Protocole de codage microsyntaxique

Sylvain Kahane

Protocole de codage microsyntaxique

Version : 23 octobre 2013

Rédacteurs : Sylvain Kahane (en collaboration avec l'ensemble de l'équipe syntaxe Rhapsodie et en particulier Kim Gerdes, Paola Pietrandrea et Christophe Benzitoun)
Version révisée par Rachel Bawden et les autres annotatrices (Marie-Amélie Botalla et Adèle Désoyer)

Nous divisons ce document en quatre sous-sections :

- analyse morphosyntaxique
- analyse microsyntaxique en dépendance
- traitement des entassements
- analyse microsyntaxique en constituants

Analyse morphosyntaxique

(Sylvain Kahane, Kim Gerdes)

L'analyse morphosyntaxique comprend la segmentation du texte en mots (appelés dorénavant lexèmes pour éviter toute confusion avec les mots orthographiques), la lemmatisation et l'étiquetage morphosyntaxique.

Un texte est découpé en mots. Sauf exception (les amalgames comme *du* ou *au*), un mot est un lexème ou une forme fléchie d'un lexème, c'est-à-dire un lexème combiné à des morphèmes grammaticaux. (Attention nous parlons ici de mots au sens linguistique du terme. Nous les distinguons des mots orthographiques, avec lesquels ils coïncident généralement néanmoins).

La lemmatisation est l'attribution à chaque mot/lexème d'un lemme (le lemme est le nom que l'on utilise conventionnellement pour désigner un lexème ; pour un verbe il s'agit par exemple de la forme infinitive). L'étiquetage morphosyntaxique est l'attribution à chaque mot de la partie du discours de son lemme assortie d'éventuels traits catégoriels et flexionnels.

Découpage en lexèmes

Mot : Par *mot lexématique* (on dit aussi *mot-forme* dans la tradition structuraliste), nous entendons une unité linguistique particulière, généralement considérée comme l'unité minimale de la syntaxe, que nous allons définir (grossièrement). Dans le projet Rhapsodie, un autre découpage en mot a également été réalisé par les prosodistes pour le calcul des groupes rythmiques. A terme, les deux découpages devront être unifiés.

Dans la suite, le terme *mot* désignera toujours un mot lexématique.

La notion de mot est directement liée à celle de *lexème* qui est l'unité minimale du lexique. Un mot est soit un lexème invariable, soit une forme fléchie d'un lexème, soit (très marginalement) l'amalgame de deux lexèmes.

Token : Nous appelons *token* (ou *mot orthographique*) tout segment de la transcription orthographique pris entre deux blancs ou un blanc et un signe de ponctuation. L'apostrophe est également considérée comme la frontière droite d'un token et *l'enfant* est donc la combinaison de deux tokens (*l'* + *enfant*), de même que *aujourd'hui* ou *quelqu'un* par conséquent. Le tiret n'est pas considéré comme une frontière de token et *dit-on* est un token que nous décomposons en deux lexèmes.

Les conventions orthographiques sont très largement motivées par des considérations linguistiques et les tokens (c'est-à-dire les mots orthographiques) correspondent en grande majorité à des mots lexématiques et vice versa.

Nous allons préciser les critères sur lesquels repose notre définition du mot (lexématique) et puis nous indiquerons les cas pour lesquels nous considérons des mots ne sont pas des tokens.

Définition du mot (lexématique)

Un segment XY est découpé en deux morceaux X et Y si X et Y commutent librement, c'est-à-dire si X et Y s'utilisent dans d'autres contextes avec d'autres éléments tout en ayant le même sens. De plus, si X' ou Y' commutent avec X et Y dans leurs autres contextes, ils doivent aussi commuter avec X et Y dans le contexte de la combinaison XY, c'est-à-dire que X'Y, XY' et X'Y' doivent être acceptables et avoir des propriétés comparables à XY.

Les mots ne sont pas les plus petits morceaux d'un tel découpage. Par exemple, dans la forme verbale *chantons*, *chant-* et *-ons* commutent librement (*chant-* avec d'autres radicaux verbaux et *-ons* avec d'autres flexions verbales). Mais *chant-* et *-ons* possèdent une très grande cohésion : il n'est pas possible de les dissocier (d'utiliser un radical verbal sans flexion et vice versa), ni de les séparer, ni encore de les modifier indépendamment l'un de l'autre.

Les mots sont les plus petites unités qui ne peuvent pas être découpées en deux morceaux commutant librement, dissociables et séparables.

Dans un mot, si celui-ci est décomposable, seul un des morceaux appartient réellement à un paradigme ouvert et est un lexème. Les autres morceaux sont des morphèmes grammaticaux associés à ce lexème. Le mot est donc alors une forme fléchie d'un lexème.

Locution : Une importante complication est due au figement sémantique : si dans la combinaison XY, on ne peut pas attribuer un sens à X et Y, le critère précédent ne peut plus être appliqué, ce qui ne veut pas dire qu'on ne veut pas découper XY en deux morceaux d'un point de vue syntaxique. Par exemple, dans *pomme de terre*, *pomme* et *terre* ne commutent pas librement (puisque *pomme* et *terre* n'ont plus de contribution sémantique propre), mais il est visible que *pomme de terre* est un figement de l'expression libre *pomme de terre* qui est construite sur le même schéma syntaxique que *corpus de français* (N *de* N), qui est lui une combinaison libre : *corpus/texte/livre… de français/chinois/syntaxe…* Nous dirons que *pomme de terre* est analogue à *corpus de français* et qu'il doit donc être découpé de la même manière. Un segment XY est dit analogue à X'Y' s'il existe des acceptions de X et Y où X et Y se comportent de la même façon que X' et Y' et où XY se comporte de la même façon que X'Y'. Un segment XY qui ne commute pas librement mais qui est analogue à un segment qui commute librement est appelé une locution ou un phrasème.

Nous avons fait le choix de rester à un niveau syntaxique et donc de décomposer les locutions et de les analyser de la même façon que les combinaisons libres auxquelles elles sont analogues.

Nous allons donner une longue liste d'exemples qui permettra d'éclaircir cette définition. Il est important de comprendre que la notion de commutation libre, comme la notion d'analogie, sont des notions graduelles et qu'ils existent des unités dont le statut de mot est flou et pour lesquels nos choix peuvent paraître arbitraire. Néanmoins le choix de traiter une unité XY comme un tout où comme la combinaison de X et de Y et donc de créer un lien entre X et Y n'a pas d'incidence sur le reste de l'analyse d'un énoncé. Même si nous avons voulu traiter le découpage en mots avec le plus de rigueur possible, les choix que nous avons fait ont essentiellement une portée locale qui ne touche que des unités qui sont problématiques quelle que soit l'analyse retenue.

Mots à l'intérieur du token

Nous avons évité de découper des tokens en mots. Par exemple, *afin* est considéré comme un seul mot même si on peut encore y reconnaître une combinaison *à* + *fin* et que les deux sont séparables comme dans *à seule fin (de faire ça)*.

Amalgame : Nous avons séparé en deux lexèmes les amalgames *au* et *aux* : *au* = *à* + *le*. Pour *des*, nous avons distingué le cas où *des* commute avec *ces* de celui où il commute seulement avec *de ces*. Dans le deuxième cas seulement, *des* a été traité comme une combinaison *de* + *les* :

- ensuite c'est **des** escaliers (M0010:2)
- **…**dans le vingtième c'est le problème des (**de les)** écoles maternelles et primaires dans lequel... (D0002:21)

Nous avons fait les mêmes choix pour *du* selon qu'il s'agit d'un déterminant partitif et qu'il commute avec *ce* ou qu'il introduit un groupe prépositionnel en *de* :

- ça j'avoue qu'on a **du** mal quand on voit que Paul Valéry passe... (D001:112)
- **…**le sherpa du (**de le)** président le porteur de valises le conseiller influent du (**de le**) prince... (D2005:6)

Par souci d'homogénéité, *de la* et *de l'* reçoivent également deux analyses, en un ou deux mots selon les cas.

- et il y a aussi de la (**de_la)** très bonne culture (D1001:26)
- sa femme est originaire **de la** région (D009:182)

Tirets : Les tokens comprenant un tiret sont considérées comme un seul mot lexématique, sauf quand il s'agit de la combinaison d'une forme verbale et d'un clitique :

- dit-on = dit + -on
- a-t-il = a + -t-il : qui il y a dans qui y **a** -**t**-**il** dans la voiture noire (D2010:186)

Enfin, *là* dans les combinaisons du type *ce N-là* est également considéré comme un lexème à part entière :

- …très difficile d' d'apprendre le français à des petits enfants de cet âge -**là** (D002:52)

Les tokens *là-bas, là-dedans, là-dessus* sont considérés comme un unique mot lexématique. On pourrait envisager d'isoler *là-* mais sa syntaxe ne serait analogue à aucun autre élément du lexique et le paradigme des éléments qui se combine avec lui reste assez restreint, contrairement au *-là* postposé qui se combine avec tous les N.

Mots formés de plusieurs tokens

Voici les listes des mots formés de plusieurs tokens que nous avons considérés :

Mots grammaticaux

à nouveau
à part
à peine
a priori
à savoir
à travers
alors que
au moins
autre chose
bien sûr
c'est-à-dire
d'abord
d'accord (quand il s'agit de l'interjection)
d'ailleurs
de nouveau
de plus
de plus en plus
du tout
eh ben
eh bien
encore que
en fait
en tant que
en tout cas
en quelque sorte
et caetera
et puis (mais pas ou bien, ou encore, …)
jusqu'à (quand c'est un Adv (jusqu'à chez moi), mais pas quand c'est Adv + Pre (jusqu'à Paris))
l'un (mais pas *l'autre*, parce qu'on a *les deux autres*)
lors de (mais pas faute de)
n'importe quel
n'importe quand
n'importe qui
parce que
petit à petit
peut-être
quand même
quelqu'un
quelque chose

quelque part
sauf que
sur ce
surtout que
tout à fait
tout de suite
vis-à-vis (de)
y compris

Tous les nombres

deux mille neuf
dix-neuvième
dix-huit cent
dix-huit cent quatre-vingt
neuf cent cinquante
quatre-vingt-douze
trois cents
vingt-deux
…

Les noms composés

Tous les noms composés orthographiés avec un tiret ont été considérés comme des mots : après-midi, arrière-grand-mère, aujourd'hui, baby-sitter, belle-mère, centre-ville, chef-d'œuvre, contre-attaques, contre-littérature, enseignant-chercheur, fauteuil-crapaud, grands-parents, mathématicien-écrivain, mi-temps, outre-mer, pâtissier-boulanger, rendez-vous, rond-point, week-end …

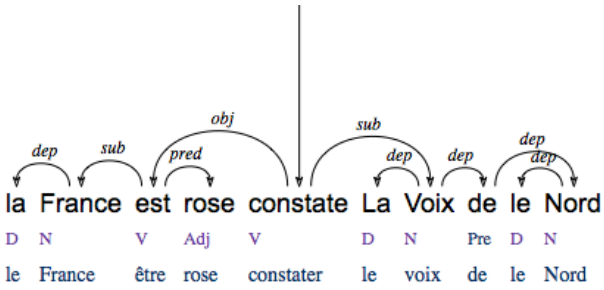
Le nom *face à face* (*un face_à_face très attendu*) est traité comme un mot, mais la locution adverbiale (*ils sont face à face*) est traitée comme trois mots distincts.

Les noms propres

General Motors
Hauts-de-Seine
Jean-Paul
Pointe-à-Pitre
Proche-Orient
Royaume-Uni
Saint-Jean
Saint-Jean-de-Maurienne
(avenue) Alsace-Lorraine

Les séquences Prénom Nom (*Françoise Giroud*) sont considérées comme la combinaison de deux mots (ne serait-ce que parce que chacun des deux est effaçable au profit de l'autre) et le premier est traité comme la tête.

Les noms composés en *N de N* sont analysés comme une combinaison de mots, y compris des noms propres comme *La Voix du Nord* :



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

Un certain nombre de composés très cohésifs ont été quand même considérés comme des combinaisons de mots car ils possèdent encore une certaine combinatoire syntaxique :
à côté (à droite, à gauche, …), mais pas à_travers qui se construit différemment : à côté **de** la maison vs à travers la maison
à la fois (??)
à le fond de, à le milieu de, à le dessus de, (tout) à le long de, à le moment de, à le sein de, à le travers de, à l'égard de
à partir de
à raison de, à cause de, à propos de
de ce que
de côté (de face, de près, de loin)
en gros, en cours, en général, en bas
en matière de, en face de, en raison de
faute de
grâce à (face à)
l'autre (un autre, deux autres, les autres …) mais pas l'un, les uns (*deux uns)
même si
par hasard (par chance)
par rapport à
petit à petit (pas à pas, face à face)
plus jamais
qui que ce soit
sans doute
un peu (un petit peu, à peu près, pour le peu que j'en sais …)

Les locutions composées sont considérées comme deux mots quand il y a commutation sur *que* :
dès que je suis arrive ~ dès mon arrivée -> dès + que
alors que je partais ≠ *alors mon départ -> alors_que

Partie du discours

Nous considérons 13 parties du discours :

- V pour les verbes (et *voici* et *voilà* lorsqu'ils prennent des compléments, comme par exemple *voici un extrait...*)
- N pour les noms
- Adj pour les adjectifs
- Adv pour les adverbes

- Pre pour les prépositions
- CS pour les conjonctions de subordination
- J pour les joncteurs : il s'agit des traditionnelles conjonctions de coordinations et d'autres éléments qui lient les couches d'un entassement, comme *c'est-à-dire* ou *y compris*. Les éléments clôtureurs d'entassement comme *et caetera* sont classés comme joncteurs également.
- D pour les déterminants
- I pour les interjections, y compris des marqueurs de discours comme *bon, ben, euh, hein,* (les formes impératives comme *allons, écoute, tiens* sont traitées comme des verbes à l'impératif et non pas comme des interjections)
- Qu pour les mots qu- que sont les relatifs et les interrogatifs
- Cl pour les clitiques, y compris les clitiques sujets (*je, tu, il, on, ce*) et l'adverbe de négation *ne*.
- Pro pour les autres pronoms
- X pour les éléments dont on ne peut déterminer la catégorie syntaxique : partie inaudible (XXX), certaines amorces (quand on ne peut pas deviner le lexème et sa partie du discours), ainsi que les positions non instanciées marquées par &, uniquement quand on ne peut pas déterminer la catégorie attendue.

Remarques :

- Les numéraux sont classés D ou Adj selon leur position. Ainsi *deux* est D dans *deux chaises* et Adj dans *les deux chaises*. Même chose pour *quelques* dans *quelques chaises* et *ces quelques chaises*. Ce choix est notamment justifié par le fait que nous ne faisons pas d'annotation fonctionnelle des déterminants (puisque tous les dépendants d'un N ont la même fonction) et que l'étiquette D est donc autant fonctionnelle que catégorielle.
- Le modifieur *tout* sont classés de la même façon que les numéraux lorsqu'il se rattache à un nom. Ainsi *toute* est D dans *de toute façon* et *tout* est Adj dans *tout le monde*
- Les déictiques comme *demain* sont classés parmi les Adv en suivant la tradition, même s'il y a de bons arguments pour en faire des noms au même titre que *lundi* : *il vient demain/lundi/lundi prochain/ce lundi*
- A l'inverse *grâce* dans *grâce à lui* est classé parmi les N même s'il fonctionne là comme un Adv.

Pro : Pro est une des catégories les plus difficiles à discerner. Nous définissons comme Pro :

- les pronoms personnels toniques (*moi*, *toi*, *soi*, *elle*, *lui*, *eux* …) y compris les pronoms possessifs (*mien*, *tien, leur* …)
- les pronoms démonstratifs *ça, cela* et *ceci* et la forme *ce* uniquement quand il est dans une construction relative indéfinie du type *ce que j'aime…* ou *ce qui me plaît…* A noter que *ce* en tant que sujet d'un verbe est considéré comme un Cl.
- les pronoms indéfinis *rien, chacun* et *tout*, sauf quelques exceptions telles que *un petit rien* où *rien* a plutôt le rôle d'un N.
- les nombres lorsqu'ils représentent des pronoms quantifiable, ex : ***l'un*** (mais pas *l'autre*), *il y en a* ***dix****, il en faut* ***un.***

  Il ne fait pas confondre des nombres représentants des concepts nominaux avec les pronoms. Par exemple les emplois suivants sont considérés non pas comme des pronoms, mais comme des noms :

    - les nombres des bus
    - des étages
    - des classes scolaires (*première, seconde*)
    - les arrondissements (*dans le **vingtième***)
    - les scores de football (***un** à **zéro***)
    - les numéros de page *(page **cent douze** de votre ouvrage*)
    - les opus (*deux préludes de Karol Szymanowski extraits de son opus **un** de mille neuf cents le **sixième** et le **septième** interprétés par…*)
    - un chiffre en tant que chiffre (*ce chiffre de **trente mille***)
    - les dates
    - etc.
- et les suivants comme des adjectifs :
    - un nombre de personnes (*tous les **deux**, les **deux** elles passent…*), suivant notre analyse des numéraux ci-dessus, même s'il y a de bons arguments pour les inclure dans la catégorie des pronoms. Par contre *l'un* et *les uns* sont traités comme des pronoms, puisque *l'un* est considéré comme un seul token et a un comportement différent de *les deux* ou *les autres*.
    - les ordinaux (*le **premier***, *la **deuxième***)

Quelques exemples de pronoms :

> il { faut | faut } compter autour de { **soixante** | **soixante-dix** } // (D2009)
> ^ parce ^ que c' était une frange de { **vingt** | **trente** } // (D2009)
> corner à **deux** (D2003)

Traits morphosyntaxiques

Les V reçoivent un trait de mode qui peut prendre 6 valeurs : *indicative, subjunctive, imperative, infinitive, past_participle, present_participle*. Seuls les V à l'indicatif varient en temps ; le trait *tense* possède 5 valeurs : *present, imperfect, future, conditional* et *perfect* (qui correspond au passé simple et qui ne figure qu'une seule fois dans notre corpus). Les temps composés sont traités au niveau syntaxique et non morphologique ; un passé composé sera donc un V *être* ou *avoir* au mode="indicative", tense ="present" dont le dépendant *pred* est un V au mode="past_participle". Il n'y a pas de marquage explicite de la distinction passé composé vs. passif pour une forme ambiguë comme *il est passé*.

Les V reçoivent aussi des traits d'accord : le trait *number* a deux valeurs *sg* et *pl*, le trait *genre* deux valeurs *fem* et *masc* et le trait *person* trois valeurs *1, 2* et *3.*

Les N, Adj et D ont les traits *number* et *genre*. Les Cl et Pro ont en plus un trait *person*.

Certains traits peuvent rester sous-spécifiés. Par exemple, les noms de villes et les emplois nominaux de nombres ont un trait number= "masc/fem".

Lemmes

Les lemmes sont comme il est d'usage la forme pour les lexèmes invariables, la forme infinitive pour les verbes, le singulier pour les noms et le masculin singulier pour les adjectifs.

Le lemme pour les articles *le, la, l', les* est *le*, le lemme pour *un* et *des* est *un* et le lemme pour *du, de_la, de_l'* est *du* (même s'il y a de bons arguments pour en faire une forme de *un* au même titre que *des*).

Le lemme pour les pronoms de 1ère et 2ème personne *je, tu, nous, vous, me, te …* est la forme. Le lemme pour les pronoms de 3ème personne est la forme du singulier : par exemple, *lui* pour *eux* ou *leur*, *elle* pour *elles*, *il* pour *ils*.

Le lemme pour les déterminants possessifs (*mon, ma, mes, ton, ta, tes …*) est toujours *son*.

Le lemme pour les mots inachevés est inachevé même si on pense pouvoir reconstruire le mot que le locuteur souhaitait produire :

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

Analyse des noms propres

Nous avons déjà vu dans l'étape de la segmentation notre choix d'analyser la structure interne de mots constitués de plusieurs tokens. Nous reproduisons ce choix par rapport aux noms propres, en analysant la structure interne si elle est suffisamment productive.

Notre corpus contient un grand nombre de titres de livres, de journaux et d'établissements qui contiennent plusieurs tokens, dont des noms propres, des noms communs, des adjectifs et des mots grammaticaux. Dans le cas où un token individuel n'est pas lui-même un nom propre, même s'il appartient à un nom propre constitué de plusieurs mots, le lemme aura la forme habituelle. Par exemple dans

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

le titre *Des épidémies et des Hommes* correspond aux lemmes 'de+le', 'épidémie', 'et', 'de+le', 'homme'. Il est à noter que la majuscule de *Hommes* n'est pas transféré à son lemme, puisqu'il ne s'agit pas d'un nom propre.

Les noms des journaux ont également nécessité des conventions de lemmatisation particulières.

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

Le nom propre *Figaro* ou *Libération* est traité comme un nom propre et garde alors sa majuscule dans son lemme, même si le mot *libération* est ailleurs un nom commun. L'article par contre est traité comme un mot grammatical est analysé selon les règles citées ci-dessus. Les lemmes correspondants au second exemple sont 'le', 'dernier', 'nouvelle', 'de' et 'Alsace'.

Analyse microsyntaxique en dépendance

(Sylvain Kahane, Kim Gerdes)

La syntaxe décrit la façon dont les unités linguistiques se combinent. La microsyntaxe décrit les relations entre mots caractérisées par une forte cohésion syntaxique.

Rection et dépendance microsyntaxique

Rection : La microsyntaxe se limite aux relations de type *rection*. On parle de rection lorsqu'un élément impose à un autre élément sa nature, ses marqueurs et/ou sa place. Par exemple, le complément d'objet d'un verbe est *régi* par ce verbe. Dans *Pierre admire le paysage*, *le paysage* est régi par la forme verbale *admire*. En effet :

- la forme est imposée : le paradigme des éléments qui peuvent commuter avec *le paysage* se limite à des groupes nominaux ;
- les marqueurs sont imposées : dans le cas du complément d'objet direct en français, il n'y a pas de marqueur explicite, mais si le complément est pronominalisé (*Pierre l'admire*), une forme particulière du pronom doit être utilisée ;
- la place est imposée : le complément d'objet direct doit suivre le verbe (sauf formes pronominales particulières ou rares cas d'antéposition (*deux euros ça coûte*)).

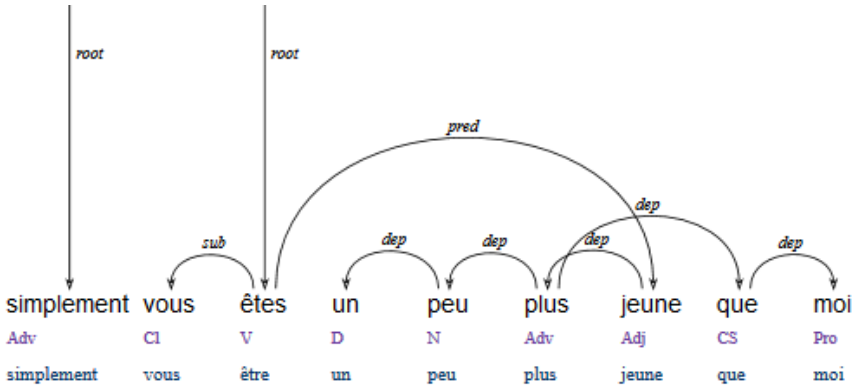
Nous retenons comme un des tests majeurs pour caractériser les éléments régis par un verbe la possibilité d'être clivé (*c'est le paysage que Pierre admire*).

Dans :

> simplement < vous êtes un peu plus jeune que moi // (D0001)
>
> ^et "euh" donc < { ces | ces } années d'école < ça a été des bonnes années // (D0001)

les syntagmes *simplement* ou *ces années d'école* sont non dépendants car considérés comme non régis, puisque non clivables :

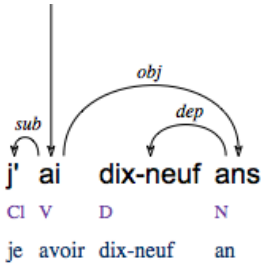
> *c'est simplement que vous êtes un peu plus jeune que moi
>
> *c'est ces années d'école que ça a été des bonnes années



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

Notons au passage que le fait de faire dépendre *que moi* de *plus* est justifié par le fait que la présence de ce syntagme est validé par la présence de *plus* (**vous êtes jeune que moi*) et que *plus et que moi* peuvent former ensemble un syntagme autonome (*vous êtes jeune //+ plus que moi > en tout cas //*).

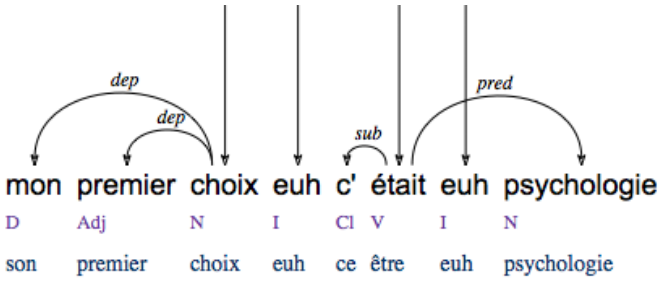
Dépendance microsyntaxique : Nous décidons d'encoder la structure microsyntaxique par un graphe de dépendance. Formellement, une dépendance est une relation orientée entre deux mots que nous représentons par une flèche : l'origine de la flèche est appelée le *gouverneur* et la cible le *dépendant*. (Dans le cas où la dépendance encode une relation de rection, on peut également utiliser les termes *recteur* et *régi*). Chaque dépendance représente une relation de rection. Dans l'exemple suivant, *dix-neuf* est un modifieur de *ans* : nous représentons cela par une dépendance de *ans* (le gouverneur) vers *dix-neuf* (le dépendant) :



De même, *dix-neuf ans* est le complément d'objet de la forme verbale *ai* : nous représentons cela par une dépendance de *ai* vers la tête du syntagme *dix-neuf ans*, c'est-à-dire *ans*. La *tête* d'un syntagme est le lexème qui domine les autres du point de vue de la dépendance. Le verbe *ai* qui est le verbe principale de cet énoncé n'est donc pas régi. Nous marquons cela par une dépendance verticale.

Unité rectionnelle : Une unité rectionnelle (UR) est une unité maximale pour la rection. Une UR possède une tête qui n'est pas régie et tous les éléments de l'UR sont dominés par cette tête, c'est-à-dire qu'ils sont régis par un lexème qui est régi par un lexème, …, qui est régi par la tête de l'UR. Autrement dit une UR est la projection maximale d'un lexème non régi.

Nous distinguons l'UR de l'unité illocutoire (UI) (cf. macrosyntaxe). Une UI peut être composée de plusieurs UR. Dans l'exemple suivant on a quatre UR, le groupe nominal *mon premier choix*, la proposition *c'était psychologie* et les deux occurrences du marquer de discours *euh*.

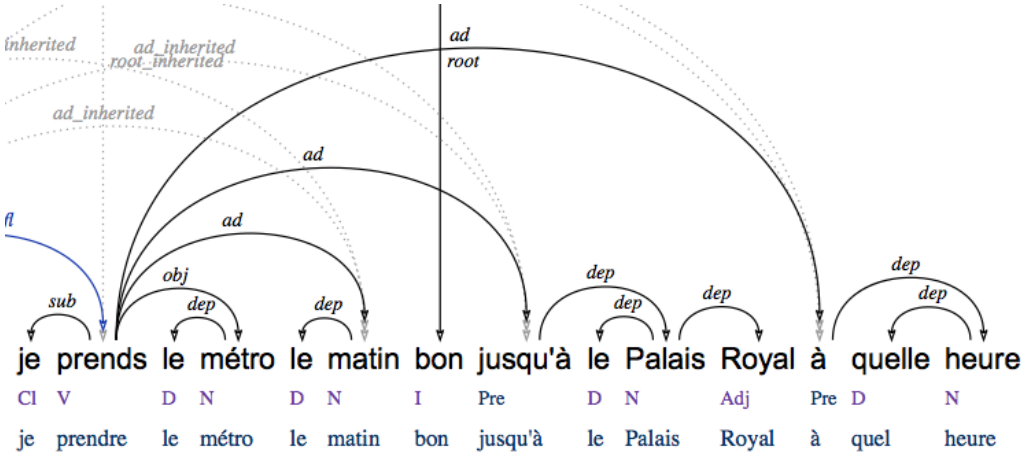


mon premier choix " euh " < c' était " euh " psychologie //

Rection au delà de l'Unité Illocutoire et du tour de parole : Rappelons que la rection ne s'arrête pas au tour de parole. Dans l'exemple suivant :

$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)

la question de $L2 (*à quelle heure*) continue la construction microsyntaxique qui précède (*je prends le métro le matin jusqu'à Palais Royal*) et la réponse de $L1 (*je prends le métro le matin à huit heures et demie*) a exactement la même structure que la concaténation des deux tours de parole qui précèdent (*je prends le métro le matin à quelle heure*).

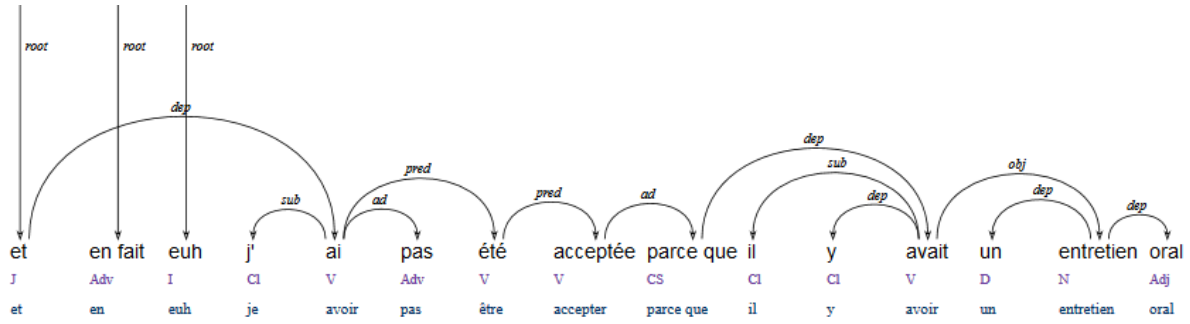


Dépendance *vs* rection : Notre structure de dépendance est plus restrictive que la structure de dépendance utilisée dans d'autres cadres théoriques (cf. par exemple Mel'čuk 1988, *Dependency Syntax*). En effet, nous ne représentons par notre structure de dépendance que les faits de microsyntaxe, c'est-à-dire ceux où il y a rection. La dépendance est un moyen formel qui permet de représenter différentes relations de « dépendance ». Nous aurions pu décider de représenter également des faits de macrosyntaxe, comme la « dépendance » des adnoyaux au noyau. Nous avons décidé d'encoder séparément micro et macrosyntaxe et de n'utiliser la dépendance que pour les faits de microsyntaxe.

Choix de la tête : La *tête* d'un groupe est intuitivement l'élément le plus important de ce groupe. Il s'agit d'une part de l'élément qui contrôle la distribution de ce groupe (la tête

externe) et d'autre part l'élément qui valide la présence des autres éléments du groupe (la tête interne). Nous allons regarder les principales configurations qui peuvent poser problème.

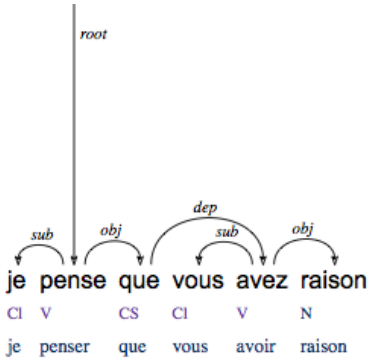
Auxiliaire : La tête d'une proposition est ce qu'on appelle traditionnellement le verbe principal. Lorsqu'il y a une forme verbale complexe nous traitons l'auxiliaire comme la tête. Dans l'exemple suivant *ai* régit *été* qui régit *acceptée* :



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

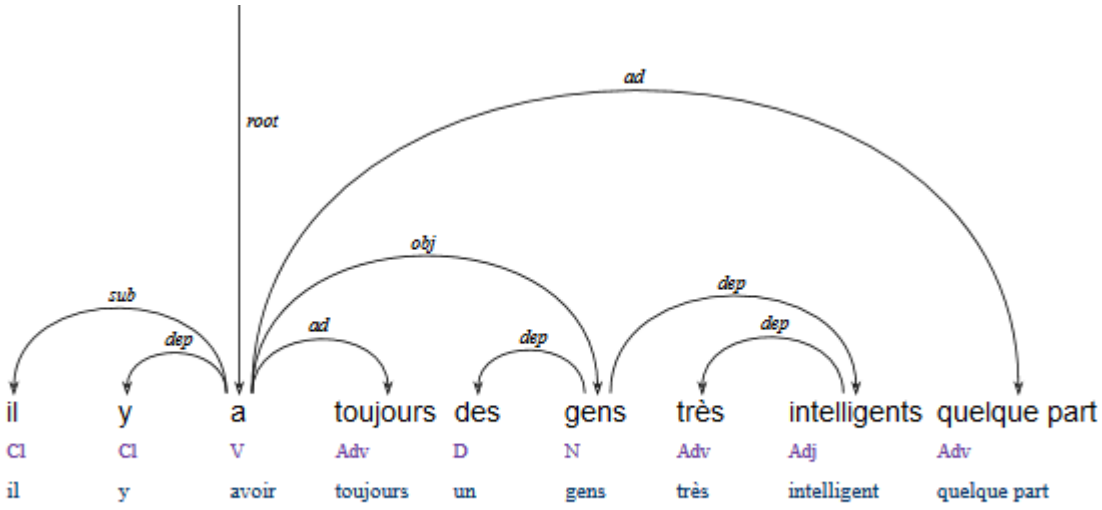
Ce choix est justifié par le fait que l'auxiliaire est la forme finie, il porte donc le mode (*elle* ***a*** *été acceptée* vs *il est étonnant qu'elle* ***ait*** *été acceptée*) et les modalités énonciatives (***a-t-elle*** *été acceptée ?*). De plus, il impose un régime (c'est-à-dire un marquage rectionnel particulier) au verbe auxilié (*elle a* ***accepté*** (participe passé) vs *elle va* ***accepter*** (infinitif)).

Marqueurs (préposition, conjonction de subordination) : Les marqueurs sont généralement traités comme des têtes du groupe qu'il marque puisqu'ils en contrôlent ainsi la distribution. Ainsi dans *Pierre parle à Marie*, *à* est la tête du groupe *à Marie* et dans *Pierre pense que Marie dort*, *que* est la tête de la proposition subordonnée *que Marie dort*.



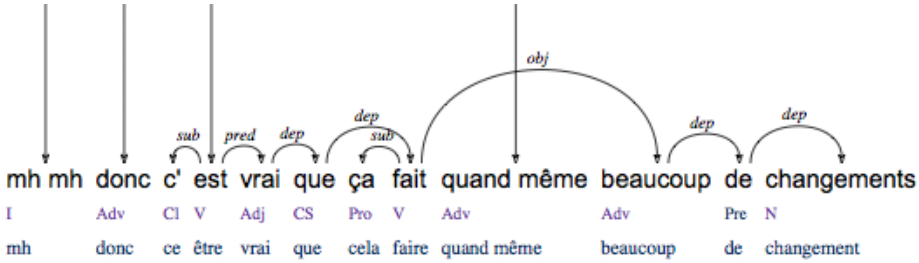
je pense que vous avez raison //

Déterminant : Nous considérons que dans le syntagme nominal, le nom est la tête est qu'il est donc le gouverneur du déterminant. Ce choix est en partie arbitraire (le déterminant possède aussi un rôle de marqueur du syntagme nominal qui justifierait d'en faire la tête), mais il reste le plus usuel en syntaxe de dépendance.



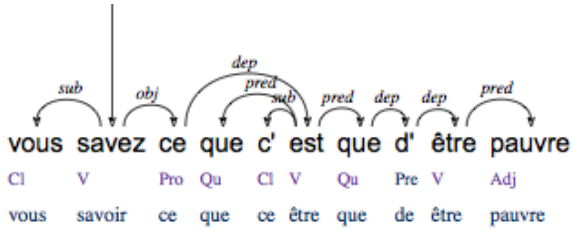
il y a toujours des gens très intelligents quelque part //

Déterminant complexe : Signalons également que, dans les syntagmes nominaux de la forme « Adv de N » (*peu de gens, trop de gras …*), nous considérons l'Adv comme la tête :



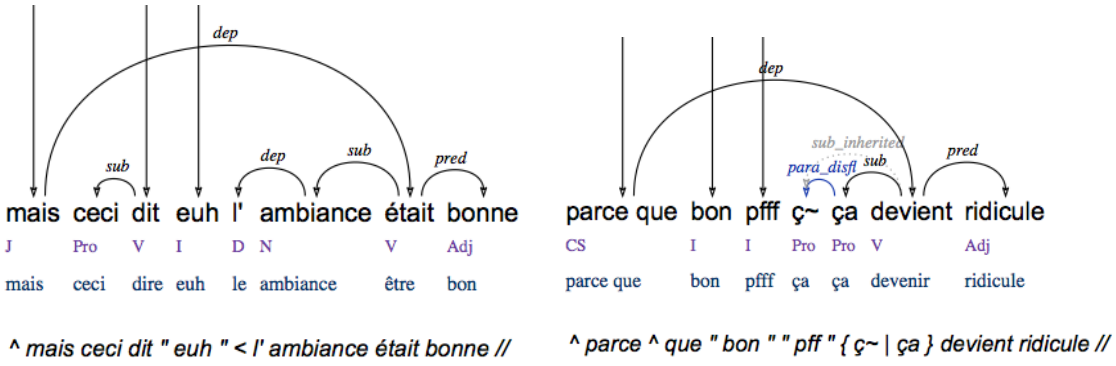
Complétive et adjectif : On notera aussi dans l'exemple précédent que dans les constructions « c'est Adj que P » nous considérons que la complétive « que P » dépend de l'Adj. Il s'agit encore une fois d'une analyse de surface qui ne tient pas compte du fait que la complétive est un sujet profond (que P est Adj), car en surface le sujet de *être* est le pronom *ce*. Elle se base sur la possibilité de faire un syntagme autonome de l'Adj et de la complétive dans certains cas (*impossible qu'il vienne*).

Un cas limite est la construction suivante, où nous analysons les deux *que* comme des pronoms en place d'un Adj (*c'est dur d'être pauvre*) :



Nos autres choix seront explicités en étudiant différentes constructions les unes après les autres.

Introducteur d'UI : Ceux-ci, qu'ils soient des joncteurs (J) ou des conjonctions de subordination (CS) sont traités comme la tête de l'UI et ont donc comme dépendant la tête du noyau :



Le cas des CS qui sont introducteurs d'UI va être expliqué dans le paragraphe suivant.

Subordonnées non intégrées : Nous considérons que dans certains cas une conjonction de subordination peut introduire une proposition qui n'est pas à proprement parler « subordonnée », puisqu'elle ne remplit pas les critères de rection. Comparons :

(1) Pierre est à la fac parce qu'il a un article à finir

(2) Pierre est à la fac parce que sa voiture est dans le parking

Ces deux exemples ont apparemment la même structure de surface *P1 parce que P2*. Pourtant la relation sémantique entre P1 et P2 n'est pas la même dans les deux cas et les propriétés qui en découlent également.

Dans (1), P2 cause P1. Dans ce cas, on peut antéposer *parce que P2*, le cliver et insérer *et cela* :

(3) **a**. parce qu'il a un article à finir Pierre est à la fac
b. *c'est* parce qu'il a un article à finir *que* Pierre est à la fac
c. Pierre est à la fac *et cela* parce qu'il a un article à finir

Dans (2), P2 ne cause pas P1, mais le fait que le locuteur pense que P1. Dans ce cas l'antéposition de *parce que P2* n'est possible que dans une structure échoïque :

(4) $L1 Pierre est à la fac, sa voiture est dans le parking
$L2 et alors parce que sa voiture est dans le parking, Pierre est à la fac ?

On ne peut pas le cliver véritablement :

(5) #c'est parce que sa voiture est dans le parking que Pierre est à la fac

même si on peut quand même le cliver avec une négation :

(6) c'est pas parce que sa voiture est dans le parking que Pierre est à la fac

Néanmoins (6) doit être contrasté avec (7) construit à partir de (1) :

(7) c'est pas parce qu'il a un article à finir que Pierre est à la fac

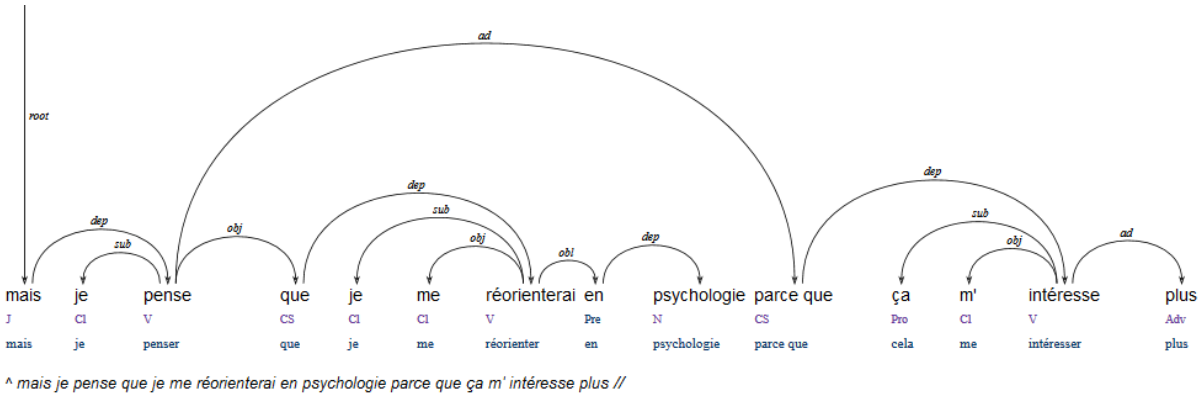
En effet, (6) n'implique pas P1 (on peut dire ça si on pense que Pierre n'est pas à la fac et c'est même surtout dans ce cas qu'on dira (6)), alors que (7) présuppose P1 (ce qu'on discute c'est la cause de P1 et donc ça n'a de sens que si P1 a lieu).

Enfin on ne peut pas ajouter *et cela*, mais on peut ajouter *et je pense cela* :

(8) **a.** #Pierre est à la fac *et cela* parce que sa voiture est dans le parking
b. Pierre est à la fac *et je pense cela* parce que sa voiture est dans le parking

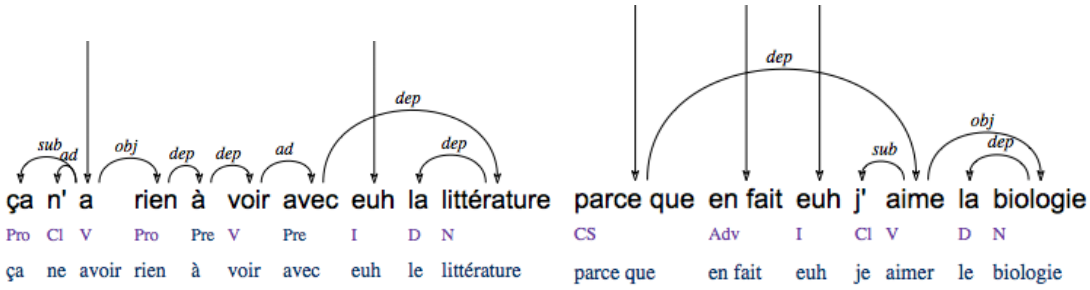
En conclusion, nous considérons que, dans le cas de (1), la subordonnée *parce que P2* est un modifieur du verbe de P1 et fait donc partie de la même UR :

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



Dans le cas de (2), nous décidons de ne pas marquer la rection de la « subordonnée » *parce que P2* :

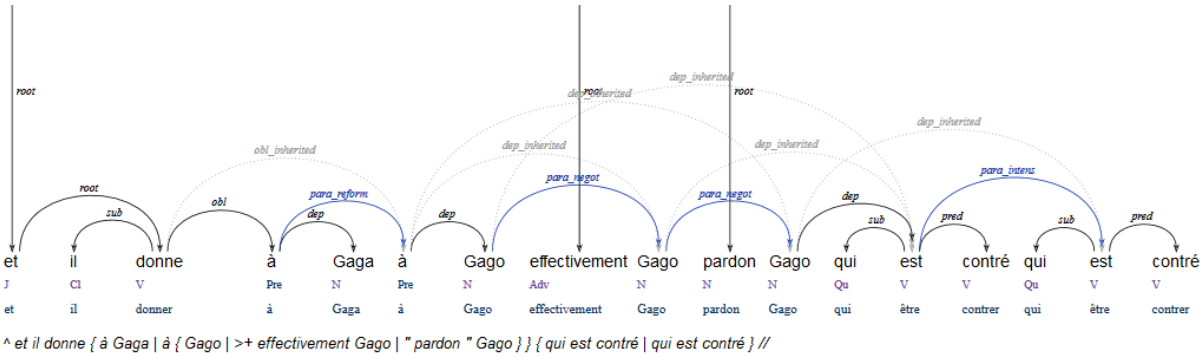
(10) c'est un bac "euh" { SMS | donc technologique } // c'est sciences médico-sociales // ça n'a rien à voir avec "euh" la littérature // ^parce ^qu' en fait < "euh" j'aime la biologie // (M1001:7)



Dans les deux cas, *parce que* est catégorisé comme une conjonction de subordination et il régit le verbe principal de la proposition « subordonnée ».

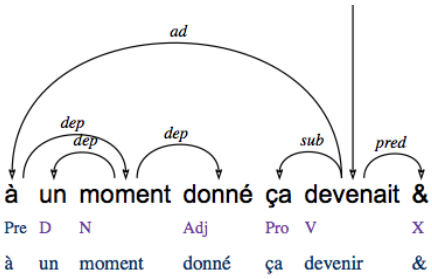
Rection au delà de l'UI : Notons que nous ne considérons pas l'unité illocutoire (UI) ou même le tour de parole comme une frontière d'UR. Par exemple, l'échange suivant sur 4 tours de parole donne lieu à une unique UR (entrecoupée d'autres UR comme *effectivement* ou *pardon*) :

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



Nous présenterons le traitement des mots *qu-* (relative, clivage …) après les fonctions syntaxiques et le traitement des entassements (coordination, reformulation …) dans un chapitre à part.

Inachèvement : Lorsqu'une UR apparaît clairement comme inachevée, c'est-à-dire qu'une position obligatoire n'est pas remplie, nous la marquons par un &. Ce symbole indique la position non instanciée, il n'est pas une instanciation par un élément vide d'une position (cf. les traces en grammaire générative).

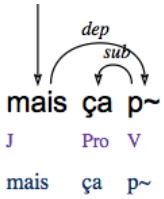


{ en vingt-cinq | en vingt-cinq } ans <+ les gens & // il y a vingt-cinq ans <+ les gens ne connaissaient rien "hein" //

$L1 ils savaient pas travailler un & // { ils sa~ | ils savaient } pas utiliser un ordinateur //

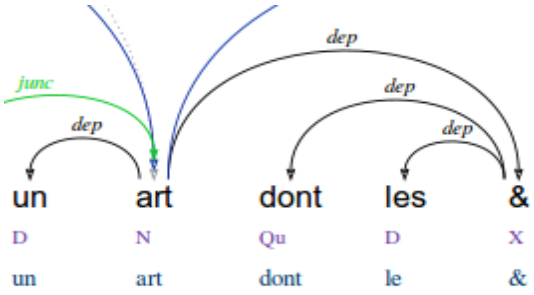
Lorsqu'il y a déjà un mot inachevé, nous n'indiquons pas en plus un inachèvement syntaxique :

^mais ça p~ // c' est pas obligatoire //



Nous limitons le marquage de l'inachèvement à une seule esperluette, même dans les cas comme le suivant où nous pourrions supposer qu'un nom et un verbe sont manquants (*art* serait lié au verbe attendu de la proposition relative et *les* au sujet nominal de ce verbe) :

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~ } } //



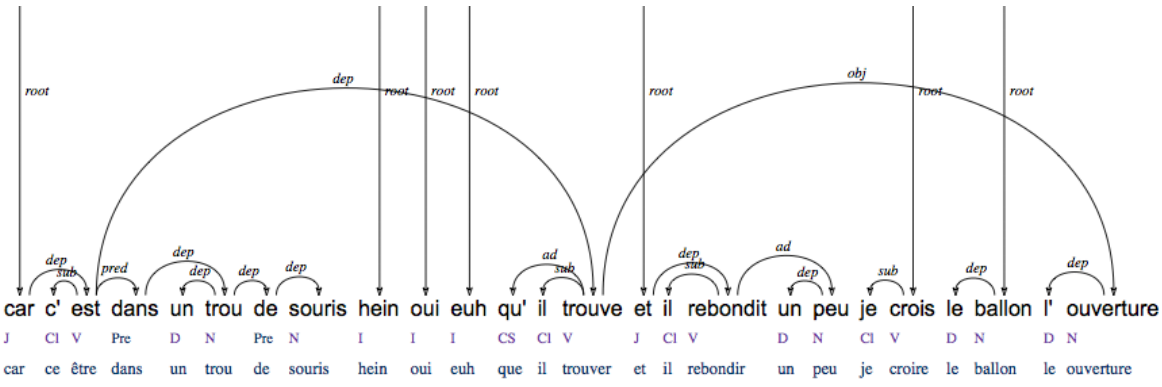
Fonctions syntaxiques

Nous avons décidé de réduire le nombre de fonctions syntaxiques au minimum. Les fonctions syntaxiques peuvent être introduites dans deux buts assez différents :

- rapprocher des dépendants qui se comportent de la même façon et distinguer ceux qui se comportent différemment. Ainsi, le complément *à Marie* de *parler à Marie* se comporte de la même façon que celui de *donner quelque chose à Marie* (*lui parler, lui donner quelque chose*), mais différemment de celui de *penser à Marie* (*penser à elle, y penser*).
- distinguer les différents dépendants d'un même mot. Par exemple, dans *Pierre a nommé Louis général*, *Louis* et *général* sont deux dépendants du même verbe, et seul le premier est cliticisable (*Louis, Pierre l'a nommé général ; *Général, Pierre l'a nommé Louis*).

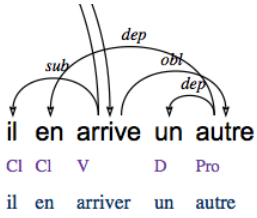
Nous nous plaçons clairement dans la deuxième optique. Seuls les dépendants du verbe ont été distingués et 7 fonctions ont été considérées :

- root : pour toutes les racines, c'est-à-dire les éléments qui ne sont pas régis par un autre élément :

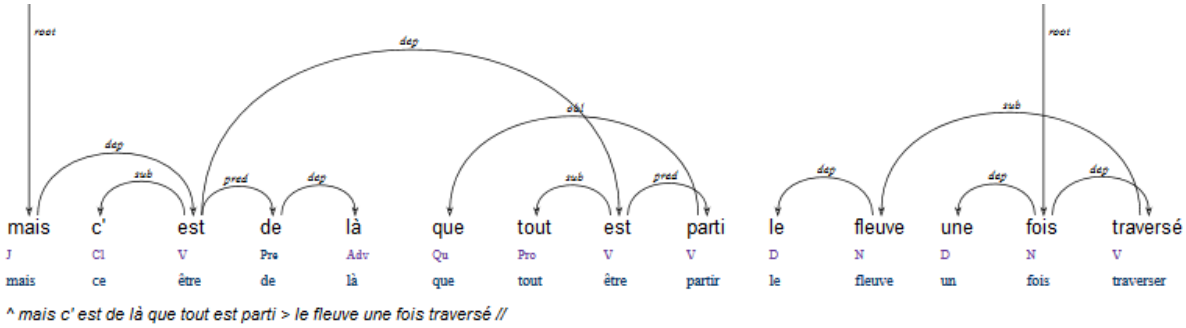


^ car c' est dans un trou de souris " hein " (oui //) " euh " qu' il trouve (^ et il rebondit un peu " je crois " > le ballon //) l' ouverture //

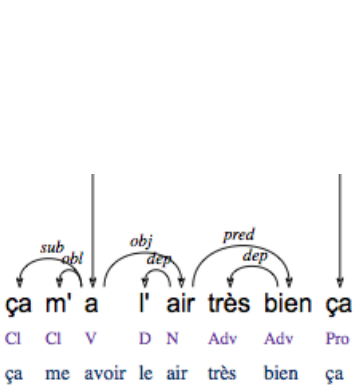
- sub : pour le *sujet* du verbe ;
  - le sujet est le sujet grammatical :

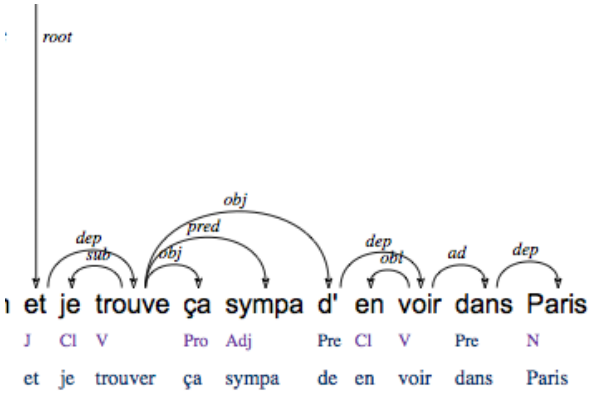


- la fonction sub est également utilisée pour des constructions prédicatives. Dans <u>*les mains*</u> *sur la tête il rigolait*, *les mains* est le sujet de *sur la tête* et il y a donc une dépendance sujet de *sur* à *mains* (cf. exemple similaire ci-dessous avec *le fleuve une fois traversé* = une fois que le fleuve a été traversé) :

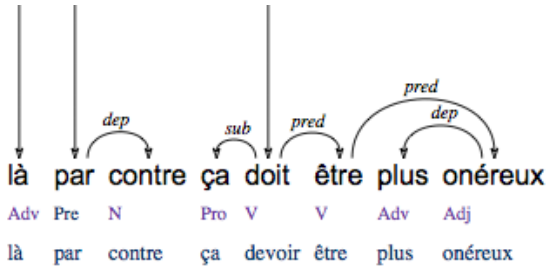


- obj : pour le *complément d'objet direct* :
  - nous incluons les compléments de mesure lorsqu'ils sont cliticisables (pas d'exemple dans le corpus) : *il a payé <u>dix euros</u> pour ce livre* (*il les a payé pour ce livre*), mais pas *il a payé ce livre dix euros* (**il les a payé ce livre*)
- pred : sont traités comme des pred, tous les éléments qui forment un *prédicat complexe* avec le verbe qui les gouverne :
  - les constructions dite attribut du sujet (*il est <u>gentil</u>*) ou de l'objet (*il trouve Marie <u>gentille</u>*) :



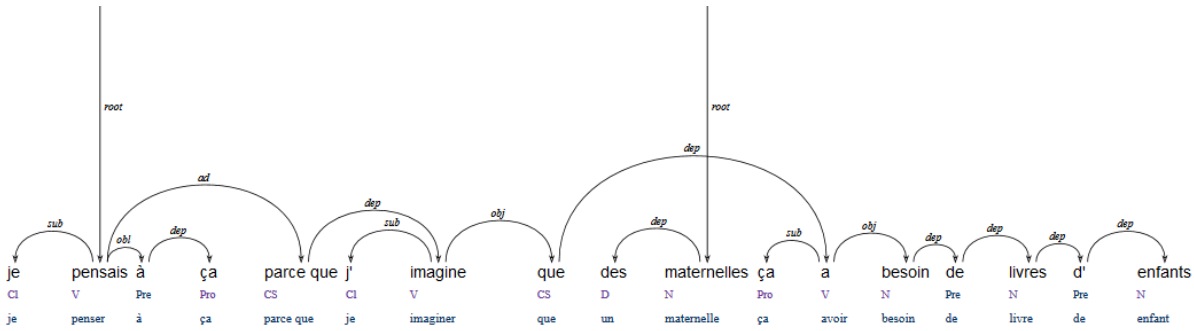


  - les formes verbales complexes (*avait mangé, est parti*)
  - les constructions avec un verbe modal (*peut venir, doit manger*), où l'infinitif ne commute pas facilement avec un syntagme nominal, y compris des constructions comme *vouloir dire* (dans *qu'est que ça veut dire ?*) où l'infinitif ne peut pas commuter avec un syntagme nominal :

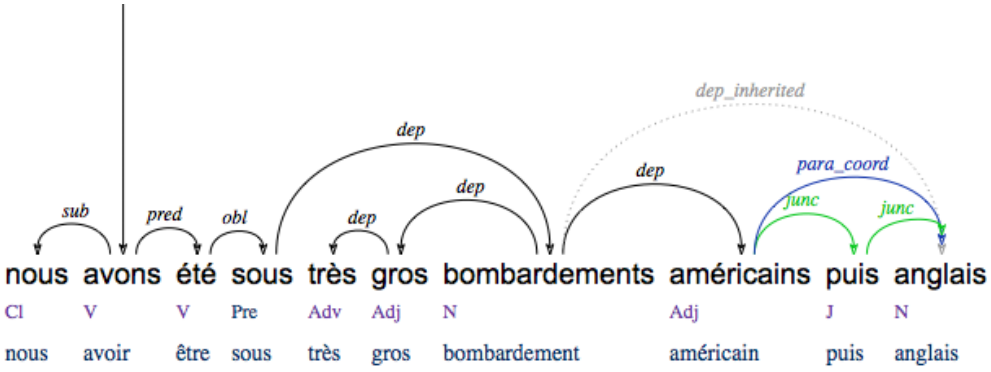


là < par contre < ça doit être plus onéreux //

- sauf les constructions à verbe support (*avoir l'intention, avoir besoin, faire peur …*), où le nom prédicatif est traité comme un objet direct :

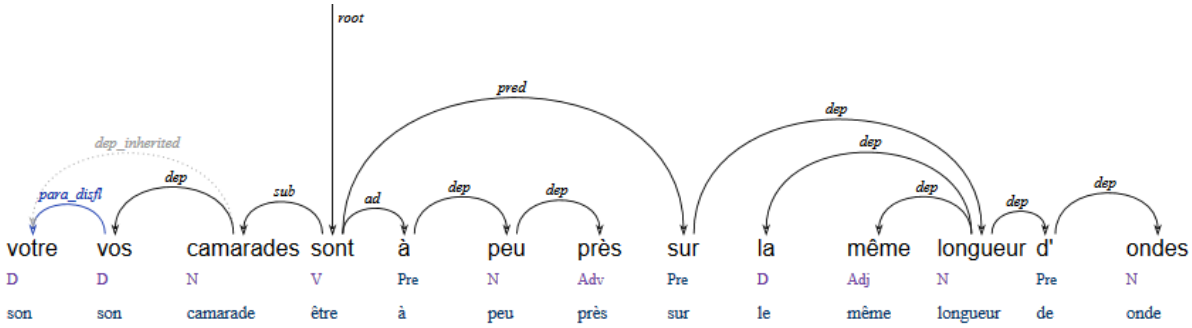


- obl : pour tous les *compléments obliques*, c'est-à-dire les compléments sous-catégorisés par le verbe qui ne sont pas des obj. Les compléments prépositionnels qui font partie d'une forme figée (*mettre <u>en doute</u>*) sont aussi traités comme des obliques. Les constructions locatives avec *être* sont aussi traitées comme des *obl* :



A noter qu'il s'agit d'un lien *obl* uniquement lorsque le locatif s'interroge par *où*. Ainsi, pour *être* + *sur/sous/en* où le sens est plutôt métaphorique, *être* est lié à la préposition qui suit par un lien *pred* :

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

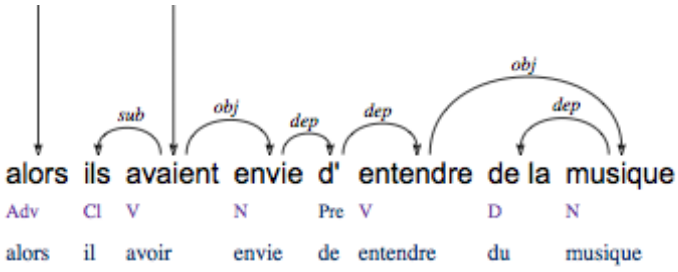


{ votre | vos } camarades sont à peu près sur la même longueur d' ondes //

Cette consigne a été élaborée a posteriori et n'a pas forcément été respectée dans les annotations.

- ad : pour les ajouts (*adjuncts*) au verbe, c'est-à-dire les compléments non sous-catégorisés.
- dep : tous les dépendants des formes non verbales.
    - certains éléments dépendants du verbe ne peuvent être considérés ni comme des ajouts, ni comme des éléments sous catégorisés. Nous en avons identifiés deux types : les éléments faisant partie d'une forme verbale figée (clitique : se souvenir, en avoir marre, il y a un problème) ; la subordonnée dans les dispositifs (voir plus loin).

Les compléments des tournures verbales sont traités comme des dépendants de l'élément prédicatif (Adj ou N), mais reçoivent une fonction comme dépendant d'une construction verbale et reçoivent donc une fonction *dep* :



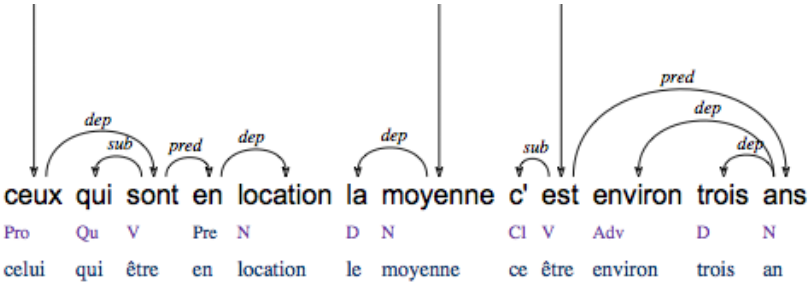
alors < ils avaient envie d' entendre de la musique //

Extraction et mots *qu-*

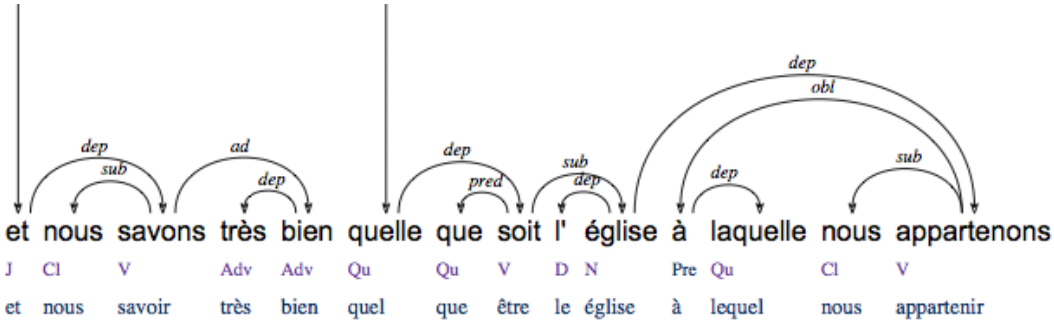
Relative : Il est possible de considérer, à la suite de Tesnière (1959) par exemple, que dans les relatives, le pronom relatif joue à la fois un rôle de pronom à l'intérieur de la proposition relative et de complémenteur en permettant à la proposition relative de modifier un nom. Cette analyse voudrait donc que le pronom relatif occupe une double position syntaxique à la fois comme tête de la proposition (en tant que complémenteur) comme dépendant dans la proposition (en tant que pronom). Certains travaux vont

même jusqu'à défendre que certains mots *qu-*, notamment *que* dans les relatives, sont avant tout des complémenteurs.

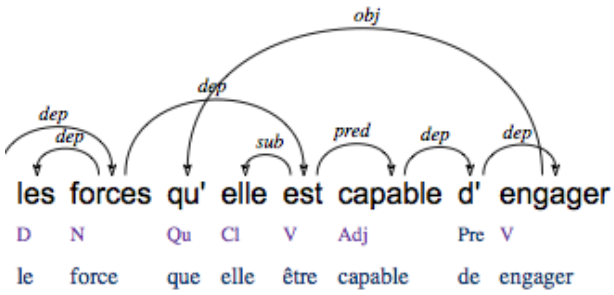
Pour notre part, nous n'encodons que la position du pronom à l'intérieur de la relative :



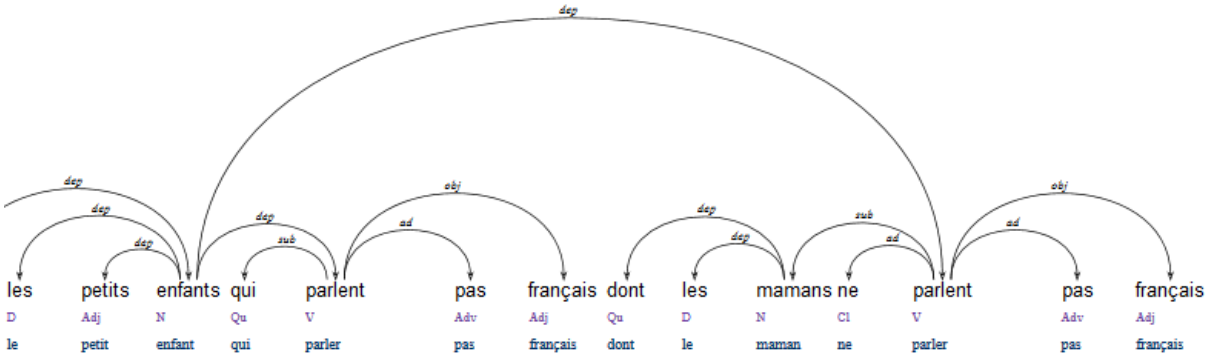
ceux qui sont en location < la moyenne < c' est environ trois ans //



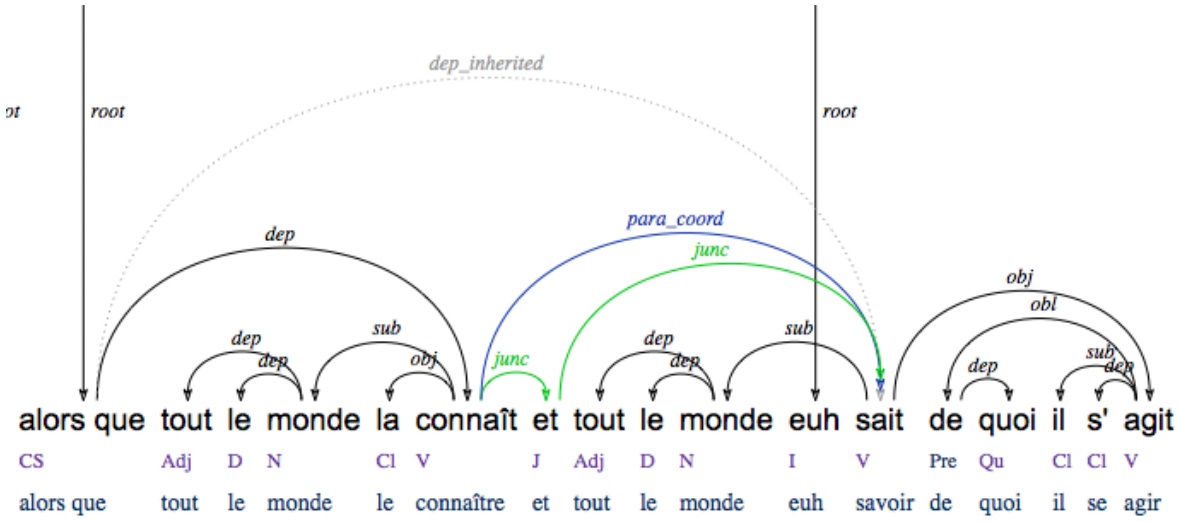
Nous ne rejetons pas du tout le fait que les pronoms relatifs et les mots *qu-* en général jouent un rôle de complémenteur, mais nous décidons, par souci de simplification de ne pas encoder cette position qui peut facilement être récupérée. A l'inverse la position pronominale du mot *qu-* et sa fonction ne peut pas être facilement reconstruite, en raison notamment des dépendances « longue distance », c'est-à-dire les cas où le pronom relatif occupe une position profonde dans la relative, dont résulte notamment une structure non projective, puisque le pronom relatif ne se positionne pas à côté du gouverneur de la position extraite :



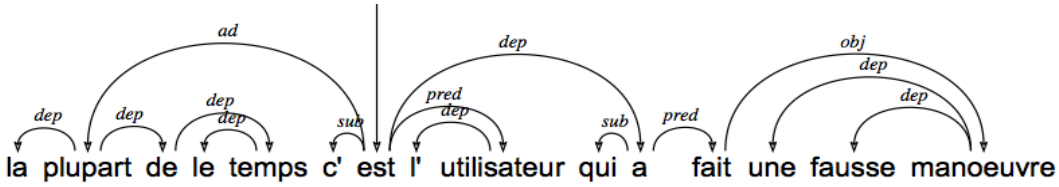
L'analyse s'étend à tous les pronoms relatifs. Dans l'exemple suivant, *dont* est analysé comme étant dépendant du *mamans*, du fait que son équivalent déclaratif serait *les mamans des enfants*.

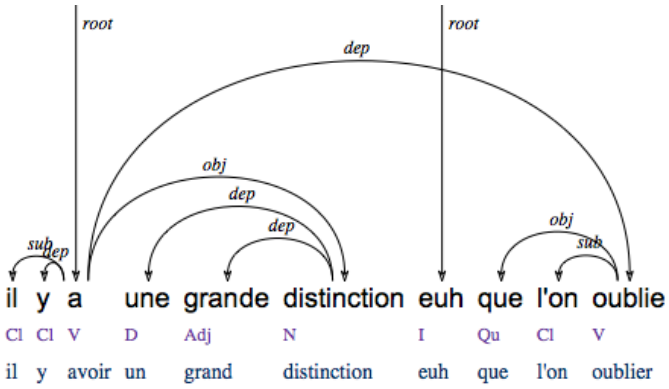


Interrogative : Les mêmes remarques que pour les pronoms relatifs valent pour les pronoms interrogatifs, notamment dans les interrogatives indirectes. Nous faisons les mêmes choix en attribuant au pronom interrogatif sa seule position comme pronom.



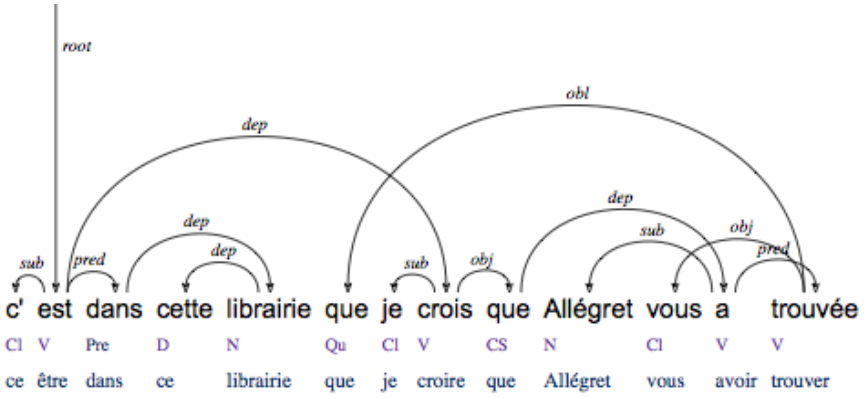
Clivage : Pour les constructions clivées en « c'est X qui/que P » (*c'est l'utilisateur qui a fait une fausse manœuvre*), nous traitons la subordonnée comme une proposition relative, mais dépendant du marqueur du clivage, c'est-à-dire du verbe *être* et non du syntagme nominal extrait par le clivage :



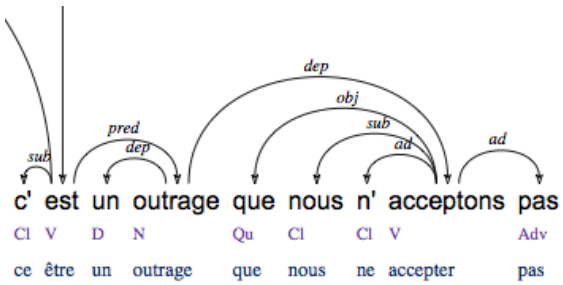


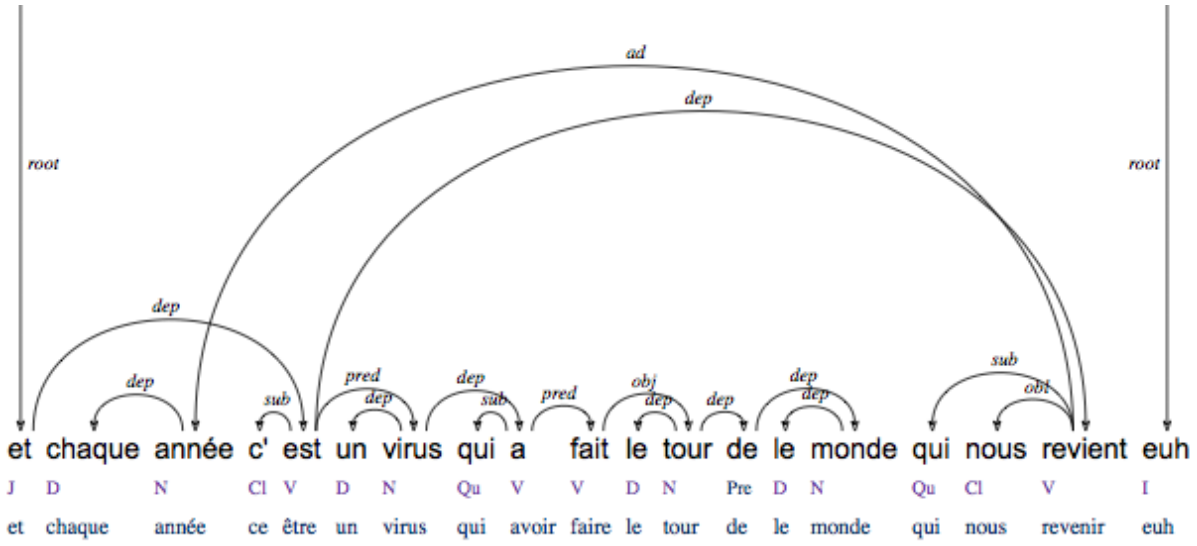
De plus, étant donné le caractère très particulier de la relation entre le clivage et la subordonnée, nous lui donnons la fonction *dep*, tandis que l'élément extrait a la fonction *pred* (avec l'idée que la construction sous-jacente est *ce qui/que X est P* : *celui qui a fait une fausse manœuvre est l'utilisateur*).

Nous étendons cette analyse également au cas où l'élément extrait est complément oblique ou adjoint. Bien que dans ce cas, plus encore que dans les précédents, il y a de bonnes raisons de considérer que le mot *qu-* est complémenteur et non pronom, nous préférons cet encodage qui a l'avantage de marquer quel est le gouverneur de l'élément extrait et notamment de repérer les extractions longue distance :



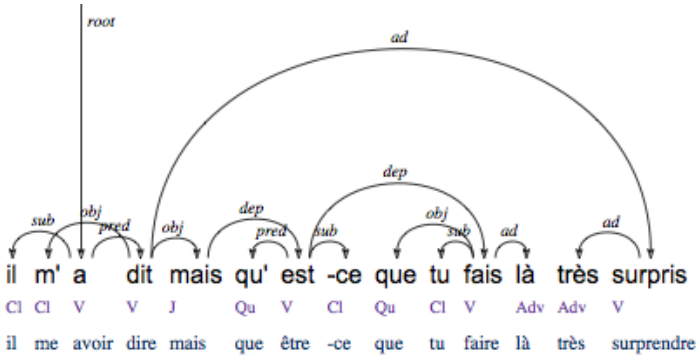
Notre représentation du clivage permet d'encoder la différence de structure entre le clivage et le cas où un syntagme nominal attribut possède une relative :





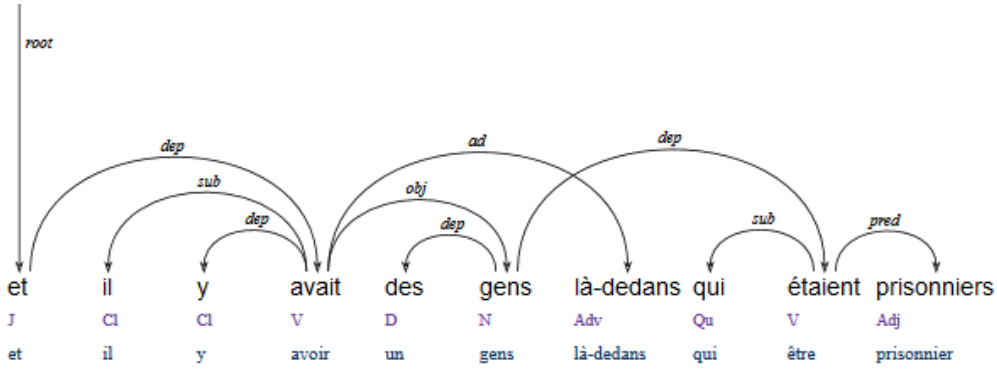
^ et chaque année <+ c' est un virus qui a fait le tour du monde qui nous revient " euh " //

Nous étendons notre analyse des clivées aux interrogatives en *qu'est-ce que*, considérant que c'est une forme interrogative d'une clivée (c'est quoi que tu fais là → qu'est-ce que tu fais là) :



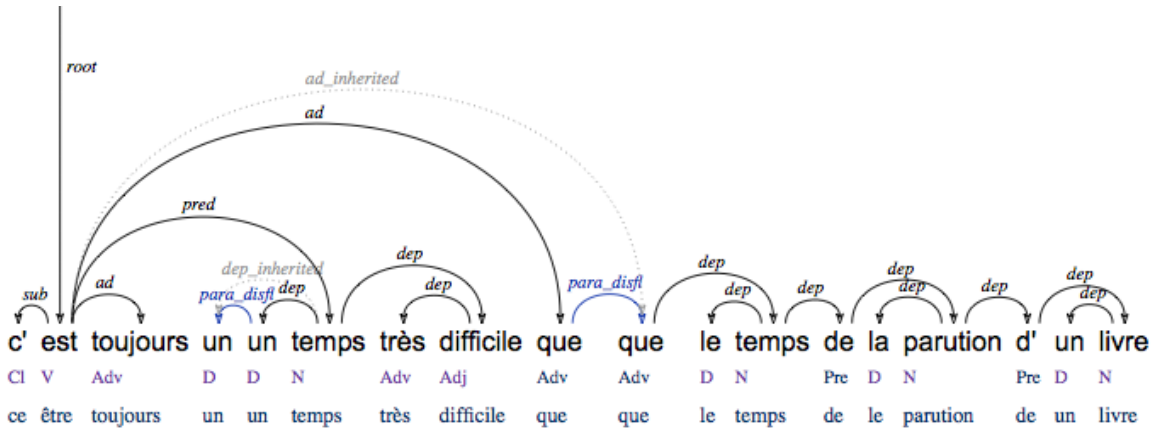
il m' a dit [^ mais qu' est-ce que tu fais là //] très surpris //

En revanche, dans les présentatives en « il y a X qui P », nous traitons la subordonnée « qui P » comme une relative dépendant de X :



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

Constructions atypiques : Nous avons aussi évidemment rencontré des constructions singulières pour lesquelles nous avons essayé de proposer des analyses sans pouvoir réellement les relier à nos autres choix. Dans l'extrait suivant par exemple, le rôle de *que* est inhabituel :

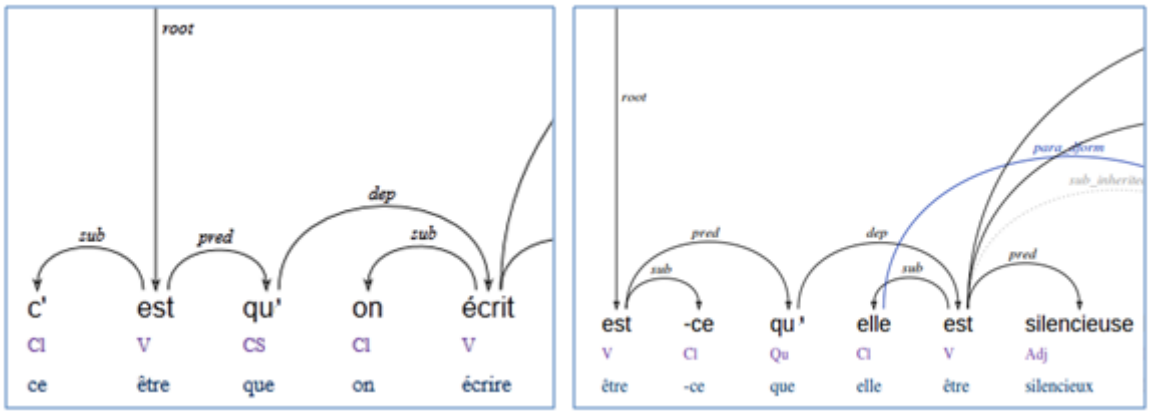


c' est toujours { un | un } temps très difficile { que | que } le temps de la parution d' un livre //

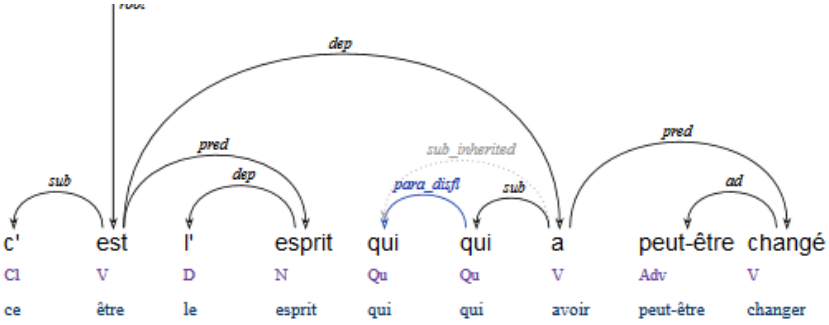
L'interrogation

Les interrogatives totales vs. les interrogatives partielles : Ce traitement particulier du clivage et de la subordonnée fait que le *que* de *est-ce que* et le deuxième *que* de *qu'est-ce que* n'ont pas la même fonction par rapport au noyau du syntagme interrogatif. Leurs structures sont analogues à celles de leurs équivalents déclaratifs :

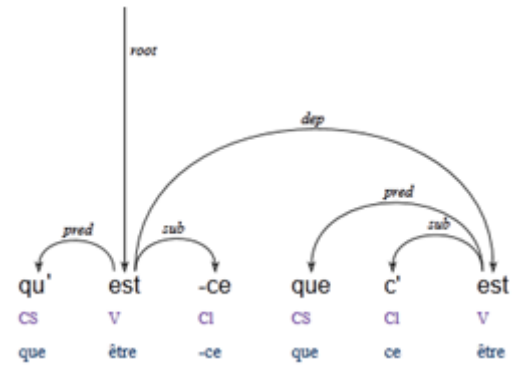
c'est que X a fait -> *est-ce que X a fait ?*



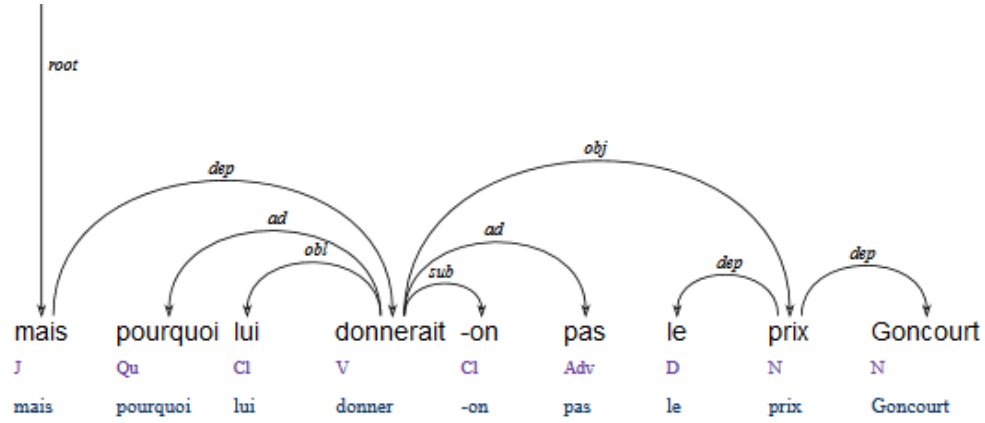
C'est quoi que X a fait -> *qu'est-ce que X a fait ?*



c' est l' esprit { qui | qui } a peut-être changé //



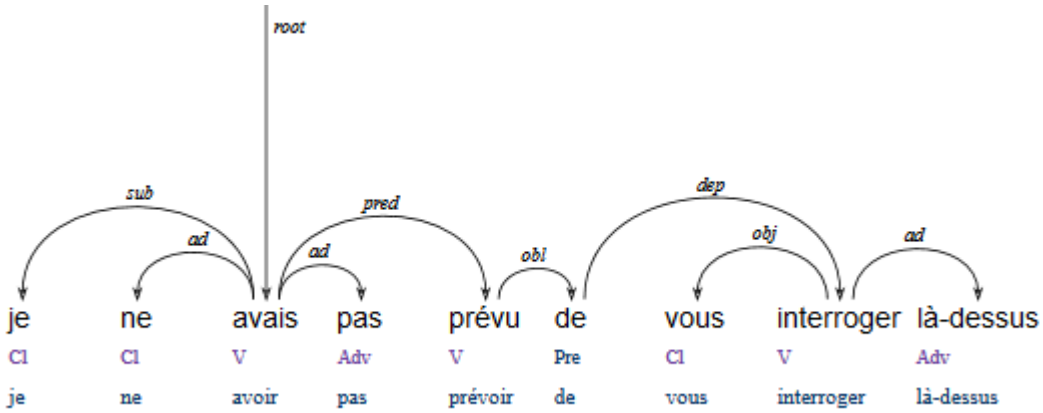
Les mots interrogatifs : Malgré le fait que les mots interrogatifs tels que *pourquoi, comment* etc. puissent être considérés d'une certaine manière comme le noyau d'une question, notre analyse va les considérer comme des ajouts au verbe principal. Ainsi, dans l'exemple suivant, *pourquoi* prend le même rôle qu'aurait *parce que* :



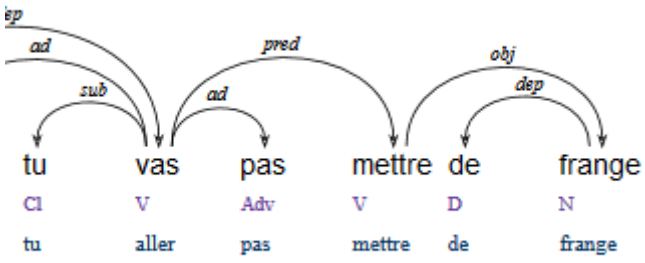
^ mais pourquoi lui donnerait-on pas le prix Goncourt //

La négation

La négation verbale : Le clitique *ne* et les adverbes de négation *pas, jamais, point, guère* etc. sont gouvernés par la forme verbale finie et sont considérés comme des ajouts de ce verbe. Ainsi, dans le cas d'un verbe au passé composé, le *ne* et *pas* sont des ajouts de l'auxiliaire :

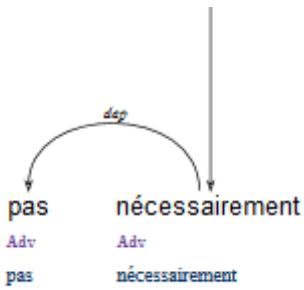


Pas de X : Bien qu'il existe des arguments pour considérer *pas de X* comme étant un déterminant complexe (voir *beaucoup de.., peu de…*), *pas de X* ne fonctionne pas toujours comme une unité cohésive, puisque il peut être divisé par des adverbes ou des formes verbales (*il met* ***pas de frange*** vs *tu vas* ***pas*** *mettre* ***de frange***) . Ainsi le *pas* est toujours traité comme un ajout au verbe et le nom comme l'objet. Le *de* négatif est traité comme le déterminant du nom.

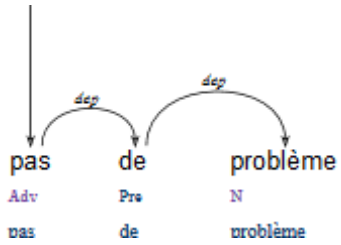


tu vas pas mettre de frange

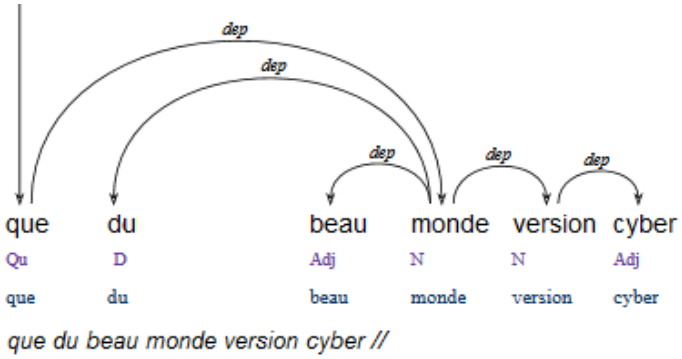
La négation averbale : Quand il s'agit d'une négation averbale (par exemple une réponse telle que *pas nécessairement*), la tête du syntagme sera l'élément sous négation et la négation est considérée comme dépendante de cet élément :



sauf quand il s'agit d'un syntagme nominal isolé du type *pas de X*, où la solution alternative citée ci-dessus est utilisée et que l'adverbe de négation est traité comme la tête du syntagme. Nous considérons cette décision plus judicieuse car *pas* n'a pas de verbe auquel s'attacher, même si nous reconnaissons que cela crée une double analyse de *pas de X* :



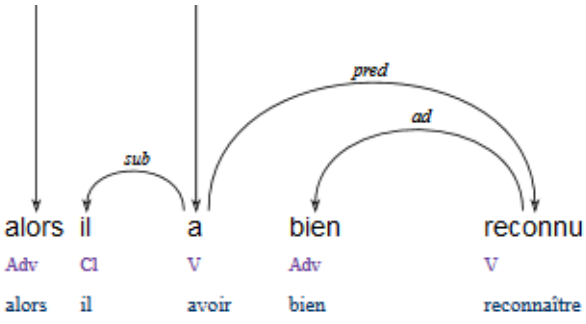
***Que* restrictif :** *que* est traité comme les autres adverbes de négation (ajout au verbe pour la négation verbale et racine pour la négation verbale du type *que de X) :*



que du beau monde version cyber //

Les adverbes

Les adverbes sont en général des ajouts au verbe. Dans le cas d'une forme verbale complexe comme le passé composé, les adverbes (excepté les adverbes de négation qui sont décrits ci-dessus) sont liés au verbe plein et non pas à l'auxiliaire :

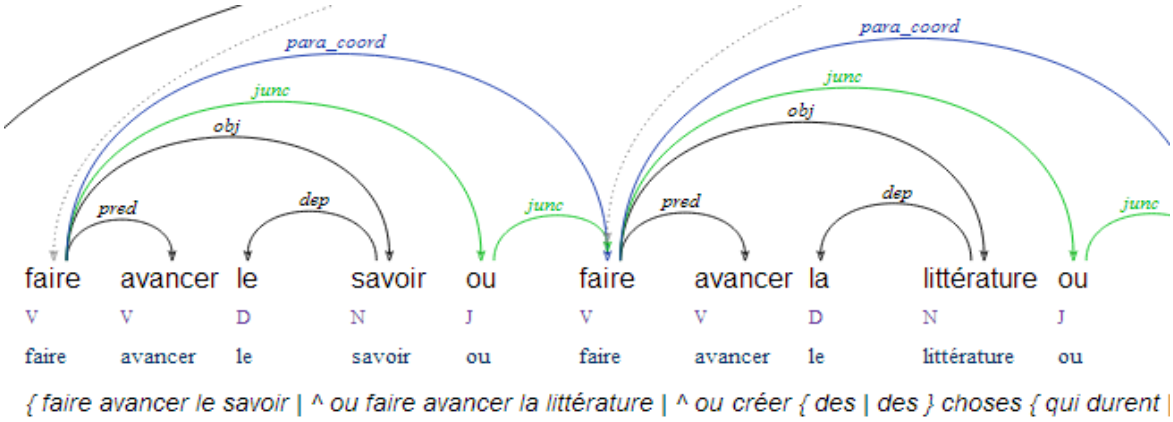


alors < il a bien reconnu //

Les constructions causatives

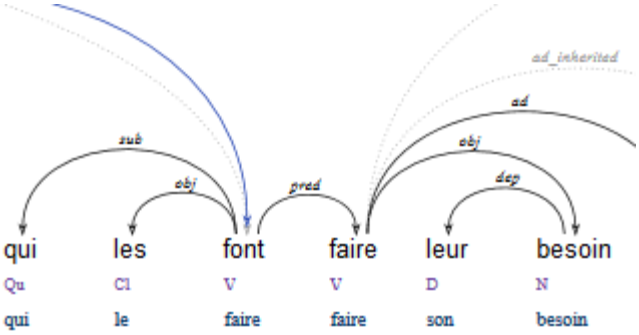
Dans le cas d'une construction causative, un nom peut être à la fois l'objet du verbe *faire* est le sujet d'un infinitif suivant. Notre choix et de marquer uniquement le lien *obj* entre le nom et le verbe *faire*, le deuxième verbe étant lié par un lien *pred* au verbe *faire*.

...**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



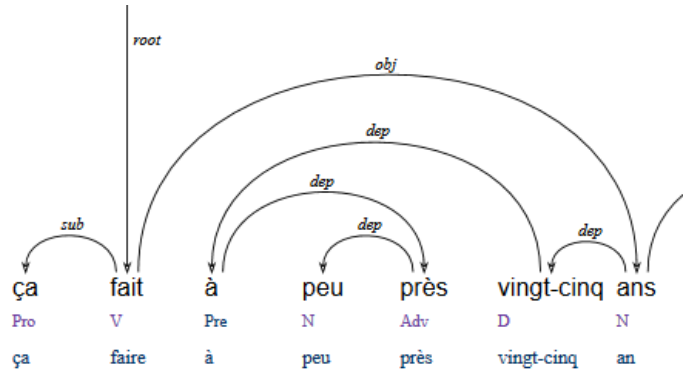
Ce choix permet de ne jamais associer un sujet à un verbe à l'infinitif et se justifie par le fait que le sujet rétrogradé se cliticise sur *faire*. On a même dans l'exemple suivant un clitique objet alors que l'infinitif a lui-même un objet :

{ qui | qui | qui } prennent leur chien | qui **les font faire** leur besoin { { dans | dans } le trottoir |}



Les expressions figées

Comme mentionné au début de ce document, nous ne traitons comme lexème que les expressions qui n'ont pas de syntaxe interne. Les autres expressions figées sont traitées sur le calque des expressions libres équivalentes. C'est le cas par exemple de *à peu près*. Bien qu'il soit possible de l'analyser comme une unité toute faite, nous avons choisi d'analyser sa structure, étant donné que *peu* peut commuter avec d'autres éléments nominaux (*à une semaine près, à deux centimètres près*…) :



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

Rappelons que *il y a* (dans *il y a un problème*) est traité comme une expression figée et donc *y* est lié à la forme verbale a par un lien *dep* et non pas par un lien *ad*.

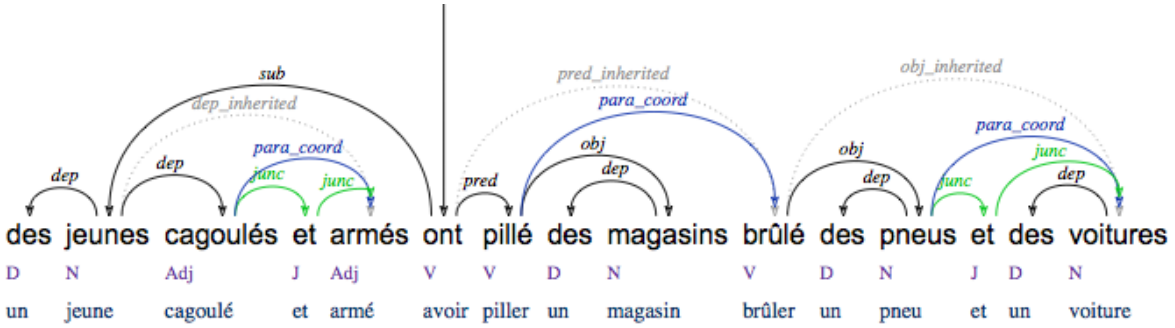
Traitement des entassements

(Sylvain Kahane, Kim Gerdes, Paola Pietrandrea)
Nous appelons entassements des constructions caractérisées par le fait que plusieurs éléments viennent occuper une même position régie. Il peut s'agir de coordinations :

1. des jeunes **{ cagoulés | ^et armés }** ont **{ pillé des magasins | brûlé { des pneus | ^et des voitures } }** // (M2006)

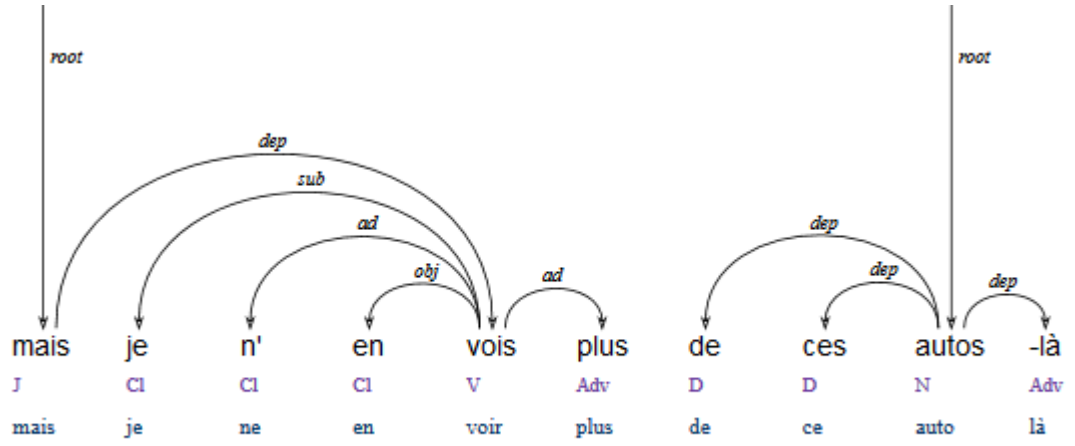
mais aussi de divers autres phénomènes que nous allons présenter, à commencer par des disfluences et des reformulations comme :

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)



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

Pas de double marquage : malgré le fait que le double marquage puisse être considéré comme une double formulation entre la forme nominale et le clitique, nous décidons de ne pas noter de double formulation dans ces cas, que ce soit la dislocation à gauche ou à droite



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

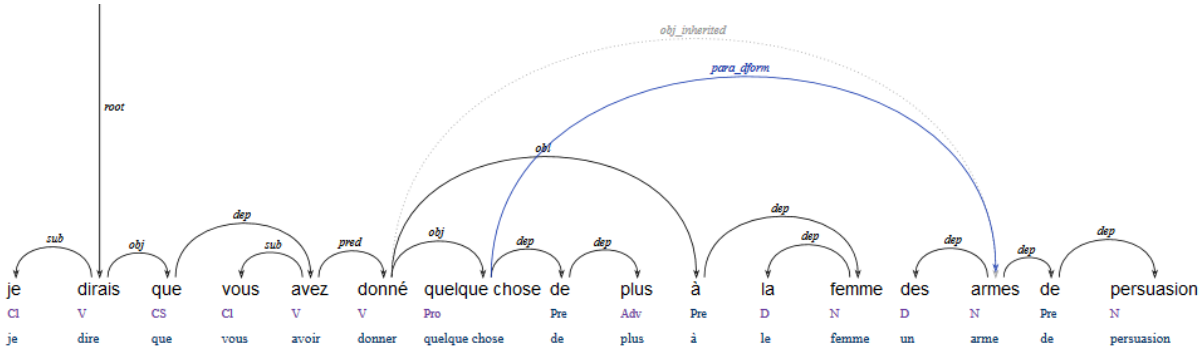
Lien paradigmatique

L'entassement est une dimension orthogonale à la rection : on dit qu'un segment Y s'entasse sur un segment X si Y vient occuper la même position régie que X. Nous notons

{ X | Y } le fait que X et Y sont entassés. Dans l'entassement { X | Y }, X et Y représentent chacun une *couche* de l'entassement. Un entassement peut avoir deux couches ou davantage.

Entassement discontinu : Un entassement est discontinu lorsque les couches de l'entassement ne sont pas contiguës. Nous utilisons alors la 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)

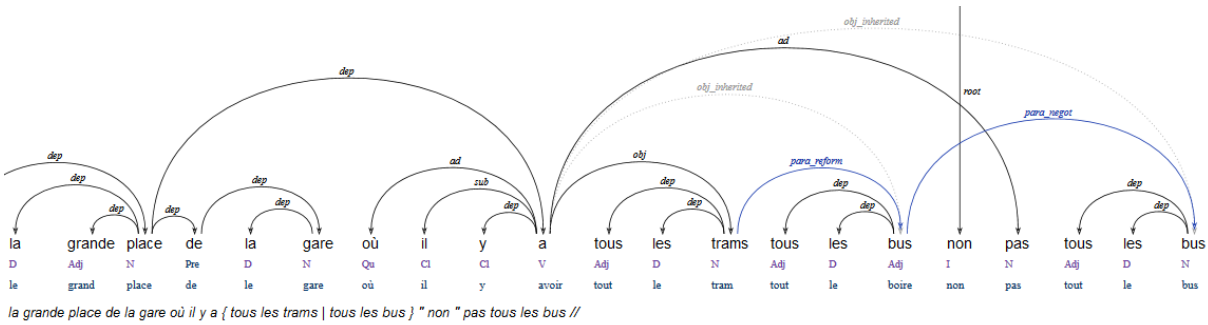


je dirais que vous avez donné { quelque chose de plus |} à la femme {| des armes de persuasion } //

Conjoint : La couche d'un entassement contient plusieurs types d'éléments. L'élément central de la couche s'appelle le conjoint. Chaque conjoint peut occuper seul la place qu'occupe l'entassement :

> 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

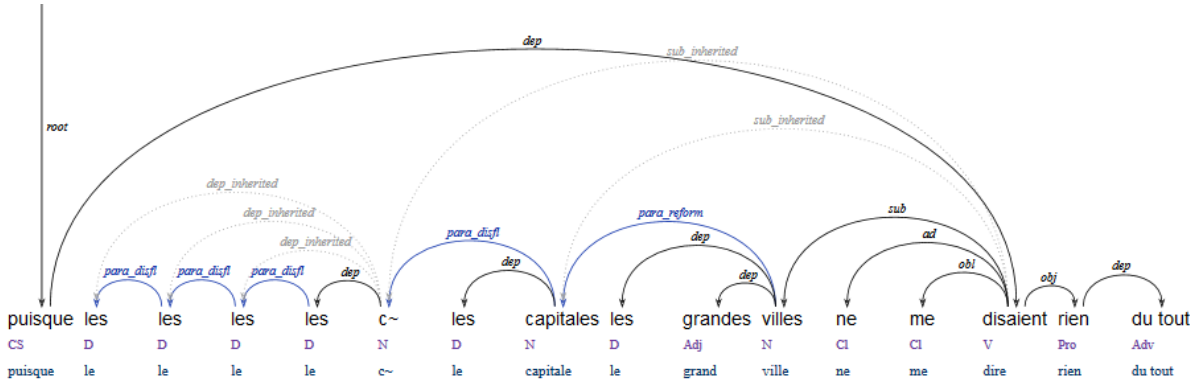
Lien paradigmatique : Les conjoints sont liés entre eux par un lien paradigmatique. Chaque conjoint est lié au conjoint qui le précède, quel que soit le nombre de couches, y compris lorsqu'il y a des entassements enchâssés (voir un cas plus loin). Les liens paradigmatiques sont notés *para* suivi d'un type (voir plus loin pour le typage des entassements). Dans l'exemple suivant, on a un entassement avec trois couches :



la grande place de la gare où il y a { tous les trams | tous les bus } " non " pas tous les bus //

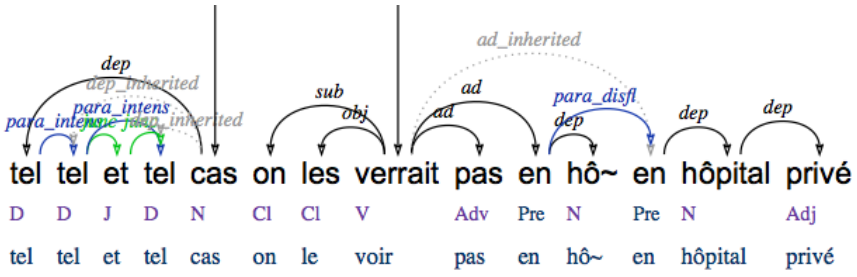
{ ^et "euh" | ^et } "ben" "voilà" j' arrive au niveau de la grande place de la gare où il y a { tous les trams | tous les bus |} //+ {| " non " pas tous les bus } //

Sens du lien paradigmatique : Pour les disfluences, les reformulations, on rattache le gouverneur de l'entassement au conjoint le plus proche. En conséquence, lorsque le gouverneur est après l'entassement, les liens paradigmatiques vont de droite à gauche (afin de préserver une structure arborescente) :

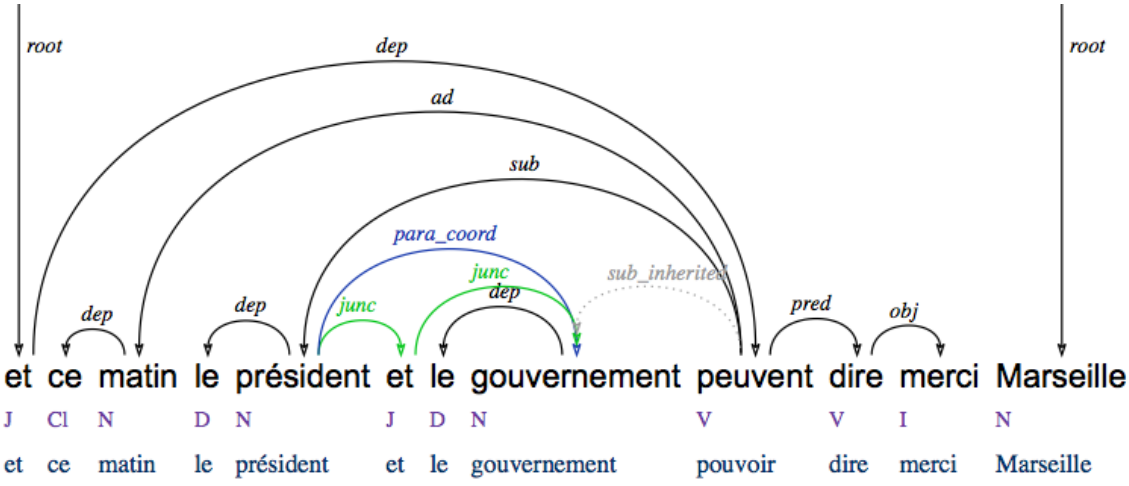


^ puisque { { { les | les | les | les } c~ | les capitales } | les grandes villes } ne me disaient rien du tout //

Pour les coordinations et les intensifications, le gouverneur est toujours attaché au premier conjoint, comme dans l'exemple où c'est la première occurrence de *tel* qui dépend de *cas* :



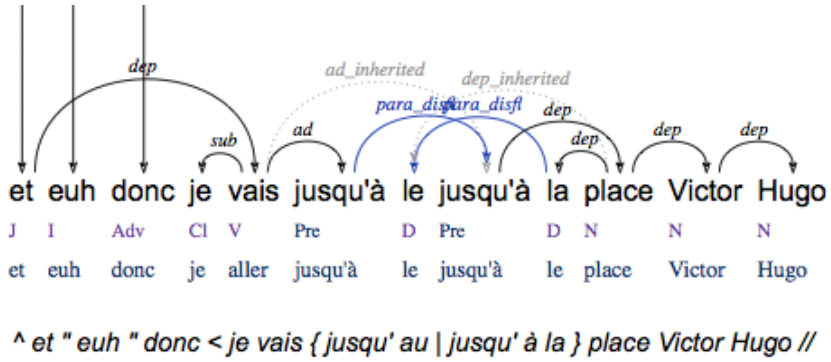
{ tel | tel | ^ et tel } cas < on les verrait pas { en hô~ | en hôpital } privé //



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

Double lien paradigmatique : En cas de disfluence, il arrive qu'un conjoint soit syntaxiquement discontinu. Dans ce cas, chacun des morceaux sera aligné par un lien

paradigmatique. Il se peut, comme dans l'exemple suivant, qu'on ait deux liens paradigmatiques en sens opposés pour le même entassement :

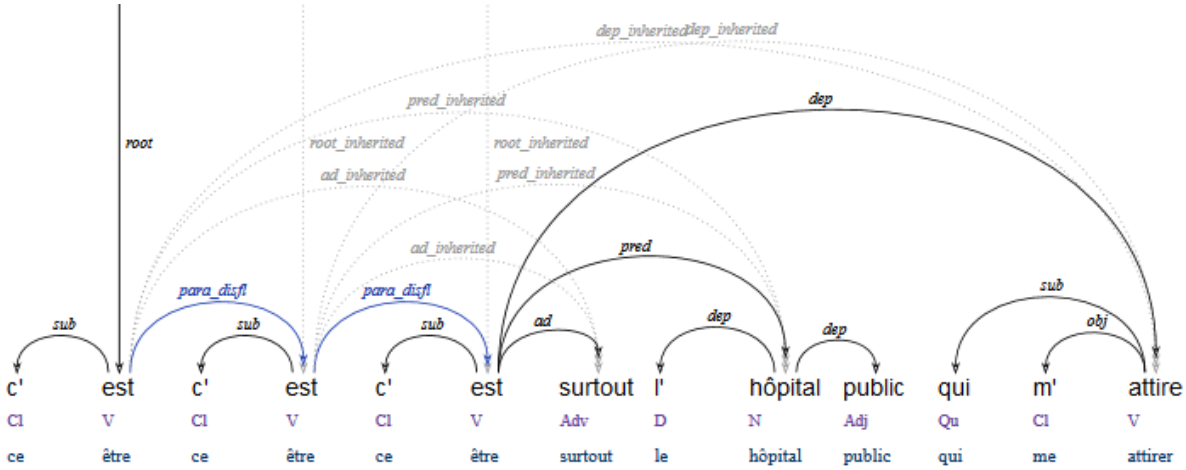


^ et " euh " donc < je vais { jusqu' au | jusqu' à la } place Victor Hugo //

Dépendance héritée

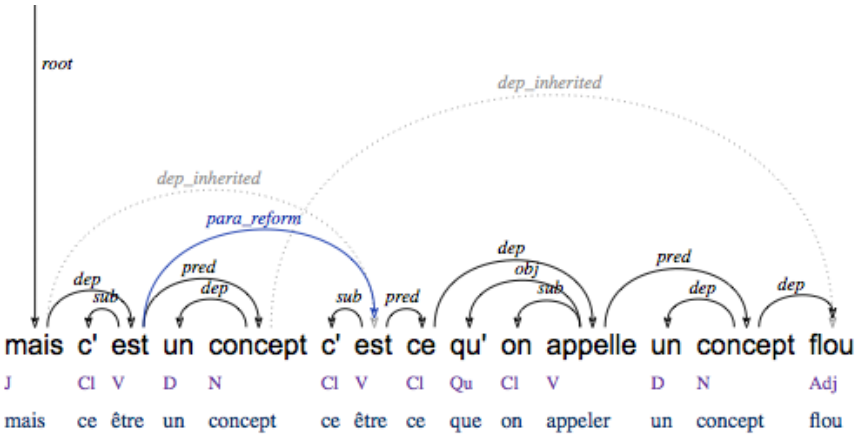
Les conjoints sont dans une relation paradigmatique : nous attachons l'un des conjoints au gouverneur de l'entassement, mais les autres conjoints peuvent commuter avec lui et héritent ainsi d'une relation du même type. Par exemple, dans l'exemple précédent, la deuxième occurrence de *jusqu'à* hérite d'un lien *ad_inherited*, tandis que *le* hérite d'un lien *dep_inherited*.

Root inherited : Lorsque des lexèmes partagent un même dépendant, nous considérons qu'ils s'entassent même s'ils n'occupent pas à proprement parler une même position régie, puisqu'ils peuvent comme dans l'exemple suivant être la racine de la structure de dépendance. Dans ce cas, comme dans les autres, les conjoints héritent de la même dépendance que le premier conjoint, qui est donc dans ce cas une dépendance *root_inherited* :



{ c' est | c' est | c' est } surtout l' hôpital public qui m' attire //

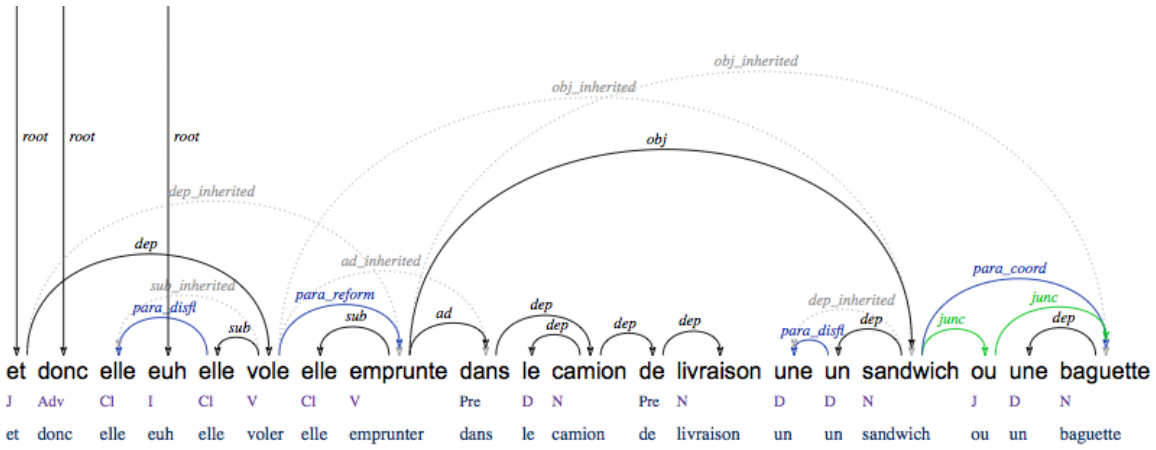
Dépendance héritée sortante : Les éléments qui dépendent de plusieurs conjoints sont en dehors de l'entassement. Ils sont dépendants du conjoint le plus proche et reçoivent une dépendance héritée des autres conjoints. Dans l'exemple suivant, *flou* modifie les deux occurrences de *concept* :



^ *mais { c' est un concept | c' est ce qu' on appelle un concept } flou //*

Remarquons au passage que notre stratégie d'annotation des entassements est de considérer qu'un élément est en dehors de l'entassement dès qu'il peut potentiellement compléter plusieurs couches. Ainsi très peu de couches s'avèrent-elles être inachevées.

Pas de double héritage : Dans l'exemple suivant, considérons les lexèmes *vole emprunte sandwich baguette* et les relations qui les unissent. Il y a une relation *obj* entre *emprunte* et *sandwich*. En raison du lien paradigmatique entre *vole* et *emprunte*, on en déduit une relation *obj_inherited* entre *vole* et *sandwich*. En raison du lien paradigmatique entre *sandwich* et *baguette*, on en déduit une relation *obj_inherited* entre *emprunte* et *baguette*. On pourrait également inférer à partir de ces deux relations une autre relation *obj_inherited* entre *vole* et *baguette*. Mais par souci de simplicité, nous n'indiquons pas ces relations :



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

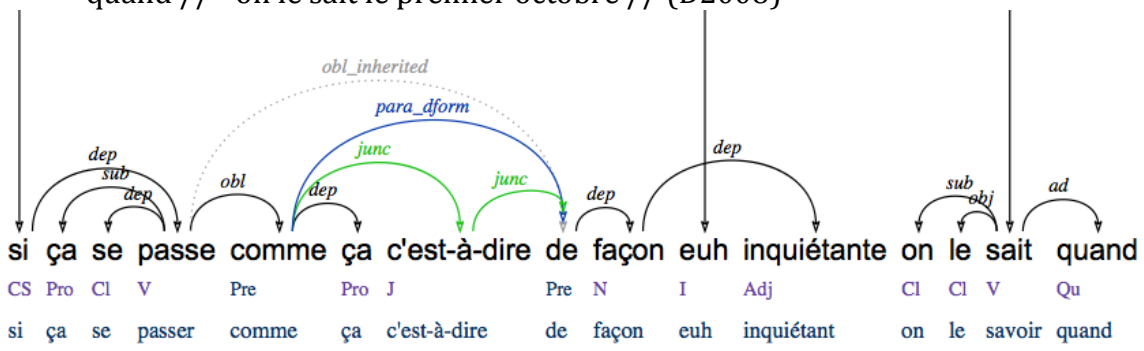
Lien de jonction

Joncteurs : Les *joncteurs* sont les éléments qui lient les conjoints (nous les marquons en les faisant précéder d'un ^). Les joncteurs sont plus ou moins les conjonctions de coordination. Nous adoptons ici, à la suite de Blanche-Benveniste et de Ndiaye (1989), une variante du terme *jonctif* utilisé par Tesnière (1959). Les joncteurs occupent un rôle uniquement à l'intérieur de l'entassement. En particulier, si l'on conserve uniquement une couche de l'entassement, les joncteurs ne peuvent pas être maintenus :

des jeunes { **cagoulés | ^et armés** }
*des jeunes **et armés**

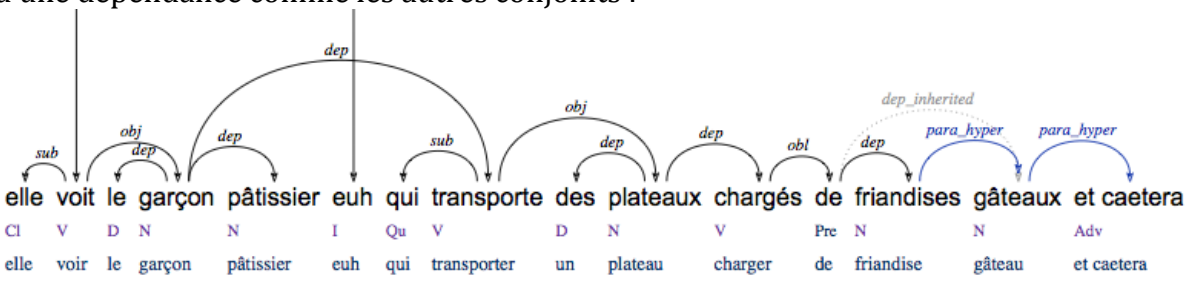
Les joncteurs apparaissent principalement dans les coordinations (*et, ou, mais, ainsi que,* etc.), mais ils sont également possibles dans les doubles formulations, notamment *c'est-à-dire* :

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)



si ça se passe { comme ça | ^ c' ^ est- ^ à- ^ dire de façon " euh " inquiétante } < on le sait quand //=

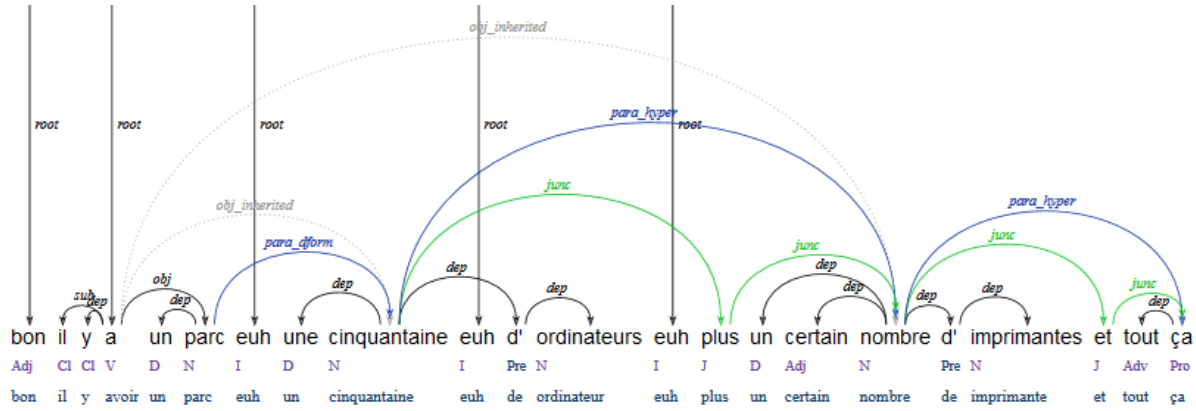
Clôtureurs : Le dernier type d'éléments, que l'on trouve dans les entassements, notamment les coordinations hyperonymiques (voir définition plus loin), ont la propriété de pouvoir clore un entassement, comme *et caetera*. Nous les appellerons des *clôtureurs*. Ils ne peuvent occuper que la dernière couche d'un entassement. C'est le cas dans l'exemple suivant de *et caetera*. Cet élément ne peut pas occuper seul la position régie : **des plateaux chargés de et caetera*. Par conséquent, cet élément n'hérite pas d'une dépendance comme les autres conjoints :



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

Mais il existe aussi de nombreux clôtureurs qui sont formés d'un joncteur et d'un type particulier de conjoint, comme *tout ça*, que nous appelons, à l'instar d'Overstreet (2005), un *extenseur* :

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)

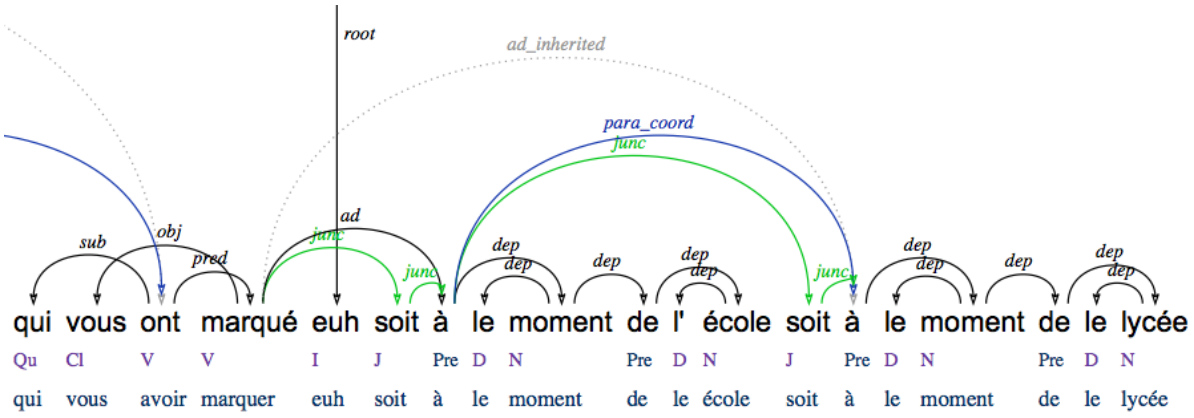


Le lexème *et caetera* fonctionne ainsi comme l'amalgame d'un joncteur et d'un extenseur, ce qu'il est à l'origine. Cela explique qu'il ne puisse pas, comme les joncteurs, apparaître en dehors d'un entassement.

Liens de jonction : Les joncteurs simples (*et, ou, mais, ainsi que,* etc.) se placent entre deux conjoints. Suivant l'analyse asymétrique de la coordination (Mel'cuk 1988), nous considérons que le joncteur forme un constituant avec le conjoint qui le suit, que ce constituant s'adjoint sur le conjoint qui précède. Comme le joncteur contrôle la distribution de ce constituant qui s'ajoute sur le conjoint précédent, il en est considéré comme la tête. Ceci nous donne donc deux dépendances : conjoint précédent --> joncteur --> conjoint suivant.
Ces dépendances sont particulières puisqu'elles sont en quelque sorte dans une dimension orthogonale à la rection. Nous étiquetons donc avec une fonction particulière *junc* (Tesnière appelle cette relation la jonction). Voir les exemples ci-dessus et ci-dessous où les liens de jonction sont en vert.
Les liens de jonctions sont toujours doublés par un lien paradigmatique, sauf dans le cas de double joncteurs et de jonction sans entassement (cf. ci-dessous), où les liens de jonctions sont doublés par une relation de rection.

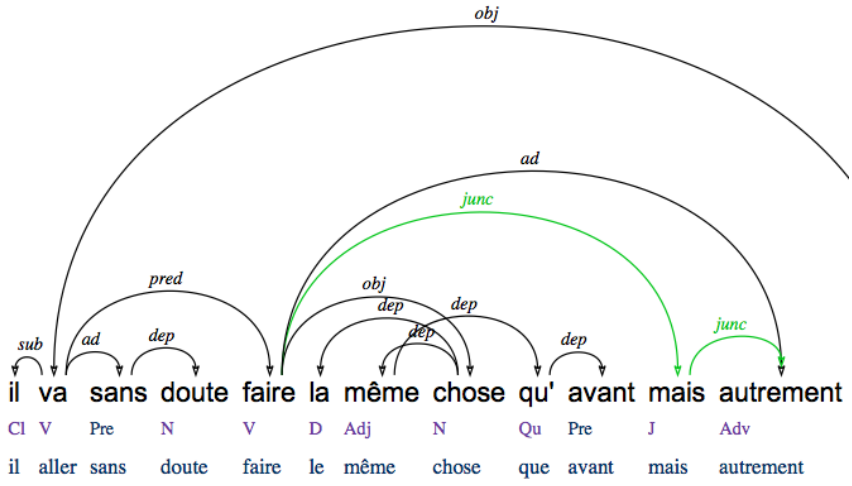
Doubles joncteurs : Lorsque le premier conjoint est introduit par un joncteur comme dans les coordinations en *soit A soit B*, nous faisons dépendre ce joncteur du gouverneur de la pile :

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)



Cette analyse permet de traiter les deux joncteurs de manière parallèle, comme gouverneur d'un conjoint. De nouveau le lien *junc* double une dépendance, mais cette fois-ci ce n'est pas un lien paradigmatique, mais un lien de rection. Néanmoins, le principe est similaire, le joncteur fonctionnant comme un marqueur du lien qu'il double.

Jonction sans entassement : Dans une construction du type *il parle anglais et bien*, le joncteur *et* ne marque pas un entassement car les conjoints n'occupent pas la même position syntaxique et qu'il n'y a pas de relation paradigmatique entre les conjoints. Nous considérons qu'il s'agit d'une coordination entre deux UI — *il parle anglais* d'une part et *et bien* de l'autre — qui forme une seule UR. Donc *bien* est un dépendant de *parle* (comme il le serait dans *il parle bien anglais*) et en même temps des liens *junc* lie le joncteur *et* à *parle* et *bien* : *parle* –junc-> *et* –junc-> *bien*. Même chose dans l'exemple suivant :



[il va sans doute faire la même chose qu' avant //+ ^mais autrement //] pronostique Francis Brochet // (D2013)

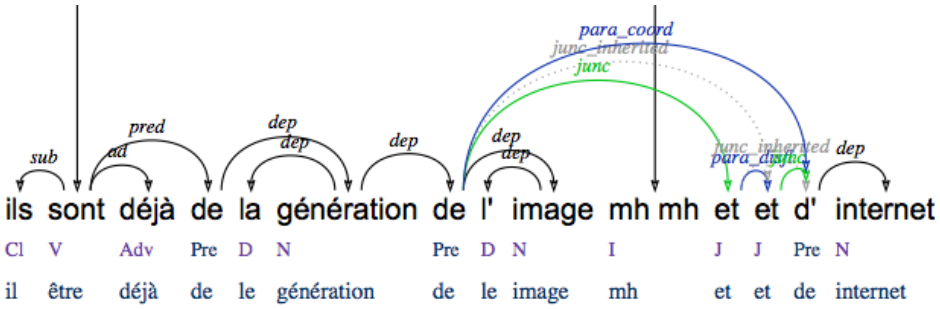
Cette construction n'est pas rare semble-t-il. On en trouve plusieurs exemples dans notre corpus :

on veut bien parler avec vous //+ mais { a~ | après } le déménagement // (D0006)

normalement < c' est du bois de hêtre dessous //+ ^ et { qui est p~ | qui est laqué noir } //

{ 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)

Entassement de joncteurs : Il est possible qu'on ait des joncteurs entassés en cas de disfluence sur le joncteur. Dans ce cas, les joncteurs entassés hériteront de liens *junc_inherited* avec les conjoints :

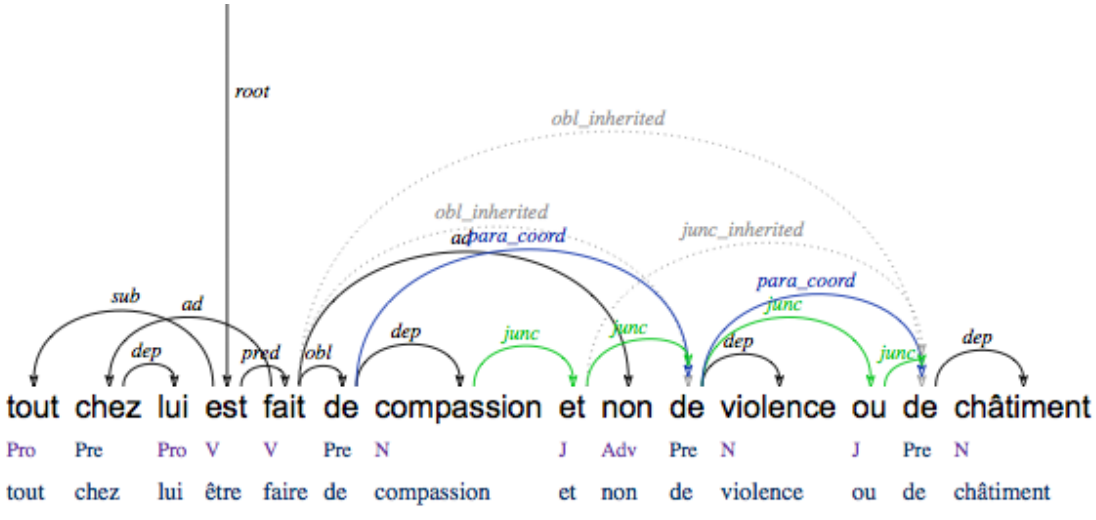


Coordinations enchâssées : On parle de coordination enchâssée lorsqu'on a une des couches d'une coordination qui est elle-même occupée par une coordination. Dans l'exemple bien connu *Nous cherchons quelqu'un qui parle anglais et allemand ou italien*, il y a deux interprétations possibles avec à chaque fois une coordination enchâssée :

a. Nous cherchons quelqu'un qui parle { { anglais | ^et allemand } | ^ou italien }
b. Nous cherchons quelqu'un qui parle { anglais | ^et { allemand | ^ou italien } }

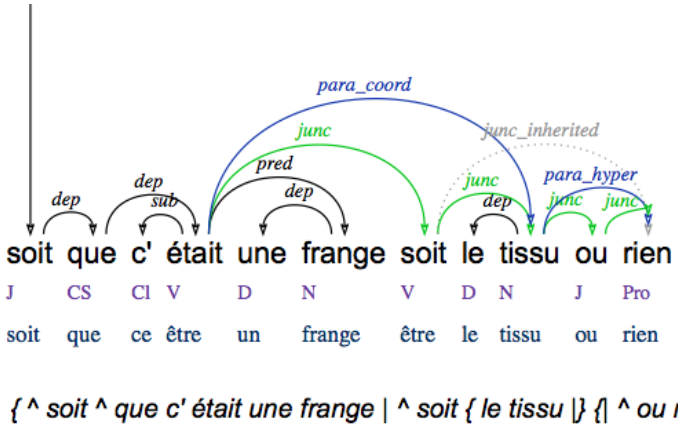
Ces deux exemples reçoivent pour nous une analyse légèrement différente. Dans les deux cas, nous avons les mêmes liens paradigmatiques anglais -> allemand -> italien et les mêmes liens *junc* anglais -> et -> allemand -> ou -> italien, mais dans (a), *anglais* est coordonné avec *allemand* et hérite d'un lien *junc_inherited* avec *ou*, tandis que dans (b), c'est *italien* qui est coordonné avec *allemand* et qui hérite d'un *junc_inherited* avec *et*.

Nous avons un exemple du type (b) dans notre corpus, où le 3ème conjoint hérite d'un lien *junc_inherited* :



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

Ou encore cet autre exemple avec double joncteurs :



Adverbe paradigmatisant

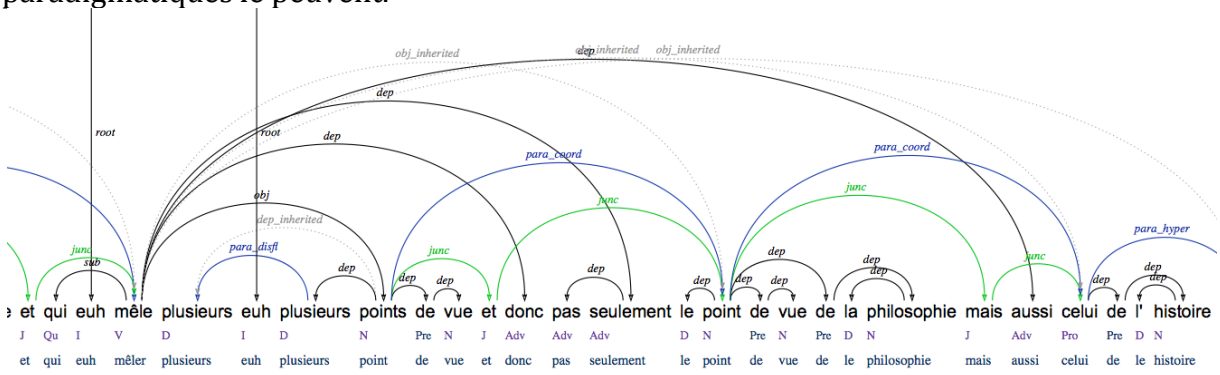
On trouve également dans les entassements des *adverbes*, notamment *paradigmatisants* (Nølke, 1983), comme *seulement* ou *aussi,* en gras dans l'exemple suivant :

> ^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)

A la différence des joncteurs, ceux-ci peuvent être maintenus si une seule couche est conservée :

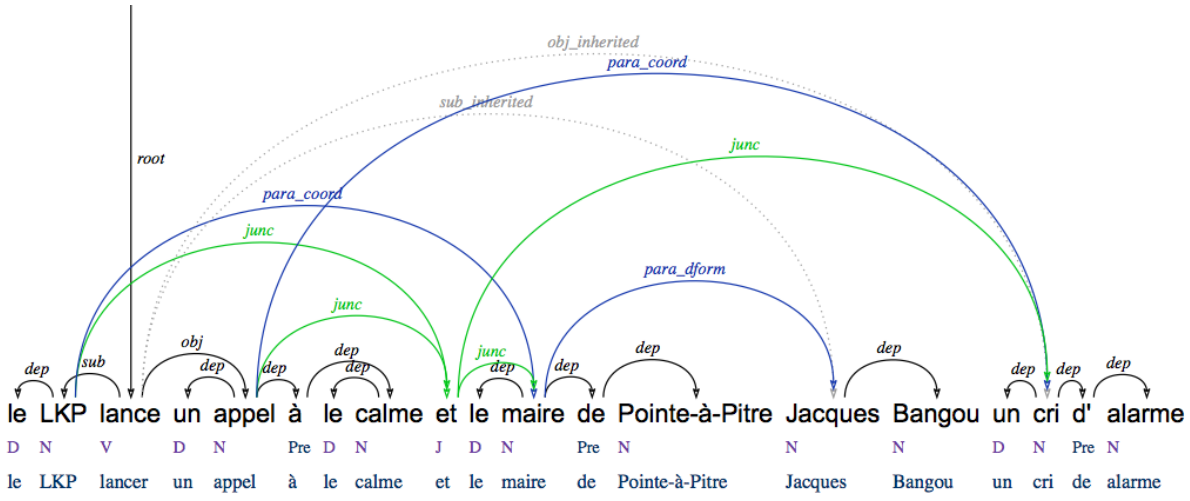
a. mon approche (ne) mêle **donc pas seulement** le point de vue de la philosophie
b. mon approche mêle **aussi** le point de vue de l'histoire.

Ils sont donc « visibles » pour la rection, mais occupent une position différente de celle du conjoint : alors que les conjoints ne peuvent être effacés, les adverbes paradigmatiques le peuvent.



Coordinations problématiques

Coordinations non standards : Nous traitons les cas dits de coordination elliptique (angl. *gaping coordination*) sans recours à une ellipse. Nous considérons qu'un joncteur peut commander deux coordinations parallèles, comme dans l'exemple suivant :



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

Joncteur hors de l'entassement : Le joncteur *ni* est particulier, puisqu'il intègre une négation, c'est-à-dire un adverbe paradigmatisant. Il peut résulter de cela qu'il occupe la position d'un tel adverbe (c'est-à-dire d'un *pas*), plutôt que celle du joncteur :

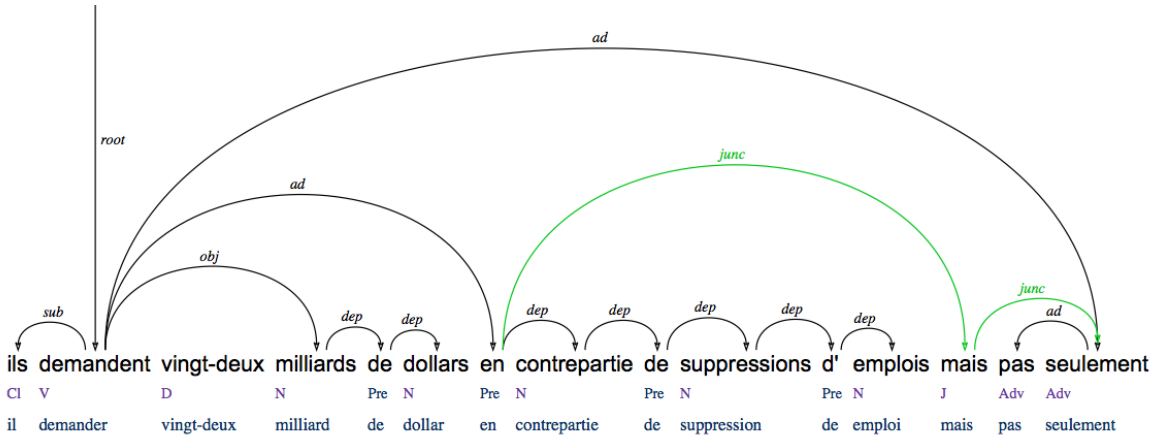
> parce qu' { on remarque "euh" { ass~ "euh" | de façon "euh" } & | on remarque bien que le salaire n'est **^ni** en adéquation **{ { avec** le nombre d'années & "euh" | avec le nombre d'études suivies } | **^ni avec** le travail { à fournir | personnel que l'on fournit } } // (M1003)

On peut alors hésiter à traiter ce *ni* comme un adverbe ou comme un joncteur. Rappelons que de toute façon notre traitement des joncteurs doubles n'est pas satisfaisant (voir plus haut).

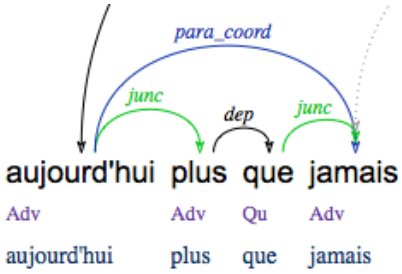
Couche sans conjoint : Il existe enfin des configurations particulières de joncteurs et d'adverbes paradigmatiques qui permettent l'omission du second conjoint (qui serait en fait ici une pure répétition du premier conjoint) :

> ils demandent vingt-deux milliards de dollars { en contrepartie de suppressions d'emplois | ^mais pas seulement } // (M2006)

Nous n'utilisons pas ici le symbole & qui marque les inachèvements puisque cet énoncé est considéré comme complet. Nous n'avons donc pas d'autre dépendant possible pour le lien *junc* sortant de *mais* que l'adverbe *seulement* (lequel est un adverbe paradigmatisant dépendant de *demandent* par une relation *ad* et non *ad_inherited* comme il devrait s'il était un conjoint).

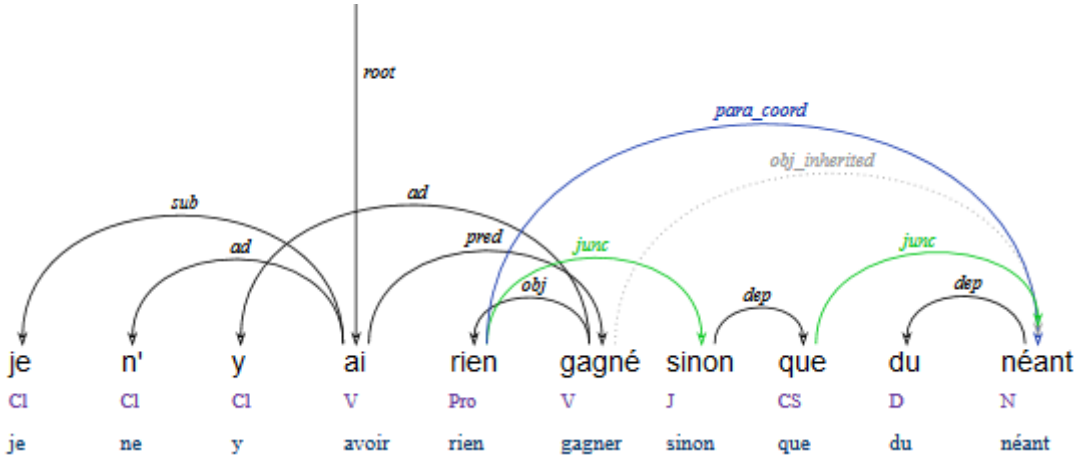


Joncteur complexe : Dans l'exemple suivant, *plus que* est traité comme un joncteur en deux lexèmes (puisqu'il y a commutation de *plus* avec *moins*), ce qui donne la configuration suivante :



Coordination discontinue : Notre annotation permet d'encoder sans difficulté des cas de coordination discontinue comme dans :

{ 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 } // (



Typage des entassements

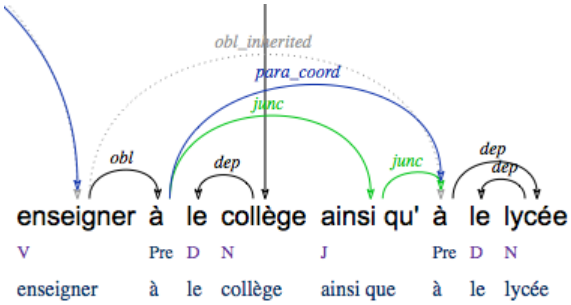
(Sylvain Kahane, Paola Pietrandrea)

Nous classons les entassements en 7 sous-catégories : coordinations, coordinations hyperonymiques, intensifications, disfluences, reformulations, doubles formulations, négociations.

Coordination (para_coord) : La coordination relationnelle (ou coordination tout court) est une coordination où chaque conjoint dénote un élément différent et où l'entassement complet possède une dénotation qui est fonction des dénotations de ses conjoints. Nous typons comme coordination deux grands types de coordinations relationnelles : les coordinations additives (c'est-à-dire les entassements dont la dénotation est la réunion des dénotations des conjoints) :

1. alors < passons maintenant { au détail des mesures discutées | **^e**t aux attentes { des syndicats | **^et** du patronat } } // (M2006)

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)



3. je travaille à la préfecture de Paris qui { n'est pas connue | **^mais** néanmoins existe } "euh" // (D0001)

et les coordinations alternatives (c'est-à-dire les entassements marquant que les éléments dénotés par les conjoints sont potentiellement substituables les uns aux autres) :

4. allez // avec Messi { qui va chercher le corner | ^et qui va trouver **{ le corner | ^ou la touche | ^ou la sortie de but }** } // (D2003)

A noter que certaines coordinations additives sont marquées par le joncteur *comme* :

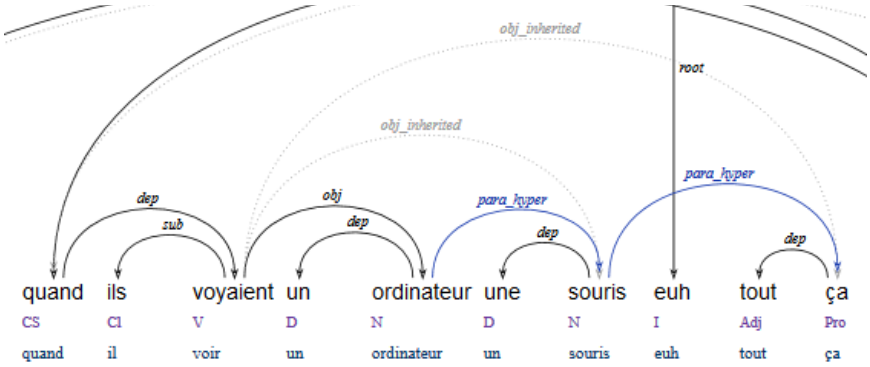
5. il y a { des | des } traditions de famille qui se transmettent { dans le domaine religieux | **^comme** dans tout le domaine du travail } // (D1001)

et que certaines coordinations alternatives sont marquées par la présence d'une conjonction comparative :

6. votre marché < c'est { Aligre | ^plus ^que Nation } // (D0001)

Coordination hyperonymique (para_hyper) : Nous appelons coordinations hyperonymiques les entassements dont la valeur renvoie à un hyperonyme des conjoints, c'est-à-dire à une classe qui contient la dénotation des conjoints, mais qui ne correspond pas à une combinaison logique des dénotations :

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)



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)

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)

Les coordinations hyperonymiques sont également possibles avec des conjoints verbaux :

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)

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)

D'un point de vue formel les coordinations hyperonymiques sont caractérisées par l'emploi de clôtureurs dédiés tels que *tout ça,* ou *quelque chose comme ça,* ou *rien du tout* et par le fait que les conjoints sont des co-hyponymes.

Nous classons parmi les coordinations hyperonymiques aussi des procédés plus lexicalisés consistant à utiliser des antonymes ou des cohyponymes suffisamment opposés pour créer un effet de parcours de l'ensemble de la classe (comme dans les expressions figées *petits et grands, jour et nuit*). Il en résulte un effet de quantification universelle de la classe dénotée. Ce procédé, assez commun à l'écrit, se rencontre dans les échantillons de notre corpus caractérisés par un registre plus soutenu.

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)

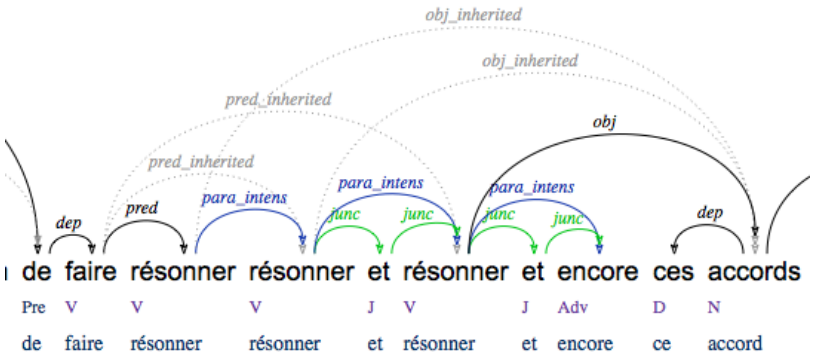
Intensification : La coordination intensive a une fonction générale d'intensification du sens exprimé par le conjoint répété. Cette fonction d'intensification se précise selon la catégorie de l'expression. Par exemple, la réitération de conjoints nominaux intensifie la quantité ('plein d'exercices', '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)

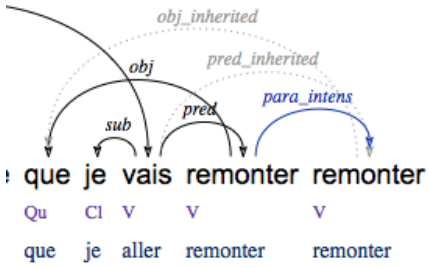
14. mais mais mais mais { les lois sociales | le droit de grève | ^et tout ça } < { ça s'est fait | ça s'est fait } sur **{ des dizaines | ^et des dizaines }** d'années // ça s'est fait sur presque "enfin" { cinquante | cent } ans // (Valibel)

La réitération de conjoints verbaux peut servir à intensifier la durée ou la fréquence de l'action et à marquer l'aspect continuatif ou itératif :

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)



16. ^ ensuite <+ " euh " je vais " euh " prendre [je crois que c' est l' avenue Alsace-Lorraine //] que je vais { remonter | remonter } //



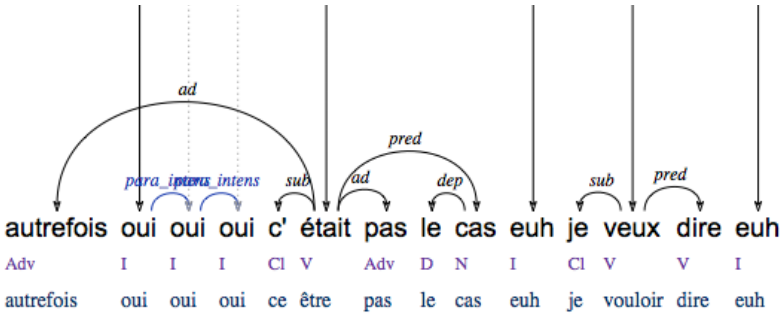
La réitération de conjoints adjectivaux ou adverbiaux peut servir à former une sorte de superlatif ('pas trop facile', 'vraiment très difficile') :

17. et puis "bon" "ben" "voilà" donc < ce qui fait que { c' est | c' est } pas **{ facile | facile }** // (D005)

18. c'est **{ très | très }** difficile à définir // (D202)

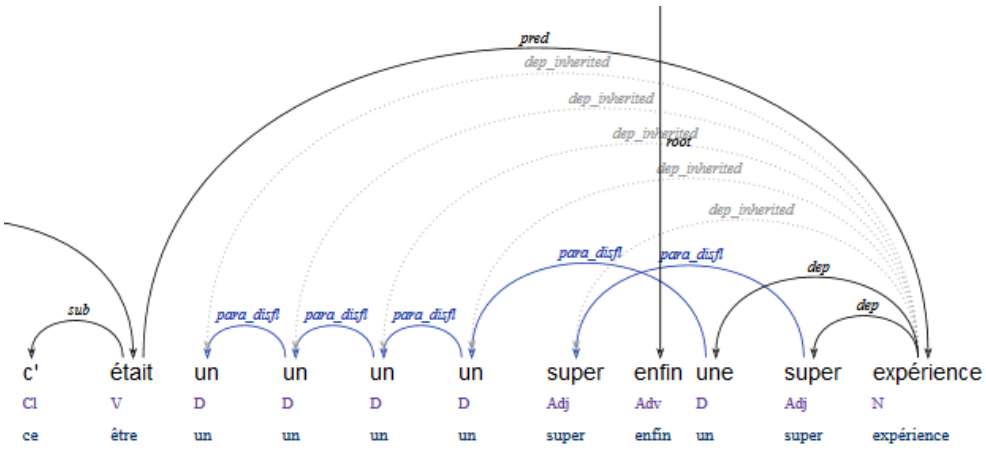
Nous notons comme coordinations intensives les suites d'adverbes telles que *oui oui oui, non non non* :

19. autrefois ({ oui | oui | oui } //) <+ c' était pas le cas " euh " " je veux dire " " euh " // (D001)

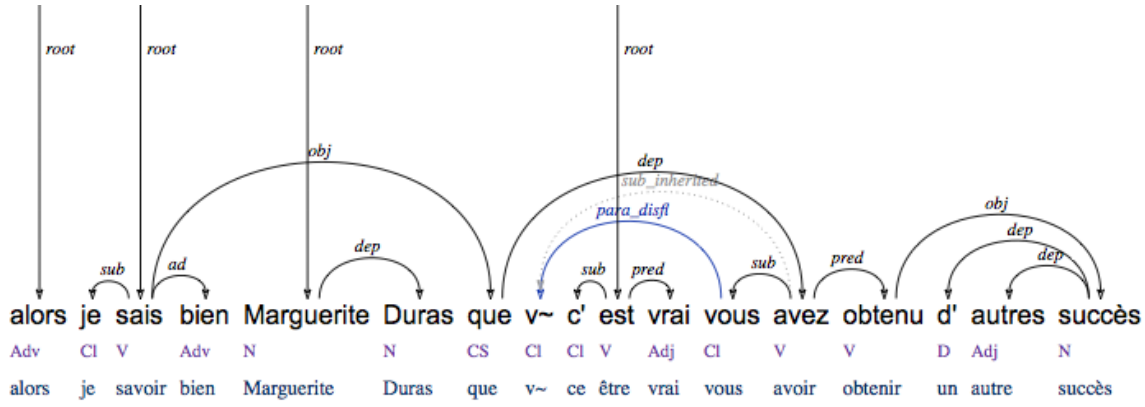


autrefois ({ oui | oui | oui } //) <+ c' était pas le cas " euh " " je veux dire " " euh "

Disfluence (para_disfl) : On parle de disfluence lorsque le locuteur piétine sur une position syntaxique afin d'ajuster sa formulation. Ce piétinement se traduit par un entassement de mots ou d'amorces de mots :



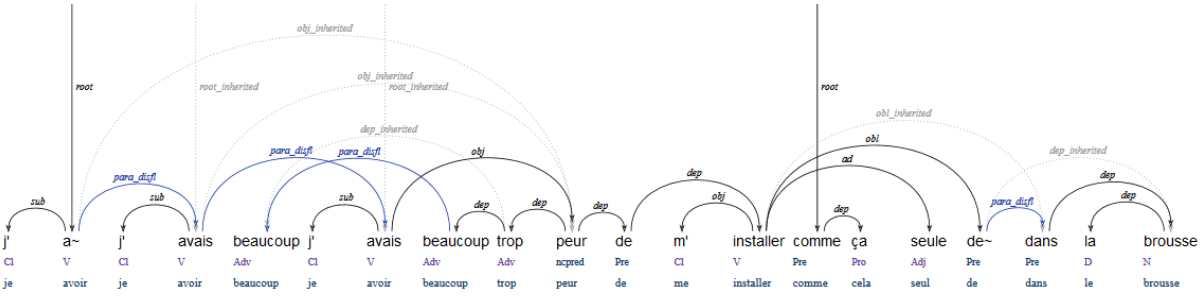
c' était { { un | un | un | un } super | " enfin " une super } expérience] //



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

Ce piétinement peut conduire à la répétition de segments assez longs :

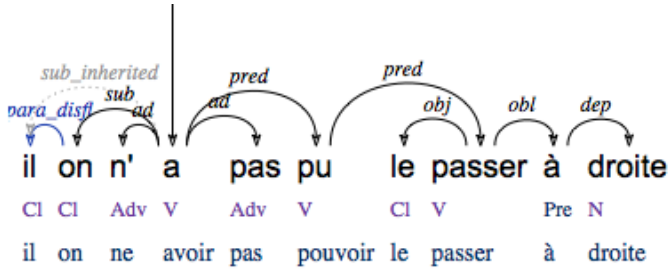
20. alors < **{ { j'a~ | j'avais } beaucoup | j'avais beaucoup }** trop peur de m'installer (comme ça) seule **{ d~ | dans }** la brousse // (D204)



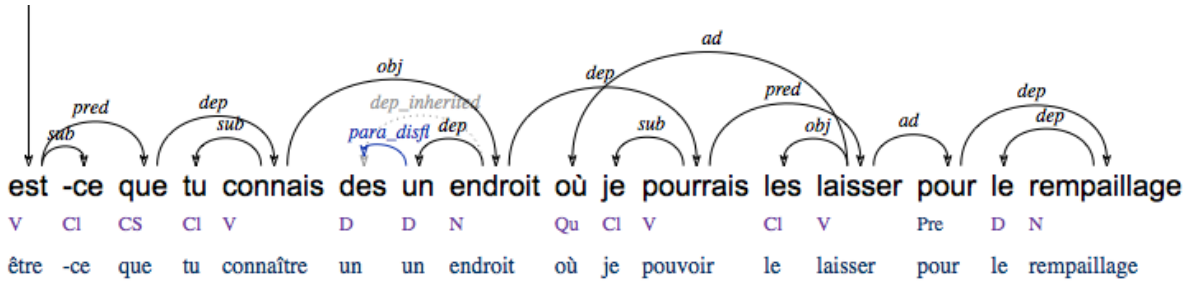
{ { j' a~ | j' avais } beaucoup | j' avais beaucoup } trop peur de m' installer (comme ça) seule { de~ | dans } la brousse //

Dans le cas des disfluences, le segment répété n'a pas d'interprétation propre, contrairement aux cas qui seront décrits dans les prochaines sections, où chaque conjoint possède une dénotation.

Nous considérons qu'il y a disfluence uniquement s'il n'y a pas de changement lexical à l'exception des mots grammaticaux :



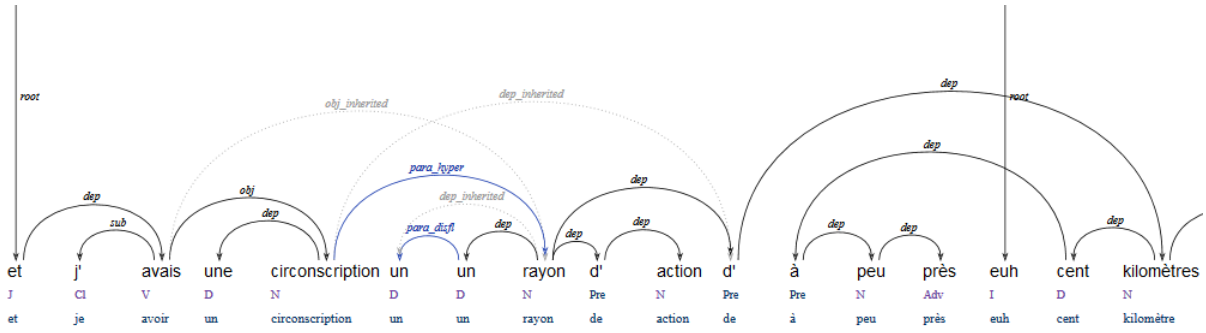
{ il | on } n' a pas pu le passer à droite //



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

Reformulation (para_reform) : Un locuteur peut proposer une première formulation dénotative et y revenir à plusieurs reprises pour la remplacer par d'autres formulations. On est dans ce cas dans un procédé que nous appelons la *reformulation dénotative*.

21. et j'avais **{ une circonscription | { un | un } rayon d'action }** d'à peu près "euh" cent kilomètres tout autour de cet endroit // (D2004)



^ 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)

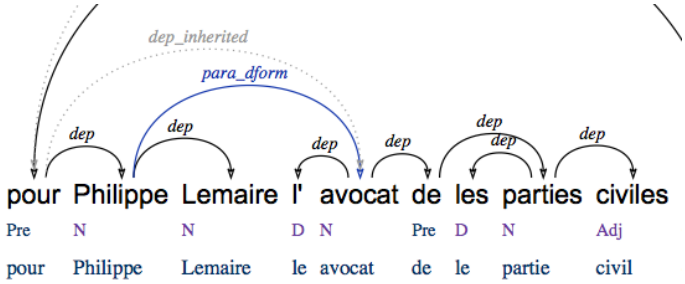
La reformulation a lieu généralement à l'intérieur de la même composante illocutoire.

Toutefois des interruptions, même très longues, sont parfois possibles. Dans l'extrait suivant par exemple, trois UI parenthétiques viennent interrompre un entassement de reformulation { qui parlent pas français | dont les mamans ne parlent pas français } dont la seconde couche est très éloignée de la première, sans que pourtant il y ait eu changement de composante illocutoire.

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)

Double formulation (para_dform) : Le procédé de double formulation consiste à cumuler plusieurs dénotations pour un même élément :

24. pour **{ Philippe Lemaire | (+ l'avocat des parties civiles)** } <+ { c'est d~ | ce sont des } procédés terroristes // (M2006)



La deuxième dénotation fonctionne alors comme une composante illocutoire, voire une unité illocutoire en soi.

Nous notons comme doubles formulations toutes les appositions reformulatives (c'est-à-dire les appositions dont le second élément remplit toutes les conditions morphologiques pour commuter avec le premier), par exemple :

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

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) }** //

A noter que nous ne notons pas comme entassements les appositions modificatives, c'est-à-dire les appositions dans lesquelles le deuxième élément ne commute pas avec le premier, mais en dépend syntaxiquement, en le modifiant :

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~ & } } //Nous notons comme doubles formulations des entassements encodant une reformulation intentionnelle, reliés par le joncteur *c'est-à-dire* et reliant des conjoints non nominaux ou autres:

27. si ça se passe **{ comme ça | (+ ^c'est-à-dire de façon "euh" inquiétante) }**

    { 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) }** } //

Nous notons comme doubles formulations des entassements discontinus qui peuvent être considérés comme des doubles formulations inclusives : la dénotation du deuxième conjoint n'est pas identique à celle du premier, mais elle est incluse dans celle-ci. Ceux-ci prennent soit la forme d'une particularisation (le premier conjoint est constitué par un mot général ou un pronom indéfini) :

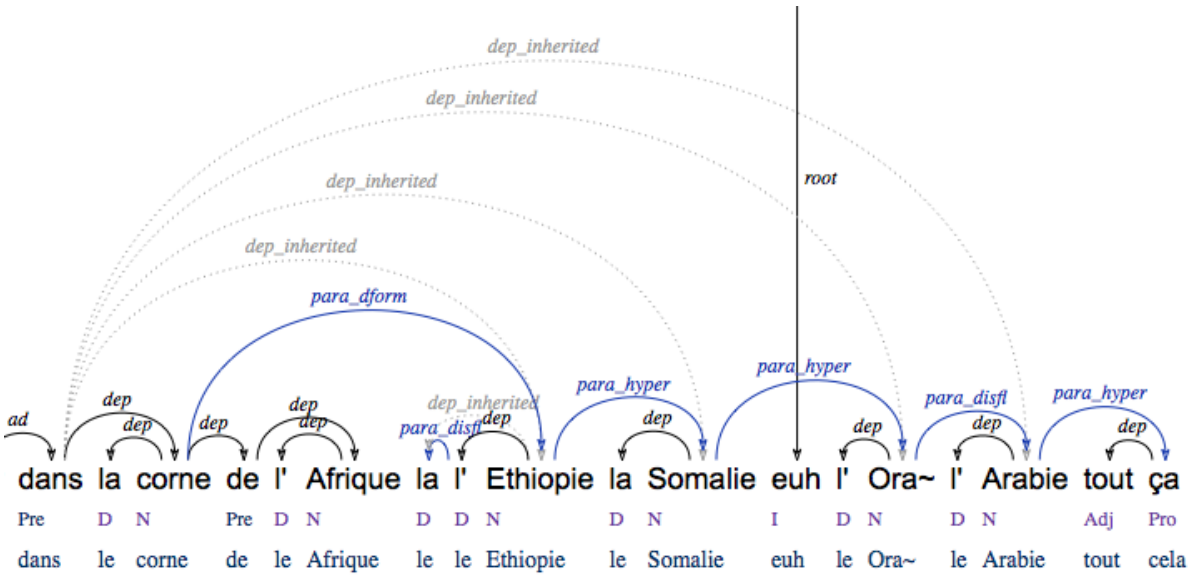
28. si je ne craignais pas d'entrer dans le jeu de certains hommes qui abusent de leur condition < je dirais que vous avez donné **{ quelque chose de plus |}** à la femme //+ **{| des armes de persuasion }** // (D2001)

29. "ben" en fait < il y a { **pas mal de choses** |} qui rentrent en compte //+ { déjà "euh" **l'ambiance du magasin ...** } // (Olive, GARS)

30. et j'ai trouvé **{ cet endroit | (+ Olkaloo) }** où ils avaient besoin d'un médecin // (D2004)

soit la forme d'une exemplification (le deuxième conjoint est constitué par un entassement de co-hyponymes du premier conjoint) :

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)

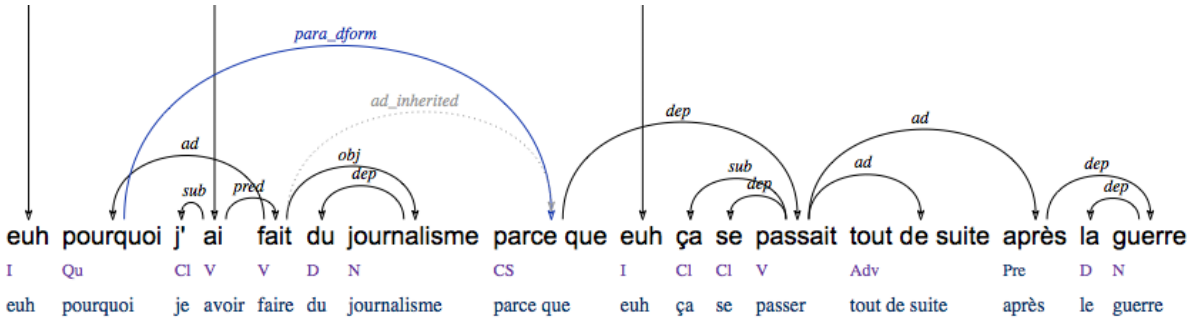


32. "euh" et sinon < les spécialités { les m~ | un { peu moins (je sais pas si c'est ça qui vous intéresse //) | petit peu moins } } prises < "bah" { c'est les | c'est **{ les } spécialités à risques |} //+ {| { la gynéco obstétrique (par exemple) | la cancérologie } }** // (D006)

Parmi les doubles formulations inclusives, nous incluons aussi les questions-réponses (partielles, c'est-à-dire avec un pronom interrogatif). Malgré leur distribution sur deux UI différentes dotées chacune de sa propre force illocutoire (l'une est une question et l'autre la réponse), ces structures remplissent toutes les conditions pour être considérées comme des doubles formulations inclusives : le pronom interrogatif et l'élément utilisé en réponse occupent la même place structurale, ils sont co-dénotationnels et la dénotation du deuxième conjoint est incluse dans la dénotation du premier conjoint :

33. $L1 et il faut compter **{ combien de temps après |}** //
$L2 très rapidement "hein" // **{| { quinze jours | quinze jours maximum } }** // (D0009)

34. "euh" **{ pourquoi |}** j'ai fait du journalisme //+ **{| parce que "euh" ça se passait tout de suite après la guerre }** // (D2001)



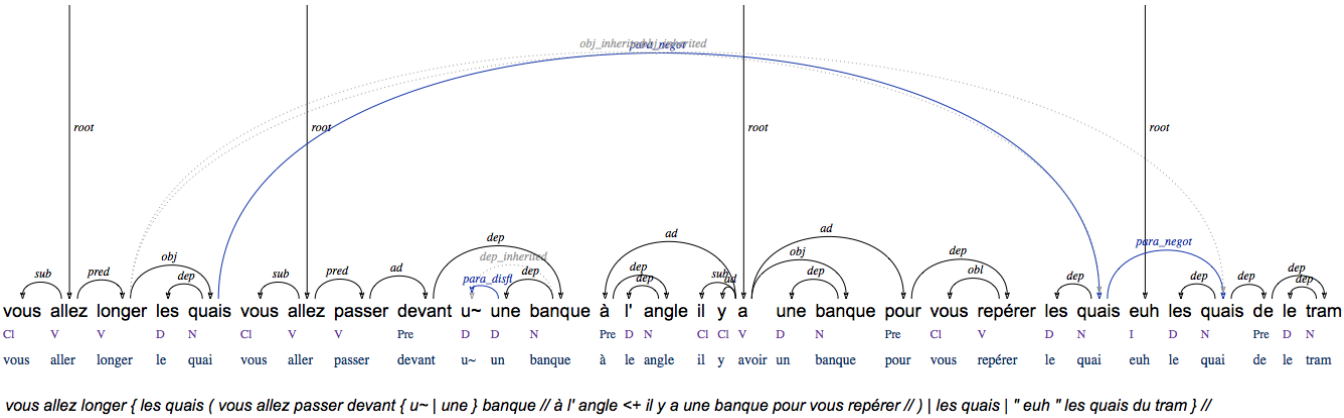
Reformulation ou double formulation : Il peut arriver d'hésiter dans le typage entre reformulation ou double formulation. Par exemple **:**

35. je ne compte **{ que des nuits de souffrance | que de nuits de souffrance }** dans notre humanité // (M2003)

Un test pour choisir entre reformulation et double formulation consiste à faire précéder le deuxième conjoint par *je veux dire* (noyau associé marquant l'intention de reformulation) ou par *c'est-à-dire* (joncteur marquant la double formulation) : dans ce cas, le test permet de typer sans hésitation l'entassement précédent comme une reformulation.

La négociation de la formulation Nous avons relevé quatre opérations de négociation opérant dans les entassements dans notre corpus : la demande de confirmation, la confirmation, la réfutation, la correction. Lors d'une négociation par entassement, il y a répétition d'un segment avec prosodie appropriée. Les exemples suivants montrent une demande de confirmation (répétition de *les quais* avec une prosodie interrogative) et confirmation (avec reformulation en *les quais du tram*) ou une confirmation directe (répétition de *quarante-huit ans* avec une prosodie assertive) :

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 //

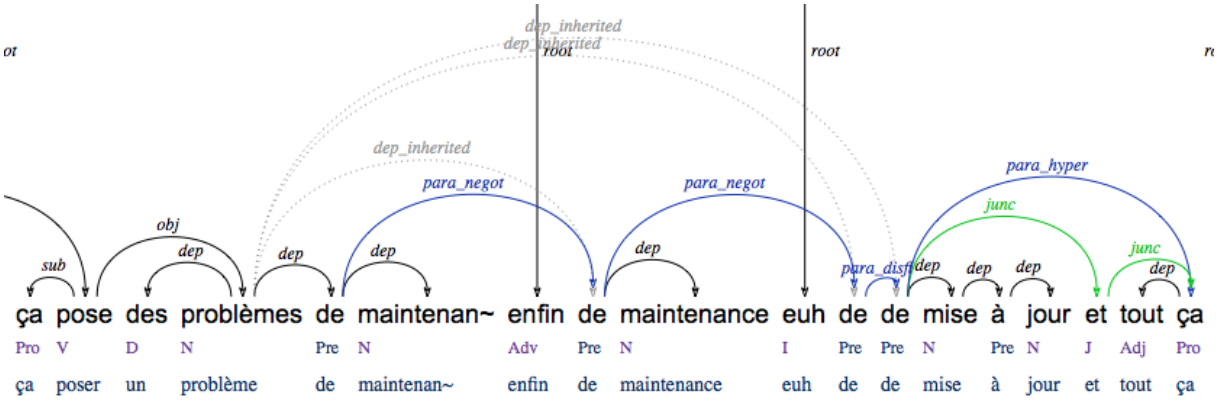


37. $L1 puisque finalement < ça fait "euh" { **quarante-huit ans** |} que vous êtes au Kenya // $L2 {| **quarante-huit ans** } // { oui | oui } // (D204)

Une répétition peut aussi être une réfutation de l'élément répété. Dans les exemples suivant la réfutation est introduite par l'élément *enfin*, qui semble spécialisé pour cet usage, et elle est suivie d'une 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)

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)



Analyse microsyntaxique en constituants

(Kim Gerdes, Sylvain Kahane)

Nous calculons automatiquement une structure de constituants (angl. *phrase structure*) à partir de la structure de dépendance.

Constituants syntagmatiques

Nous appelons *projection maximale* d'un lexème X l'ensemble des lexèmes qui sont dominés par X, c'est-à-dire X, les dépendants de X, les dépendant de ceux-ci et ainsi de suite.

Chaque lexème du corpus donne deux constituants : un constituant syntagmatique qui est sa projection maximale et un constituant lexical (éventuellement fusionnés par soucis de simplicité). Par exemple, un lexème nominal donne une projection de catégorie NP et constituant lexical de catégorie N. Plus généralement :

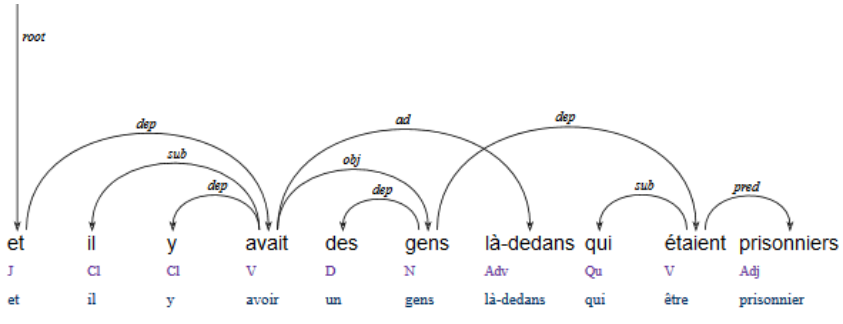
- Les N projettent un NP (nominal phrase)
- Les V finis projettent un S (sentence)
- Les V non finis (infinitifs et participes) projettent un VP (verbal phrase)
- Les CS projettent un CP (complementizer phrase)
- Les Pre projettent un PP (prepositional phrase)
- Les joncteurs projettent un JP (junctor phrase)

Pour ne pas alourdir inutilement les structures, les Adj, les Adv, les D et les pronoms (Pro, Cl, Qu) projettent des AdjP, AdvP, etc., que lorsqu'ils ont des dépendants.

Arbres de constituants microsyntaxiques

Les relations d'emboitement entre constituants microsyntaxiques donnent une structure d'arbre que nous appelons l'*arbre de constituants microsyntaxiques*.

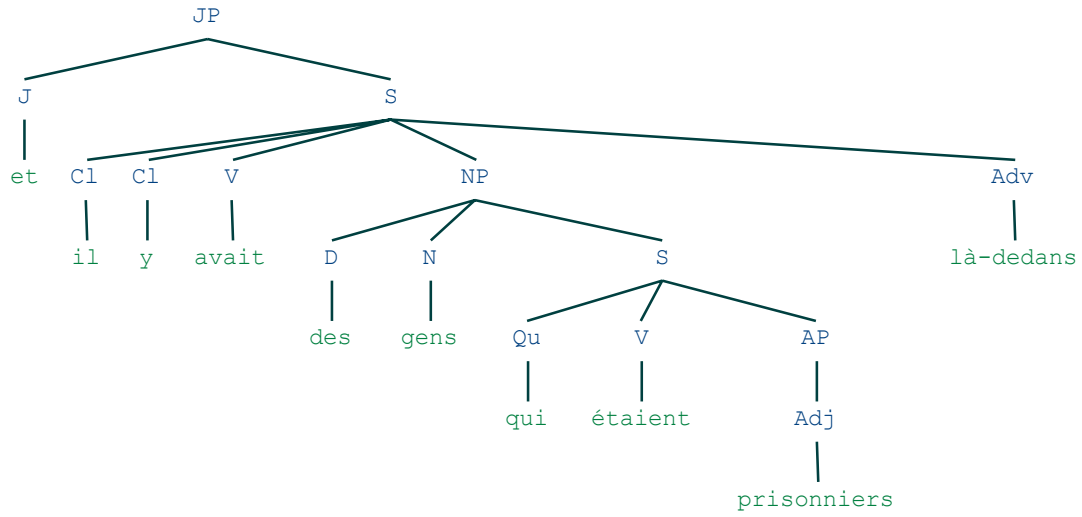
L'existence de dépendances non projectives ajoute une complication que nous illustrons par l'exemple suivant (D0003) :



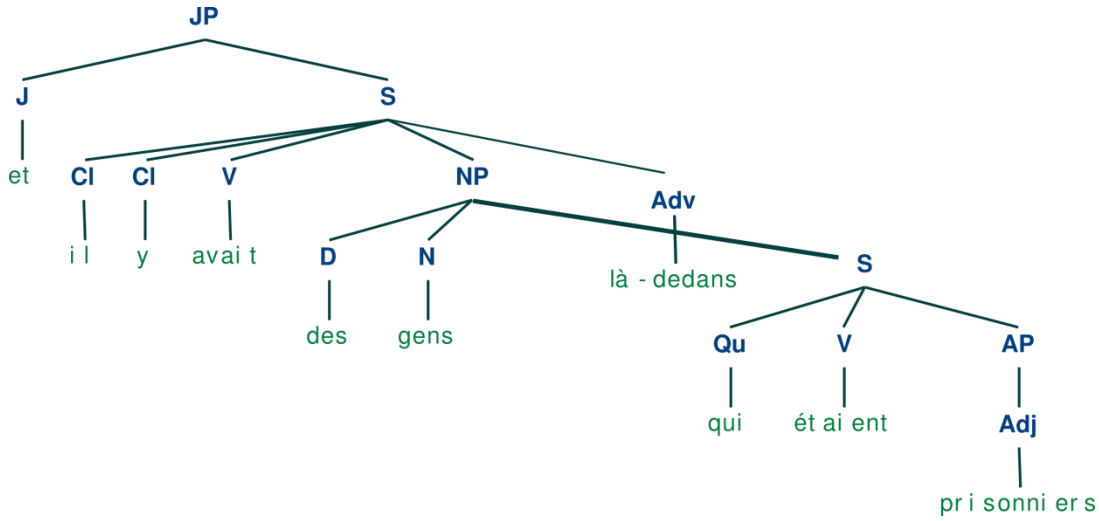
^ *et il y avait des gens là-dedans qui étaient prisonniers //*

La projection de *gens* est le constituant *des gens qui étaient prisonniers*. Il résulte de la non-projectivité de la structure de dépendance (le fait que *là-dedans* viennent interrompre le groupe nominal) que :

- soit l'ordre des mots n'est que partiellement conservé dans la structure de constituants microsyntaxiques :



- soit il faut considérer des arbres de constituants dont les branches se coupent :



Fonction

Chaque constituant reçoit une étiquette fonctionnelle qui est l'étiquette de la relation de dépendance qui gouverne sa tête. Ainsi dans l'exemple qui précède le NP *des gens qui étaient prisonniers* reçoit le trait func="obj".

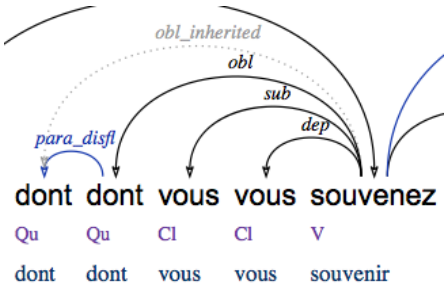
Les constituants reçoivent également tous les traits morphosyntaxiques de leur tête. Les constituants lexicaux qui projettent un constituant syntagmatique reçoivent le trait func="head". Ainsi dans l'exemple précédent, le N *gens* reçoit le trait func="head", le D *les*, qui ne projette pas de constituant syntagmatique, reçoit le trait func="dep".

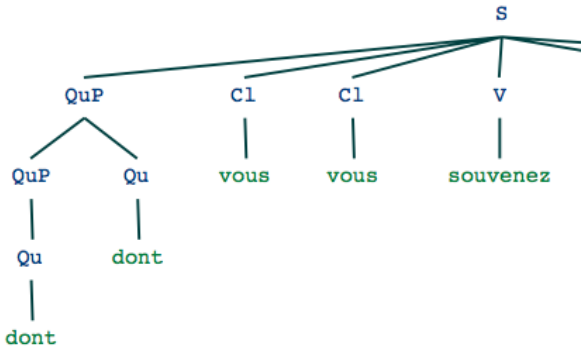
Les *unités rectionnelles* (*UR*) sont les constituants avec le trait func="root".

Analyse des entassements

Pour les entassements, nous faisons le choix de l'analyse asymétrique dans la structure de constituants, c'est-à-dire que les dépendances héritées ne sont pas considérées.

Dans l'exemple qui suit, l'entassement { *dont* | *dont* } est traité comme un QuP car il s'agit de la projection du deuxième *dont* qui est un Qu. Afin d'indiquer que c'est bien le deuxième *dont* qui a été traité comme la tête de ce constituant, le premier *dont* projette également un QuP. Ces deux QuP reçoivent respectivement les traits func="obl" et func="para_disfl".





Lorsqu'il y a un joncteur, celui-ci est considéré comme la tête de la couche, qui devient donc un JP. Par contre, la fonction de ce JP est donnée par le lien que les liens *junc* doublent. Dans l'exemple suivant, le premier JP a donc le trait func="ad", tandis que le deuxième JP a le trait le trait func="para_coord".