten of them', *il en faut* ***un*** 'one is needed'.

  It is important not to confuse numbers representing nominal concepts and

pronouns. For example the following uses are not considered as pronouns but as nouns:

- bus numbers
- floor numbers
- school classes (*première* 'year 12 (sixth year of secondary school)', *seconde* 'year 11 (fifth year of secondary school)'
- arrondissements (*dans le* ***vingtième*** 'in the twentieth')
- football scores (***zéro*** *à* ***un*** '**zero** to **one**')
- page number (*page* ***cent douze*** *de votre ouvrage* 'page **one hundred and twelve** of your work')
- opus numbers (*deux préludes de Karol Szymanowski extraits de son opus* ***un*** *de mille neuf cents le* ***sixième*** *et le* ***septième*** *interprétés par…* 'two preludes by Karol Szymanowski from his opus one from nineteen hundred the **sixth** and the **seventh** interpreted by …')
- a figure as a figure (*ce chiffre de* ***trente mille*** 'this figure of three thousand')
- dates
- etc.

And the following as adjectives:

- a number of people (*tous les* ***deux*** 'both of them', *les* ***deux*** *ells passent…* 'both of them pass …', following our analysis of numbers above, even if there are good arguments to include them in the pronoun category. However *l'un* 'one' and *les uns* 'some of them' are treated as pronouns, since *l'un* is considered a single token and has has a different behavior from *les deux* or *les autres* 'the others'
- ordinal numbers (*le* ***premier*** 'the first', *la* ***deuxième*** 'the second')

Some examples of pronouns :

il { faut | faut } compter autour de { **soixante** | **soixante-dix** } // (D2009)
'you { should | should } expect around **{ around | sixty seventy }** (euros) //'

^ parce ^ que c' était une frange de { **vingt** | **trente** } // (D2009)
'^because it was a fringe of **{ twenty | thirty }** (centimeters) //'

corner à **deux** // 'corner kick with two (players) //'(D2003)

Morpho-syntactic features

Vs are assigned a modal feature which can take 6 values: *indicative, subjunctive, imperative, infinitive, past_participle, present_participle.* Only indicative Vs vary for the 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 and not the morphological level: a compound past is therefore a V *être* or *avoir* 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 and the passive for an ambiguous form such as *il est passé*.

Vs are also assigned agreement features: the feature *number* has deux values *sg* and *pl*, the feature *genre* deux 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 et 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”.

Lemmas

Lemmas are, as traditionally defined, the form for invariable lexemes, the infinitive form for verbs, the singular for nouns and the masculine singular for adjectives.

The lemma for the articles *le, la, l’, les* is *le,* the lemma for *un* and *des* is *un* and the lemma for *du*, *de_la* and *de_l’* is *du* (even if there are good arguments to assign the lemma *un* as is the case for *des*).

The lemma for the 1st and 2nd person pronouns *je, tu, nous, vous, me, te …*, is the form. The lemma for the 3rd person pronouns is the singular form; for exemple, *lui* for *eux* or *leur*, *elle* for *elle, il* for *ils*.

The lemma for the possessive determiners (*mon, ma, mes, ton, ta, tes …*) is always *son*.

The lemma for uncompleted words is uncompleted even if it is possible to reconstruct the word that the speaker was going to produce:

ils	sav~	ils	savaient	pas	ce	que	c'	était
Cl	V	Cl	V	Adv	Pro	Qu	Cl	V
il	sav~	il	savoir	pas	ce	que	ce	être

‘they didn’t kno~ they (didn’t) know what it was’

Analysis of proper nouns

We have already seen in the segmentation stage our decision to analyse the internal structure of words composed of several tokens. We repeat this choice for proper nouns, by analysing their internal structures if they are sufficiently productive.

Our corpus contains a number of titles of books, newspapers and institutions that contain several tokens, including proper nouns, common nouns, adjectives and grammatical words. When an individual token is not itself a proper noun, even if it belongs to a proper noun composed of several tokens, its lemma has the usual form. For example in

> " euh " je rappelle que votre livre { **Des épidémies | ^ et des Hommes** } vient de paraître aux Editions de la Martinière // (D2008)

> ‘erm I’ll add that you book { **Epidemies | ^ and Men }** has just been published by Les Editions de la Martinière //’

the title *Des épidémies et des Hommes* ‘Epidemies and Men’ corresponds to the lemmas ‘de+le (of the)’, ‘épidémie (epidemy)’, ‘et (and)’ and ‘homme (man)’. Note that the capital letter of *Hommes* is not transfered to its lemma, since it is not a proper noun in itself.

Newspaper titles also necessitate particular lemmatisation conventions.

> …répond Étienne Mougeotte dans **Le Figaro** (D2013)
> …écrivent **Les Dernières Nouvelles d' Alsace** (D2013)

The proper noun *Figaro* or *Libération* is treated as a proper noun and therefore keeps its capital letter in its lemmatised form, even if the word *Libération* is a common noun in other contexts. The article however is treated as a grammatical word and analysed according to the rules above. The lemmas corresponding to the second example are ‘le (the)’, ‘dernier (last)’ ‘nouvelle (news/word)’, ‘de (of)’, and ‘Alsace’.

Micro-syntactic dependency analysis

(Sylvain Kahane, Kim Gerdes)

Syntax describes the way in which linguistic units combine. Micro-syntax describes the relations between words that are characterised by a strong syntactic cohesion.

Government and micro-syntactic dependency

Government: Micro-syntax is limited to *government* relations. We speak of government (Fr. *rection*) when an element imposes on another element its nature, its markers and/or its position. 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 not 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*, lit. 'two euros that costs')).

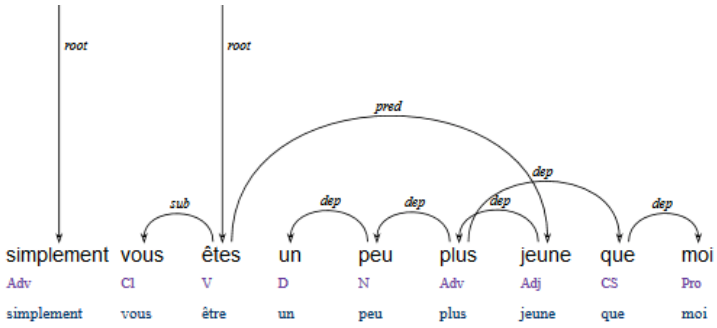
We use of the possibility of clefting as one of the major tests to characterise elements governed by a verb (*c'est le paysage que Pierre admire* 'it's the scenery that Pierre admires').

In :

> simplement < vous êtes un peu plus jeune que moi // (D0001)
> 'in short, you are a little younger than me'
>
> ^et "euh" donc < { ces | ces } années d'école < ça a été des bonnes années // (D0001)
> 'and erm, so those those schools years, they were good years'

the phrases *simplement* 'simply' and *ces années d'école* 'those schools years' are not dependent as they are considered to be non-governed, since they are not cleftable:

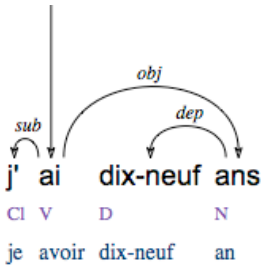
> *c'est simplement que vous êtes un peu plus jeune que moi
> '*it's in short that you are a bit younger than me'
>
> *c'est ces années d'école que ça a été des bonnes années
> '*it's those school years that they were good years'



simplement < vous êtes un peu plus jeune que moi //

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 phrase (*vous êtes jeune //+ plus que moi > en tout cas //* 'you are young //+ more so than me > in any case //').

Micro-syntactic dependency: We choose to encode the micro-syntactic structure by a dependency graph. Formally, a dependency is a directional relation between two words, which we represent by an arrow: the origin of the arrow is called the *governor* and the target the *dependent*. Each dependency represents a government relation. In the following example, *dix-neuf* modifies *ans*: we represent this by a dependency from *ans* (the governor) towards *dix-neuf* (the dependent):



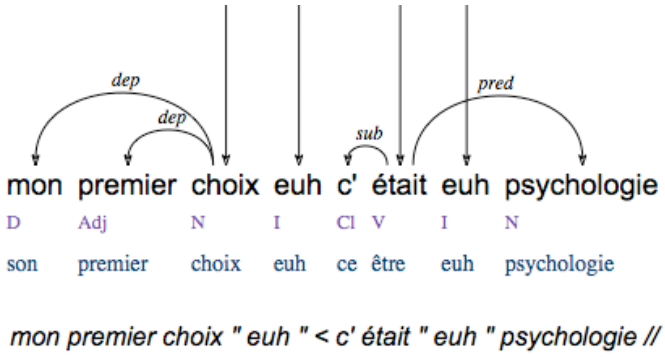
'I'm nineteen years old'

In the same way, *dix-neuf ans* 'nineteen years' is the direct object of the verb form *ai* 'have': we represent this by a dependency from *ai* towards the head of the phrase *dix-neuf ans*, i.e. *ans*. The head of the phrase is the lexeme that dominates the others in terms of dependency (see definition below). The verb *ai*, which is the main verb of this utterance, is therefore not governed. This is noted by a vertical dependency.

Government unit: A government unit (GU) is a maximal unit for government. A GU has a head, which is not governed, and all the elements of the GU are dominated by this head, i.e. 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 illocutory unit (IU) (cf. macrosyntax). An IU can be composed of several GUs. In the following example there are four GUs; the noun phrase

mon premier choix ‘my first choice’, the clause *c’était psychologie* ‘it was psychology’ and two occurrences of the discourse marker *euh* ‘erm’.



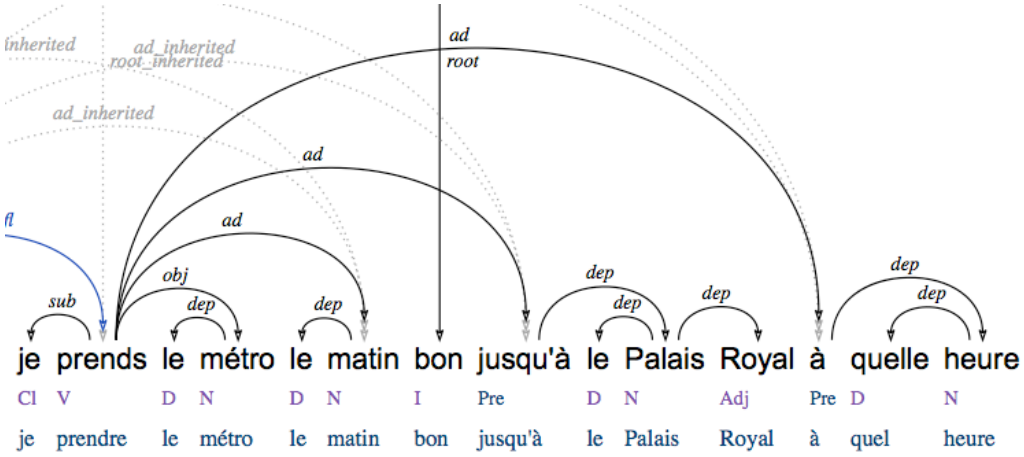
‘my first choice "erm" < it was "erm" psychology //’

Government beyond the Illocutory Unit and turn taking: Note that government does not necessarily stop at the end of a turn. In the following example:

$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)

‘$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 take the metro in the morning at eight thirty** //’

$L2’s question (*à quelle heure* ‘at what time’) continues the preceding microsyntactic construction (*je prends le métro le matin jusqu’à Palais Royal*) and $Ls’s reply (*je prends le métro le matin à huit heures et demie*) 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’).

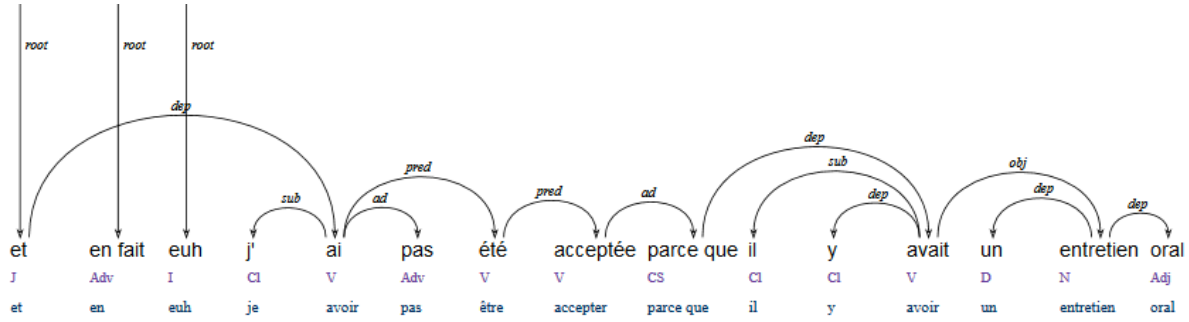


Dependence vs government: Our dependency structure is more restrictive than the dependency structure used in other theoretical frameworks (cf. for example Mel’čuk

1988, *Dependency Syntax*). Indeed, we only represent microsyntactic information by our structure, i.e. information relevant to government. Dependency is a formal means of representing different dependency relations. We could also have decided to represent microsyntactic information, such as the dependency of post or pre-nuclei to the nucleus. We have decided to encode micro and macrosyntax separately and to only use dependency for microsyntactic information.

Choice of the head: The *head* of a phrase is intuitively the most important element of the phrase. On the one hand it is the element that controls the distribution of the phrase (the external head) and on the other hand it is the element that validates the presence of the other elements of the phrase (the internal head). We shall now look at the main configurations which 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. In the following example *ai* governs *été* which governs *acceptée*:

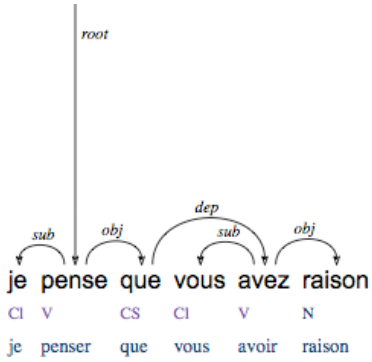


^ et en fait " euh " < j' ai pas été acceptée parce qu' il y avait un entretien oral //

'^and well "erm" < I wasn't accepted because there was an oral interview //'

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 enunciative modalities (***a-t-elle*** *été acceptée ?* 'was she accepted?'). Moreover, it imposes the case (i.e. a particular government marker) of the main verb (*elle a* ***accepté*** 'she was accepted' (past participle) vs *elle va* ***accepter*** 'she will accept' (infinitive)).

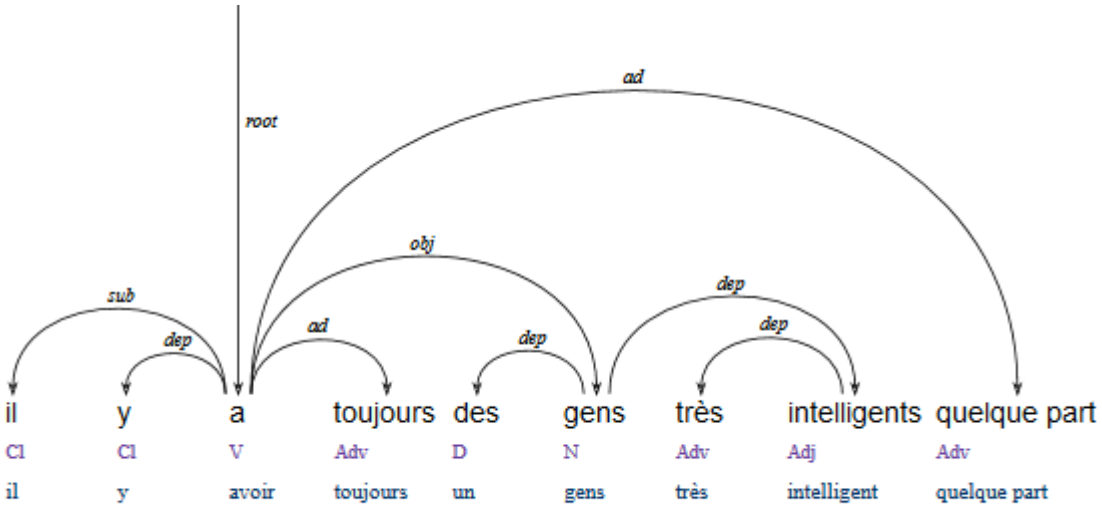
Markers (preposition, subordinating conjunction): Markers are generally treated as heads of the phrase that they mark, since they control its distribution. Therefore in *Pierre parle à Marie* 'Pierre speaks to Marie', *à* 'to' is the head of the phrase *à Marie* 'to Marie' and in *Pierre pense que Marie dort* 'Pierre thinks that Marie is sleeping', *que* 'that' is the head of the subordinate clause *que Marie dort* 'that Marie is sleeping'.



je pense que vous avez raison //

'I think you are right //'

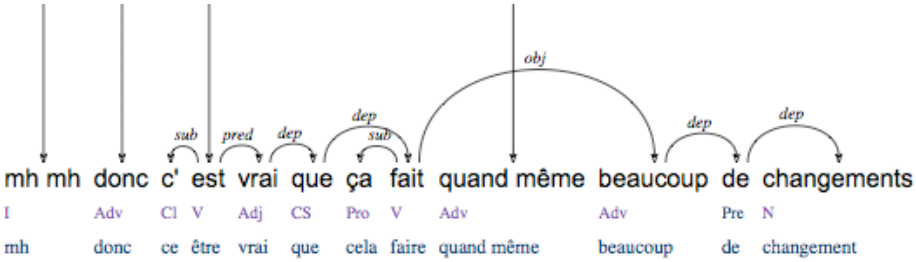
Determiner: We consider that in noun phrases, the noun is the head and governor of the determiner. This choice is partly arbitrary (the determiner also has the role of marker of the noun phrase, which would justify it being the head), but is the most common in dependency syntax.



il y a toujours des gens très intelligents quelque part //

'There are always very intelligent people somewhere //'

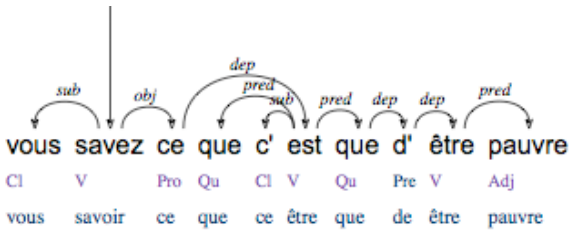
Complex determiner: Note that, in noun phrases of the form "Adv de N" (*peu de gens* 'few people'*, trop de gras* 'too much fat' ...), we consider the Adv to be the head:



'"mm mm" so < it's true that that makes actually a lot of changes'

Completive subordinate clauses and adjectives: We would also like to point out that in the previous example, constructions "c'est Adj que P" 'it's Adj that P' are analysed such that the completive clause "que P" 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 P est Adj), since on the surface the subject of *être* is the pronoun *ce*. It 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').

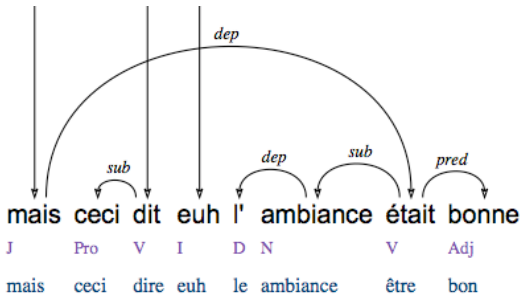
A borderline case is the following construction, where we analyse the two *que* as pronouns in the place of an Adj (*c'est dur d'être pauvre*):



'you know what it's like to be poor'

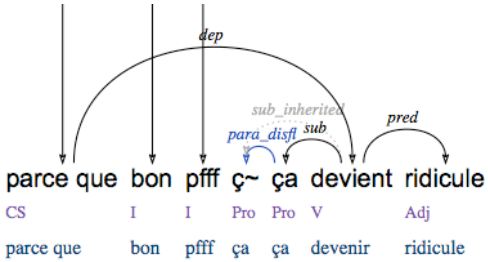
Our other choices are justified by studying the different constructions in turn.

IU initiator: Whether junctors (J) or subordinating conjunctions (CS), we treat IU initiators as the head of the IU, the head of the nucleus being treated as dependent on the initiator:



^ mais ceci dit " euh " < l' ambiance était bonne //

'^but having said that "erm" < the atmosphere was good //'



^ parce ^ que " bon " " pff " { ç~ | ça } devient ridicule //

'^because "well" "pff" { i~ | it } becomes ridiculous //'

The case of CSs which are IU initiators will be explained in the following paragraph.

Non-integrated subordinate clauses: We consider that in certain cases a subordinating conjunction may introduce a clause which is not strictly speaking "subordinating", when it fulfils the criteria for government. Compare:

(1) Pierre est à la fac parce qu'il a un article à finir
'Pierre is at uni because he has an article to finish'

(2) Pierre est à la fac parce que sa voiture est dans le parking
'Pierre is at uni because his car is in the car park'

These two examples apparently have the same surface structure *P1 parce que P2*. However the semantic relation between P1 and P2 is not the same in the two cases and neither are the resulting properties.

In (1), P2 causes P1. In this case, *parce que P2*, can be anteposed and clefted and *et cela* inserted:

(3) **a**. parce qu'il a un article à finir Pierre est à la fac
'because he has an article to finish Pierre is at uni'
b. *c'est* parce qu'il a un article à finir *que* Pierre est à la fac
'it's because he has an article to finish *that* Pierre is at uni'
c. Pierre est à la fac *et cela* parce qu'il a un article à finir
'Pierre is at uni and that is because he has an article to finish'

In (2), P2 does not cause P1, but the fact that the speaker thinks that P1. In this case, anteposition of *parce que P2* is only possible in the echoic structure:

(4) $L1 Pierre est à la fac, sa voiture est dans le parking
'Pierre is at uni, his car is in the car park'
$L2 et alors parce que sa voiture est dans le parking, Pierre est à la fac ?
'and so because his car is in the car park, Pierre is at uni?'

It cannot really be clefted:

(5) #c'est parce que sa voiture est dans le parking que Pierre est à la fac
'it's because his car is in the car park that Pierre is at uni'

even if it can be clefted with negation:

(6) c'est pas parce que sa voiture est dans le parking que Pierre est à la fac
'it's not because his car is in the car park that Pierre is at uni'

Nevertheless (6) must be contrasted with (7) formed from (1):

(7) c'est pas parce qu'il a un article à finir que Pierre est à la fac
'it's not because he has an article to finish that Pierre is at uni'

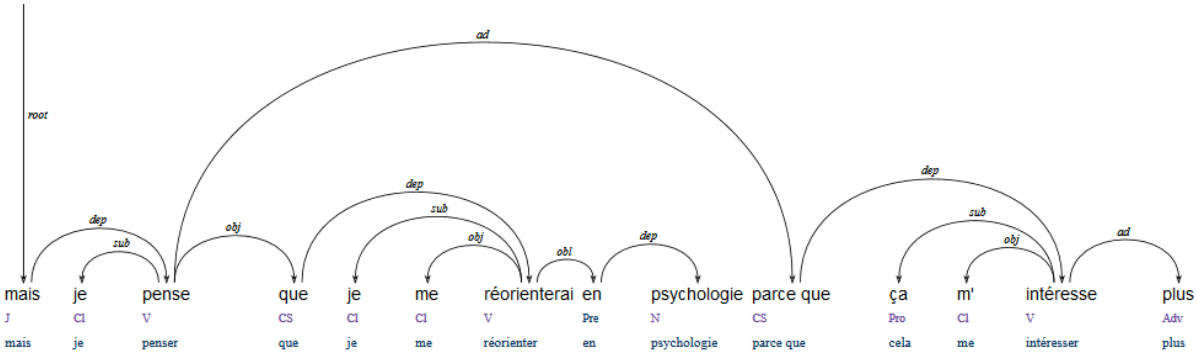
Indeed (6) does not entail P1 (we could say that if we thought that Pierre was not at uni and it is most likely to be the case for (6)), whereas (7) presupposes P1 (what is being discussed is the cause of P1 and therefore it only has meaning if P1 takes place).

Finally *et cela* cannot be added, but *et je pense cela* can:

(8) **a.** #Pierre est à la fac *et cela* parce que sa voiture est dans le parking
'Pierre is at uni and that is because his car is in the car park'
b. Pierre est à la fac *et je pense cela* parce que sa voiture est dans le parking
'Pierre is at uni and I think so because his car is in the car park'

In conclusion, we consider that in the case of (1), the subordinate *parce que P2* is a modifier of the verb of P1 and therefore is part of the same GU:

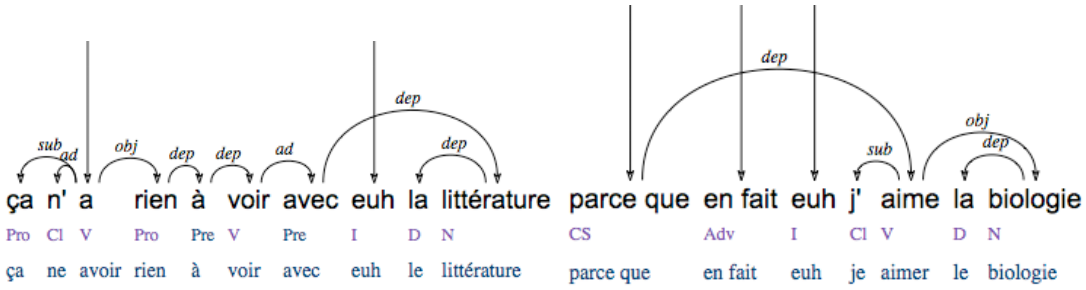
(9) ^mais je pense que je me réorienterai en psychologie parce que ça m'intéresse plus // (M1001:29)



'^but I think that I will shift to psychology because that interests me more //'

In the case of (2), we decide not to mark the goverment of the subordinate *parce que P2*:

(10) c'est un bac "euh" { SMS | donc technologique } // c'est sciences médico-sociales // ça n'a rien à voir avec "euh" la littérature // ^parce ^qu' en fait < "euh" j'aime la biologie // (M1001:7)



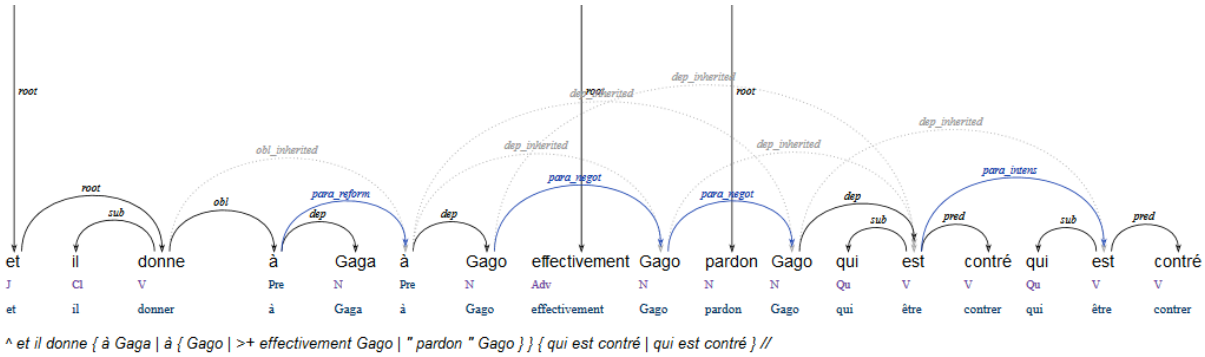
'that has nothing to do with "erm" literature // ^because well < "erm" I like biology //'

In both cases, *parce que* is categorised as a subordinating conjunction and governs the main verb of the subordinate clause.

Government beyond the IU: Note that we do not consider the illocutory unit (IU) or turn-taking as boundaries of the GU. For example, the following exchange over four turns gives a single GU (interspersed with other GUs such as *effectivement* and *pardon*):

$L1 ^et il donne à Gaga //+
$L2 à Gago >+ effectivement //+
$L1 Gago "pardon" //+

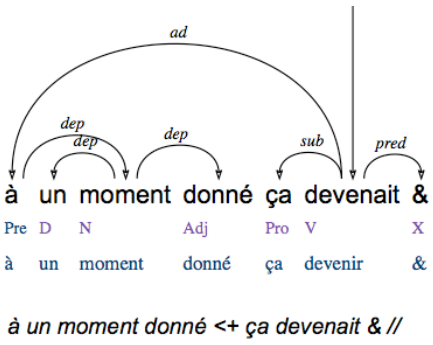
$L2 Gago qui est contré | qui est contré //



‘^and he passes { to Gaga | to Gago | >+ indeed Gago | "sorry" Gago } } { who is blocked | who is blocked } //’

We will present the analysis of *qu-* words (relative clauses, clefting …) after the syntactic functions and the analysis of pile constructions (coordination, reformulation …) in a separate chapter.

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



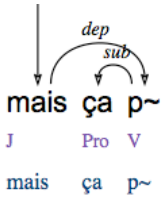
‘at one <+ point it was becoming & //’

{ en vingt-cinq | en vingt-cinq } ans <+ les gens & // il y a vingt-cinq ans <+ les gens ne connaissaient rien "hein" //
‘{ in twenty five | in twenty five } years <+ people & // five years ago <+ people don’t know anything "eh" //’

$L1 ils savaient pas travailler un & // { ils sa~ | ils savaient } pas utiliser un ordinateur //
‘they didn’t know how to work a & // { they didn’t kno~ | they (didn’t) know } how to use a computer’

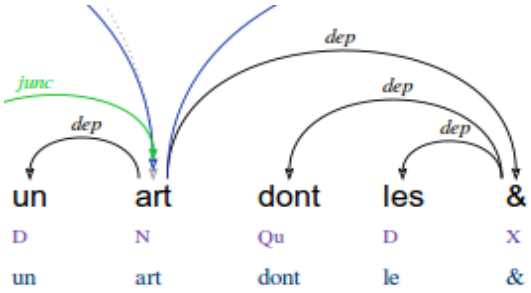
When there is already an incomplete word, we do not also indicate syntactic incompletion:

^mais ça p~ // c' est pas obligatoire //
‘but that ca~ // it’s not necessary //’



We limit the marking of incompletion to a single ampersand each time, even in cases where, as in the following example, we could say that a noun and a verb are expected for syntactic completion (*art* would be linked to the unpronounced verb of the following relative proposition and *les* to the verb's unpronounced subject noun):

c' est un philosophe " euh " américain " euh " (+ disciple du philosophe anglo { aus~ | autrichien } " euh " Wittgenstein) qui a " euh " avancé cette idée de { l' art comme concept flou | ^c'est-à-dire { un art dont les & | un a~ } } //
'it's an "erm" American "erm" philosopher (+ disciple of the Anglo-aust~ | Austrian philosopher } "erm" Wittgenstein) who "erm" put forward this idea of { art as a fuzzy concept | ^that ^is { an art whose & | an ar~ } } //'



Instead we link both dependents to the same ampersand, which therefore represents the two unpronounced positions.

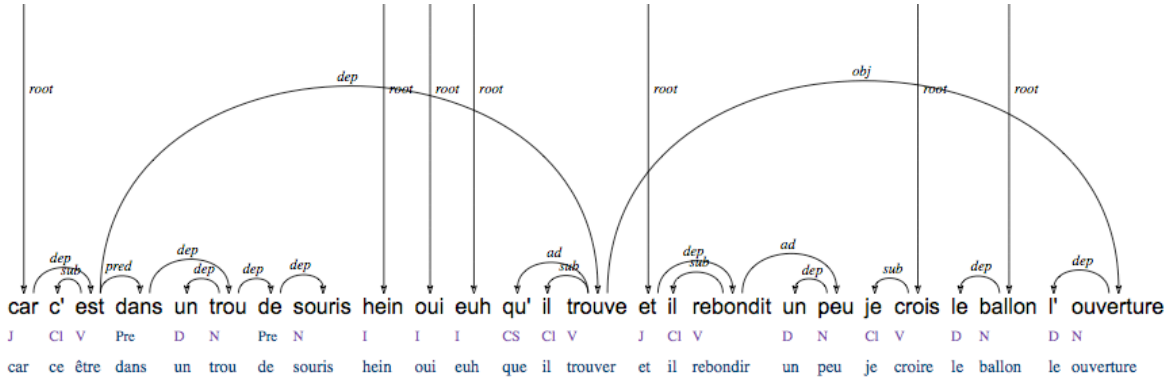
Syntactic functions

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

- to rephrase 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 dependent 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. Only the dependents of the verb have been distinguished and 7 functions considered:

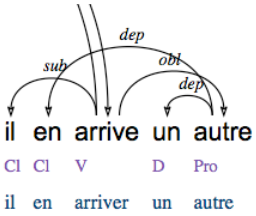
- root: for all roots, i.e. elements which are not governed by another element:



^ car c' est dans un trou de souris " hein " (oui //) " euh " qu' il trouve (^ et il rebondit un peu " je crois " > le ballon //) l' ouverture //

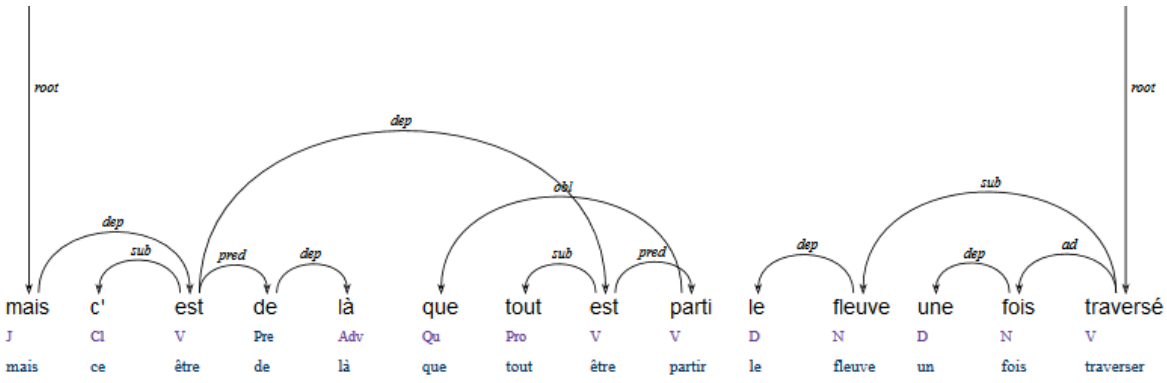
‘^since it’s a tiny hole "eh" (yes //) "erm" that he finds (^and he comes back a bit "I think" > the ball //) the opening //’

- sub: for the subject of a verb;
  - the subject is the grammatical subject:



‘there arrives another one’

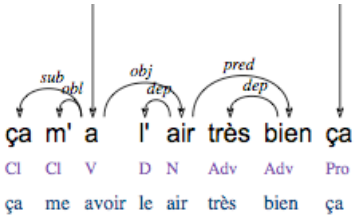
  - the function sub is also used for predicative constructions. In <u>*les mains*</u> *sur la tête il rigolait* ‘his hands on his head, he was laughing’, *les mains* ‘his hands’ is the subject of *sur la tête* ‘on his head’ and there is therefore a subject dependency of *sur* to *mains* (cf. a similar example below with *le fleuve une fois traversé* = une fois que le fleuve a été traversé) :



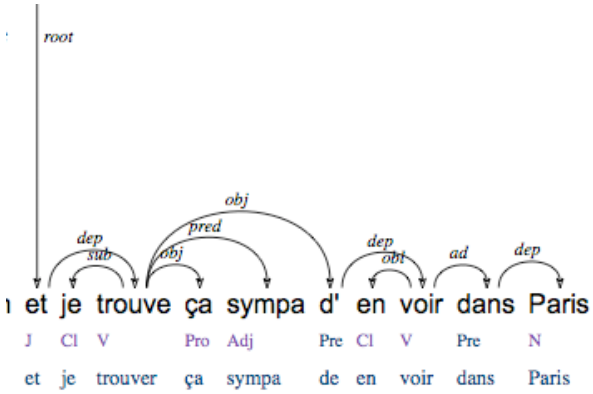
^ mais c' est de là que tout est parti > le fleuve une fois traversé //

‘^but it’s from that everything started > the river once crossed //’

- obj: for the direct object:
    - we include measurements when there can be cliticised (no examples in the corpus): *il a payé <u>dix euros</u> pour ce livre* 'he paid <u>ten euros</u> for this book' (*il les a payé pour ce livre* 'he paid that for this book'), mais pas *il a payé ce livre dix euros* (**il les a payé ce livre*)
- pred: for all the elements that form a complex predicate with the verb they govern:
    - constructions known as predicative complement of the subject (*il est <u>gentil</u>*) or of the object (*il trouve Marie <u>gentille</u>*):

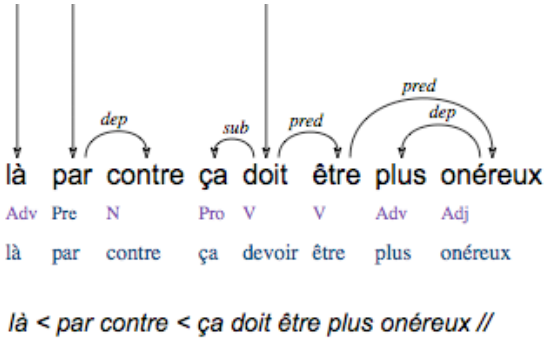


'that seems very good to me'



'and I find it nice to see them in Paris'

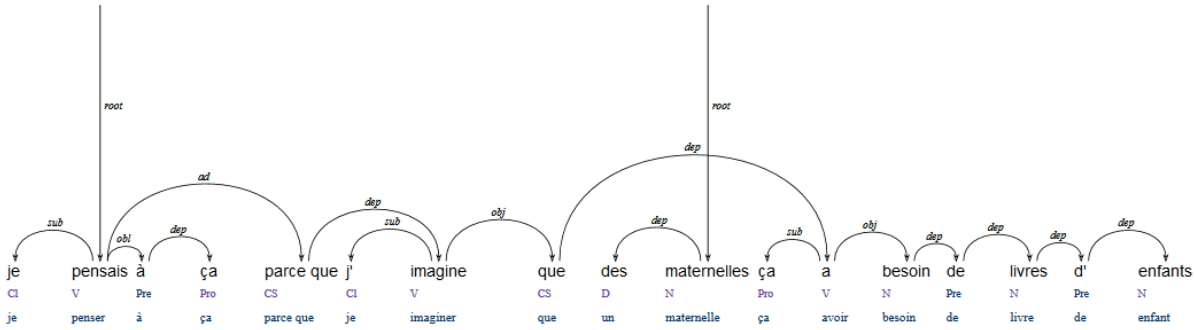
- complex verb forms (*avait mangé* 'had eaten', *est parti* 'has left')
- fixed verb forms such as *vouloir dire* (only when the meaning is 'to mean')
- Constructions with a modal verb (*peut venir* 'can come', *doit manger* 'has to eat'), where the infinitive does not easily commute with a noun phrase:



là < par contre < ça doit être plus onéreux //

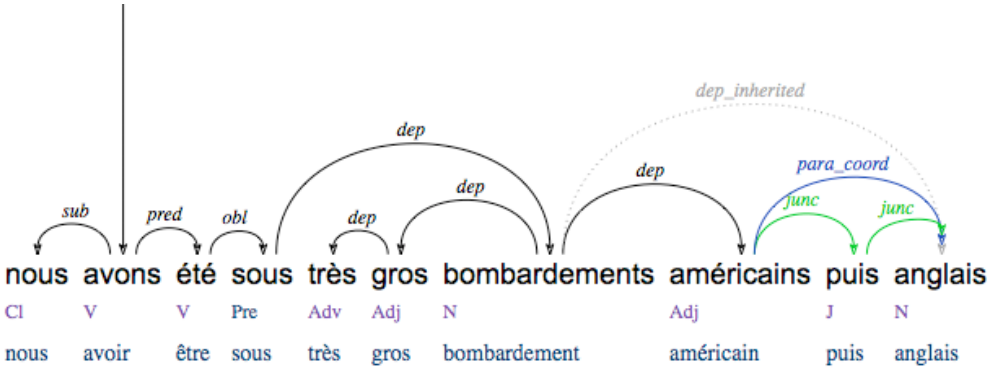
'there < on the other hand < it must be more expensive //'

- except constructions with a supporting verb (*avoir l'intention* 'have the intention', *avoir besoin* 'need', *faire peur* 'scare' ...), where the predicative noun is treated as a direct object:



'I thought of that because I imagine that nurseries need children's books'

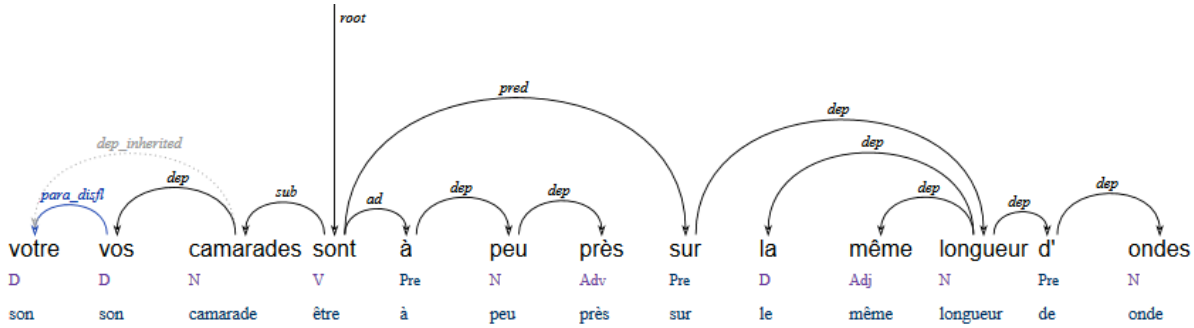
- obl: for all indirect objects, i.e. subcategorised complements of the verb which are not obj. Prepositional complements which are part of a fixed expression (*mettre <u>en doute</u>* 'put in doubt') are also treated as obliques. Locative constructions with *être* 'be' are also treated as obl:



'we were under very heavy American, then English fire'

Note that it is only an *obl* relation when the locative can be questioned via the term *where.* Therefore, for *être* + *sur/sous/en* where the meaning is metaphorical, *être* governs a following preposition by a *pred* relation:

votre vos camarades **sont** à peu près **sur la même longueur d' ondes**



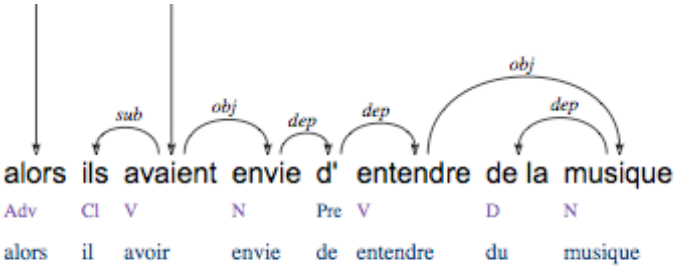
{ votre | vos } camarades sont à peu près sur la même longueur d' ondes //

'{ your | your } comrades are practically on the same wavelength //'

This indication was developed post-annotation and may therefore not have been fully respected by all annotators.

- ad: for adjuncts to the verb, i.e. non-subcategorised complements.
- dep: all dependents of non-verbal forms.
    - some elements dependent on the verb cannot be considered as adjuncts or as sub-categorised elements. We can identify two types: elements belonging to a fixed verb form (clitic: se souvenir 'remember', en avoir marre 'be fed up', il y a un problème 'there is a problem'); some subordinate clauses (see below, clefting).

Complements of verbal expressions are treated as dependents of the predicative element (Adj or N), but receive a function as dependent of the verbal construction:



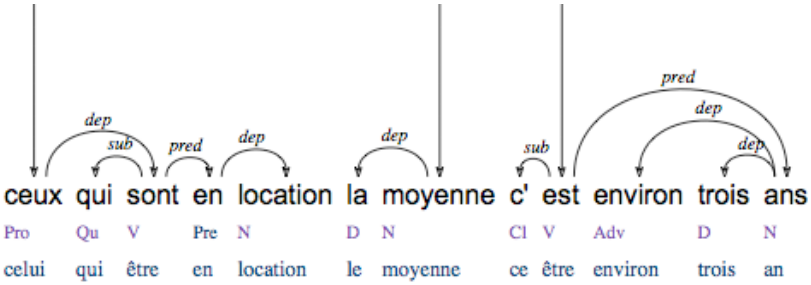
alors < ils avaient envie d' entendre de la musique //

'so < they wanted to hear some music //'

Extraction and qu- words

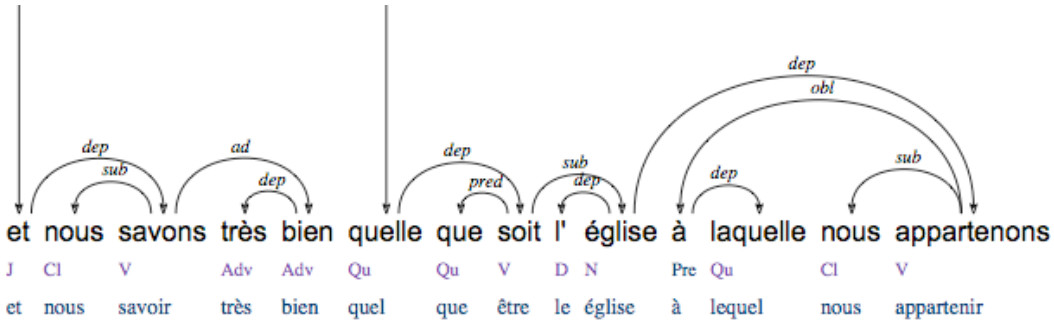
Relative: It is possible to consider, following Tesnière (1959) for example, that in relative phrases, the relative pronoun's role is as a pronoun within the relative clause and as a complementiser allowing the relative clause to modify a noun. This analysis implies that the relative pronoun should occupy a double syntactic position, as head of the clause (as complementiser) and dependent within the clause (as pronoun). Certain theories even go as far to say that certain *qu-* word, notably *que* in relative clauses, are above all complementisers.

In our analysis we only encode the position of the pronoun within the relative clause:



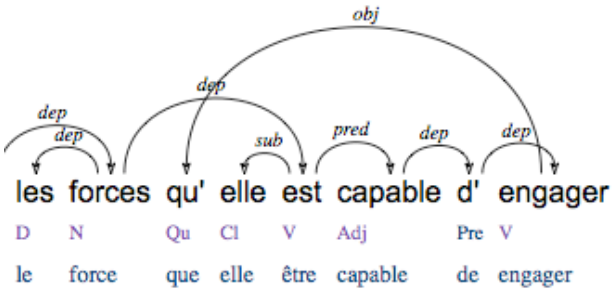
ceux qui sont en location < la moyenne < c' est environ trois ans //

'those who rent < the average < it's about three years //'



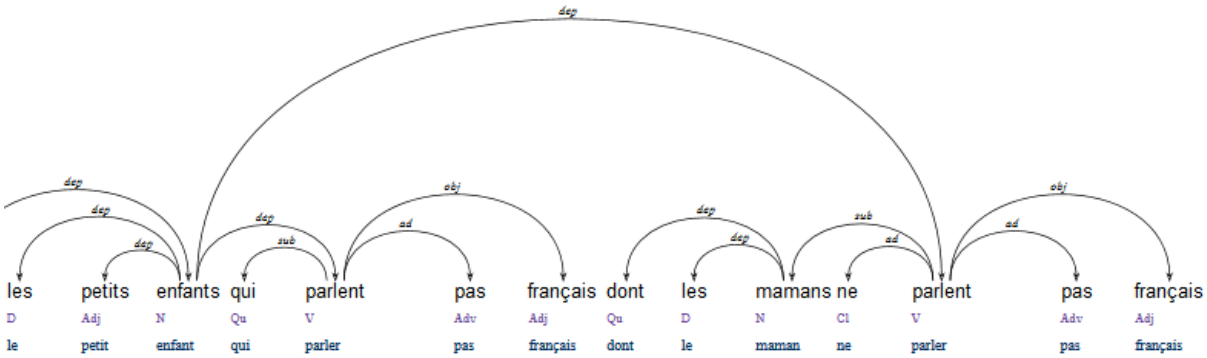
'and we know well whatever church we belong to'

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 easily 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, i.e. 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:



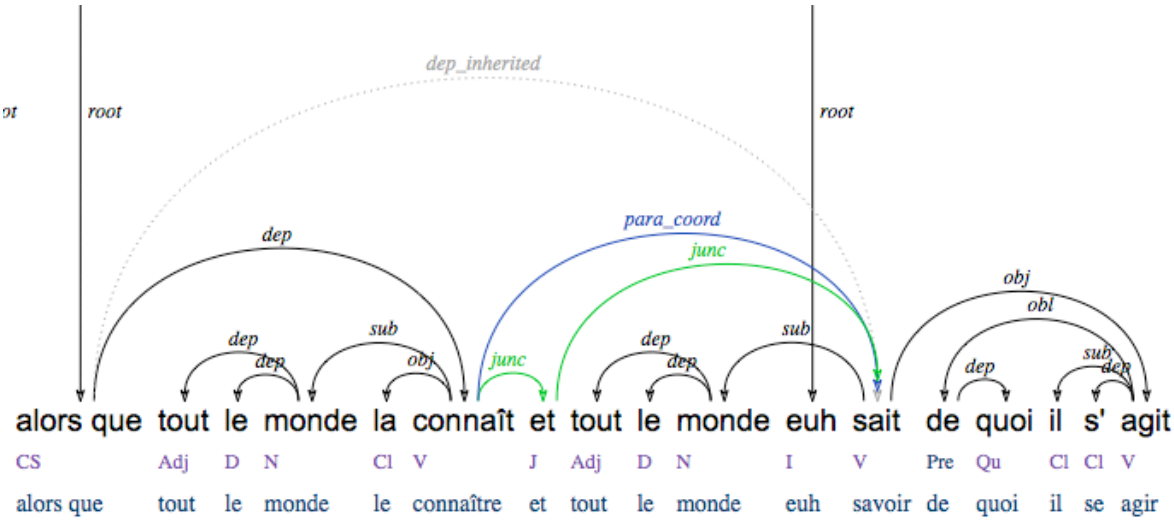
'the forces it is capable of taking on'

The analysis is extended to all relative pronouns. In the following example, *dont* is analysed as the dependent of *mamans*, given that the equivalent declarative expression would be *les mamans des enfants*.



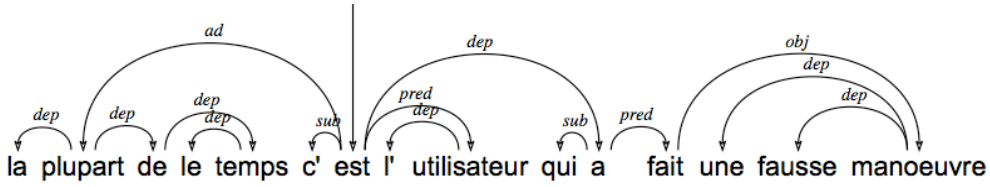
'the small children who don't speak French whose mums don't speak French'

Interrogative: The same remarks as for the relative pronouns can be made for the interrogative pronouns, notably in indirect questions. The same decision is made to attribute the interrogative pronouns a single position as a pronoun.

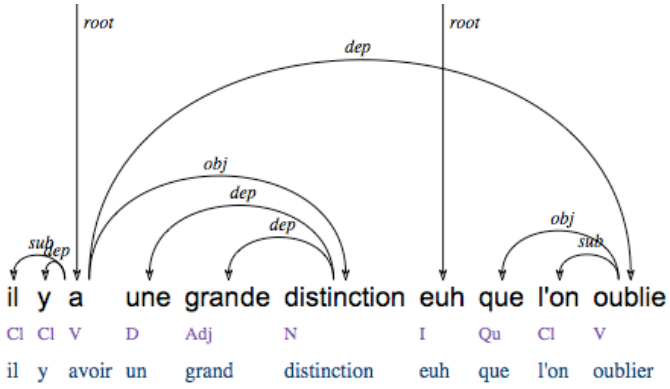


'even though everyone knows her and everyone erm knows what it's about'

Clefting: For cleft constructions of the form "c'est X qui/que P" (*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, i.e. the verb *être* rather than the noun phrase extracted by clefting:



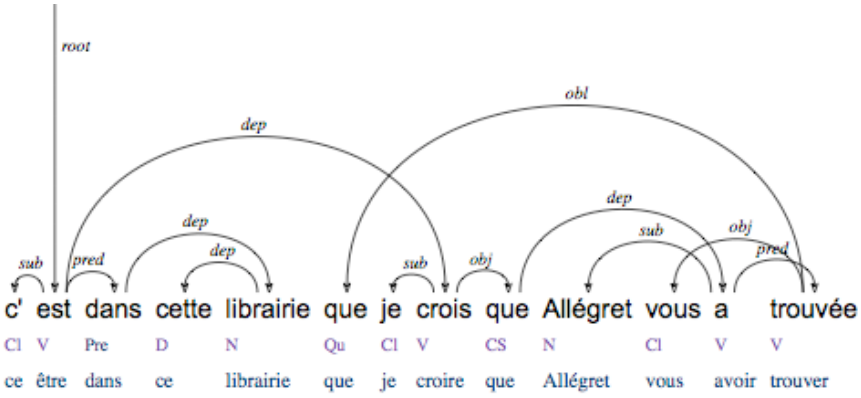
'most of the time it's the use who made a mistake'



'there is a big distinction erm that we forget'

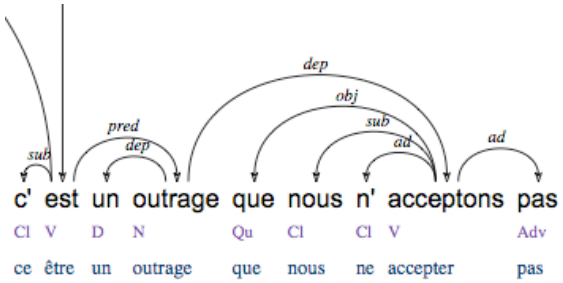
Moreover, given the particular nature of the relation between the cleft construction and the subordinate clause, we assign it the function *dep*, whereas the extracted element has the function *pred* (the idea is that the underlying construction is *ce qui/que X est P: celui qui a fait une fausse manoeuvre est l'utilisateur* 'the one who made a mistake is the user').

This analysis is also extended to the case where the extracted element is an indirect objet or an adjunct. Although in these cases, even more so than before, there are good reasons to consider that the *qu* word is a complementiser rather than a pronoun, we prefer this encoding which has the advantage of marking the governor of the extracted element and especially illustrating to long distance extractions.

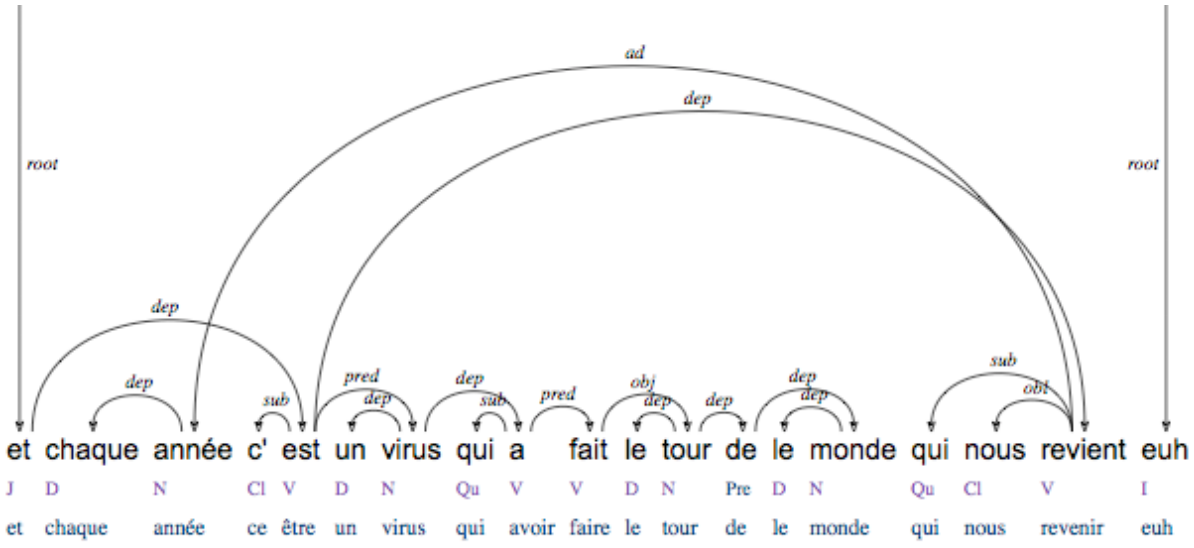


'it's in that library that I think Allégret found you'

Our representation of cleft constructions enables us to encode the difference in structure between cleft constructions and those where an attributive noun phrase possesses a relative clause:



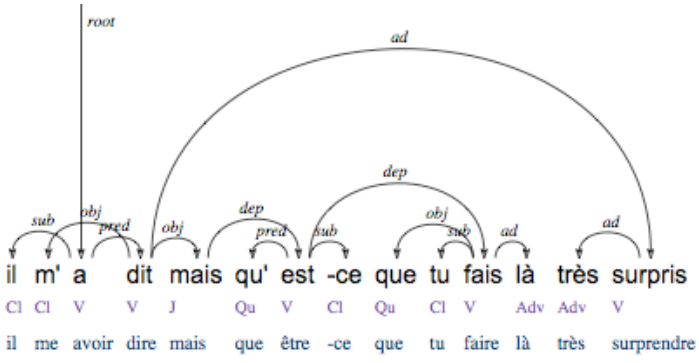
‘it’s an outrage which we will not accept’



^ et chaque année <+ c' est un virus qui a fait le tour du monde qui nous revient " euh " //

‘^and each year <+ it’s a virus that has travelled round the world that comes back to us, “erm” //’

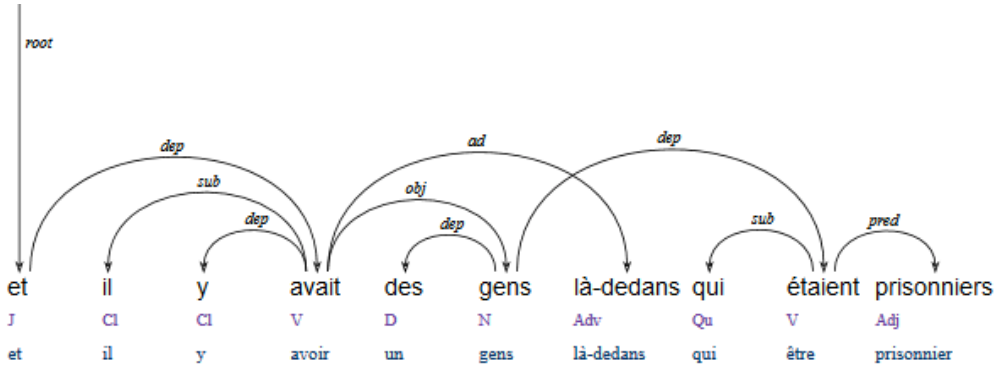
The analysis is extended to interrogative cleft constructions of the form *qu’est-ce que*, bearing in mind that it is the interrogative form of a cleft constructions (c’est quoi que tu fais là → qu’est-ce que tu fais là):



il m' a dit [^ mais qu' est-ce que tu fais là //] très surpris //

‘he said to me [^but what are you doing there //] very surprised //’

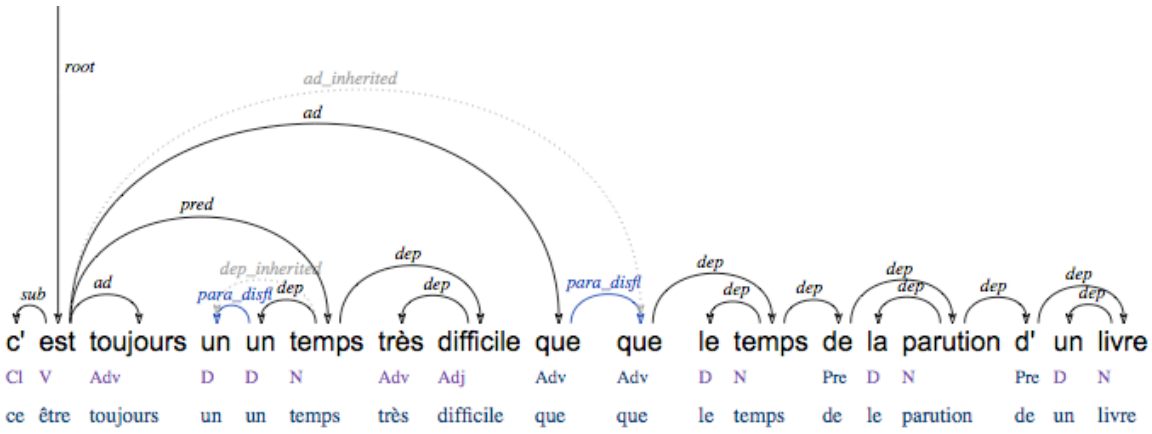
However, in presentational constructions such as “il y a X qui P” ‘there are X who P’, we treat the subordinate clause “qui P” as a relative clause, dependent of X:



^ et il y avait des gens là-dedans qui étaient prisonniers //

'^and there were people in there who were prisoners //'

Atypical constructions: We have evidently also come across certain constructions for which we have tried to propose analyses without being able to link them to our previous choices. In the following extract for example, the role of *que* is unusual:



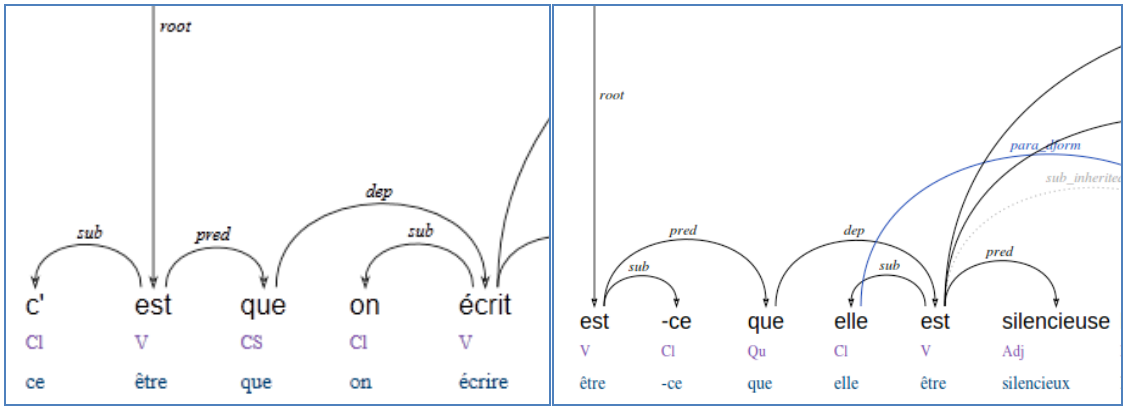
c' est toujours { un | un } temps très difficile { que | que } le temps de la parution d' un livre //

'it's always { a | a } very difficult time the moment of a book's publication // '

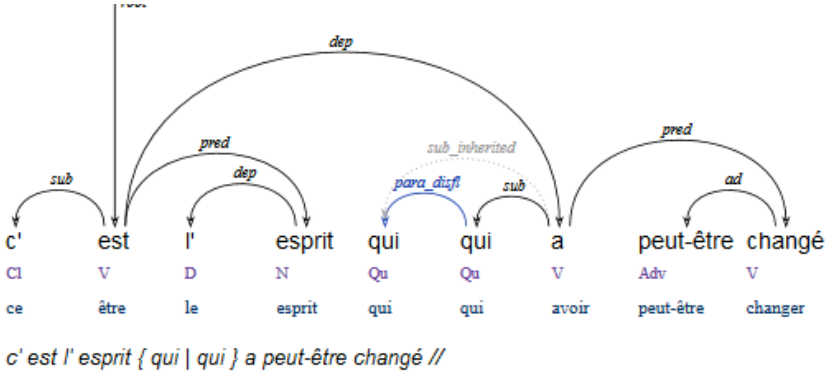
Interrogative constructions

Yes/no questions and Qu- questions: The particular treatment of cleft constructions and subordinate clauses means that the *que* of *est-ce que* and the second *que* of *qu'est-ce que* do not have the same function in relation to the head of the interrogative phrase. Their structures are analogous to their equivalent declarative structures:

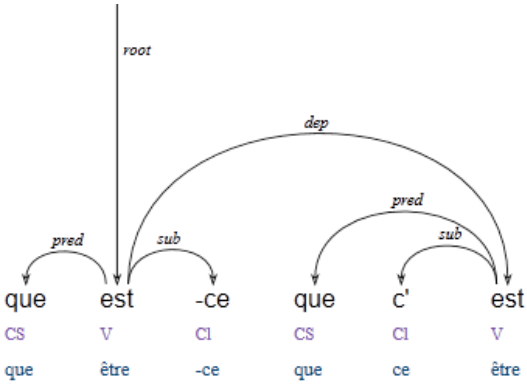
est-ce que X a fait… ?	->	*c'est que X a fait*
'Did X do…?	->	'it's that X did…'



qu'est-ce que X a fait ? -> *c'est quoi que X a fait*
'what did X do? -> 'it's what that X did'

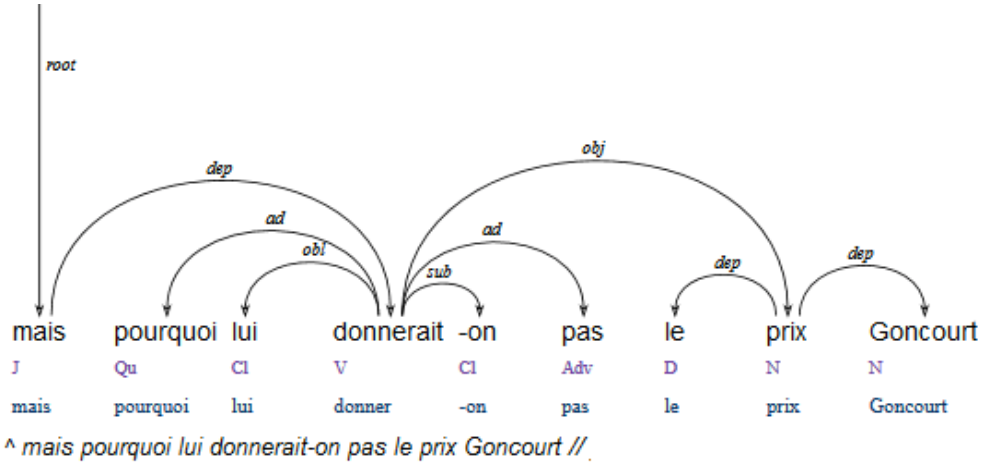


'it's the spirit { that | that } has possibly changed //'



'what is it ?'

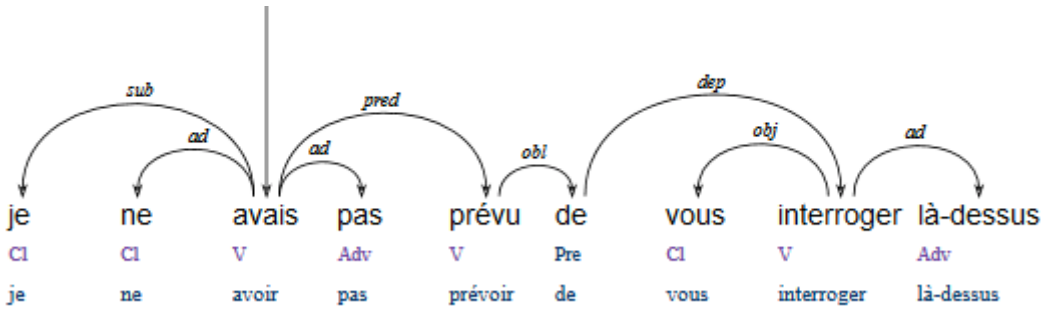
Qu- words: Despite the fact that Qu- words such as *pourquoi* 'why', *comment* 'how' etc. could be considered to be the head of an interrogative phrase, our analysis considers them adjuncts to the main verb. *Pourquoi* has the same role as *parce que* 'because', *comment* describes the way in which something is done, *quand* 'when' situates an action in time etc.



^'but why not give her the Goncourt prize //'

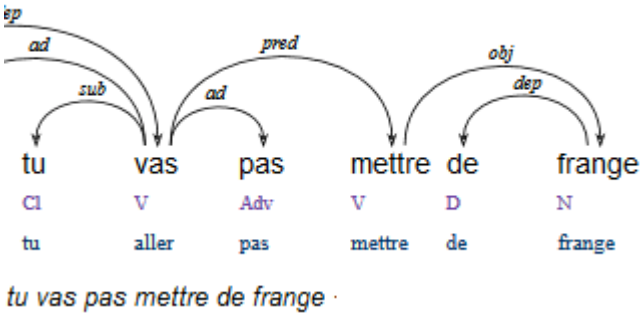
Negation

Verbal negation: The clitic *ne* and the adverbs of negation *pas* 'not', *jamais*' never', point 'not', *guère* 'hardly' etc. are governed by the first verb form of a complex verb form and are considered to be linked as adjuncts of this verb. Therefore, in the case of a verb in the compound past, the *ne* and *pas* are adjuncts of the auxiliary:



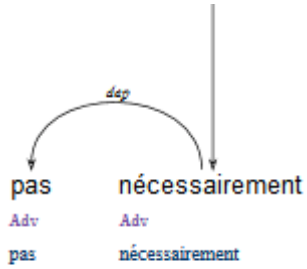
'I had not planned to question you on that'

Pas de X: Although there are good reasons to consider *pas de X* as being a complex determiner (see *beaucoup de…* 'lots of', *peu de…* 'few'), *pas de X* does not always function as a cohesive unit, since it can be separated by adverbs ou verb forms. Therefore *pas* is always considered the adjunct to the verb and the noun the object. The partitive *de* is treated as the determiner:



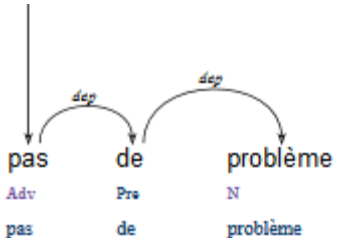
'you are not going to add a fringe'

Averbal negation: In cases of averbal negation (Ex: *pas nécessairement*), the head of the phrase is the negated element and the negation is dependent on this element:



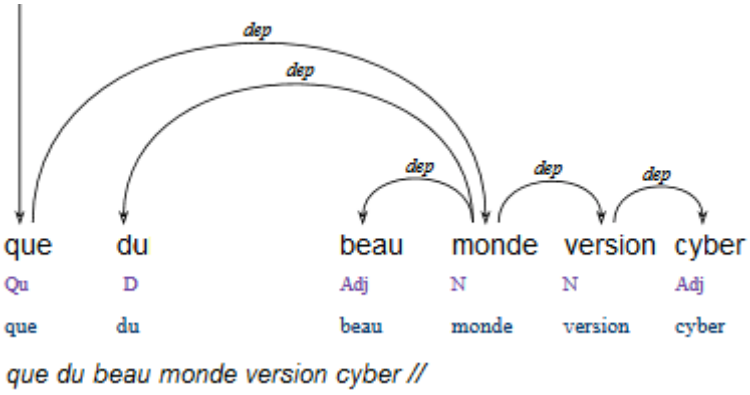
'not necessarily'

except when there is an isolated noun phrase of the form *pas de X*, where the alternative solution described above is used, the adverb of negation being treated as the head of the phrase. We consider this decision more appropriate since *pas* has not verb to which it can attach, even if we acknowledge that this creates a double (and potentially incoherent) analysis of *pas de X*.



'no problem'

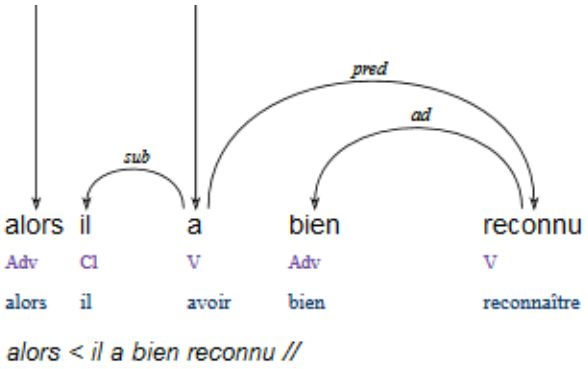
Restrictive *que*: *que* is treated like other adverbs of negation (adjunct to the verb in verbal negation and root in averbal negation of the form *que de X*):



'just A-list cyber-style //'

Adverbs

We analyse adverbs generally as adjuncts to verbs. In the case of a complex verb form such as the compound past, adverbs (except adverbs of negation which are described above) are linked to the full verb and not to the auxiliary:

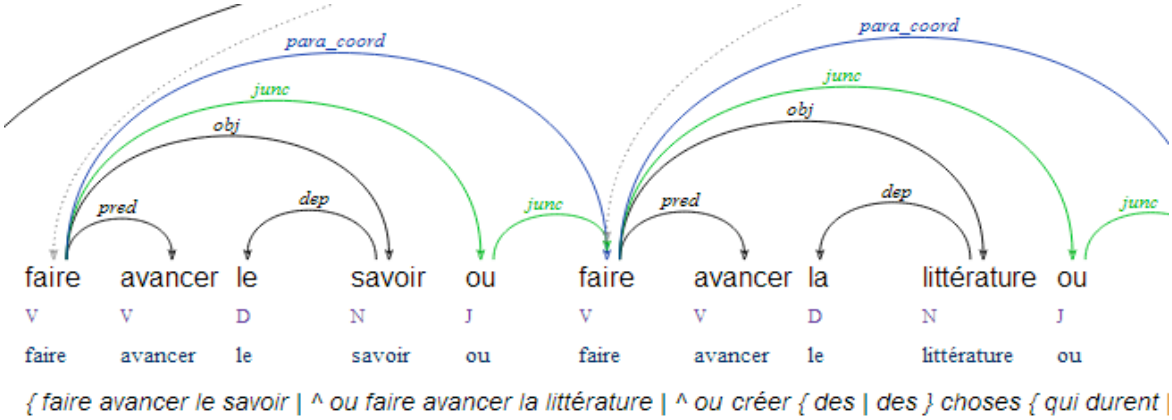


'so < he did acknowledge it //'

Causative constructions

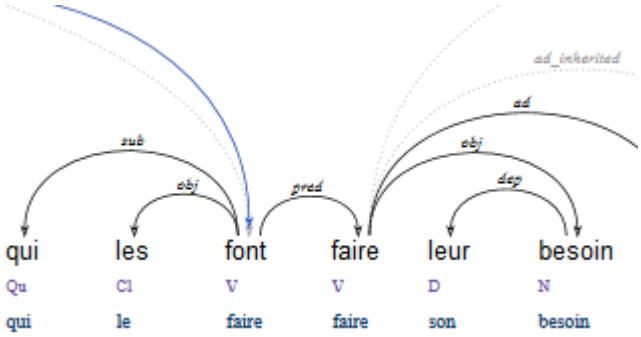
In causative constructions, a noun can simultaneously be the object of the verb *faire* and the subject of a following infinitive. Our choice of analysis is to exclusively mark the *obj* relation between the noun and the verb *faire*, the second verb being dependent on the verb *faire* by a *pred* relation.

...**et faire avancer le savoir ou faire avancer la littérature** ou créer des des choses qui durent et qui servent euh m' a paru être essentiel
'**... and furthering knowledge furthering literature** or creating new things which last and which are useful erm seemed to me to be essential'



The choice means that an infinitive can never be associated with a subject and this is justified by the fact that the demoted subject is cliticised in front of *faire*. In the following example, we even have the case of an object clitic, even though the infinitive already has an object:

{ qui | qui | qui } prennent leur chien | qui **les font faire** leur besoin { { dans | dans } le trottoir |}
'{ who | who | who } take their dog | who **have them** do their business { { on | on } the pavement |}'



Pile constructions

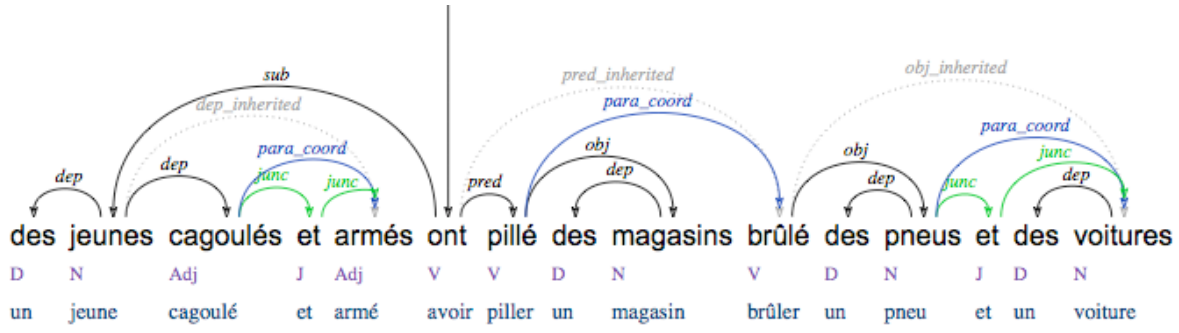
(Sylvain Kahane, Kim Gerdes, Paola Pietrandrea)

Pile constructions are constructions characterised by the fact that several elements occupy the same governed position. This is true of coordination constructions:

1. des jeunes **{ cagoulés | ^et armés }** ont **{ pillé des magasins | brûlé { des pneus | ^et des voitures } }** // (M2006)
   '**{ hooded | ^ and armed }** youths **{ looted shops | burned { tyres | ^ and cars } }**'

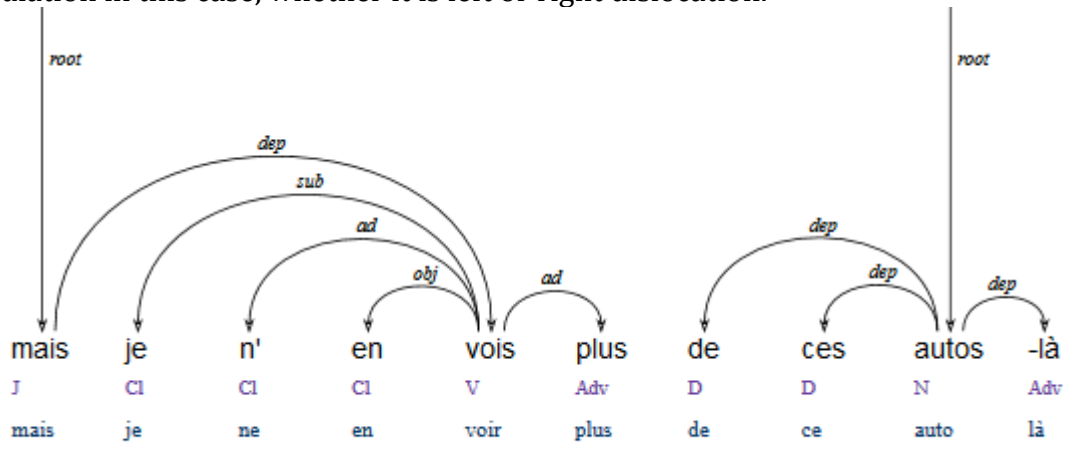
but also of many other phenomena, which we will present here, starting with disfluencies and reformulations such as:

2. et je voulais pas aller à Addis Abeba // puisque **{ { { les | les | les | les } c~ | les capitales } | les grandes villes }** ne me disaient rien du tout // (D2004)
   'and I didn't want to go to Addis Abeba // since **{ { { the | the | the | the } c~ | capitals } | the big towns }** didn't interest me in the slightest//'



des jeunes { cagoulés | ^ et armés } ont { pillé des magasins | brûlé { des pneus | ^ et des voitures } } //

No double marking: despite the fact that double marking could be considered as a double formulation of a noun and clitic form, we decide not to indicate double formulation in this case, whether it is left or right dislocation.



^ mais je n' en vois plus > de ces autos-là //

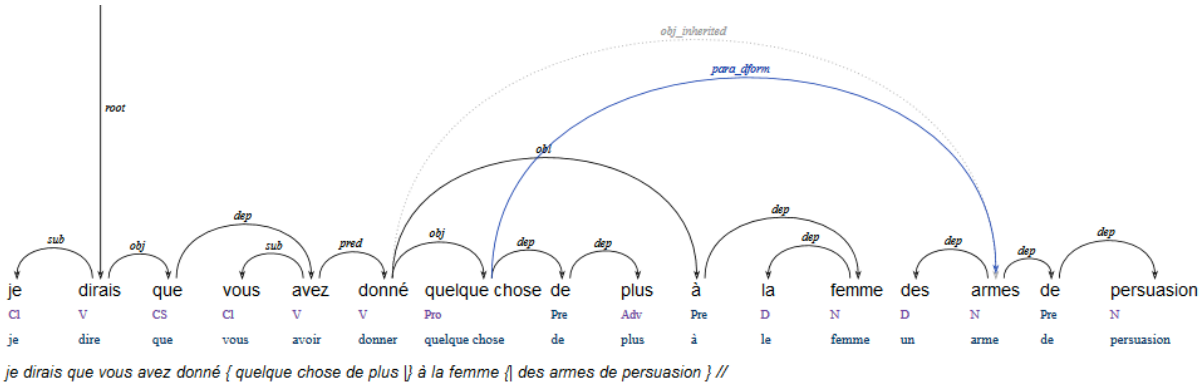
'^ but I don't see them anymore > those cars //'

Paradigmatic link

A pile is a dimension orthogonal to government: a segment Y is said to pile onto a segment X if Y occupies the same position governed by X. We note the fact that X and Y are in a pile relation by { X | Y }. In the pile { X | Y }, X and Y each represent a *layer* of the pile. A pile can have two or more layers.

Discontinuous pile: A pile is discontinuous if the layers of the pile are not adjacent. In this case we use the notation { X |} … {| Y }:

> si je ne craignais pas d'entrer dans le jeu de certains hommes qui abusent de leur condition < je dirais que vous avez donné **{ quelque chose de plus |}** à la femme //+ **{| des armes de persuasion }** // (D2001)
> 'if I wasn't scared of being drawn into the game of certain men who take advantage of their condition < I would say that you have given **{ something extra |}** to the woman //+ **{| weapons of persuasion }**'

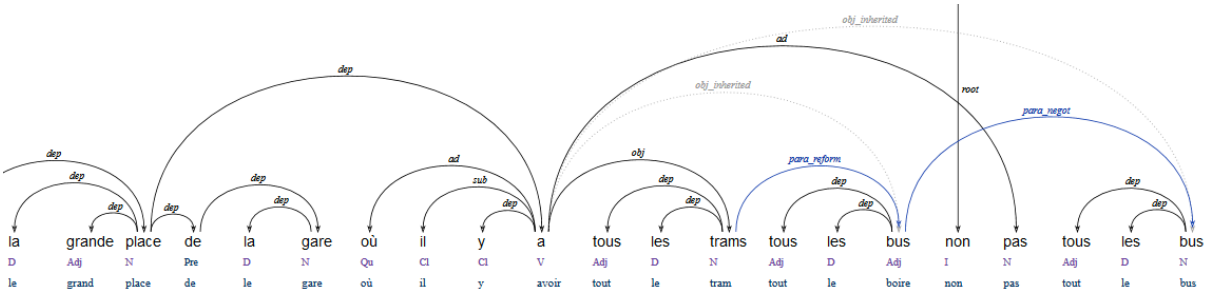


je dirais que vous avez donné { quelque chose de plus |} à la femme {| des armes de persuasion } //

Conjuncts: The layer of a pile contains several types of elements. The layer's central element is called the conjunct. Each conjunct can only occupy the place occupied by the pile:

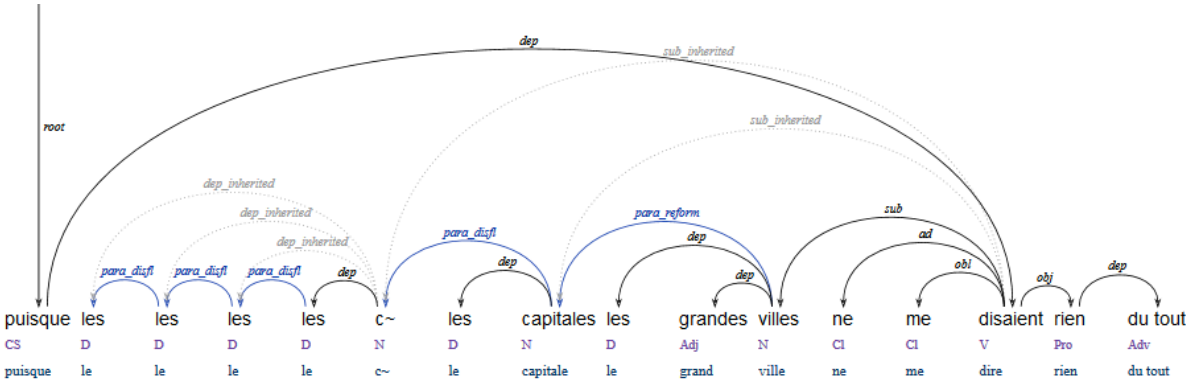
> des jeunes { **cagoulés** | ^et **armés** }
> des jeunes **cagoulés**
> des jeunes **armés**
>
> vous avez donné **quelque chose de plus** à la femme
> vous avez donné **des armes de persuasion** à la femme

Paradigmatic Link: Conjuncts are linked by a paradigmatic link. Each conjunct is linked to the preceding conjunct, regardless of the number of layers, including when there are nested piles (see below). Paradigmatic links are noted *para* followed by a type (see below for classification of piles). In the following example, there is a pile with three layers:

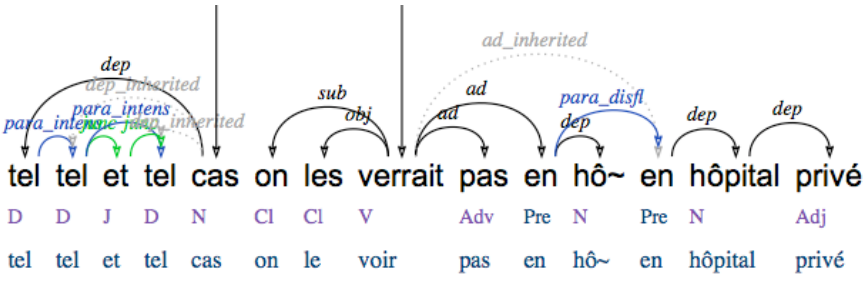


{ ^et "euh" | ^et } "ben" "voilà" j' arrive au niveau de la grande place de la gare où il y a { tous les trams | tous les bus |} //+ {| " non " pas tous les bus } // '{ ^ and "erm" | ^ and } "well" "so" I arrive at the big square of the railway station where there are { all the trams | all the buses |} //+ {| "no" not all the buses } //'

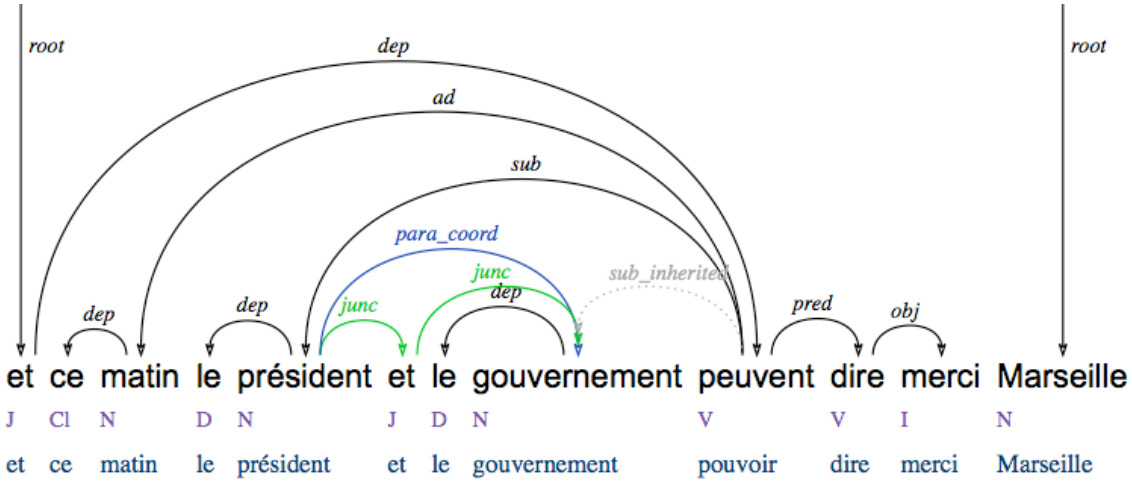
Direction of the paradigmatic link: For disfluencies and reformulations, the governor of the layer is attached to the nearest conjunct. Consequently, when the governor is after the pile, the paradigmatic links go from right to left (in order to preserve the arborescent structure):



For coordinations and intensifications, the governor is always attached to the first conjunct, as in the following example, where it is the first occurrence of *tel* which depends on *cas*:



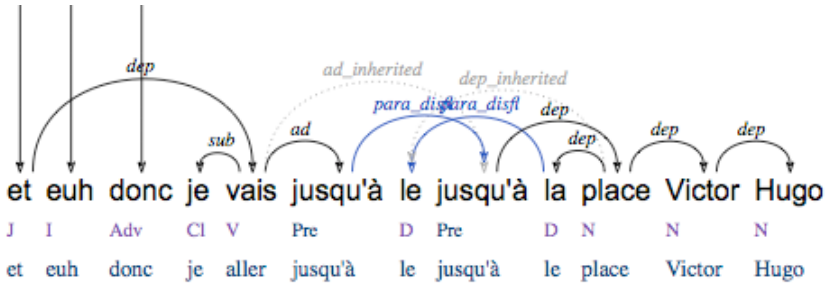
' { such | such | ^ and such } a case < we didn't see them { in the ho~ | in the private hospitals } //'



^ et ce matin <+ { le président | ^ et le gouvernement } peuvent dire [merci > Marseille //] //

'^ and this morning <+ { the president | ^ and the government } can say [thank you > Marseille //] //'

Double paradigmatic link: When there is disfluency, it may happen that a conjunct is syntactically discontinuous. In this case, each chunk is aligned by a paradigmatic link. It is possible, as in the following example, for two paradigmatic links to follow two different directions within the same pile:



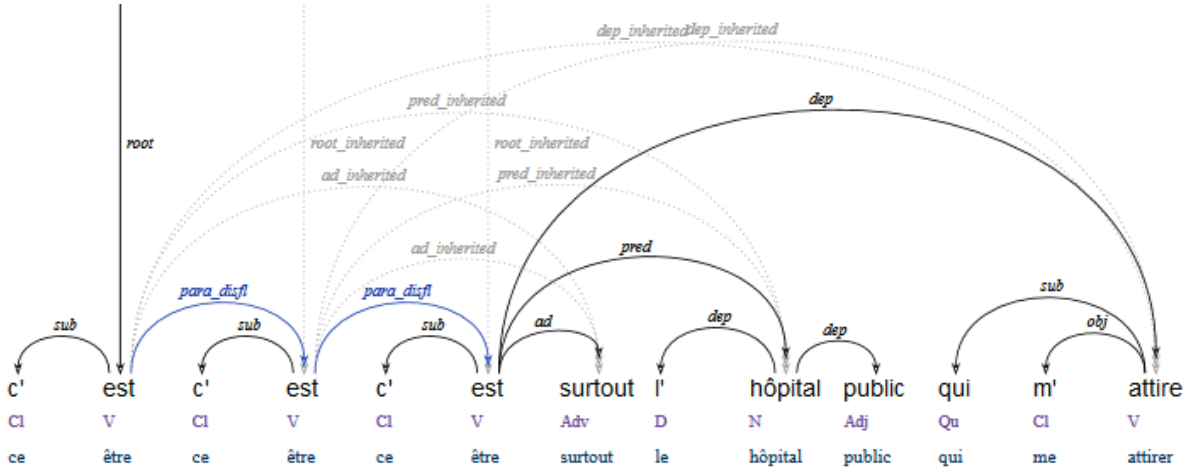
^ et " euh " donc < je vais { jusqu' au | jusqu' à la } place Victor Hugo //

'^ and "erm" so < I go { to the | to the } Victor Hugo square //'

Inherited dependency

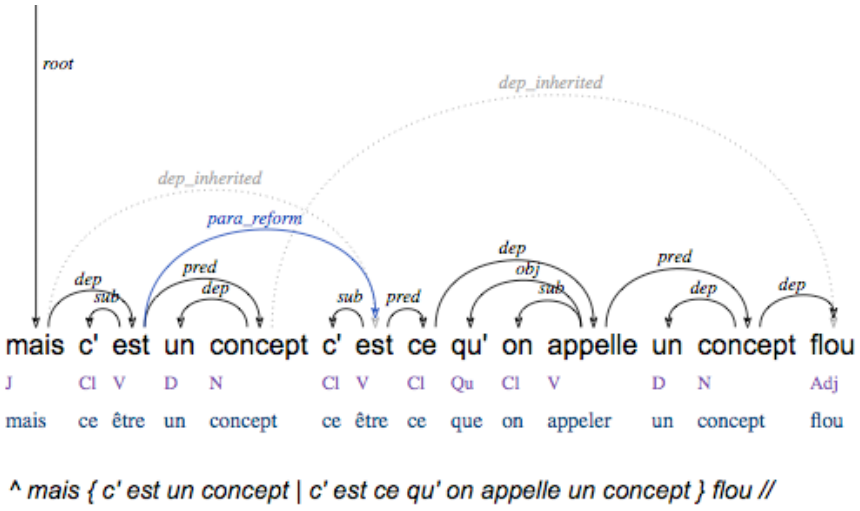
Conjuncts are in a paradigmatic relation. We attach one of the conjuncts to the governor of the pile, but the other conjuncts can commute with it and therefore inherit a relation of the same type. For example, in the previous example, the second occurrence of *jusqu'à* inherits an *ad_inherited* link, whereas the *le* inherits a *dep_inherited* link.

Root inherited: When the lexemes share a dependent, we consider that they form a pile construction even if they do not strictly speaking occupy the same governed position, since, as in the following example, they can be the root of the dependency structure. In this case, as in the others, the conjuncts inherit the same dependency as the first conjunct, which is therefore in this case a *root_inherited* dependency:



'{ it's | it's | it's } especially the public hospitals that attract me //'

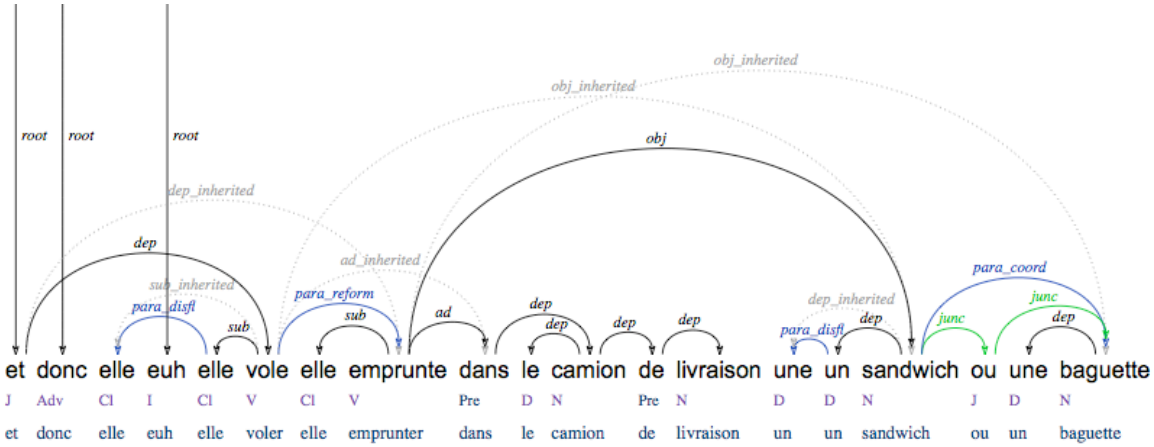
Exiting inherited dependency: Elements that depend on several conjuncts are outside the pile. They are dependent on the nearest conjunct and receive an inherited dependency from the other conjuncts. In the following example, *flou* modifies the two occurrences of *concept*:



'but { it's a concept | it's what we call a concept } which is fuzzy //'

Note meanwhile that our pile annotation strategy is to consider that an element is outside a pile from the moment when it can potentially complete several layers. Consequently, very few layers turn out to be incomplete.

No double inheritance: In the following example, consider the lexemes *vole emprunte sandwich baguette* and the relations uniting them. There is an *obj* relation between *emprunte* and *sandwich*. Because of the paradigmatic link between *vole* and *emprunte*, we deduce that there is an *obj_inherited* relation between *vole* and *sandwich*. Because of the paradigmatic link between *sandwich* and *baguette* we deduce that there is an *obj_inherited* relation between *emprunte* and *baguette.* We could also infer from these two relations another *obj_inherited* relation between *vole* and *baguette*. But for the sake of simplicity, we do not indicate these relations:



^ et donc < { { elle " euh " | elle } vole | elle emprunte } dans le camion de livraison { { une | un } sandwich | ^ ou une baguette } //

'^ and so < { { she "erm" | she } steals | she borrows } from the delivery van { { a | a } sandwich | ^ or a baguette} //'

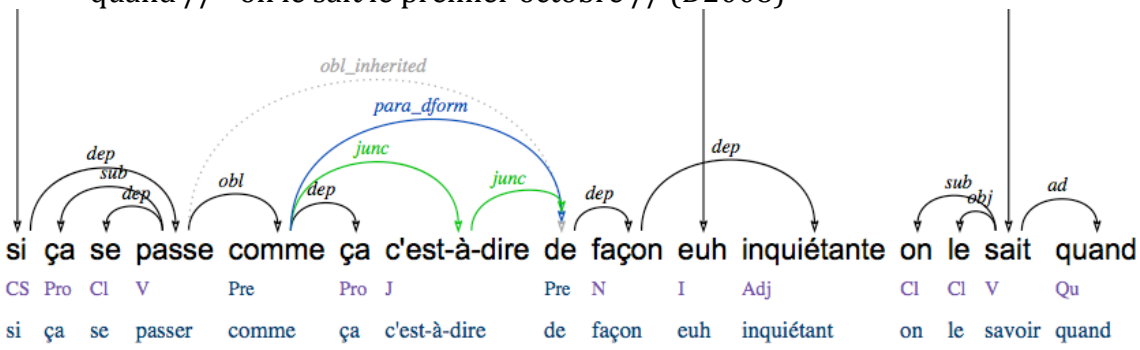
Junction relations

Junctor: *Junctors* are elements that link conjuncts (they are prefixed by a ^). Junctors are more or less coordinating conjunctions. We adopt, following the works of Blanche-Benveniste and de Ndiaye (1989), a variant of the term *jonctif* used by Tesnière (1959). Junctors only occupy a role inside the pile. If we were to conserve a single layer of the pile only, the junctors would not be kept:

des jeunes { **cagoulés | ^ et armés** }	'**{ hooded | ^ and armed }** youths'
*des jeunes **et armés**	'***and armed** youths'

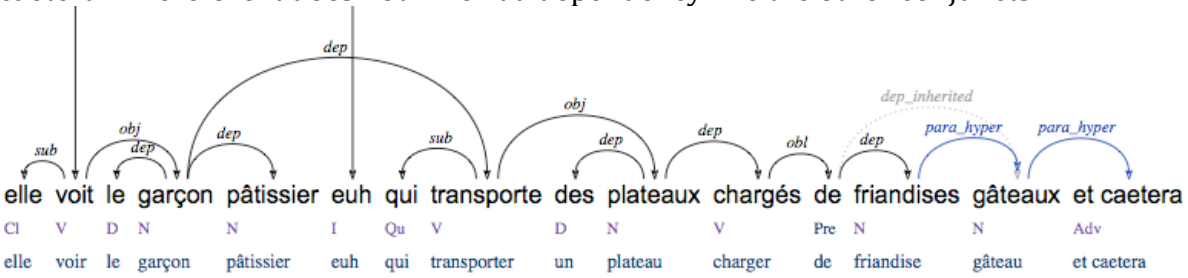
Junctors mainly appear in coordinations (*et, ou, mais, ainsi que,* etc.), but they are also possible in reformulations, notably *c'est-à-dire* :

7. si ça se passe { comme ça | ^c'est-à-dire de façon "euh" inquiétante } < on le sait quand //= on le sait le premier octobre // (D2008)



'if it goes { like that | ^ that is "erm" worryingly } < when will we know //'

General extenders: The last type of element found in pile constructions, notably in hypernymic coordinations (see definition below), are characterised by the fact that they close the pile, lie *et caetera*. We call them *general extenders.* They can only occupy the last layer of a pile. In the following example it is the case of *et caetera*. Alone it cannot occupy the governed position **des plateaux chargés de et caetera* '*trays loaded with et caetera'. Therefore it does not inherit a dependency like the other conjuncts:

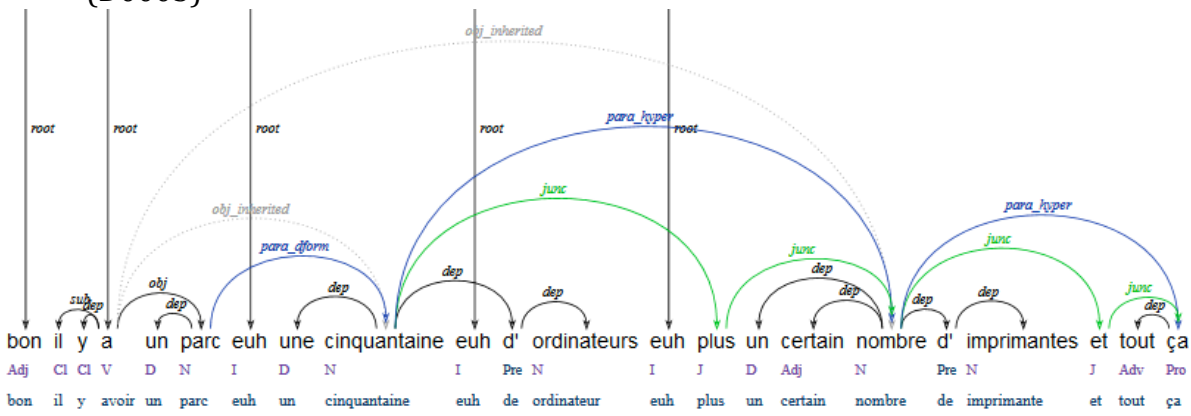


elle voit le garçon pâtissier " euh " qui transporte des plateaux chargés de { friandises | gâteaux | et caetera } //

'she sees the young baker "erm" who is transporting trays loaded with { sweets | cakes | et caetera } //'

But there are also many general extenders that are formed of a junctor and a specific type of conjunct such as *tout ça*, which we call, following the example of Overstreet (2005), an *extender*:

8. et "euh" "bon" "ben" ça pose des problèmes { de maintenan~ | "enfin" de maintenance "euh" | { { de | de } mise à jour | **^et tout ça** } "euh" } // voilà // (D0005)



" bon " il y a { un parc " euh " | { une cinquantaine " euh " d' ordinateurs | " euh " ^ plus un certain nombre d' imprimantes | ^ et tout ça } } //

'"well" there's { a park "erm" | { fifty odd "erm" computers | "erm" ^ plus a certain number of printers | ^ and all that } }'

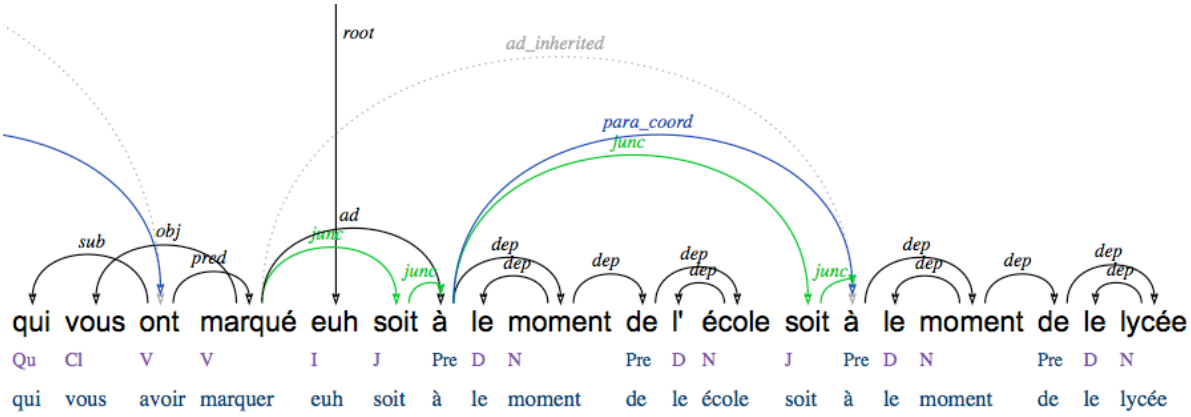
The lexeme *et caetera* functions as the amalgam of a junctor and of a extender, which is at its origin. This explains why it cannot, like other junctors, appear outside a pile.

Junctor links: Simple junctors (*et, ou, mais, ainsi que,* etc.) appear between two conjuncts. Following the asymmetrical analysis of coordination (Mel'cuk 1988), we consider that the junctor forms a constituent with the following conjunct, that the consituent joins up with the preceding conjunct. As the junctor controls the distribution of the consituent which is added to the preceding conjunct, it is considered its head. This gives two dependencies: preceding conjunct → junctor → following conjunct.

These dependencies are particular since they are to some extent in an orthogonal dimension to government. Therefore we label them with a particular function *junc* (Tesnière calls this relation *junction*. See the examples above and below, in which the junctors are marked in green.
Junctor links are always doubled by a paradigmatic link, except in the case of double junctors and of a junctor without a pile construction (cf. below), where junctor links are doubled by a government relation.

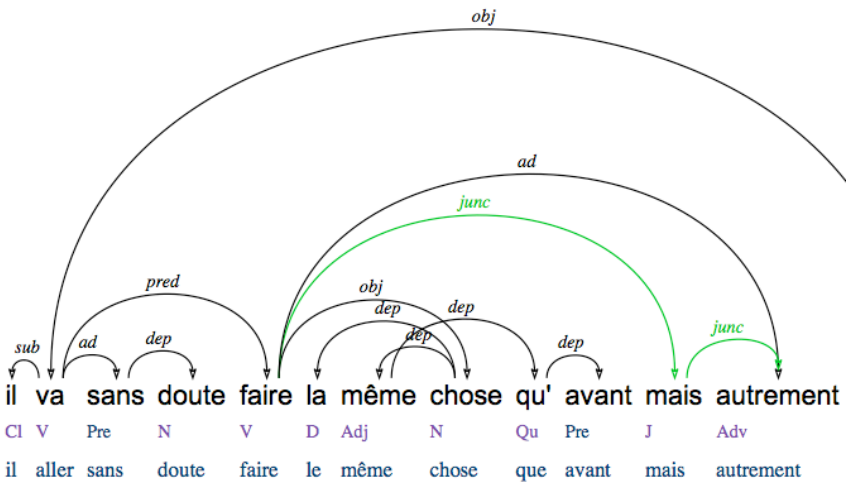
Double junctors: When the first conjunct is introduced by a junctor as in the coordinations of the form *soit A soit B* 'either A or B', the junctor depends on the governor of the pile:

1. est-ce que vous avez des enseignants { { dont | dont } vous vous souvenez particulièrement | qui vous ont marqué "euh" { ^soit au moment de l'école | ^soit au moment du lycée } // (D0001)
   'do you have any teachers { { that | that } you remember in particular | who left their mark on you "erm" { ^either at school | ^or at college }'



This analysis allows us to treat the two junctors in the same way, as a governor of a conjunct. Once again, the link *junc* doubles a dependency, but this time it is not a paradigmatic link, but a government link. Nevertheless, the principle is similar; the junctor acts as a marker of the link it doubles.

Junctor without a pile construction: In a construction of the form *il parle anglais et bien* 'he speaks English and well', the junctor *et* does not indicate a pile construction because the conjuncts do not occupy the same syntactic position and there is no paradigmatic relation between the conjuncts. We consider it to be coordination between two IUs – *il parle anglais* on the one hand and *et bien* on the other – which form a single GU. Therefore *bien* is dependent on *parle* (as it would be in *il parle bien anglais*) and at the same time the *junc* links link the junctor *et* to *parle* and to *bien*: *parle* –junc-> *et* –junc-> *bien*. The same phenomenon can be seen in the following example:



[il va sans doute faire la même chose qu' avant //+ ^mais autrement //] pronostique Francis Brochet // (D2013)
'[he is without a doubt going to do the same thing as before //+ ^but differently //] analyses Franci Brochet //'

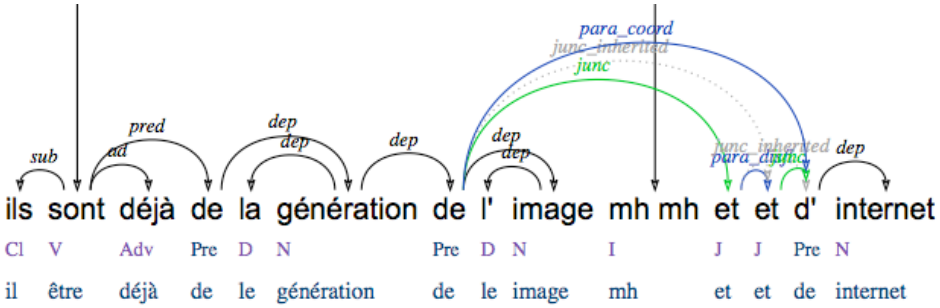
This construction is not rare, it seems. Several examples can be found in our corpus:

on veut bien parler avec vous //+ mais { a~ | après } le déménagement // (D0006)
'well will speak to you //+ but { a~ | after } the move //'

normalement < c' est du bois de hêtre dessous //+ ^ et { qui est p~ | qui est laqué noir } //
'normally < it's beech wood on top //+ ^ and { which is p~ | which is varnished black } //'

{ ce | ce } chiffre de trente mille < c' est finalement une extrapolation que vous faites //+ ^mais à partir de scénarios passés (entre guillemets) raisonnables //+ ^mais en tenant compte { de ce qu' on a pu faire | ^ou surtout de ce qu' on n' a pas pu faire } // (D2008:43)
'{ this | this } figure of thirty thousand < it turns out to be an extrapolation that you make //+ ^ but on the basis of past (quote-unquote) reasonable scenarios //+ ^but taking into account { what could be done | ^or rather what could not be done }' //

Pile constructions of junctors: It is possible that junctors form a pile construction in the case of disfluency of the junctor. In this case, the piled constructions inherit *junc_inherited* links with the conjuncts:



‘they are already from the generation { of the image “mm mm” | { ^and | ^and } of the internet }’

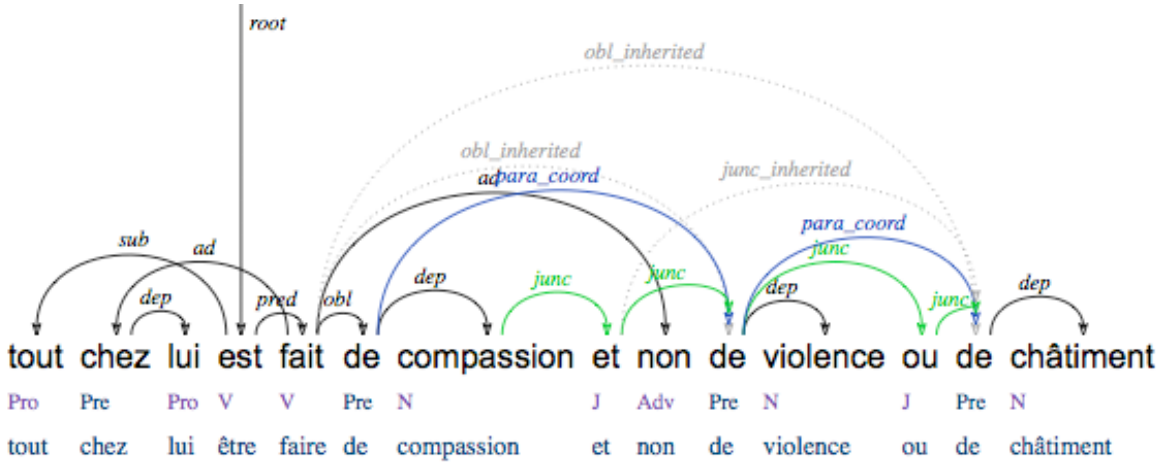
Nested coordinations: We speak of nested coordinations when a layer of coordination is itself occupied by a coordination structure. In the well known example *Nous cherchons quelqu’un qui parle anglais et allemand ou italien* ‘We are looking for someone who speaks English and German or Italian’

, there are two possible interpretations, each with a nested coordination:

a. Nous cherchons quelqu’un qui parle { { anglais | ^et allemand } | ^ou italien }
‘We are looking for someone who speaks { { English | ^and German } | ^or Italian }’
b. Nous cherchons quelqu’un qui parle { anglais | ^et { allemand | ^ou italien } }
‘We are looking for someone who speaks { English | ^and { German | ^or Italian } }’

These two examples benefit from a slightly different analysis. In the two cases, we have the same paradigmatic links anglais -> allemand -> italien and the same *junc* links anglais -> et -> allemand -> ou -> italien, but in (a), *anglais* is coordinated with *allemand* and inherits a *junc_inherited* link with *ou*, whereas in (b) it is *italien* which is coordinated with *allemand* and which inherited a *junc_inherited* with *et*.

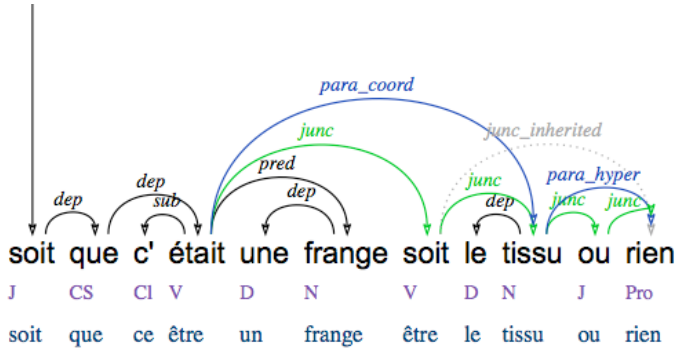
We have an example of the same form as (b) in our corpus, where the third conjunct inherits a *junc_inherited* link:



tout chez lui est fait { de compassion | ^ et non { de violence | ^ ou de châtiment } } //

‘with him everything is made { of compassion | ^and not { of violence | ^or punishment } }’

Or this second example with double junctors:



{ ^ soit ^ que c' était une frange | ^ soit { le tissu |} {| ^ ou rien } } //

'{ ^ either it was a fringe | ^ or { the material|} {| ^ or nothing } } //'

Paradigmatising adverbs

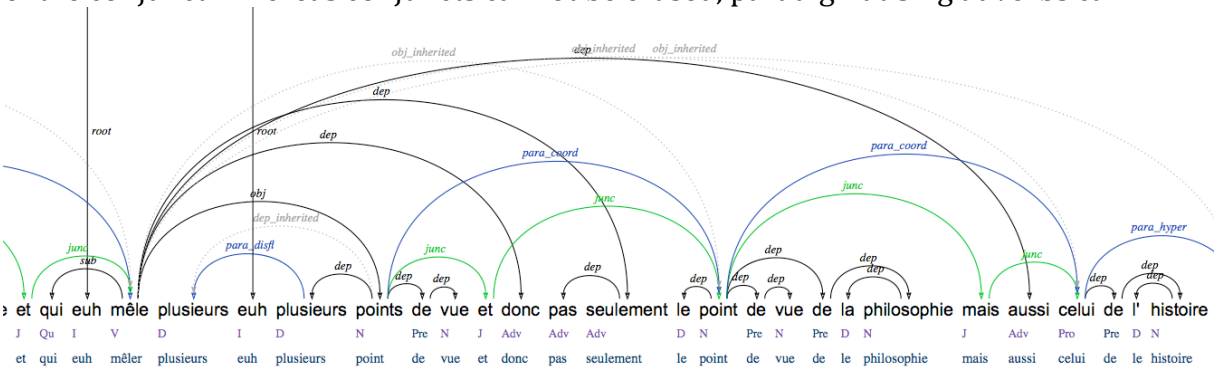
We also find in pile constructions *adverbs* known as *paradigmatising adverbs* (Nølke, 1983), such as *seulement* 'only' and *aussi* 'also', in bold in the following example:

> ^mais c'est aussi { une conférence { de | d' } histoire de l'art | une conférence d'esthétique } parce que [comme vous verrez < j'ai une approche { de | de } l'art { qui est { assez | assez } généraliste | ^et qui "euh" mêle { { plusieurs | "euh" plusieurs } points de vue | ^et **donc** { **pas seulement** le point de vue de la philosophie | ^mais **aussi** { celui de l'histoire | celui de la sociologie | et caetera | et caetera } } } }] // (M2002)

Unlike junctors, these adverbs can be kept even if there is only a single layer:

a. mon approche (ne) mêle **donc pas seulement** le point de vue de la philosophie
b. mon approche mêle **aussi** le point de vue de l'histoire.

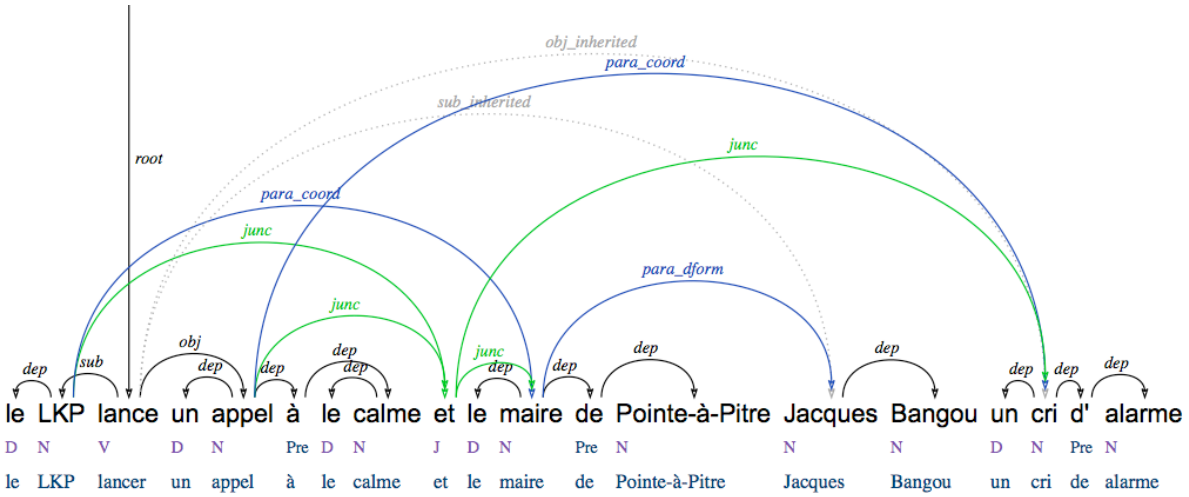
There are therefore "visible" for government, but occupy a different position from that of the conjunct: whereas conjuncts cannot be erased, paradigmatising adverbs can.



'^and who "erm" blends { { several | "erm" several } points of view | ^and so { not **only** the philosophical point of view | ^but **also** { the historical one | ...'

Problematic coordinations

Non-standard coordinations: We treat cases of gaping coordinations without resorting to ellipsis. We consider that a junctor can command two parallel coordinations, as in the following example:



{ le LKP lance un appel au calme | ^ et { le maire de Pointe-à-Pitre | Jacques Bangou } un cri d' alarme } //

'{ the LKP launches an appeal for calm | ^and the Mayor of Pointe-à-Pitre | Jacques Bangou } a cry of alarm } //'

Junctor outside the pile: The junctor *ni* 'neither/nor' is particular, since it integrates a negation, i.e. a paradigmatising adverb. This may mean that it occupies the position of such an adverb (i.e. of a *pas* 'not'), rather than that of a junctor:

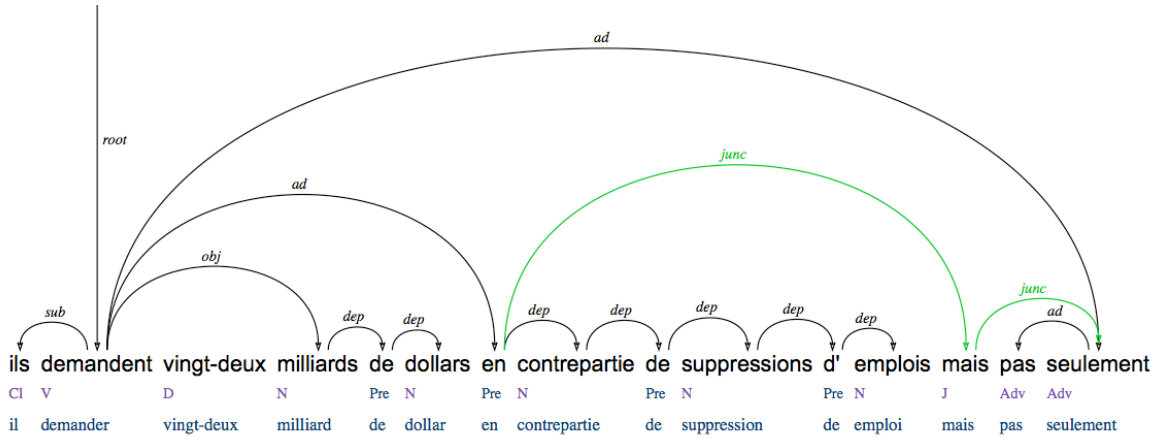
> parce qu' { on remarque "euh" { ass~ "euh" | de façon "euh" } & | on remarque bien que le salaire n'est **^ni** en adéquation **{ { avec** le nombre d'années & "euh" | avec le nombre d'études suivies } | **^ni avec** le travail { à fournir | personnel que l'on fournit } } // (M1003)
> 'because { you notice "erm" { qui~ "erm" | in such a way "erm" } & | you notice that the salary is **^neither** in keeping { { **with** the number of years & "erm" | with the number of courses followed } | **^nor with** the work { to be provided | personal *(work)* that you provide } }'

It is possible to hesitate between treating this *ni* as an adverb or as a junctor. Note that in any case our analysis of double junctors is not satisfactory (see above).

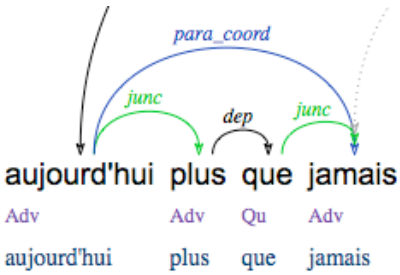
Layer without a conjunct: Finally there exist particular configurations of junctors and paradigmatising adverbs which allow for the omission of the second conjunct (which is in fact a simple repetition of the first conjunct here):

> ils demandent vingt-deux milliards de dollars { en contrepartie de suppressions d'emplois | ^mais pas seulement } // (M2006)
> 'they are asking for twenty two billion dollars { as compensation for the job cuts | ^but not just that } //'

Here we do not use the symbol &, which indicates incompletions, since this utterance is considered as being complete. We have no other possible dependent for the link *junc* from *mais* than the adverb *seulement* (which is a paradigmatising adverb dependent on *demandent* via an *ad* relation and not *ad_inherited* as it would be if it had been a conjunct).



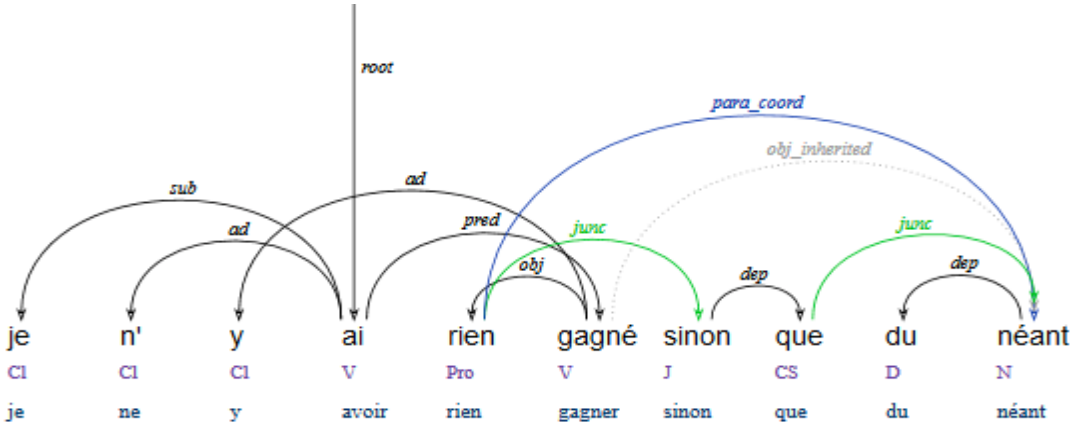
Complex junctor: In the following example, *plus que* 'more than' is treated as a junctor made up of two lexemes (since there is a commutation of *plus* with *moins*), which gives the following configuration:



'today more than ever'

Discontinuous coordination: Our annotation supports without any difficulty cases of discontinuous coordination such as in:

{ comme l' esclave qui désire un peu d' ombre | comme le manoeuvre qui attend sa paye depuis des mois } (+ dit Job) < je n' y ai { rien |} gagné {| ^sinon ^que du néant } // (M2003)
'{ like the slave who longs for some shadow | like the hired hand who has been awaiting his pay for months } (+ says Job) < I didn't earn { anything|} there {| ^except nothingness } //'



Classifying pile constructions

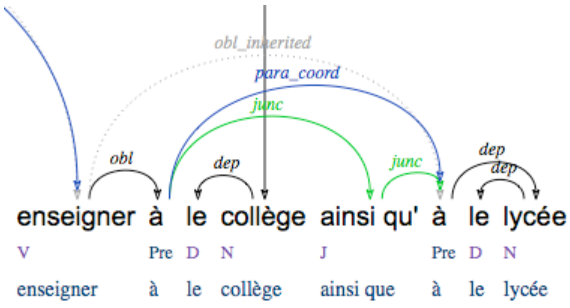
(Sylvain Kahane, Paola Pietrandrea)

We class piles into 7 sub-categories: coordinations, hypernymic coordinations, intensifications, disfluencies, reformulations, double formulations and negotiations.

Coordination (para_coord): Relational coordination (or simply coordination) is a coordination in which each layer denotes a different element and the entire pile has a denotation which is the function of the denotations of its layers. We class two types of relational coordinations as coordination: additive coordinations (i.e. piles whose denotation is the reunion of the denotations of its layers):

1. alors < passons maintenant { au détail des mesures discutées | **^et** aux attentes { des syndicats | **^et** du patronat } } // (M2006)
   'so < moving on now { to the detail of discussed measures | **^and** to the expectations { of the trade unions | **^and** of the employers } }'

2. alors < ce que je souhaiterais faire de ma vie < c'est { devenir professeur d'italien à savoir certifié | donc "euh" enseigner { au collège | **^ainsi ^qu'** au lycée } } // (M1003)
   'so < what I would like to do with my life < it is to { become a certified Italian teacher | so "erm" teach { in secondary school | ^as ^well ^as in sixth form } } //'



3. je travaille à la préfecture de Paris qui { n'est pas connue | **^mais** néanmoins existe } "euh" // (D0001)
   'I work at the prefecture of Paris, which { is not well-known | **^but** nevertheless exists } "erm" //'

and alternative coordinations (i.e. piles where the elements denoted by the layers can be potentially substituted for each other):

4. allez // avec Messi { qui va chercher le corner | ^et qui va trouver **{ le corner | ^ou la touche | ^ou la sortie de but }** } // (D2003)
   'go on // with Messi { who is going for the corner | ^and who is going to find **{ the corner | ^or the kick-in | ^or goal clearance }** } //'

Note that certain additive coordinations are marked by the junctor *comme*:

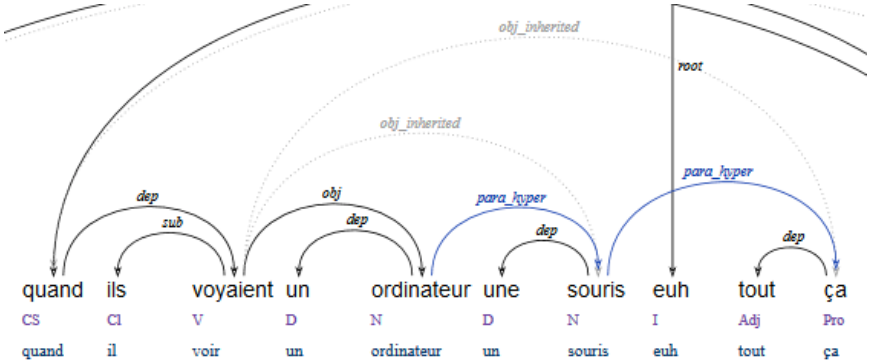
5. il y a { des | des } traditions de famille qui se transmettent { dans le domaine religieux | **^comme** dans tout le domaine du travail } // (D1001)
   'there are { some | some } family traditions which are passed on { in religious practices | ^as in all work practices } //'

and that certain alternative coordinations are marked by the presence of a comparative junctor :

6. votre marché < c'est { Aligre | ^plus ^que Nation } // (D0001)
   'your market < it's { Aligre | ^more ^than Nation } //'

Hypernymic coordination (para_hyper): We use the term hypernmic coordinations to refer to piles whose value refers to a hypernym of the layers, i.e. to a class containing the denotation of the layers, but which does not correspond to a logical combination of the denotations:

7. les gens < au début <+ quand ils voyaient **{ un ordinateur | une souris "euh" | tout ça }** <+ { ils sav~ | ils savaient } pas ce que c' était // (D0005)
   'people < at the beginning <+ when they saw **{ a computer | a mouse "erm" | all that }** <+ { they *(didn't)* kno~ | they didn't know } what it was //'



quand ils voyaient { un ordinateur | une souris " euh " | tout ça }

8. donc < { on voit | on le voit } sortir "euh" **{ du pain | des brioches | des trucs comme ça }** // (M0023)
   'so < { you see | you see him } take out "erm" **{ bread | buns | stuff like that }** //'

9. et petit à petit < ils ont essayé d'avoir { quelque chose |} à Paris ({| **{ un petit studio | ^ou quelque chose comme ça }** }) { pour revenir voir leurs amis | ^et ^puis pour y & } // (D0003)
   'and little by little < they tried to have { something |} in Paris ({| **{ a little studio**

flat | ^or something like that } }) { to return to see their friends { ^ and then to & } //'

Hypernymic coordinations are also possible with verbal elements:

10. parce que il a dit [{ elles corromp~ | elles corrompront } tous mes petits "euh" officiers de district //] "euh" { sans | sans } **{ me connaître | ^ni rien du tout }** // (D204)
'because he said [{ they corrup~ | they corrupt } all my little "erm" district officers //] "erm" { without | without } **{ knowing me | ^or anything like that }** //'

11. mais "euh" "euh" { { { se | se | se } gourer | ^et ^puis chauffer comme ça } | ^c'est-à-dire **{ dragouiller la mère | ^ou draguer | ^ou faire une déclaration d'amour à la mère }** } < non // (D207)
'but "erm" "erm" { { { slipping up[2] | ^and ^then getting all excited like that } | ^that ^is **{ hitting on the mother | ^or coming on to | ^or declaring his love for the mother }** } < no //'

From a formal point of view, hypernmyic coordinations are characterised by the use of designated closing expressions such as *tout ça* or *quelque chose comme ça* or *rien du tout* or by the fact that two elements are co-hyponyms.

We also class amongst hypernymic coordinations more lexicalised processes consisting of sufficiently opposed cohyponyms or antonyms which create the effect of covering a whole class (as in the fixed expressions *petits et grands* 'big and small'*, jour et nuit* 'day and night'). The result is an effect of universal quantification of a denoted class. This process, quite common in writing, is present in the samples of our corpus that are characterised by a higher register.

12. { à chacune | ^et à chacun d'entre vous } <+ **{ Françaises | ^et Français } { de métropole | d'outre-mer | de l'étranger }** <+ je souhaite très chaleureusement { une bonne | ^et une heureuse } année deux mille // (M2004)
'{ to each | ^and every one of you } <+ { French (women) | French (men) } { from the mainland | from overseas | from abroad } <+ I wish you with all my heart { a good | ^and a happy } year two thousand //'

Intensification: Intensive coordination has a general function of intensification of the meaning expressed by the repeated element. This function of intensification is defined precisely according to its category of expression. For example, the reiteration of nominal elements intensifies quantity ('plein d'exercices', 'plusieurs dizaines d'années'):

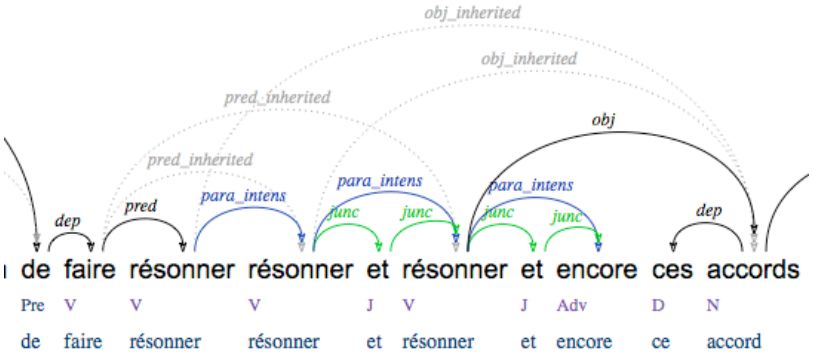
13. { le | la | le | le | la } grosse recette de Sarah "tu vois" < c'était de de faire **{ des exercices | des exercices | des exercices }** par exemple "tu vois" pour un point de grammaire // (Valibel)
'{ the | the | the | the | the great method of Sarah's "you see" < it was to to do **{ exercises | exercises | exercises }** for example "you see" on a specific grammar point //'

[2] The preceding pile construction { se | se | se } cannot be represented in the English translation.

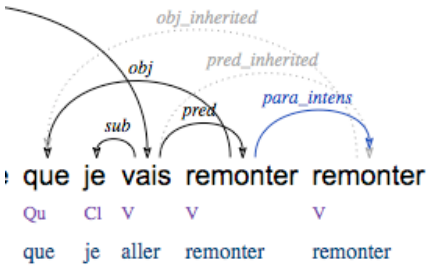
14. mais mais mais mais { les lois sociales | le droit de grève | ^et tout ça } < { ça s'est fait | ça s'est fait } sur **{ des dizaines | ^et des dizaines }** d'années // ça s'est fait sur presque "enfin" { cinquante | cent } ans // (Valibel)
'but but but but { social laws | the right to strike | ^and all that } < { that was achieved | that was achieved } over { **dozens | ^and dozens }** of years // that was achieved over almost "well" { fifty | a hundred } years //'

The reiteration of verbal elements can intensify the duration or the frequency of action and mark the iterative or continuative aspect:

15. on pouvait pas s'empêcher à la fin de { Mort | ^et transfiguration } de faire **{ résonner | résonner | ^et résonner | ^et encore }** ces accords qui nous enchantaient // (D212)
'we couldn't help ourselves at the end of { Death | ^and transfiguration } to let those enchanting chords **{ resonate | resonate | ^and resonate | ^and still more }** //'



16. ^ ensuite <+ "euh" je vais "euh" prendre [je crois que c' est l' avenue Alsace-Lorraine //] que je vais { remonter | remonter } //
'^ then <+ "erm" I go "erm" take [I think it's the avenue Alsace-Lorraine //] that I go **{ up | up }** //'



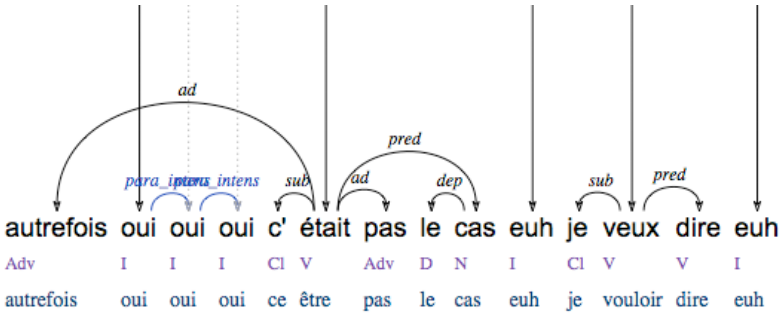
Reiteration of adjectival or adverbial elements can form a sort of superlative ('pas trop facile', 'vraiment très difficile'):

17. et puis "bon" "ben" "voilà" donc < ce qui fait que { c' est | c' est } pas **{ facile | facile }** // (D005)
'and then "well" "well" "you see" so < which means that { it was | it was } n't **{ easy | peasy }** //'

18. c'est **{ très | très }** difficile à définir // (D202)
'it's **{ very | very }** difficult to define //'

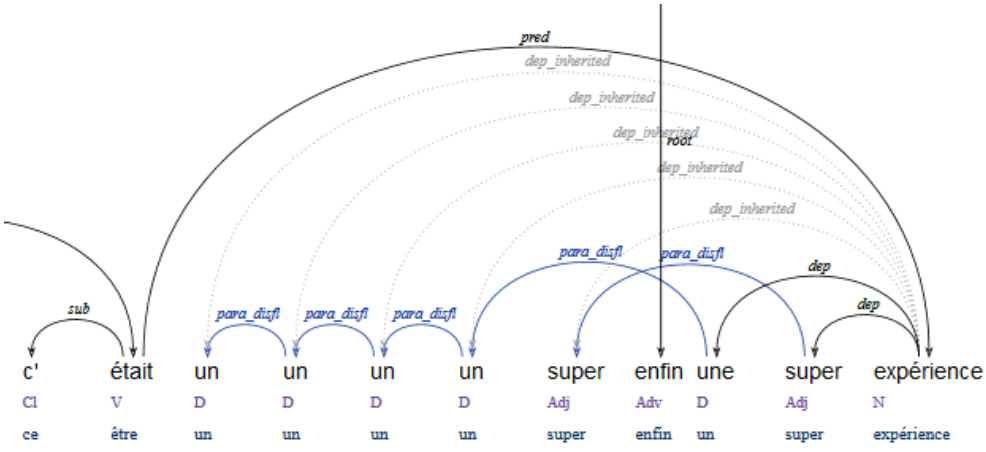
We note as intensive coordinations sequences of adverbs such as *oui oui oui, non non non*:

19. autrefois ({ oui | oui | oui } //) <+ c' était pas le cas "euh" "je veux dire" "euh" // (D001)
'in the past ({ yes | yes | yes } //) <+ it wasn't the case "erm" "I mean" "erm" //'



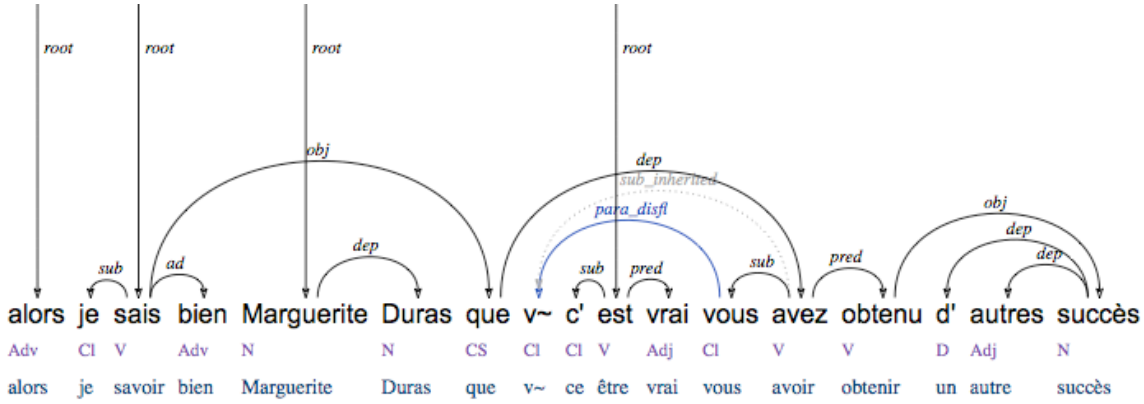
autrefois ({ oui | oui | oui } //) <+ c' était pas le cas " euh " " je veux dire " " euh "

Disfluency (para_disfl): We speak of disfluency when the speaker hovers over a syntactic position in order to adjust its formulation. This hovering is translated by a pile of words or of false starts:



c' était { { un | un | un | un } super | " enfin " une super } expérience] //

'it was { { a | a | a | a } super "well" a super } experience'

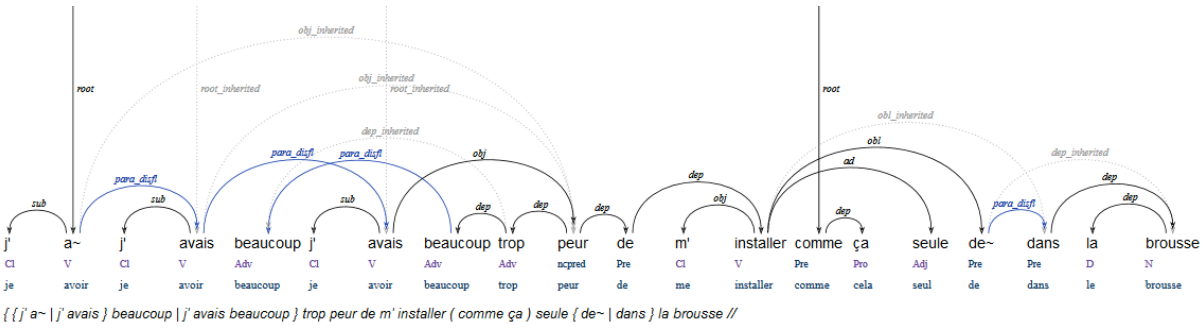


alors < je sais bien (Marguerite Duras) que { v~ | " c' est vrai " vous } avez obtenu d' autres succès //

'so < I know (Marguerite Duras) that { y~ | "it's true" you } have had other successes //'

This stalling can lead to the repetition of quite long segments:

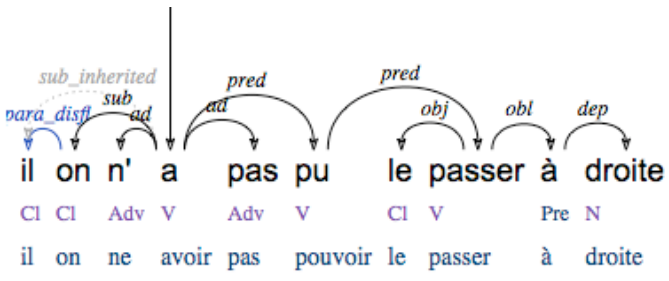
20. alors < **{ { j'a~ | j'avais } beaucoup | j'avais beaucoup }** trop peur de m'installer (comme ça) seule **{ d~ | dans }** la brousse // (D2004)
'so < **{ { I w~ | I was } much | I was much }** too scared to settle down (like that) alone **{ i~ | in }** the bush //'



{ { j' a~ | j' avais } beaucoup | j' avais beaucoup } trop peur de m' installer (comme ça) seule { de~ | dans } la brousse //

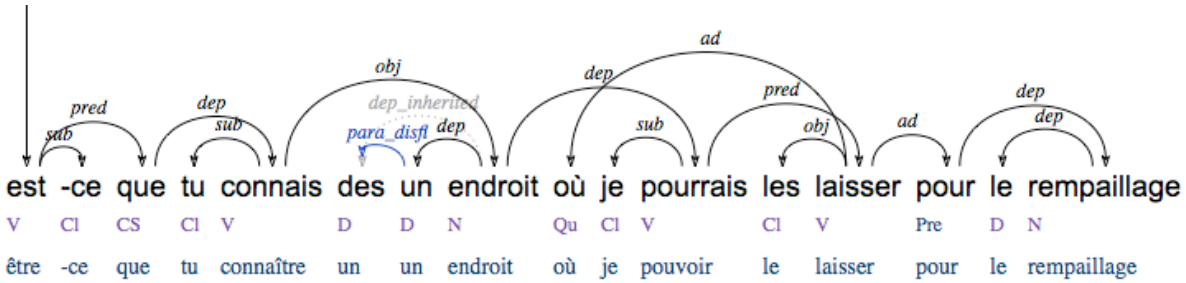
In the case of disfluencies, the repeated segment has no real interpretation, unlike the cases which will be described in the following sections, where each conjunct possesses a denotation.

We only consider there to be disfluency if no lexical change, excluding grammatical words, takes place.



{ il | on } n' a pas pu le passer à droite //

‘{ he | we } couldn’t pass it to the right //’

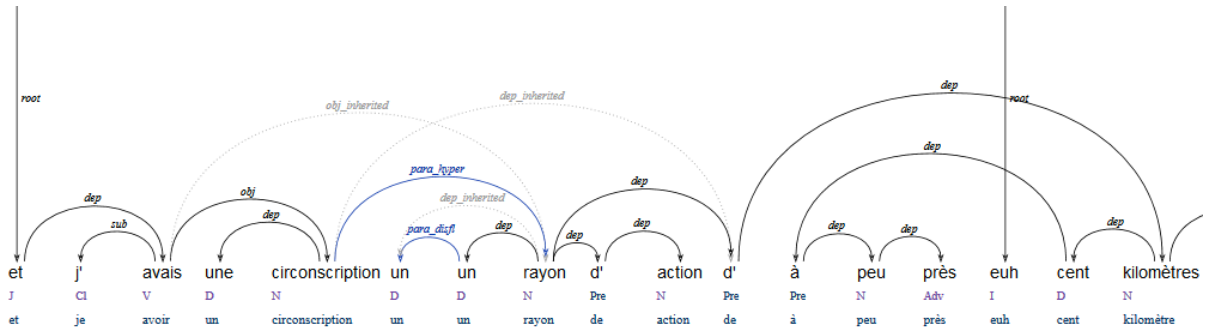


est-ce que tu connais { des | un } endroit où je pourrais les laisser pour le rempaillage //

‘do you know of { any | a } place where I could leave them to be reseated //’

Reformulation (para_reform): A speaker can propose an initial denotative formulation and then return several times to replace it with other formulations. This is a process known as *denotative reformulation*.

21. et j'avais **{ une circonscription | { un | un } rayon d'action }** d'à peu près "euh" cent kilomètres tout autour de cet endroit // (D2004)
‘and I had **{ a district | { a | a } scope of action }** of about "erm" a hundred kilometres all around that area //’



^ et j' avais { une circonscription | { un | un } rayon d' action } d' à peu près " euh " cent kilomètres tout autour de cet endroit //

22. tu arrives place aux Herbes avec { une | une } sorte **{ de halle | "quoi" { de | de | de } structure métallique }** // (M0001)
‘you arrive at the place aux Herbes with { a | a } sort **{ of covered market | "yeah" { of | of | of } metallic structure }** //’

Reformulation generally takes places within the same illocutory component.

Sometimes interruptions, even very long ones, are possible. In the following extract for example, three parenthetic IUs interrupt a reformulative pile { qui parlent pas français | dont les mamans ne parlent pas français }, of which the second layer is at a considerable distance from the first, without there being a change in the illocutory component.

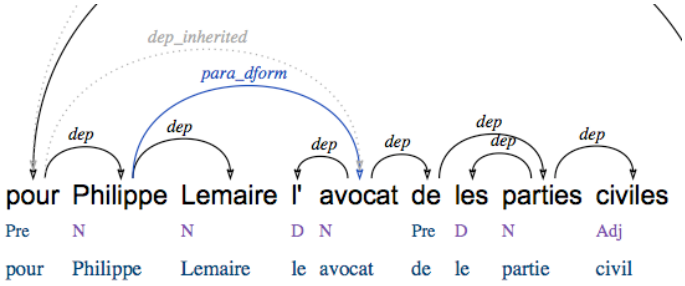
23. dans le vingtième <+ il faudrait { qu'il y ait & | qu'on sépare & | "enfin" qu'il y ait des cours de français pour les petits enfants **{ qui parlent pas français |}** }
(c'est pas compliqué quand même //

c'est pas très difficile d'apprendre le français à des petits enfants de cet âge-là // { ça | ça | ça } se fait assez facilement //)
{| dont les mamans ne parlent pas français } // (D0002)

'in the twentieth <+ it is necessary { that there be & | that we separate & | "well" that there be french classes for the small children **{ who speak French |}** }
(it's not complicated however) //
it's not very difficult to teach French to small children of that age //
{ it | it | it } is very easily done //)
{| whose mothers don't speak French } //'

Double formulation (para_dform): The process of double formulation is the accumulation of several denotations for a single element:

24. pour **{ Philippe Lemaire | (+ l'avocat des parties civiles) }** <+ { c'est d~ | ce sont des } procédés terroristes // (M2006)
'for **{ Philippe Lemaire | (+ the lawyer of the civil parties) }** { it was | these were } terrorist methods //'



The second denotation functions as an illocutory component, if not an illocutory unit in itself.

We note as double formulations all reformulative appositions (i.e. appositions where the second element fulfils all the morphological conditions to commute with the first), for example:

25. **{ le président de l'Unef | (+ Jean-Baptiste Prévot) }** au micro de Sonia Bourane // (M2006)
'**{ the president of the Unef | (+ Jean-Baptiste Prévot) }** on Sonia Bourane's microphone //'

26. il y a eu en mille neuf cent dix huit sur l'ensemble de la planète on dit **{ quarante millions de décès | (+ ^c'est-à-dire une mortalité effroyable) }** //
'in nineteen eighteen, across the the planet there were it is said **{ forty million deaths | (+ ^that ^is ^to ^say a horrifying mortality) }** //'

Note that we do not consider modifying appositions to be pile constructions, i.e. appositions in which the second element does not commute with the first, but depends on it syntactically, modifying it:

27. pour **Philippe Lemaire (+ avocat des parties civiles)** <+ { c'est d~ | ce sont des } procédés terroristes //

'for **{ Philippe Lemaire | (+ lawyer of the civil parties) }** { it was | these were } terrorist methods //

We note as double formulations pile constructions encoding an intentional reformulation, linked by the junctor *c'est-à-dire* 'that is to say' and linking non-nominal or other elements:

si ça se passe **{ comme ça | (+ ^c'est-à-dire de façon "euh" inquiétante) }**
'if it goes { like that | ^ (+ ^that is "erm" worryingly) }

{ nous | nous } étions tous les deux { d'origine bourgeoise | élevés un peu **{ de la même manière "euh" | (+ ^c'est-à-dire "disons" d'une façon un peu britannique) }** } //
'{ we | we } were both { from a middle class background, | brought up a bit **{ in the same way "erm" | (+ ^that ^is "say" a bit British) }** }'

We note as double formulations discontinuous pile constructions which may be considered as inclusive double formulations: the denotation of the second element is not identical to that of the first, but is included in it. They either take the form of a particularisation (the first element is made up of a general word or an indefinite pronoun):

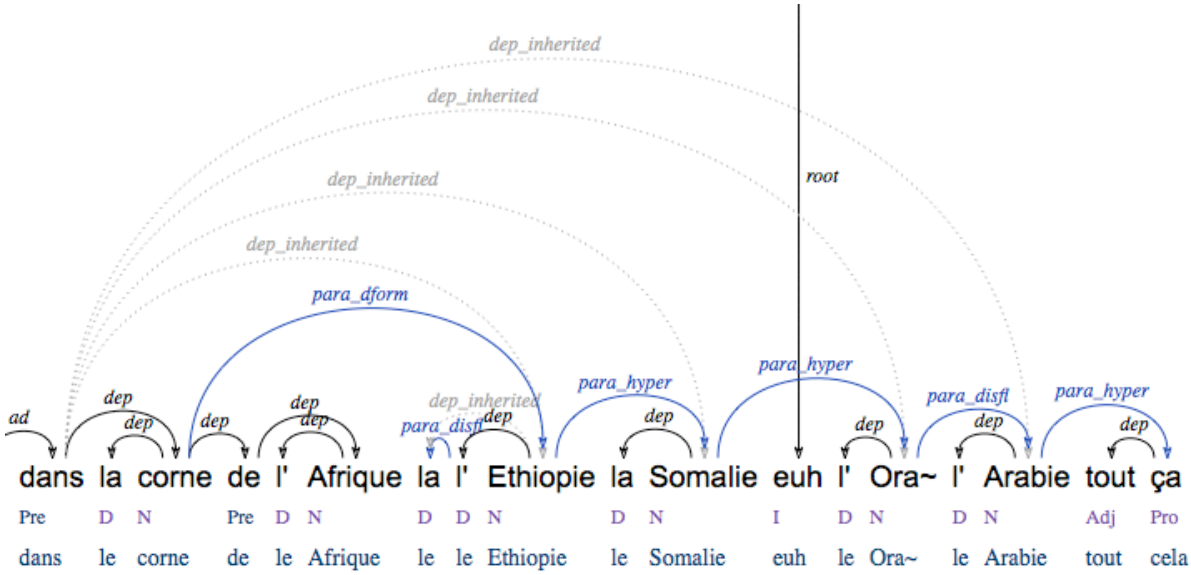
28. si je ne craignais pas d'entrer dans le jeu de certains hommes qui abusent de leur condition < je dirais que vous avez donné **{ quelque chose de plus |}** à la femme //+ **{| des armes de persuasion }** // (D2001)
'if I wasn't scared of being drawn into the game of certain men who take advantage of their condition < I would say that you have given **{ something extra |}** to the woman //+ **{| weapons of persuasion }**'

29. "ben" en fait < il y a { **pas mal de choses** |} qui rentrent en compte //+ { déjà "euh" **l'ambiance du magasin ...** } // (Olive, GARS)
'"well" so < there are **{ quite a few things |}** that come into play //+ { for a start "erm" **the atmosphere of the shop** ... }'

30. et j'ai trouvé **{ cet endroit | (+ Olkaloo) }** où ils avaient besoin d'un médecin // (D2004)
'and I find **{ this place | (+ Olkaloo) }** where they needed a doctor //'

or the form of an exemplification (the second element is made up of a pile of co-hyponyms of the first element):

31. et j'avais absolument envie d'aller dans **{ la corne de l'Afrique |} //+ {| { { la | l' } Éthiopie | la Somalie "euh" | { l' Ora~ | l' Arabie } | tout ça } }** // (D2004)
'and I really wanted to do to **{ the Horn of Africa |}** //+ **{ Ethiopia | Somalia "erm" | { Ora~ | Arabia } | all that } }** //'

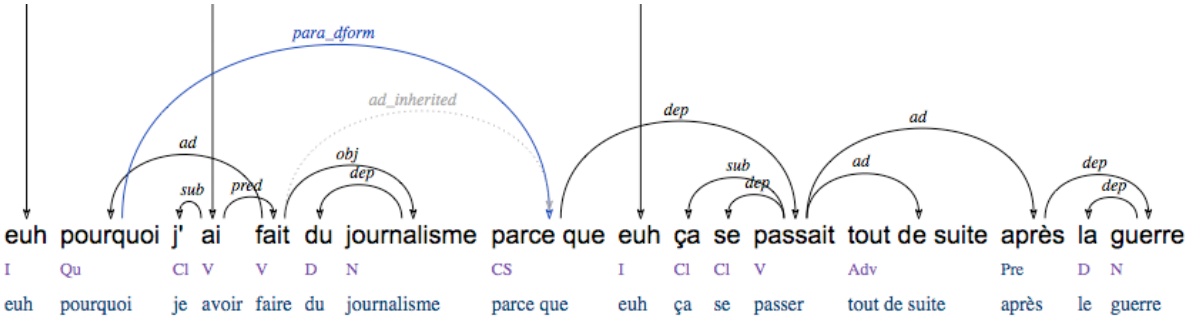


32. "euh" et sinon < les spécialités { les m~ | un { peu moins (je sais pas si c'est ça qui vous intéresse //) | petit peu moins } } prises < "bah" { c'est les | c'est **{ les spécialités à risques |} //+ {| { la gynéco obstétrique (par exemple) | la cancérologie } }** // (D006)
'"erm" and otherwise < the specialities { the leas~ | a { little less (I don't know whether if it's that that interests you //) | little less } } popular < "well" { they're the | they're **{ the risky specialities |} //+ {| { obstetrics and gynaecology (for example) | oncology } }**'

Among inclusive double formulations, we also include partial question-responses (i.e. with an interrogative pronoun). In spite of their distribution over two different IUs benefiting from their own illocutory force (one is a question and the other is the response), these structures fulfil all the conditions to be considered to be inclusive double formulations: the interrogative pronoun and the element used in response occupy the same structural position, they are co-denotational and the denotation of the second element is included in the denotation of the first:

33. $L1 et il faut compter **{ combien de temps après |}** //
'and it takes **{ how long after |}** //'
$L2 très rapidement "hein" // **{| { quinze jours | quinze jours maximum } }** //
(D0009)
'very quick "eh" // **{| { five days | five days maximum } }** //'

34. "euh" **{ pourquoi |}** j'ai fait du journalisme //+ **{| parce que "euh" ça se passait tout de suite après la guerre }** // (D2001)
'"erm" **{ why |}** I did journalism //+ **{| because "erm" it happened right after the war }**'



Reformulation or double formulation: It is possible to hesitate between reformulation and double formulation when assigning a type. For example:

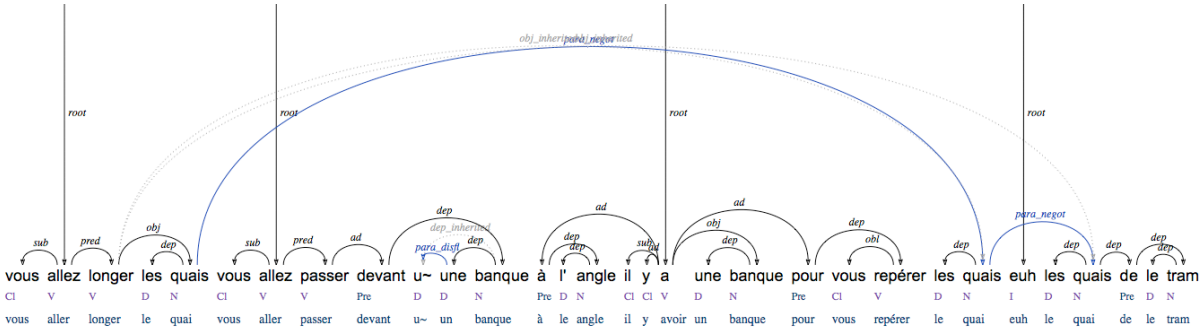
35. je ne compte **{ que des nuits de souffrance | que de nuits de souffrance }** dans notre humanité // (M2003)
'I count **{ only nuits of suffering | only nights of suffering }** in our humanity //'

A test for choosing between reformulation and double formulation is to prefix the second element by *je veux dire* 'I mean' (IU showing the intention of reformulation) or by *c'est-à-dire* 'that is to say' (junctor showing double formulation): in the preceding example, the test allows us to identify the pile construction as a reformulation.

Negotiation of the formulation: We have established four negotiation constructions operating on pile constructions in our corpus: request for confirmation, confirmation, refutation and correction. When there is negotiation within a pile, a segment is repeated with appropriate prosody. The following examples show a request for confirmation (repetition of *les quais* with an interrogative prosody) and confirmation (with reformulation with *les quais du tram*), as well as a direct confirmation (repetition of *quarante-huit ans* with assertive prosody):

36. $L1 vous allez longer { les quais |} //+
(vous allez passer devant { u~ | une } banque //
à l' angle <+ il y a une banque pour vous repérer //)
$L2 {| les quais |} //+
$L1 " euh " {| les quais du tram } //
$L2 "ah" d'accord //

'$L1 you go along the { platform |} //+
(you will pass infront of { a | a } bank //
at the angle <+ there's a bank to help you get your bearings //)
$L2 {| the platforms |} //+
$L1 "erm" {| the tram platforms } //
$L2 "ah" ok //'

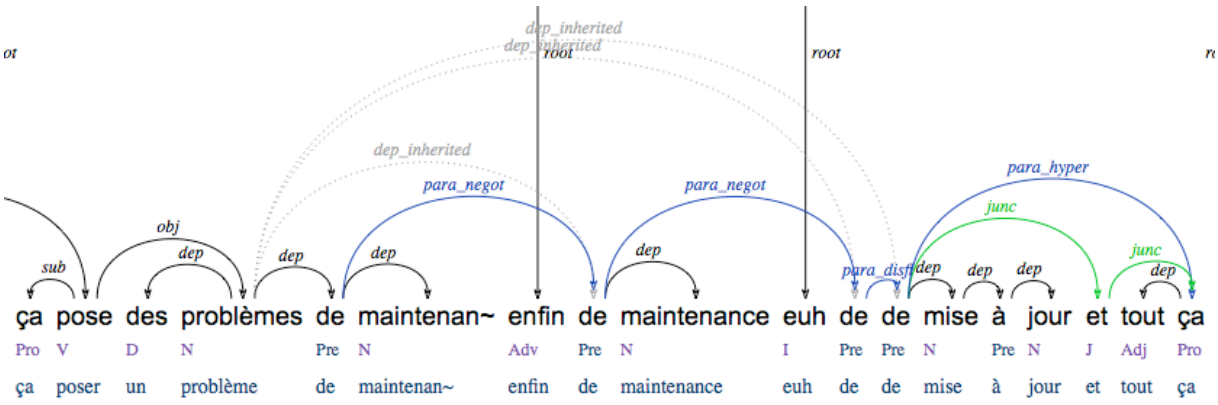


37. $L1 puisque finalement < ça fait "euh" { **quarante-huit ans** |} que vous êtes au Kenya //
‘since after all < you’ve been in Kenya for "erm" **{ forty eight years |}**’
$L2 {| **quarante-huit ans** } // { oui | oui } // (D204)
‘**{| forty eight years }** // { yes | yes } //’

A repetition can also be a refutation of a repeated element. In the following examples, refutation is introduced by the element *enfin*, which seems to be specialised for this usage, and which is following by a correction:

38. c'est la crise générale { { **des | des } Français | "enfin" des Français** | pas simplement des Français "hein" | { des | de } l'humanité | ^et de la lecture } } // (D0004)
‘it’s the general crisis { { **of the | of the } French | "well" the French** | not just the French "eh" | { of | of } humanity | ^and of reading } } //’

39. et "euh" "bon" "ben" ça pose des problèmes { **de maintenan~ | "enfin" de maintenance "euh"**) | { { de | de } mise à jour | ^et tout ça } "euh" } // voilà // (D0005)
‘and "erm" "well" "well" that poses problems **{ with mainten~ | "well" with maintenance "erm")** | { with | with } updating | ^and all that } "erm" } // there you go //’



Micro-syntactic constituent analysis

(Kim Gerdes, Sylvain Kahane)

We automatically calculate a phrase structure on the basis of the dependency structure.

Phrasal constituents

We call *maximal projection* of a lexeme X the set of lexemes dominated by X, i.e. X, the dependents of X, the dependents of the dependents of X, and so on.

Each lexeme in the corpus gives two constituents: a phrasal constituent, which is its maximal projection and a lexical constituent (potentially merged for the sake of simplicity). For example, the nominal lexeme gives a projection of the category NP and a lexical constituent of the category N. More generally:

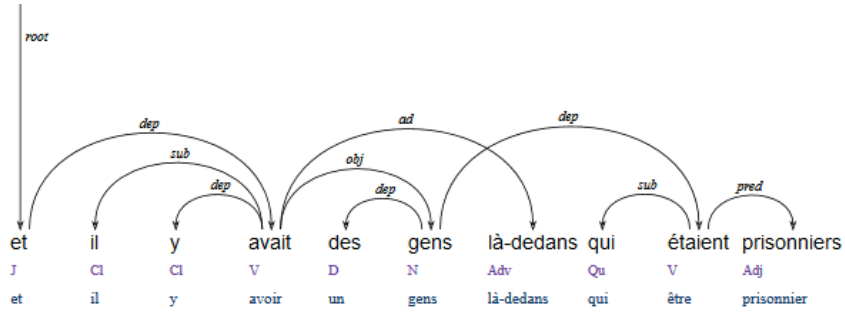
- Nouns (N) project an NP (noun phrase)
- Inflected verbs (V) projects an S (sentence)
- Uninflected verbs (V) (infinitives and participles) project a VP (verb phrase)
- Conjunctions of subordination (CS) project a CP (complementiser phrase)
- Prepositions (P) project a PP (prepositional phrase)
- Joncteurs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

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

Micro-syntactic constituent trees

Overhanging relations between micro-syntactic constituents give a tree structure which we call the *micro-syntactic constituent tree*.

The existence of non-projective dependencies adds a complication illustrated by the following example (D0003):

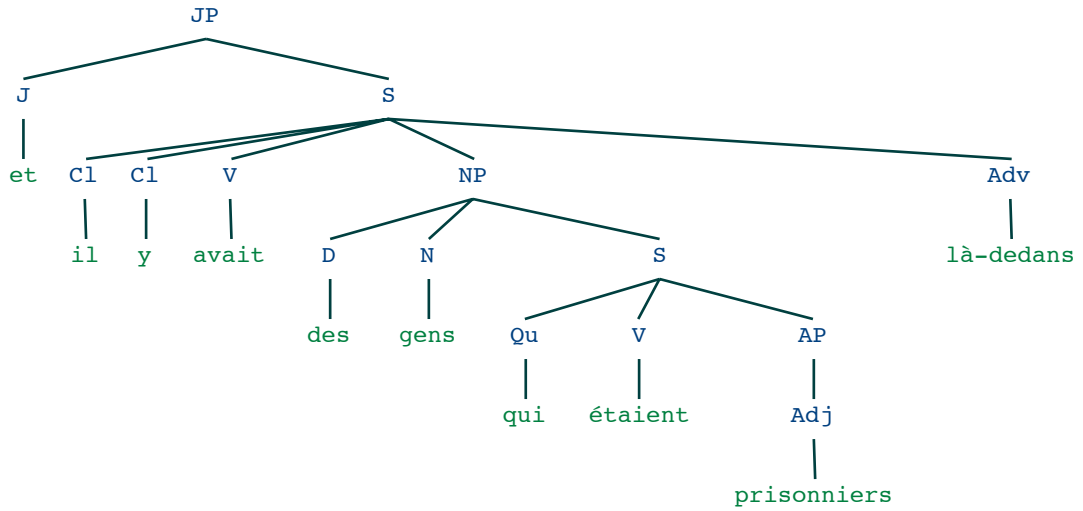


^ *et il y avait des gens là-dedans qui étaient prisonniers //*

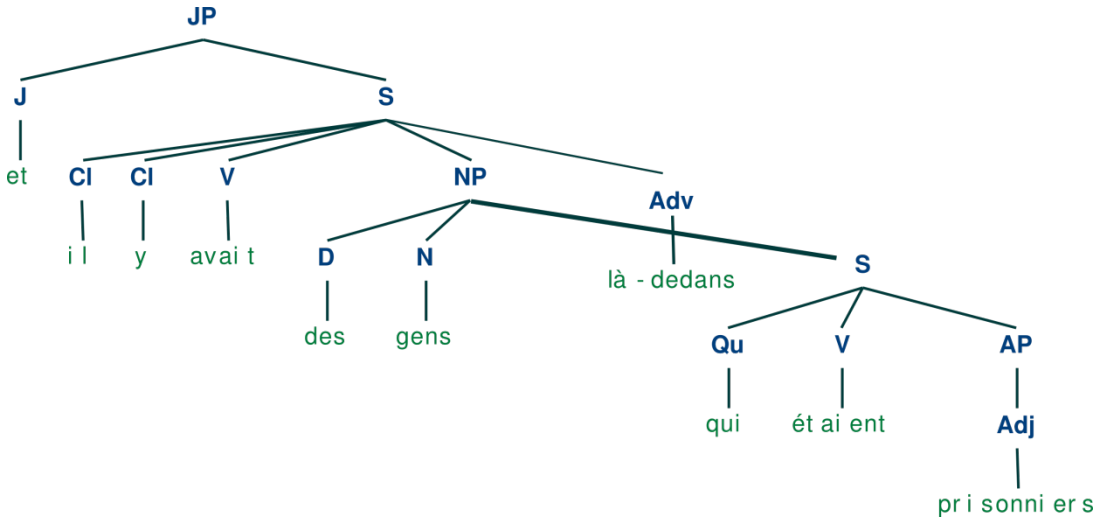
‘^and there were people in there who were prisoners //’

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 *là-dedans* interrupts the noun phrase) means that:

- either the order of the words is only partially conserved in the micro-syntactic constituent structure:



- or it is necessary to consider phrase structure trees with overlapping branches :



Function

Each constituent is given a functional label which is the label of the dependency relation which governs its head. In the preceding example the NP *des gens qui étaient prisonniers* ‘people who were prisoners’ receives the feature func = “obj”.

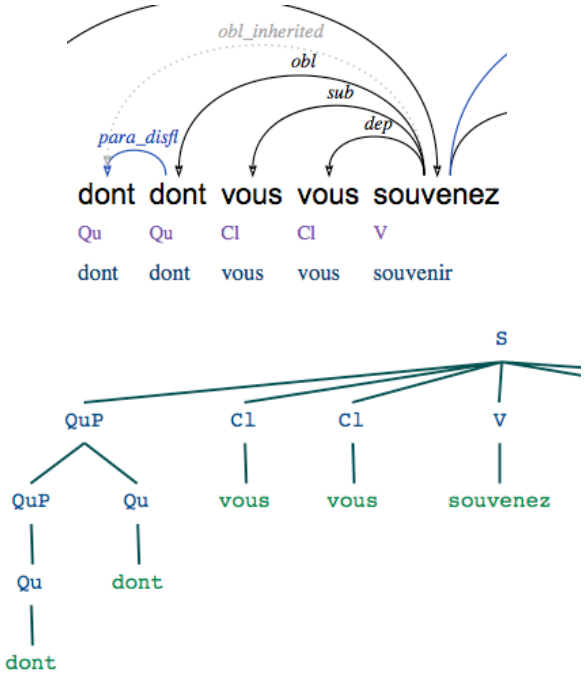
Constituents also receive the morpho-syntactic features of their head. Lexical constituents that project a phrasal constituent receive the feature func = “head”. In the preceding example, the N *gens* receives the feature func = “head”. The D *les* ‘the’, which does not project a phrasal constituent receives the feature func = “dep”.

Government units (GU) are constituents with the feature func = “root”.

Analysis of pile structures

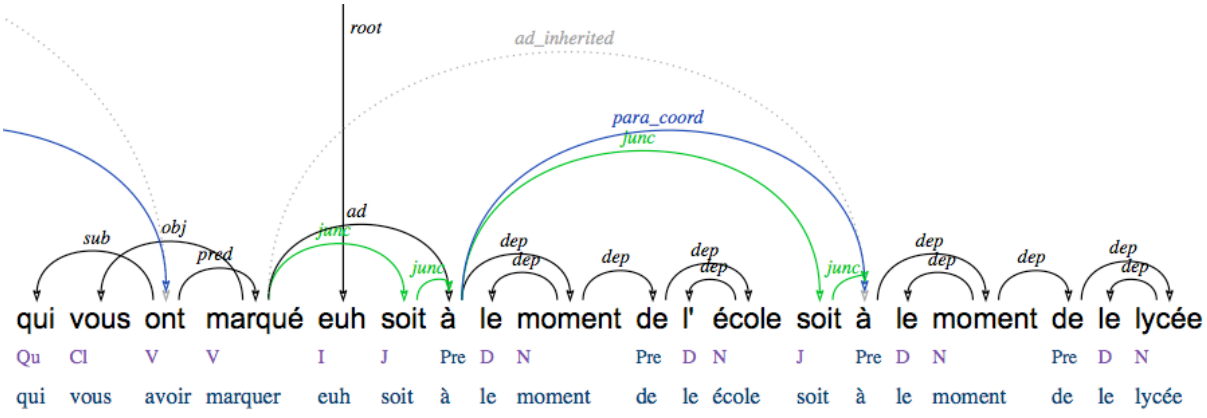
We choose to analyse the phrase structure of pile structures asymmetrically, i.e. such that inherited dependencies are not considered.

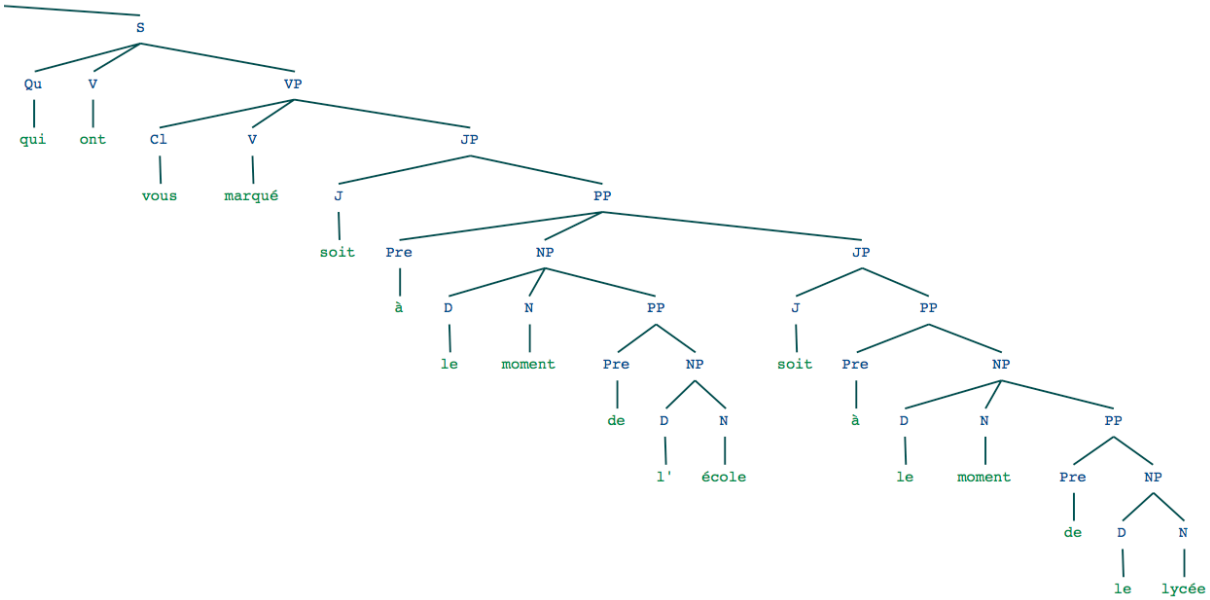
In the following example, the pile { *dont* | *dont* } is treated as a QuP (qu-word phrase) since it is the projection of the second *dont*, which is a Qu. So as to indicate that it is indeed the second *dont* that is treated as the head of the constituent, the first *dont* also projects a QuP. The two QuPs respectively receive the features func = “obl” and func = “para_disfl”.



‘that that you you remember’

When there is a junctor, it is considered the head of the layer, which becomes a JP (junctor phrase). However, the function of the JP is assigned by the link that is doubled up by the *junc*. In the following example, the first JP has the feature func = “ad”, whereas the second JP has the feature func = “para_coord”.





‘that left their mark on you, either at school or at college’