Guide d'annotation syntaxique Orféo (version Platinum)

Sylvain Kahane, Henri-José Deulofeu, Kim Gerdes, André Valli, Alexis Nasr

Guide d'annotation syntaxique du corpus Orfeo (version platinum)

Version : novembre 2016

Rédacteurs : Sylvain Kahane, José Deulofeu, Kim Gerdes, Alexis Nasr, André Valli
Avec la collaboration des annotateurs :
Marion Bernard, Anaïs Chanclu, Fanny Lafontaine,
Marie Marcia, Chloé Monnin, Rafaël Poiret

Note préalable : certaines analyses sont arbitraires et visent à minimiser les erreurs d'analyse automatique. C'est le cas de l'analyse des clivées (*c'est Marie qui vient*) qui ne sont pas distinguées des constructions avec une relative ordinaire (*c'est la fille qui devait venir*).

1. Segmentation

1.1 Unité maximale

Selon le guide de segmentation, « *les unités maximales de segmentation (US) sont basées sur des constructions verbales, mais aussi nominales, adjectivales ou adverbiales regroupant un élément tête ainsi que toutes les séquences qui sont régies par elles. Les US sont en fait des énoncés, constitués de la séquence tête + éléments régis étendue aux éléments dits « associés » dans le cadre de l'AP ou périphériques dans d'autres cadres.* »

L'unité maximale est également appelée ***énoncé***.

Nous associons à chaque énoncé un arbre complet dont le nœud racine reçoit la fonction *root*. La structure de dépendance d'un énoncé est donc toujours connexe et intègre aussi bien des relations micro- que macrosyntaxiques.

1.2 Unité minimale

Les énoncés sont découpés en tokens, lequel ***token*** constituant l'unité minimale de l'analyse en dépendance.

Nous appelons mots orthographiques les segments de textes maximaux comprenant des lettres et l'un des deux autres symboles utilisés dans les transcriptions orales : l'apostrophe et le tiret.

Le guide doit être complété pour la segmentation de l'écrit, mais en première approximation on peut dire que les séquences de chiffres et les signes de ponctuation (utilisés comme ponctuation, cf. les différents usages du point) forment des tokens.

Les mots comportant une apostrophe sont décomposés en deux tokens avec l'apostrophe à gauche (*l' enfant*), à l'exception de *aujourd'hui* et de mots grammaticaux comme *quelqu'un* ou *l'un*, qui figurent dans notre lexique des unités grammaticales.

Les mots comportant un tiret sont décomposés en deux tokens avec le tirets à droite lorsque le token ainsi formé appartient à notre lexique de mots grammaticaux : *dit -on*, *a -t-il*, *maison -là, …* Les autres mots comportant un tiret forment un token : *avant-hier, au-dessus, soutien-gorge, ...*

En dehors de ces cas, un token ne peut jamais être une partie de mot. En particulier, les amalgames (*au, du, des, ...*) ne sont jamais décomposés.

Les mots lexicaux forment un token même lorsqu'ils font partie d'une locution. Ainsi *pomme de terre* forme trois tokens.

Les locutions (expressions multi-mots) grammaticales répertoriées dans le lexique Orféo (adapté du Lefff) forment des tokens. Cela concerne les catégories suivantes :

- ADV : *à coup sûr, belle lurette, bien entendu, dans ce cas, ...*
- COO : *ainsi qu', c'est-à-dire, et puis, y compris, plutôt que,...*
- CSU : *est-ce que, parce que, sous réserve qu', toujours est-il que, ...*
- DET : *Dieu sait quelle, le moins de, n'importe quels, tel et tel, une drôle de, ...* (mais pas les ADV *de* comme *beaucoup de, combien de, moins de*, *plein de, …*)
- INT : *à bientôt, mh mh, ouh là, ...*
- PRE : *aux côté de, d'après, de façon à, en face d', ...*
- PRO : *autre chose, elles-mêmes, le mien, l'autre, n'importe lequel, ...*
- PRQ : *où est-ce que, qui est-ce qui*, *…*

Note technique : Les locutions grammaticales qui possèdent par ailleurs une analyse compositionnelle sont décomposées lors du pré-traitement et reliées par un lien *morph* lors de l'analyse en dépendance. Cette analyse sera revue en post-traitement en vue d'effacer la fonction *morph* dans le treebank final.

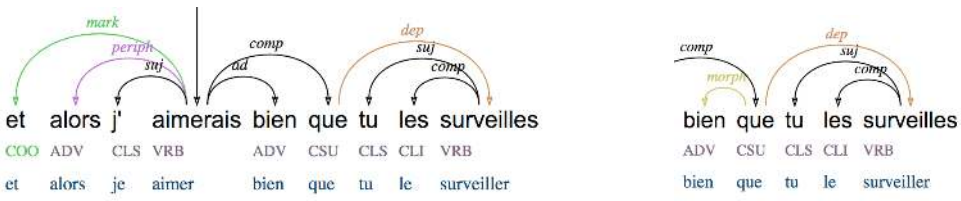


Figure 1. Les deux analyses de *bien que*

Afin d'optimiser la reconnaissance des *morph* ceux-ci suivent l'analyse syntaxique qu'aurait la locution. Lorsqu'un élément possède la même catégorie que la locution, il est privilégié comme tête (c'est cas de *que* pour *bien que* ci-dessus).

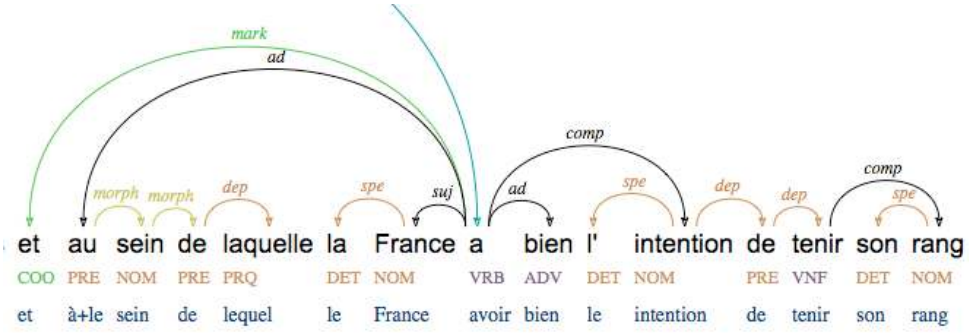


Figure 2. Exemple de liens *morph*

2. Morphosyntaxe

Chaque token est pourvu d'une étiquette morphosyntaxique (donnée par le lexique).

2.1 Liste des catégories morphosyntaxiques

- ADJ (adjectifs qualificatifs) : *méchant*, *petit*, *long*, *gigantesque*, *drôle*, *rouge*, etc.
- ADN (adverbes de négation) : *pas*, *jamais*, *nullement*, *guère*, *plus*, etc.
- ADV (adverbes) : *savamment*, *peut-être*, *in extremis*, *très*, *environ*, etc.
- CLI (autres clitiques) : *te*, *lui*, *-le*, *-y*, *en*, *-leur*, *nous*, etc.
- CLN (clitique de négation) : *ne*
- CLS (clitiques sujets) : *tu*, *elles*, *vous*, *-vous*, *c',* etc.
- COO (conjonctions de coordination) : *et, ou, alias*, *mais encore*, *voire*, *puis*, etc.
- CSU (conjonctions de subordination) : *au fur et à mesure qu', alors que*, *lorsque*, etc.
- DET (déterminants) : *cette, certains, quelques*, *un*, etc.
- INT (interjections) : *hein, ben, allô*, *pfff*, *no comment*, *niark*, *okidoki*, *parbleu*, etc.
- NOM (noms) : *diplodocus*, *Montastruc-la-Conseillère*, *topinambour*, *Google*, etc.
- NUM (nombres) : *six*, *treize*, *milliard*, *quatorze*, *mille*, *billion*, *dix-sept, quatre-vingt-onze, vingt-cinq,* etc. (mais pas *soixante et un*)
- PCT (signes de ponctuation) : *!*, *?, !*, *etc.*, *(*, *»*, etc.
- PRE (prépositions) : *de, des, nonobstant*, *parmi*, *pour cause de*, *par delà*, *outre*, etc.
- PRO (pronoms) : *moi, celles*, *les tiens*, *plusieurs*, *vous-mêmes*, *nul*, *pas grand-chose*, etc.
- PRQ (pronoms interrogatifs-relatifs) : *combien est-ce que*, *lequel*, *pourquoi*, *que*, etc.
- VNF (verbes à l'infinitif) : *tenir*, *poindre*, *jouer*, *entendre*, etc.

- VPP (verbes au participe passé) : *tenu*, *point*, *joué*, *entendu*, etc.
- VPR (verbes au participe présent) : *tenant*, *poignant*, *jouant*, *entendant*, etc.
- VRB (verbes à la forme finie) : *tiens, poignent, joueraient, entendissions*, etc.
- X (mot inconnu, étranger ou tronqué de catégorie indécidable) : *El Paìs, fuck you,* etc.

Remarques diverses :

- *des* est toujours analysé comme PRE, qu'il s'agisse de l'article indéfni ou de l'amalgame de *de* et de l'article défini *les*.

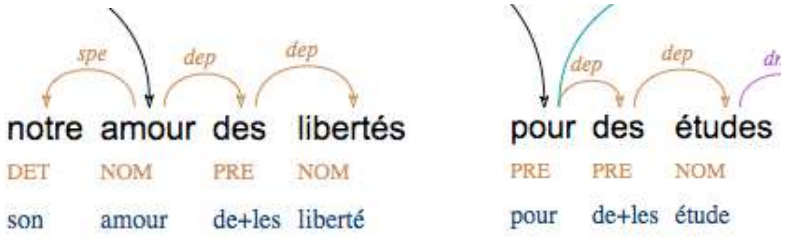


Figure 3. *des* comme PRE

- *deux* est toujours analysé comme NUM qu'il commutent avec des DET (*deux chaises*), des ADJ (*les deux chaises*), des PRO (*j'en ai deux*) ou des NOM (*la deux*).
- Par contre, *million* est NUM dans *deux millions cinq cent mille* et NOM dans *deux millions de personnes*.

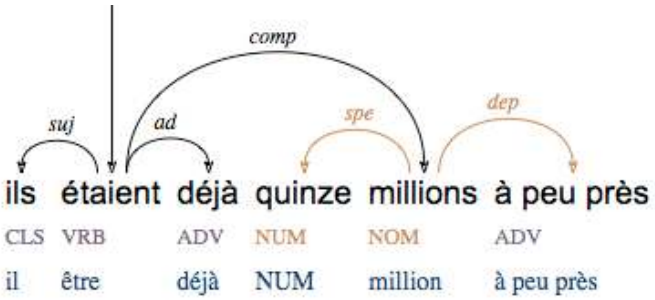


Figure 4. NUM vs NOM

- *quelques* est toujours ADJ, avec fonction *spe* dans *quelques chaises* et *dep* dans *ces quelques chaises*. De même pour *tout* : il est ADJ et *spe* dans *toute autre solution* ; ADJ et *dep* du nom *jours* dans *tous les jours* ; mais PRO dans *je sais tout* et ADV dans *tout jaune*.
- Les déictiques comme *demain* sont classés parmi les ADV en suivant la tradition, même quand il commute des NOM comme *lundi* : *il vient demain/lundi/lundi prochain/ce lundi/le lundi de mon anniversaire*.

2.2 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 (DET) et les pronoms clitiques (CLI) *le, la, l', les* est *le*, le lemme pour *du* et *des* est *de+le*.

Le lemme pour les pronoms clitiques de 1ère et 2ème personne *je, tu, nous, vous, me, te ...* est la forme non élidées (*me* pour *m'*). Le lemme pour les clitiques sujet (CLS) de 3ème personne (*il, ils, elle, elles*) est la forme du masculin singulier *il*. Le lemme pour les pronoms forts personnels (PRO) (*toi, lui, eux, elle, elles*) est la forme de 1ère personne singulier *moi*.

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

Le lemme pour les mots tronqués est le token lui-même, même si on pense pouvoir reconstruire le mot que le locuteur souhaitait produire :

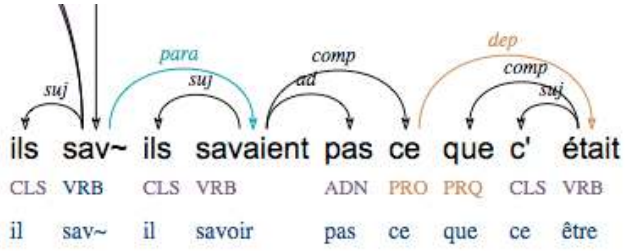


Figure 5. Lemmes

3. Microsyntaxe

Les relations de dépendance dont nous avons besoin en microsyntaxe sont :

- *root* (racine)
- *suj* (sujet)
- *comp* (complément)
- *ad* (ajout)
- *aux* (auxiliaire)
- *spe* (spécifieur)
- *dep* (autres dépendants)
- *disflink* (segment non analysable)

3.1 La racine *root*

root désigne l'élément racine de l'énoncé, qui est la tête du noyau de l'énoncé. Cet élément ne dépend d'aucun autre élément aussi bien à l'échelle microsyntaxique qu'à l'échelle macrosyntaxique. Lorsqu'un élément forme à lui seul l'intégralité de l'énoncé, il s'agit alors d'un élément *root*.

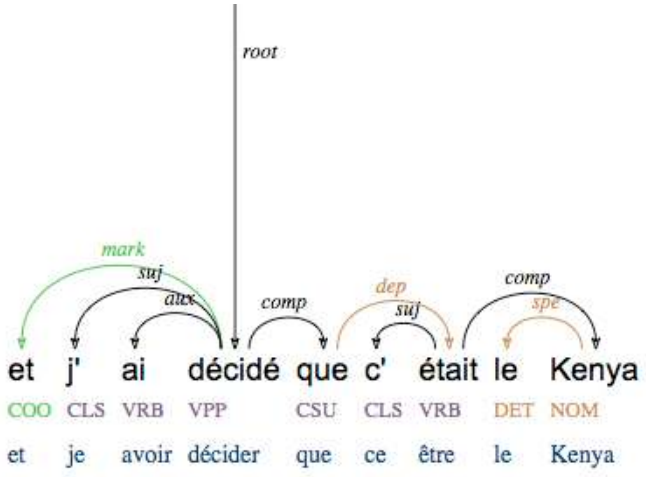


Figure 6. *root*

Convention : *oui* est un élément *root* uniquement quand il constitue une réponse à une question. Dans les autres cas, il est marqueur de discours (relation *dm*). Il peut aussi être *comp* dans *il pense que oui*.

NB : Toutes les catégories peuvent être *root*, sauf DET et COO.

Le participe passé (VPP) est la tête d'un forme verbale complexe (voir *aux*).

Les constructions introduites par une CSU sont normalement rattachées, même lorsqu'il s'agit a priori de subordonnées non régies « périphériques » (*il doit être à la fac parce que sa voiture est dans le parking*) (que le segmenteur automatique ne pourrait pas distinguer de complément de causation : *il doit être à la fac parce qu'il a cours aujourd'hui*).

3.2 Dépendants du verbe

3.2.1 Le sujet : *suj*

suj désigne le sujet du verbe.

En cas de construction impersonnelle, c'est le sujet grammatical, c'est-à-dire le pronom qui porte la fonction *suj* :

Figure 7. Sujet impersonnel

Certaines participiales prennent un sujet (on fait ici l'équivalence avec *une fois que le fleuve est traversé*, ce qui justifie également le rôle de tête de *une fois*).

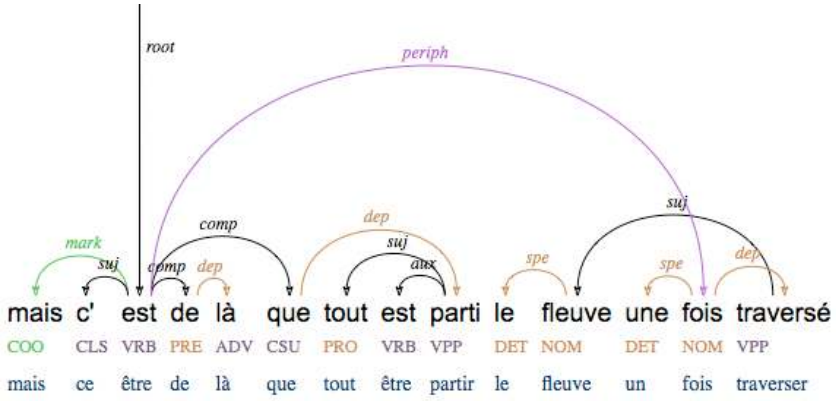


Figure 8. Sujet d'une participiale

Dès que le verbe porte un enclitique sujet, celui-ci est déclaré comme sujet. En conséquence, un verbe peut exceptionnellement avoir deux sujets :

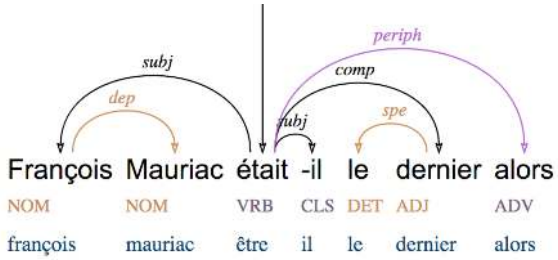


Figure 9. Double sujet

Cette situation est néanmoins exceptionnelle. En cas de dislocation gauche du sujet, seul le pronom clitique occupant la position microsyntaxique de sujet portera la fonction *suj* (voir *periph*) :

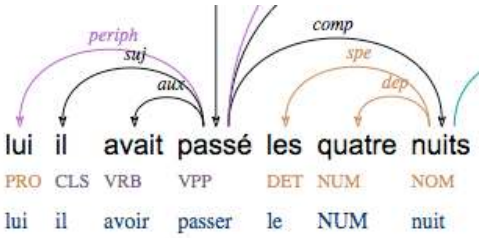


Figure 10. Dislocation du sujet

3.2.2 L'auxiliaire :*aux*

La racine d'un proposition finie est toujours le verbe fini, et donc l'auxiliaire dans le cas d'une forme verbale complexe. Le verbe au participe passé (VPP) dépend de l'auxiliaire par la relation *aux*. Seul le sujet et le CLN *ne* sont attachés à l'auxiliaire. Tous les autres dépendants de la forme verbale sont attachés au VPP.

Remarque : Dans les figures hors de cette section, le VPP gouverne l'auxiliaire, comme décidé dans un précédent schéma d'annotation.

Convention : Sont seulement considérés comme auxiliaires ÊTRE et AVOIR.

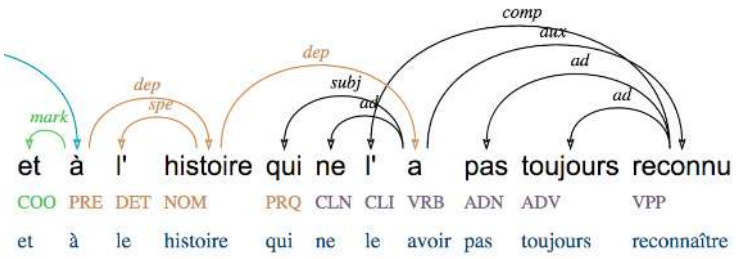


Figure 11. Auxiliaire et clitique

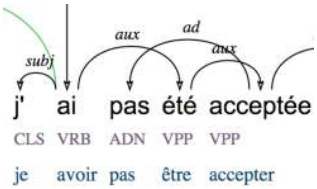


Figure 12. Double auxiliaire

3.2.3 Les compléments du verbe : *comp*

comp désigne les compléments régis du verbe par la valence verbale (objets direct et indirect, attributs du sujet ou de l'objet, complément oblique, complément locatif régi, etc.).

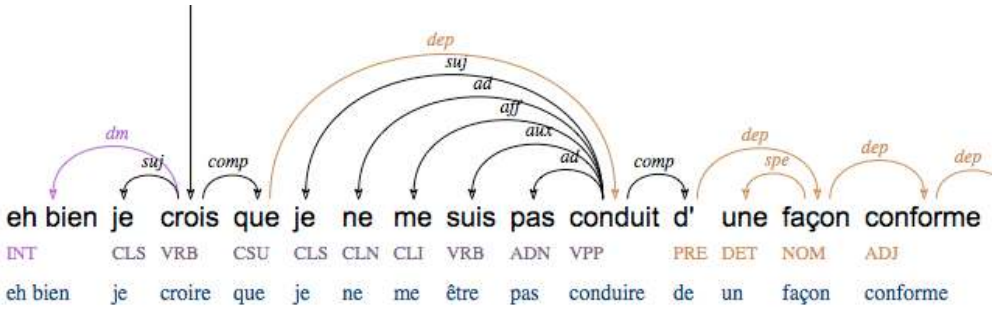


Figure 13. Compléments

A terme, la distinction sera faite en référence à un dictionnaire de valences, si possible automatiquement. En l'état, les annotateurs se réfèrent aux critères habituels.

Conventions :

- Tout adjectif qualificatif en fonction attribut du sujet est un *comp* du verbe.

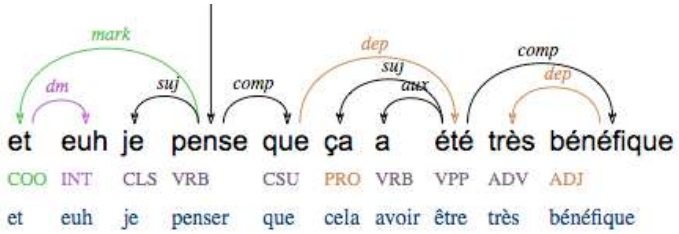


Figure 14. Attribut

Un cas d'inversion : *tel* ←comp- *était* -suj→ *mon cas à cette époque*.

- De même, dans les constructions du type *La valise pèse lourd* ou encore *Théo a froid*, les adjectifs qualificatifs se trouvent être *comp* du verbe.
- En revanche, il faut distinguer les constructions faisant apparaître des adjectifs modifieurs du verbe donc notés *ad*: telles que *La voisine rit jaune*, *Il chante faux* ou encore *Elle dort nue*.
- En ce qui concerne les pronoms relatifs-interrogatifs (PRQ), certains sont *comp* du verbe tel que *qu'est-ce que* (un seul token).

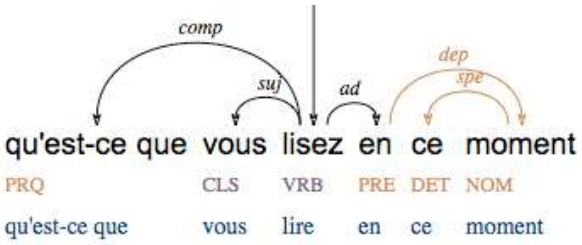


Figure 15. Pronom interrogatif

- A contrario, d'autres se trouvent être des *ad* comme en témoigne *Où as-tu dormi*.
- Les discours direct introduit par un verbe de dire sont traités comme des compléments de ce verbe, sauf lorsque celui-ci se trouve en incise (*je viendrais dit-il*, cf. *dm*).

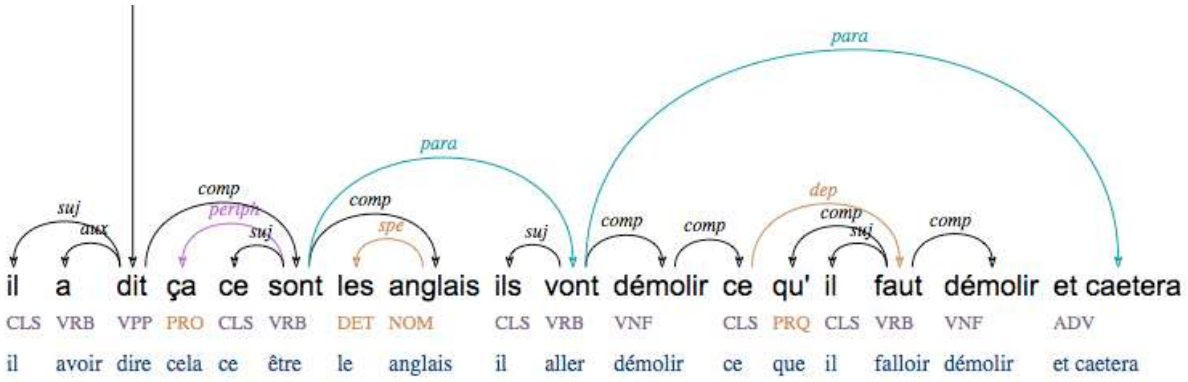


Figure 16. Discours rapporté

- *comp* est utilisé uniquement pour les dépendant d'un verbe. Les dépendant d'un nom sont toujours *dep* même lorsqu'il s'agit d'une construction à verbe support :

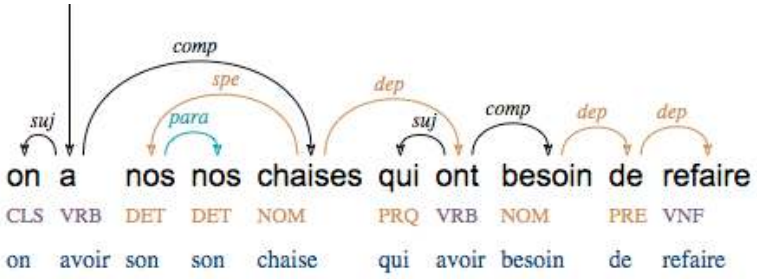


Figure 17. Construction à verbe support

Cas limites

Il existe de nombreux cas limites entre *comp* et *ad*. En l'absence d'un lexique de référence, nous laissons une certaine zone de flou. Notons quand même que seront traités comme *comp* :

- les compléments d'agent : *il a été retrouvé par Marie* ;
- les compléments locatifs des verbes de mouvement ou de localisation : *elle habite à Paris, elle travaille à Paris, elle travaille dans la finance, elle voyage en France, elle a mis son doigt dans le trou, elle l'a attrapé par le col, elle lui a marché sur le pied*
- les compléments locatifs qui n'indique pas la lieu du procès mais son résultat : *elle a cassé les œufs dans un bol*
- les datifs bénéficiaires : *elle lui repeint sa chambre, elle lui lave les mains, elle lui ouvre la porte*
- Les pronoms réfléchis, y compris lorsqu'ils sont figés (*se souvenir*) ou marque le moyen (*ce livre se vend bien*). De même, le *y* de *il y a* (*il y a pas de problèmes*) ou le en de *s'en aller* (*va-t-en d'ici*) est *comp*.

Seront traités comme *ad* :

- les compléments en *avec* : *il est venu avec son frère, il le coupe avec son canif, il dort avec son doudou*.
- Les datifs éthiques se trouvent être *ad* du verbe : *Je **te** lui foutrais une baffe*.

3.2.4 Les adjoints : *ad*

Lorsque qu'un modifieur du verbe se trouve au sein du syntagme verbal, il est alors *ad* de ce verbe.

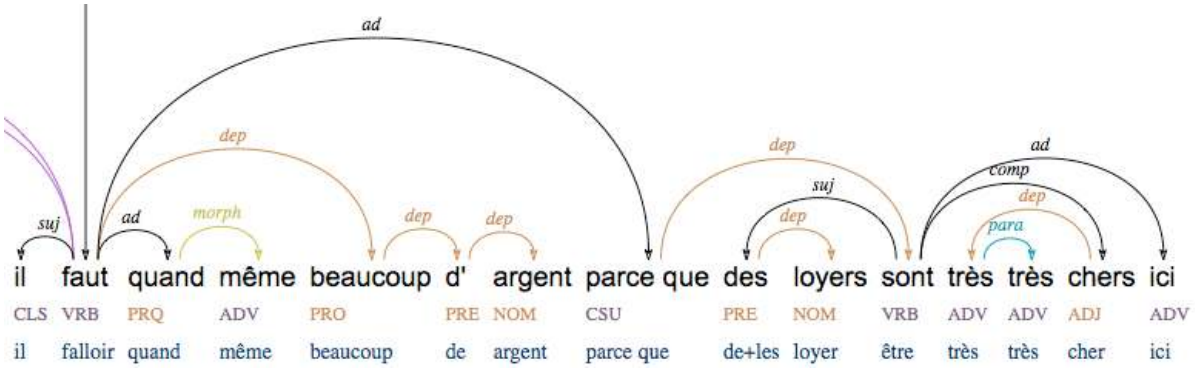


Figure 18. Adjoints

La plupart des adverbes et groupes adverbiaux modifiant le verbe sont susceptibles d'être *ad* lorsqu'ils se trouvent au sein du syntagme verbal sauf :

- quand ils sont régis par la valence verbale : *Il se comporte -comp→ bien*, complément locatif des verbes de mouvement :

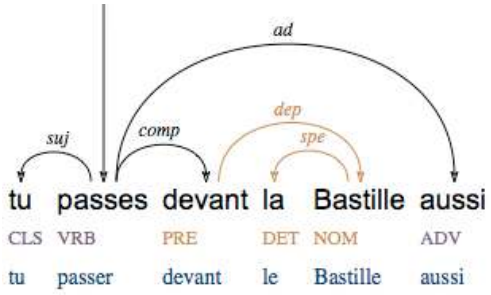


Figure 19. Complément locatif

- ceux qui sont en position détachée, notamment les prénoyaux (voir *periph*) :

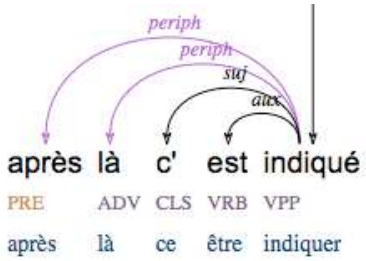


Figure 20. Modifieurs hors noyau

3.4 Le spécifieur : *spe*

spe désigne le spécifieur du nom. Un seul des éléments rattaché au nom peut porter cette étiquette. Les autres éléments porteront alors la fonction *dep*.

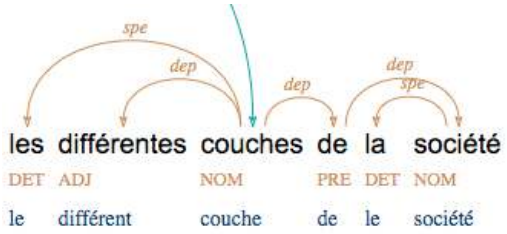


Figure 21. Spécifieur et dépendants du nom

Rappel : *de, du* et *des* sont toujours PRE et toujours analysés comme gouverneur du NOM qu'ils introduisent. Ils ne sont donc jamais *spe* :

Figure 22. *de, du, des*

3.5 Les autres dépendants : *dep*

dep regroupe tous les éléments dépendant microsyntaxiquement d'une tête qui ne font pas l'objet d'une fonction spéciale (*subj, aux, spe, mark*) ou ne sont pas traités au niveau macrosyntaxique (*periph*). Sont également traités comme *dep* les adverbes intégrés au noyau (*il est **quand même** venu*) et les subordonnées, même lorsque leur staut comme complément régi est discutable (*il est à la fac **parce que sa voiture est dans le parking***).

NB : dans les exemples la dépendance très générale dep apparaît parfois décomposée en deux relations quand le gouverneur est un verbe (VRB) : *comp* (pour les compléments sous catégorisés par le verbe et *ad* pour les autres compléments. Cette distinction n'a pas été retenue dans l'annotation, mais reste présente dans les images que l'on n'a pas eu le temps de modifier. Il faut donc lire dans les images *dep* à la place de *comp* et de *ad*

Parmi les relations de dépendance *dep*, on notera :

- les compléments régis par le verbe qu'ils fassent partie ou non de la valence verbale (objets direct et indirect, attributs du sujet ou de l'objet, complément oblique, ajouts ou modifieurs).

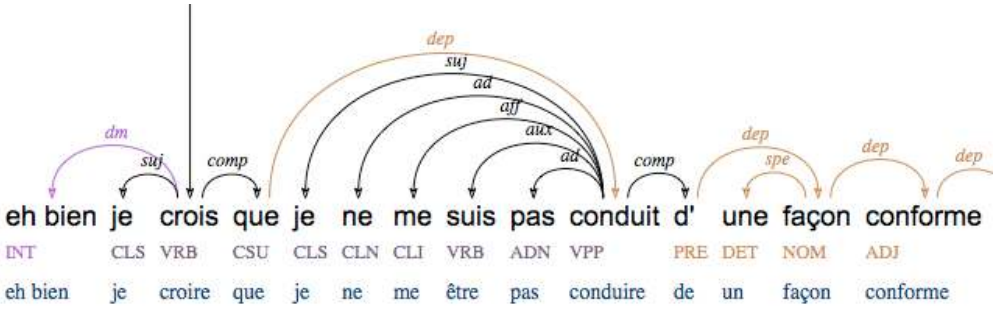


Figure 23. Compléments d'un verbe

- le complément de la préposition, de l'adjectif, de l'adverbe
- le verbe introduit par une conjonction de subordination (pour les conjonctions de coordination, voir *mark*) :

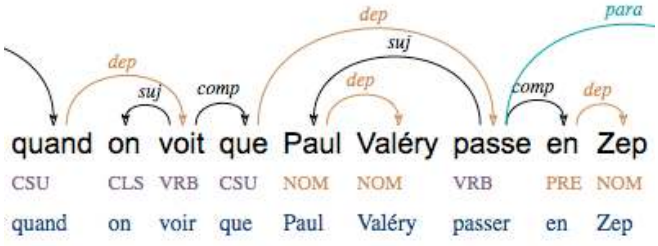


Figure 24. Complément de CSU et de PRE

- les pronoms relatifs-interrogatifs (PRQ) qui ne sont pas sujets, y compris lorsqu'ils précèdent le sujet :

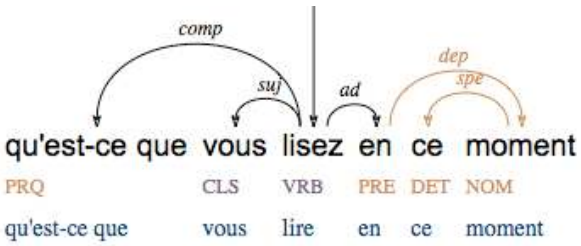


Figure 25. Pronom interrogatif

- les discours direct introduit par un verbe de dire, sauf lorsque celui-ci se trouve en incise (*je viendrais dit-il*, cf. *insert*).

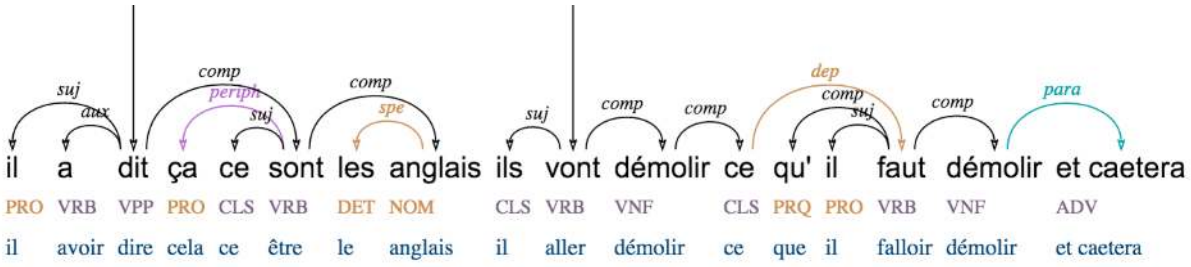


Figure 26. Discours rapporté

- Les modifieurs adverbiaux du verbe qui se trouvent au sein du syntagme verbal sont notés *dep* (cf. *quand même, parce que, ici* dans l'exemple suivant)

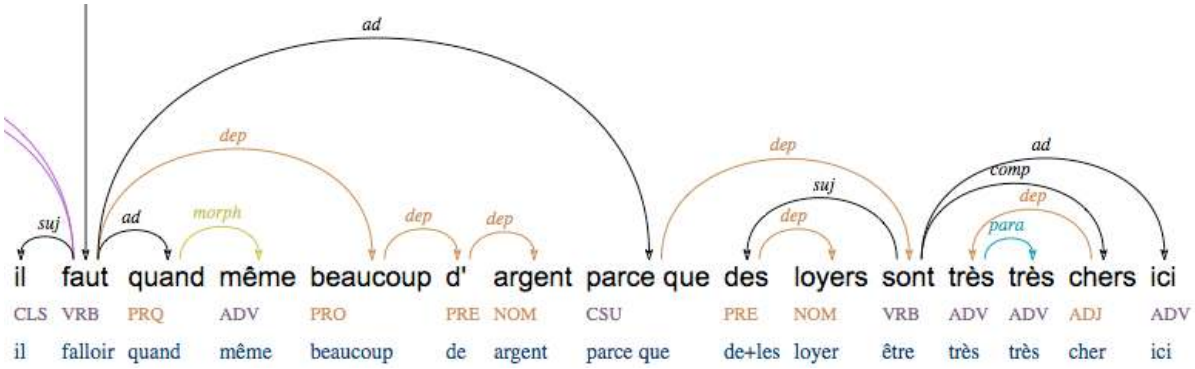


Figure 27. Adverbiaux

Mais pas ceux qui sont en position détachée à gauche, qui sont *periph* (voir plus loin) :

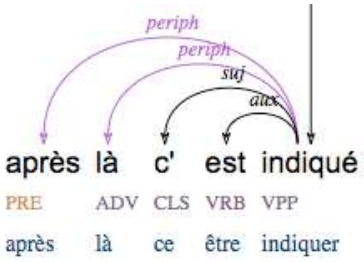


Figure 28. Modifieurs hors noyau

3.6 Eléments disfluents : *disflink*

On utilise la fonction *disflink* pour rattacher un élément en l'absence de son gouverneur : *Il mange* -disflink→ *quelques*. La tête du groupe nominal étant absente et l'adjectif *quelques* ne dépendant pas normalement d'un verbe, il est rattaché par le lien *disflink*.

Un élément *disflink* est rattaché au mot qui précède.

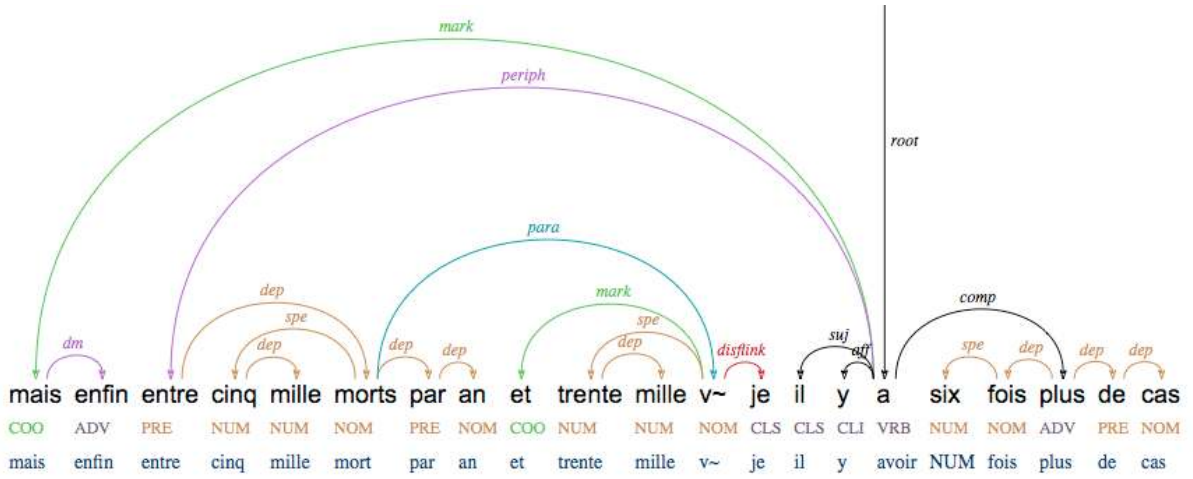


Figure 29. *disflink*

Les amorces disfluentes qui figurent dans une liste sont normalement gérées par le lien *para*. On évite le lien *disflink* autant que possible. Celui-ci n'est utilisé que quand aucune autre analyse correcte ne semble possible. Dans les exemples qui suivent, l'amorce en tant que telle est gérée par le lien *para* (*qu' qu'*, *j' j'*, *de de, la la*). L'utilisation du lien *disflink* est rendue nécessaire par le fait que l'amorce contient plusieurs éléments qui ne sont pas liés par une dépendance régulière (*qu' on*, *j' en*, *de la*).

Dans la figure 21 on ppropose de mettre disflink entre les deux clitiques sujet (répétition simple

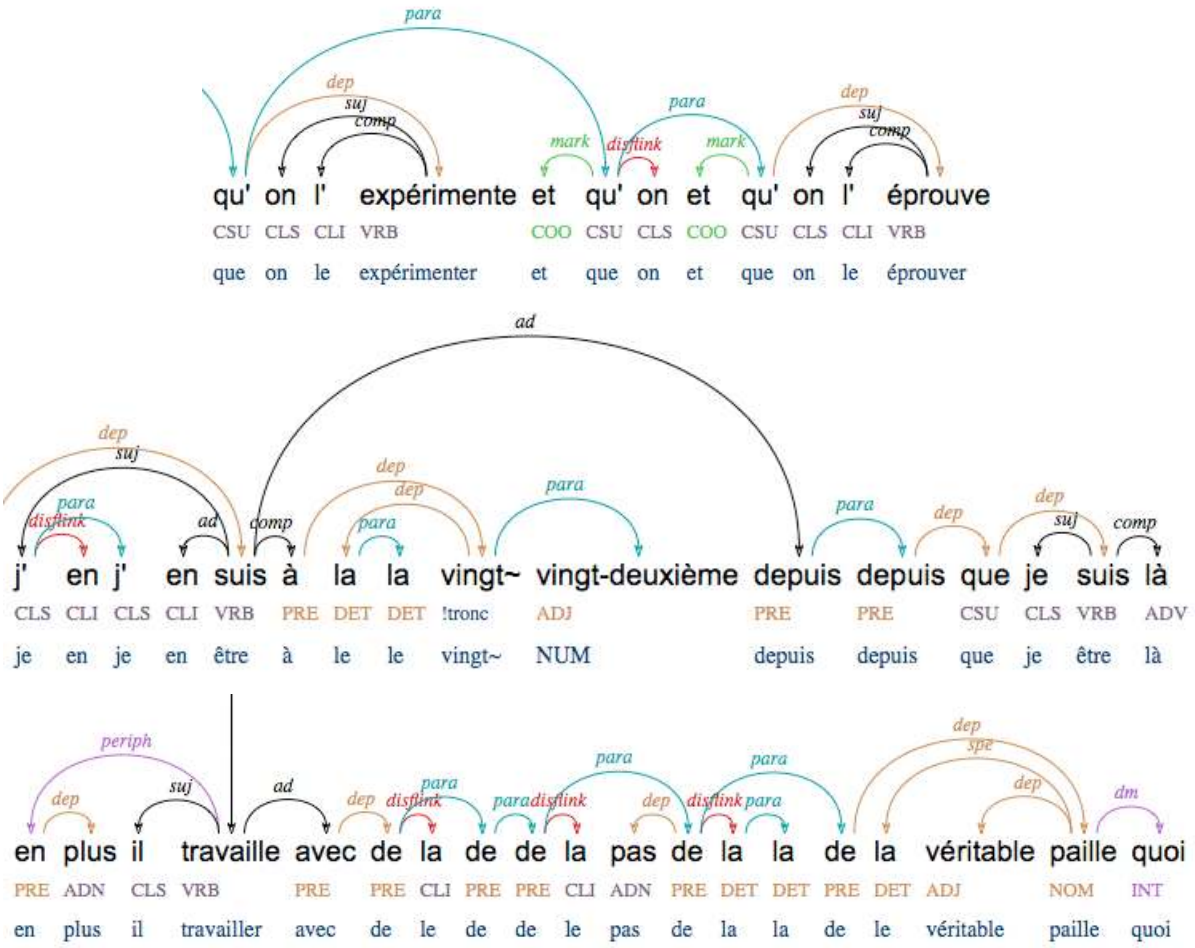


Figure 30. Amorces disfluentes

Les prépositions sans leur complément obligatoire sont analysées normalement :

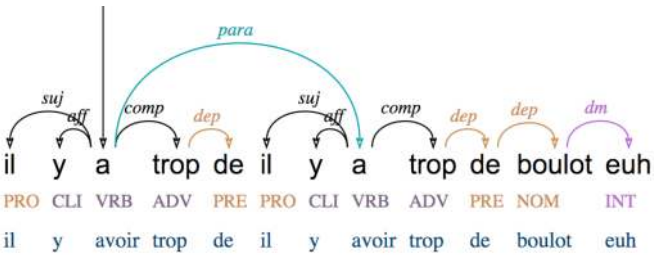


Figure 31. Préposition sans complément

Qu'il s'agisse des amorces d'auxiliaire ou de verbe pour les cas de répétition simple nous proposons de maintenir *para* ne serait-ce que pour faciliter les requêtes:

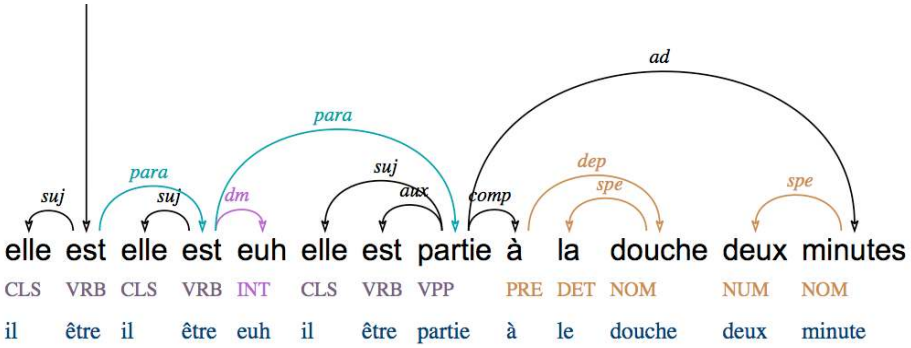


Figure 32. Amorces d'auxiliaires

3.5 Constructions microsyntaxiques particulières

3.5.1 Propositions relatives et interrogatives indirectes

Bien que le pronom relatif possède un double rôle de complémenteur et de pronom, seul son rôle de pronom au sein de la relative est pris en compte. En conséquence, la tête d'une relative est le verbe principal qui est lui-même *dep* de l'antécédent.

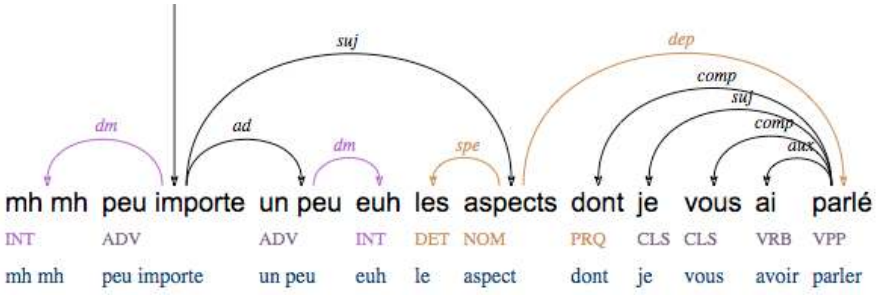


Figure 33. Relative

Il en va de même pour les interrogatives indirectes, mais ici le verbe de l'interrogative est dep du verbe de la principale.

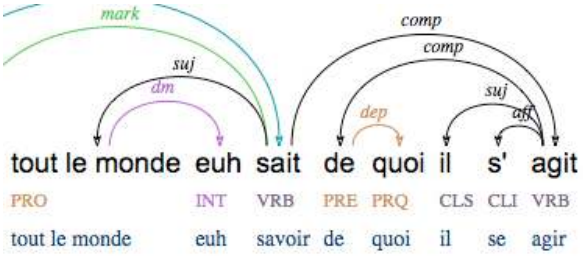


Figure 34. Interrogative indirecte

Même analyse pour les relatives sans antécédent.

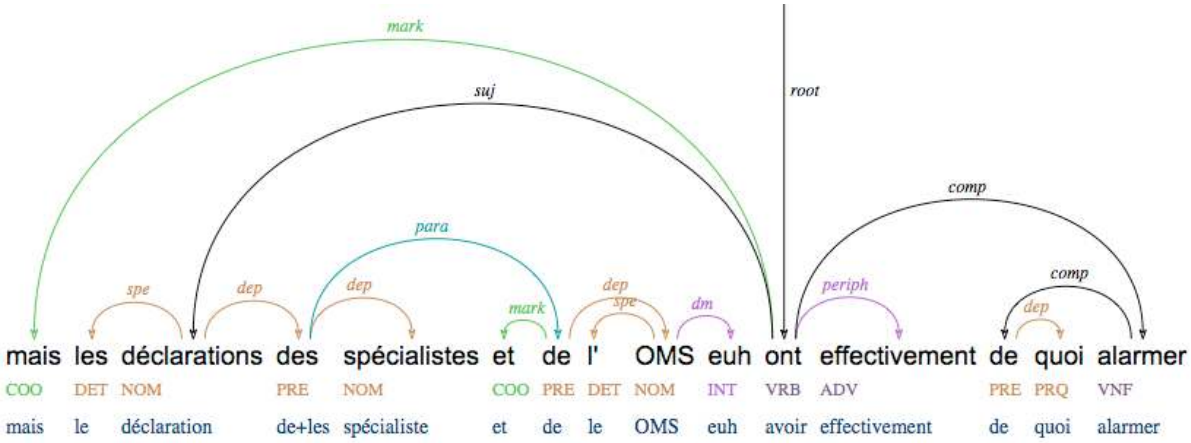


Figure 35. Relative sans antécédent

NB les relatives « de libre choix :

On analyse les séquences comme : quel qu'il soit, où qu'il soit, de quelque manière qu'il agisse, comme des relatives où le premier PRQ est antécédent ou contenu dans l'antécédent , le second PRQ (qu') dépendant du verbe:

Quel - dep→ soit

Soit ---dep-→ qu'

La relative et son antécédent sont PERIPH d'un VRB du contexte

Quelle (PERIPH de agirons) que soit sa réponse nous agirons

NB. *qui que ce soit*, *quoi que ce soit* sont traités comme des PRO composés dans le lexique *où que ce soit*, comme ADV composé . Ces formes se comportent en effet comme des lexèmes uniques:

Je ne parlerai pas à qui que ce soit

Le à dépend de parler comme dans *à n'importe qui.*

3.5.2 Constructions clivées

Pour chaque construction clivée qui possède la forme *c'est X qui* ou *il y a X qui*, la proposition subordonnée dépendra de X. Aucune différence n'est faite entre une construction clivée et une construction "relative présentative". Par conséquent, *c'est un ami qui m'a aidé* et *c'est l'ami qui m'a aidé* seront analysés de façon identique. (La raison en est qu'il ne nous semble pas possible pour un analyseur automatique de discriminer entre les deux situations sans indices prosodiques et pragmatiques.). Cette analyse vaut pour l'objet direct clivé également.

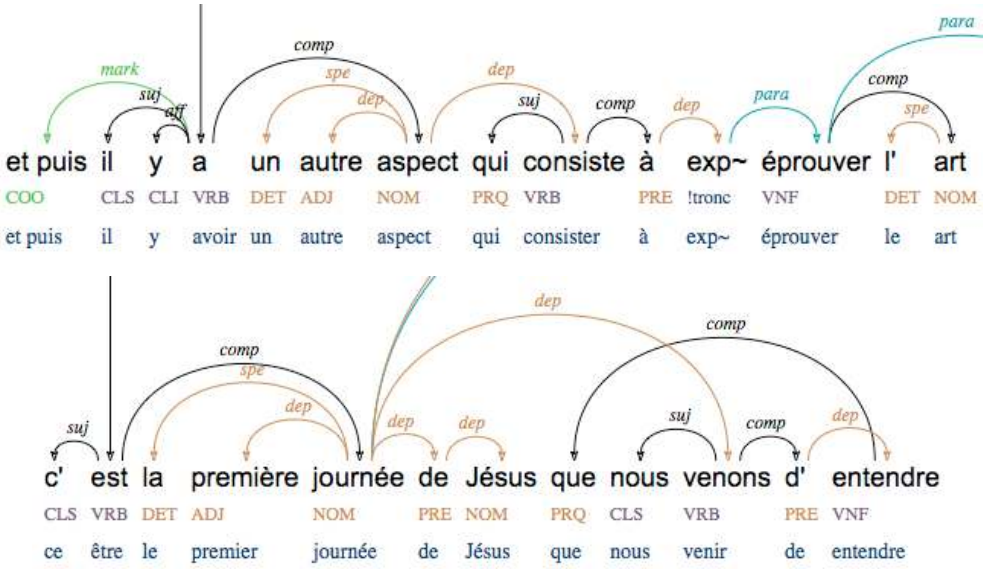


Figure 36. Clivage d'un sujet ou d'un objet

Construction clivée avec syntagme prépositionnel

Lorsque les propositions clivées présentent un syntagme prépositionnel dans la proposition principale, la proposition subordonnée, qui n'a plus la forme d'une relative standard, est alors *comp* du verbe de la proposition principale et *que* est analysé par convention CSU :

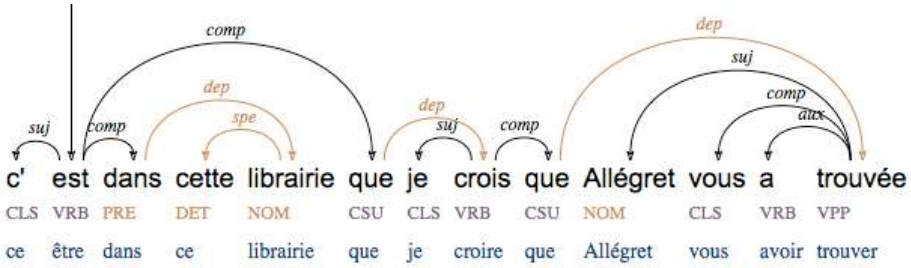
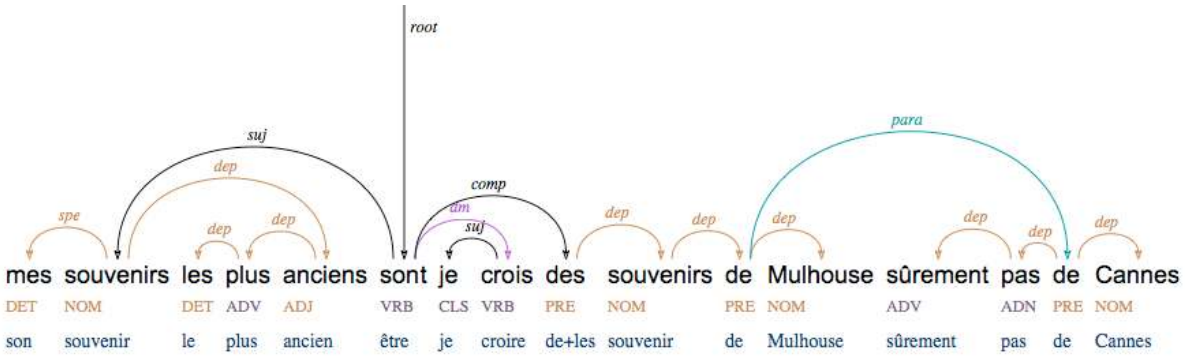


Figure 37. Clivage d'un groupe prépositionnel

3.5.3 Négations averbales

Lorsque la négation *pas* forme un syntagme avec une tête non verbale, il dépend de cette tête :



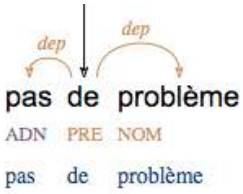


Figure 38. *Négation averbale*

3.5.4 Adverbes dans des entassements paradigmatiques

Les adverbes sont normalement dépendants d'un verbe. Il est néanmoins courant que des adverbes apparaissent dans des entassements paradigmatiques, où ils forment un syntagmes avec les conjoints. Dans ce cas l'adverbe sera marqué comme un dépendant de la tête du conjoint. On aura ainsi, selon la position de l'adverbe et du syntagme sur lequel il « porte », deux analyses possibles :

- S'il n'y a pas d'entassement paradigmatique, l'adverbe dépend du verbe : il est -*dep*→ surtout ennuyeux.
- Si l'adverbe forme un syntagme avec un conjoint, il dépend de ce conjoint : il est triste et surtout ←*dep*- ennuyeux.

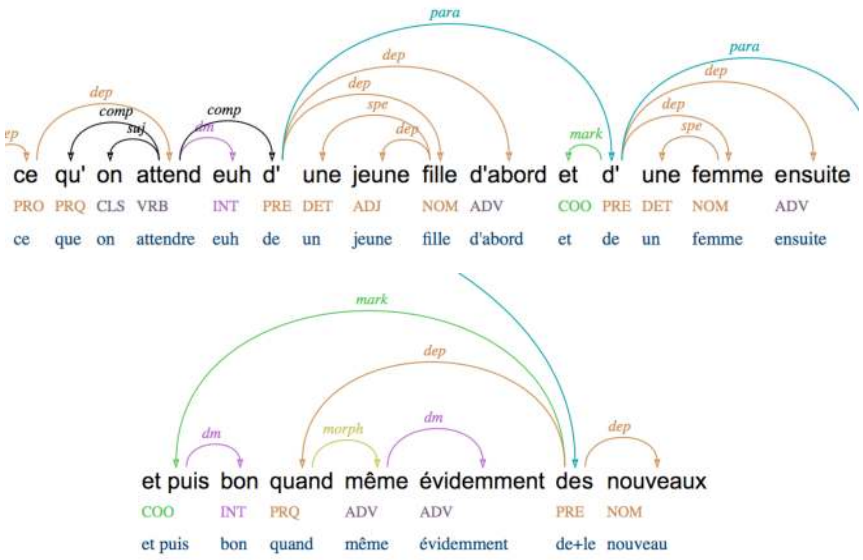


Figure 39. Adverbes dans entassements paradigmatiques

Voir section 5.2 pour le traitement de *évidemment* comme marqueurs de discours dans l'exemple précédent.

3.5.5 Adv *de* N

Une construction de la forme *J'ai mangé trop de sushis* ou *trop de sushis sont vendus sans label* est analysée de la même manière que *J'ai mangé des tonnes de sushis* : *trop* = tête adverbiale.

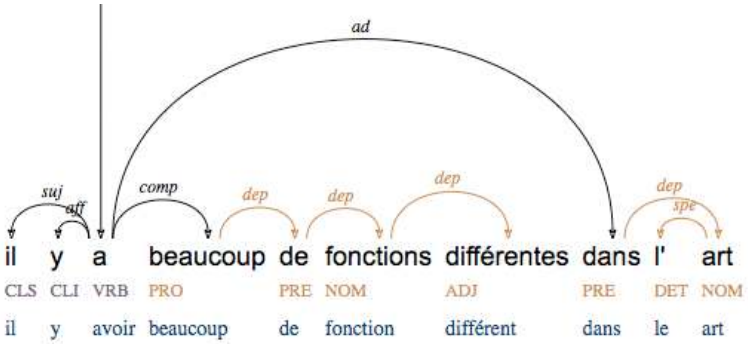


Figure 40. Adv *de N*

Ceci concerne *trop de N, peu de N, beaucoup de N, tant de N, combien de N, plein de N,* etc.

Par contre, les constructions de la forme *J'ai trop mangé de sushis*, où *trop* n'est pas contigu à *de N*, est analysé de la même manière que *Je n'ai pas mangé de sushi*.

3.5.6 *Que* + S et *Comme* + S

Dans cette construction, *que* et *comme* sont considérés comme PRQ et dépendent du verbe :

oh que c'est moche : que ←*dep*- est

Mais *que* est ADV et *root* dans *que de N* :

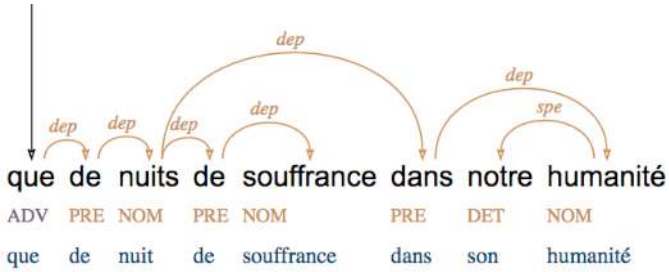
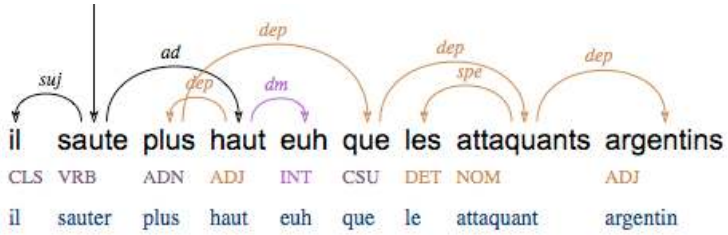


Figure 41. *que de N*

3.5.7 comparatives (*Plus* + ADJ + *que* , *plus* + ADV + *que) et consécutives*

Le complément du comparatif en *que* dépend bien du comparatif :



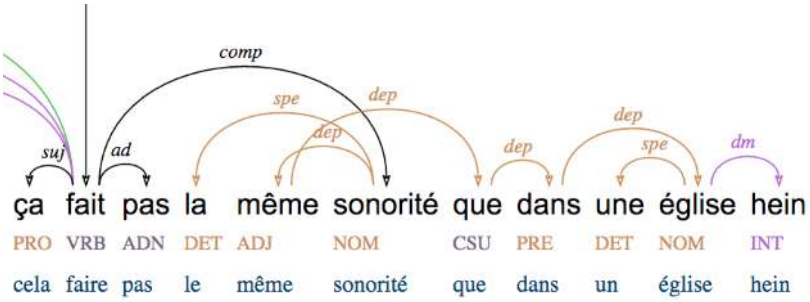


Figure 42. Comparatives

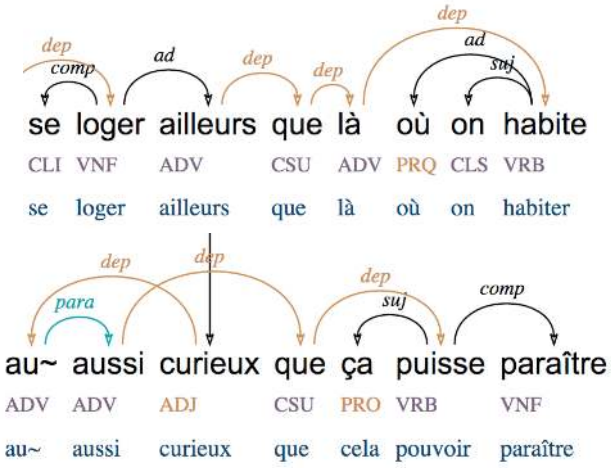


Figure 43. Consécutives

3.5.8 Greffes

Les greffes sont des propositions qui viennent occuper une place où un syntagme d'une autre catégorie est attendu (*je vais prendre je crois que c'est l'avenue AL* au lieu de *je vais prendre l'avenue AL*), le verbe greffé est considéré comme *dep* du verbe hôte, malgré la rupture de sous-catégorisation :

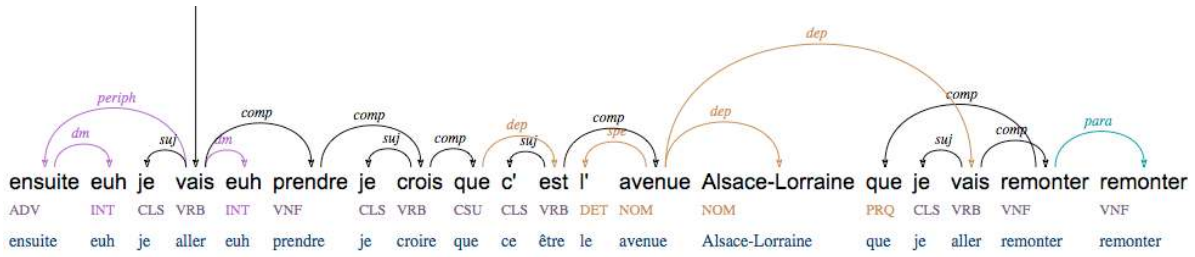


Figure 44. Greffe

3.5.9 *l'un l'autre*

l'un est traité comme un dépendant dans cette construction :

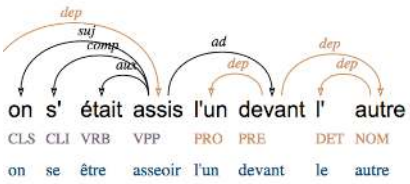


Figure 45. *l'un* PRE *l'autre*

4. listes (ou entassements) paradigmatiques : coordinations, énumérations, reformulation, apposition, disfluences

Le terme *liste paradigmatique* rassemble les configurations de termes unis par des liens paradigmatiques (conjoints occupant une même fonction syntaxique par rapport à une tête) . Les étiquettes *para* et *mark* sont spécialement conçues pour gérer les listes.

Para entre les têtes des séquences en liste *(cf 4.1)*

Mark entre la tête d'un terme de la liste et une éventuelle conjonction de coordination (4.2)

Les listes regroupent les phénomènes suivants :

- la coordination et les énumérations : *Hier, j'ai mangé avec Pierrot, Paulo, et Jacquot ; Pour faire de bonnes crêpes, il faut de la farine, du lait, du sucre, et caetera. ; Les gens ont peur des souris, des écrans, des trucs comme ça.*

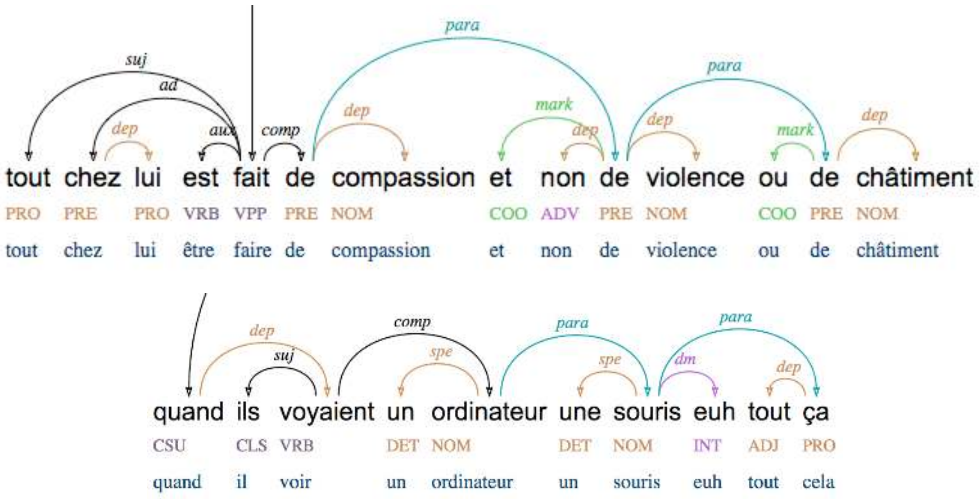


Figure 46. Coordination

- l'intensification : *Il était vraiment très très drôle.*

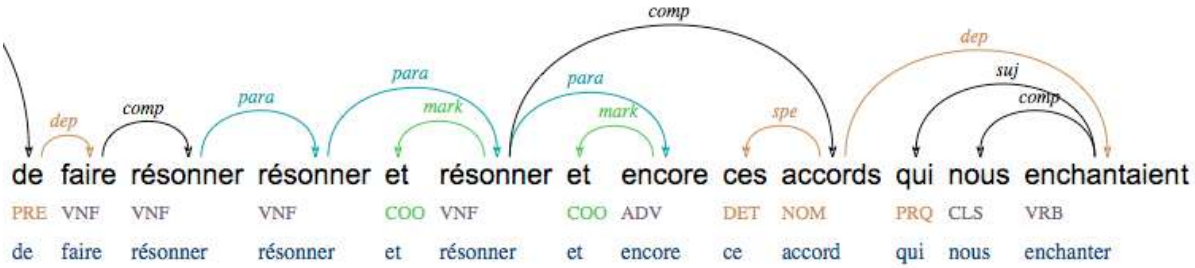


Figure 47. Intensification

- la disfluence : *C'était très ch~ très chouette ! ; C'est du de la framboise.*

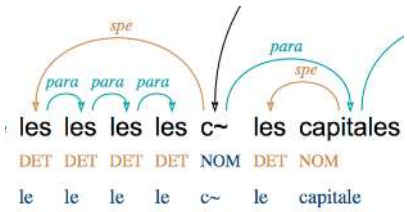


Figure 48. Disfluence

- la reformulation : *J'ai acheté un lit, un petit lit rose, pour ma fille*

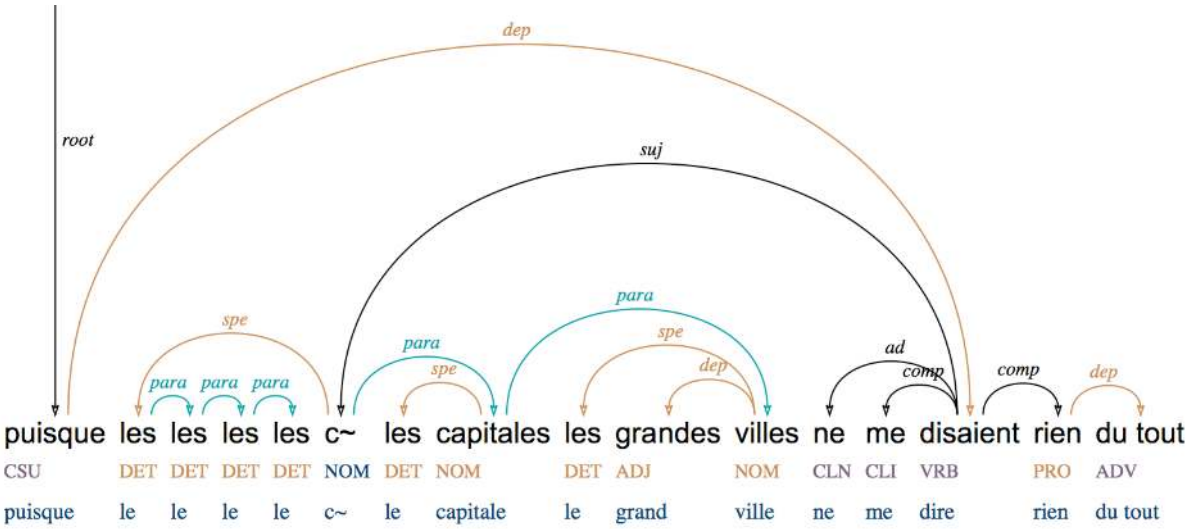


Figure 49. Reformulation

- la double formulation ou effet deux points : *J'ai acheté quelque chose d'original : un bouquin sur les axolotls ; Elle m'a fait un beau cadeau : un chèque en bois ; Un de mes amis, un gars que j'ai rencontré à la fac, m'a aidé.*

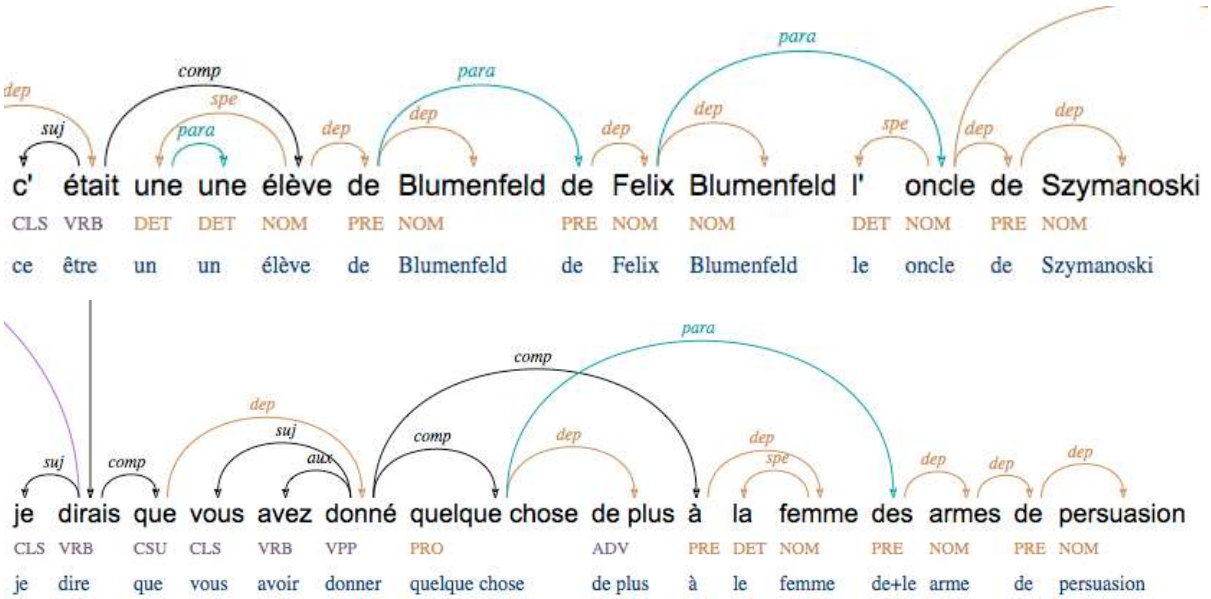


Figure 50. Double formulation

- question-réponse : *il vient quand ? demain* (néanmoins les questions-réponses sont généralement segmentées en deux énoncés et la tête de la réponse est alors *root*), en particulier quand elles sont sur deux tour de paroles. Quand elle ne le sont pas, elles donnent une liste paradigmatique :

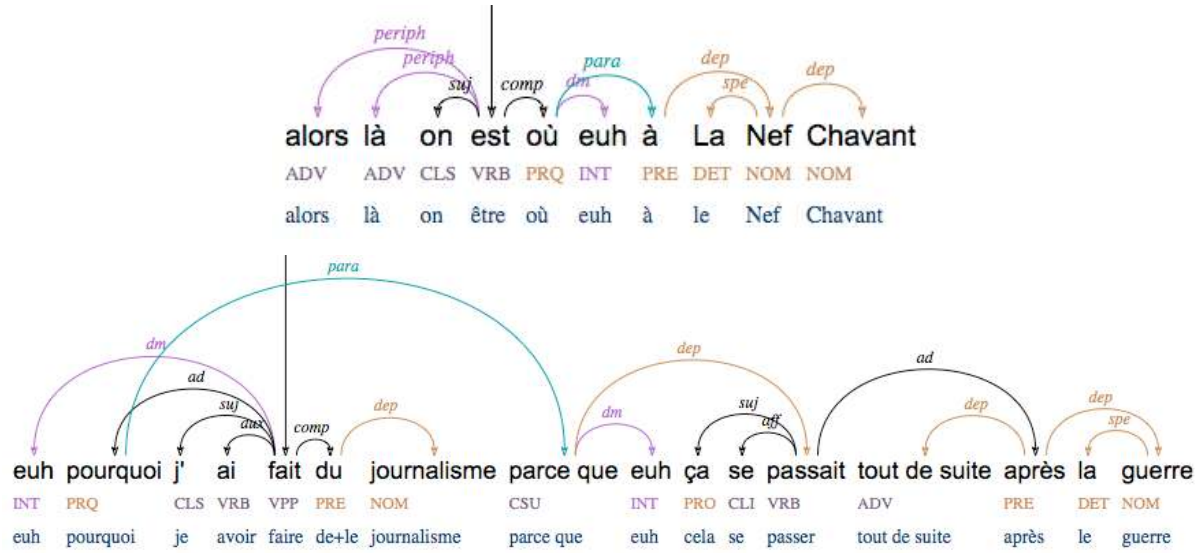


Figure 51. Question-réponse

4.1 Lien paradigmatique : *para*

para représente un lien paradigmatique qui rattache un élément à son conjoint le plus proche au sein des listes. Le premier conjoint d'une liste est la tête (et le gouverneur de la liste s'y rattache). Un dépendant commun se rattache au conjoint le plus proche. Les liens paradigmatiques vont toujours de gauche à droite.

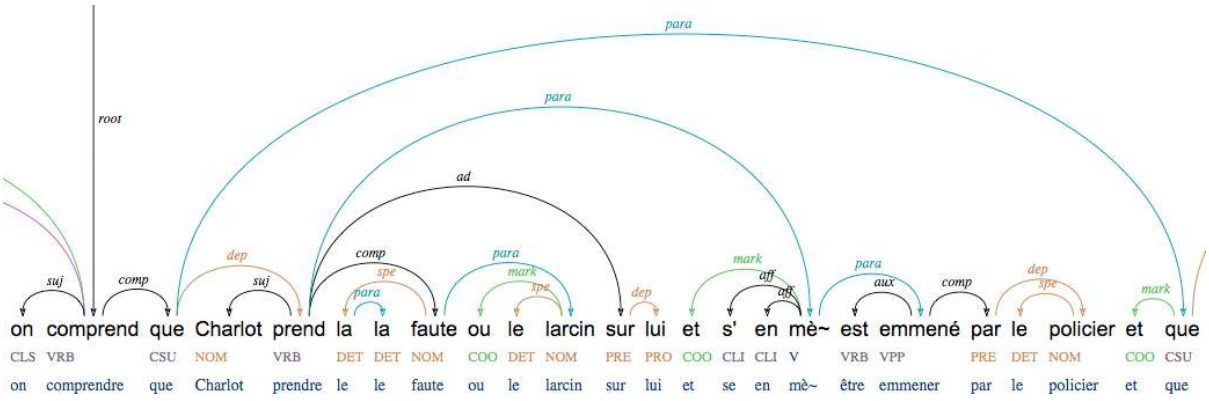


Figure 52. Liens *para*

Par défaut, il n'y a pas de lien *para* entre des verbes principaux, même lorsque ceux-ci sont dans un discours direct (voir guide de segmentation). Par contre, deux verbes principaux qui partagent un dépendant seront liés par un lien *para*. L'exemple suivant illustre les deux cas :

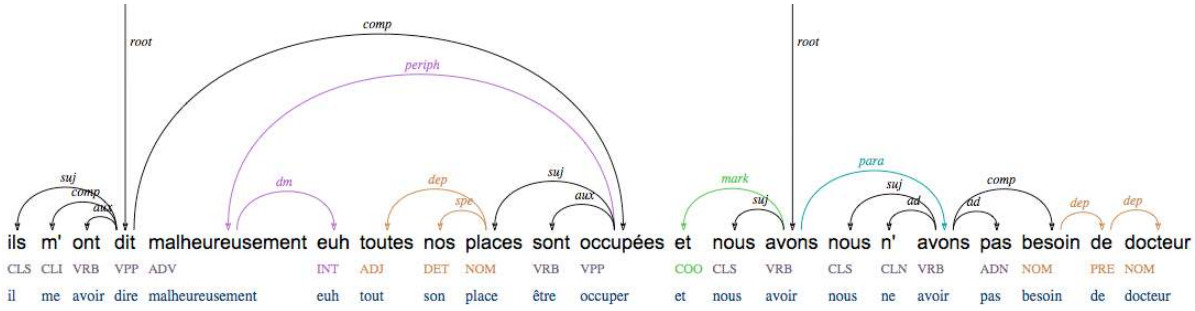


Figure 53. Liens *para* entre verbes principaux

On utilisera aussi un lien *para* pour la construction *de X à Y*, où on a une forme de coordination (ordre fixe *à Y de X) : *de trois à quatre personnes, du début à la fin, le train de Paris à Marseille*.

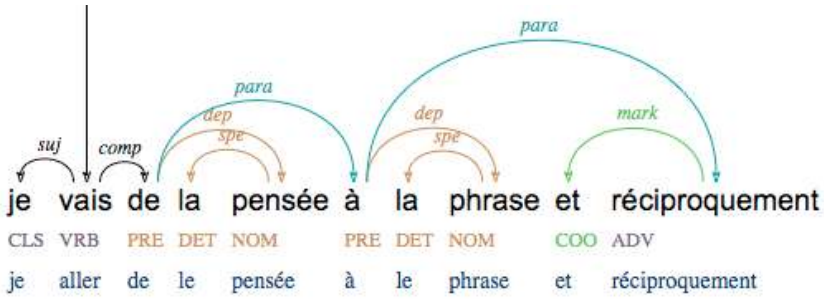


Figure 54. *de X à Y*

4.2 Lien marqueur : *mark*

Les conjonctions de coordination (COO) sont analysées comme dépendant du conjoint qui suit par un lien *mark*. Cette analyse permet de privilégier le lien *para* entre les deux conjoints et de rendre compte de l'asymétrie de la construction (la conjonction forme un syntagme avec le conjoint qui suit et pas celui qui précède).

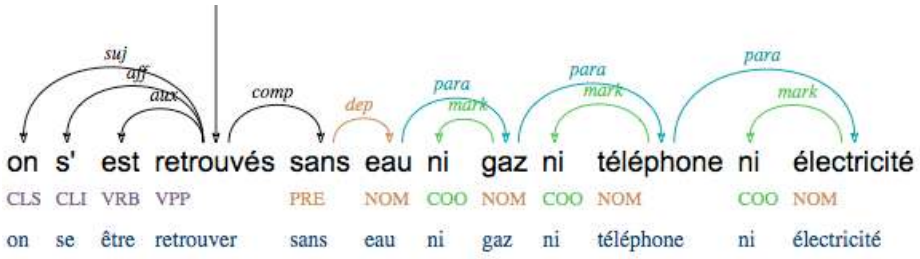


Figure 55. *mark*

Les COO en début d'énoncé dépendent de la racine par un lien *mark*.

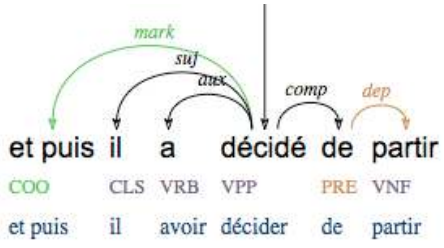


Figure 56. CCO introducteur

Les épexégèses (ou compléments différés) peuvent être introduits pas une conjonction de coordination sans qu'il y ait alors de lien *para* :

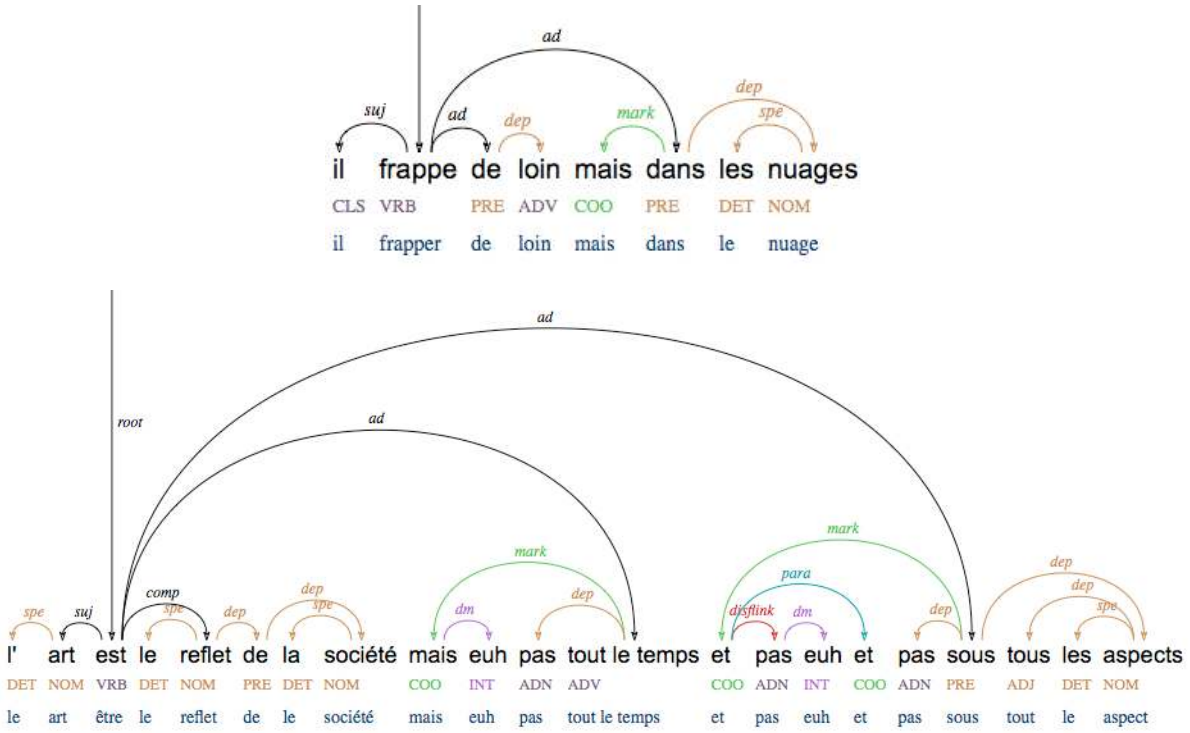
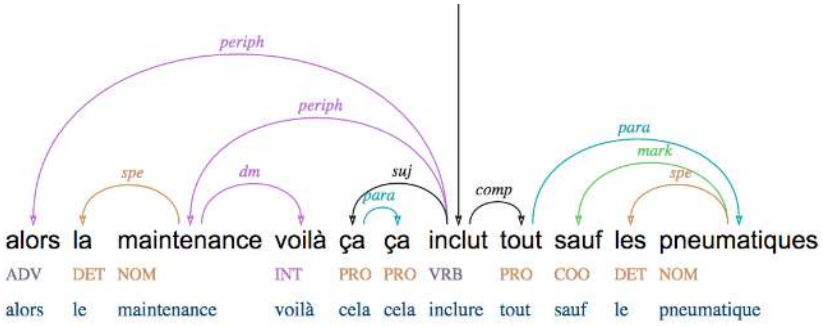


Figure 57. Epexégèse

Les exceptives sont traitées comme des cas de listes paradigmatiques, avec *sauf, excepté, hormis* ou *à part* comme COO et *mark*. Ce traitement est justifié par le fait que les éléments comme *sauf* peuvent être suivi de syntagmes que n'autorisent pas les PRE (*sauf à Paris*, *hormis quand il pleut*) et qu'ils ne sont jamais précédé de *et*. (Par contre *sauf que* n'a rien à voir au niveau syntaxique et est traité comme une CSU figée.) Il s'agit souvent d'épexégèses, sans lien *para* (2e exemple ci-dessous).



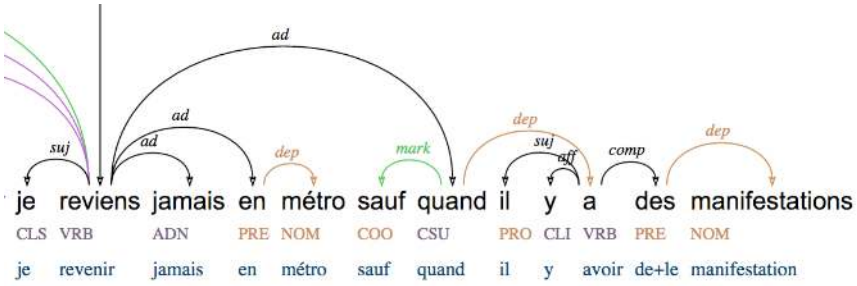


Figure 58. Les exceptives (*sauf, ...*)

Certains adverbes paradigmatisants se comportent de manière proche des COO, mais le fait qu'il puisse cooccurer avec *et* ne permet pas d'en faire des *mark*. C'est le cas de *y compris* :

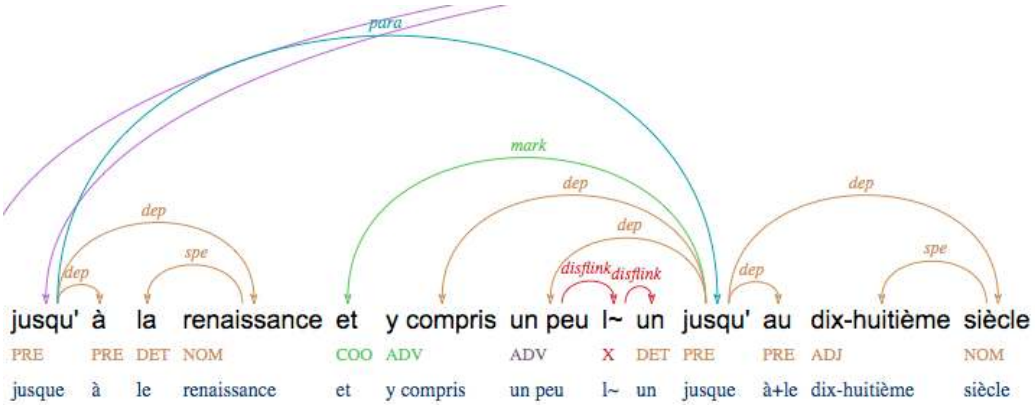


Figure 59. L'adverbe *y compris*

5. Macrosyntaxe

L'analyse en macrosyntaxe prend en compte le rattachement des éléments non régis ainsi que, par convention dans ce guide, les éléments en position détachée, régis ou pas. Afin de pouvoir les analyser comme il convient, nous utilisons les relations *periph*, *dm*, *insert* et *parenth*.

5.1 Eléments périphériques : *periph*

La relation *periph* relie les éléments périphériques, en position détachée, à la tête de l'énoncé (c'est-à-dire l'élément *root*). Les constituants à gauche du sujet seront systématiquement traités comme *periph*, même lorsqu'on pourrait considérer qu'ils sont dans la valence du verbe. Cela est également vrai pour des constituants dépendant d'un verbe, qui sont à la périphérie de la construction régie par ce verbe sans être à la périphérie de l'énoncé. Il est beaucoup plus complexe de repérer les *periph* lorsqu'ils se trouvent à droite du noyau, sauf avec certains lexèmes (puisque, de sorte que, adverbes comme *heureusement, franchement*…) et pour les cas de dislocation avec reprise par clitique *(comment ça marche , les autres fig. 52*). C'est pour cela qu'en cas d'ambiguïté, il est plus sûr d'utiliser la fonction *dep*.

Certaines constructions verbales sans introducteur qui ne peuvent fonctionner seules et jouent le rôle de présentatifs d'un thème (*il y a N, j'ai N)* sont aussi annotées *periph* (dernier exemple de la figure52)

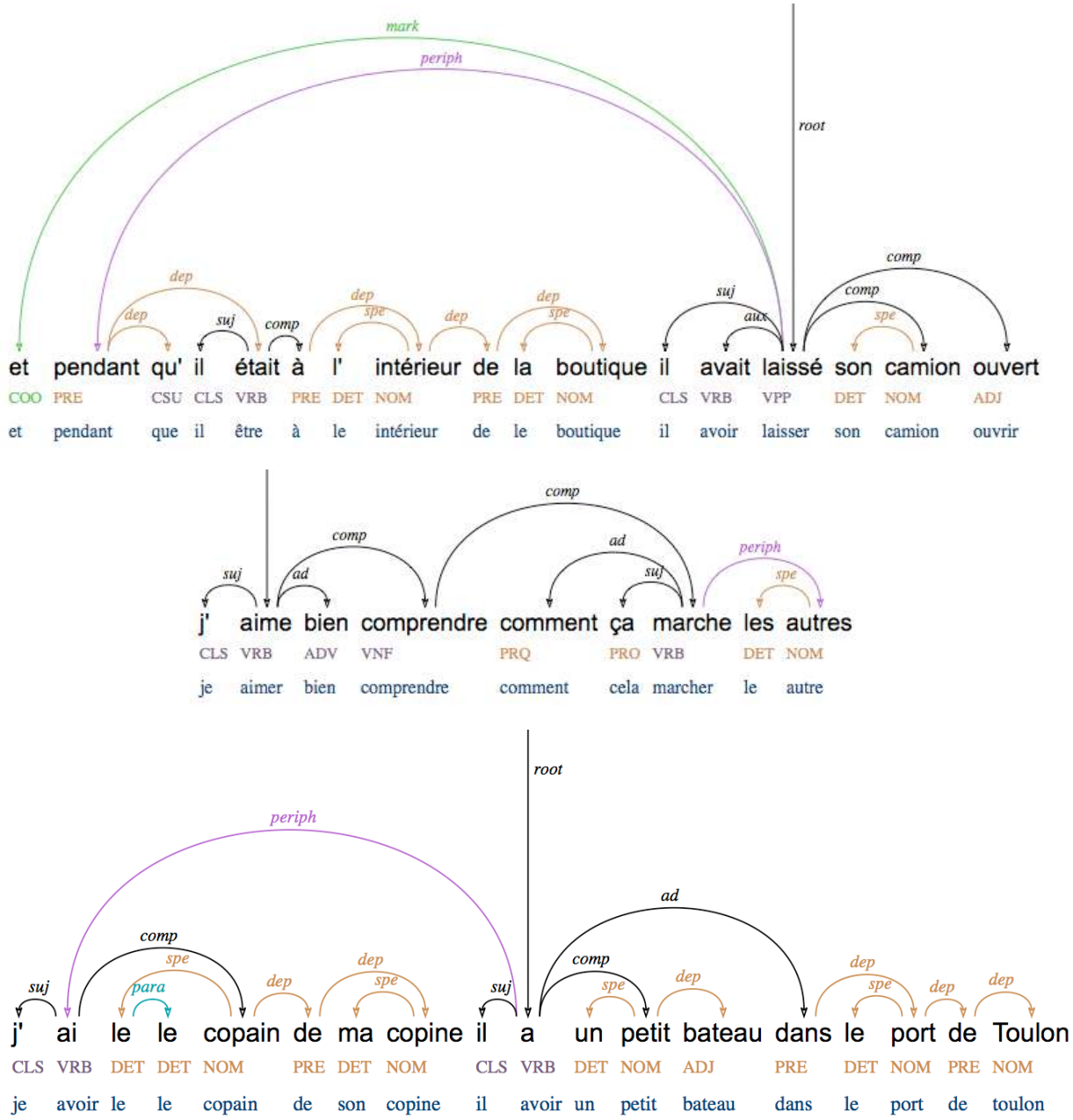


Figure 60. *periph*

- Les pronoms interrogatifs ou relatifs (qui peuvent se trouver facilement à gauche du sujet) ne sont jamais *periph* : *A qui cette personne peut-elle s'adresser ?*
- Sauf lorsqu'ils portent sur un énoncé entier ou sont détachés et commutent avec d'autres périph : *d'ou, sur quoi, après quoi* : *il refusa de répondre après quoi/ça il se fit un grand silence*

- En cas d'inversion du sujet (*De ceci découle cela*, *Au plafond pendaient des lanternes*, *Sur la place se dresse une cathédrale*), l'élément se trouvant immédiatement avant le verbe est un *dep* en général :

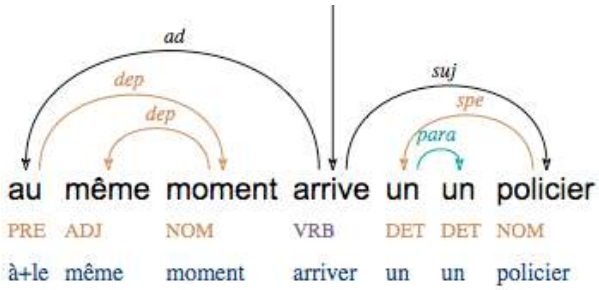


Figure 61. Inversion du sujet

Les adverbes (y compris les dits « adverbes de phrase ») sont *dep* quand ils sont intégrés au noyau et *periph* quand ils sont en périphérie du noyau à gauche:

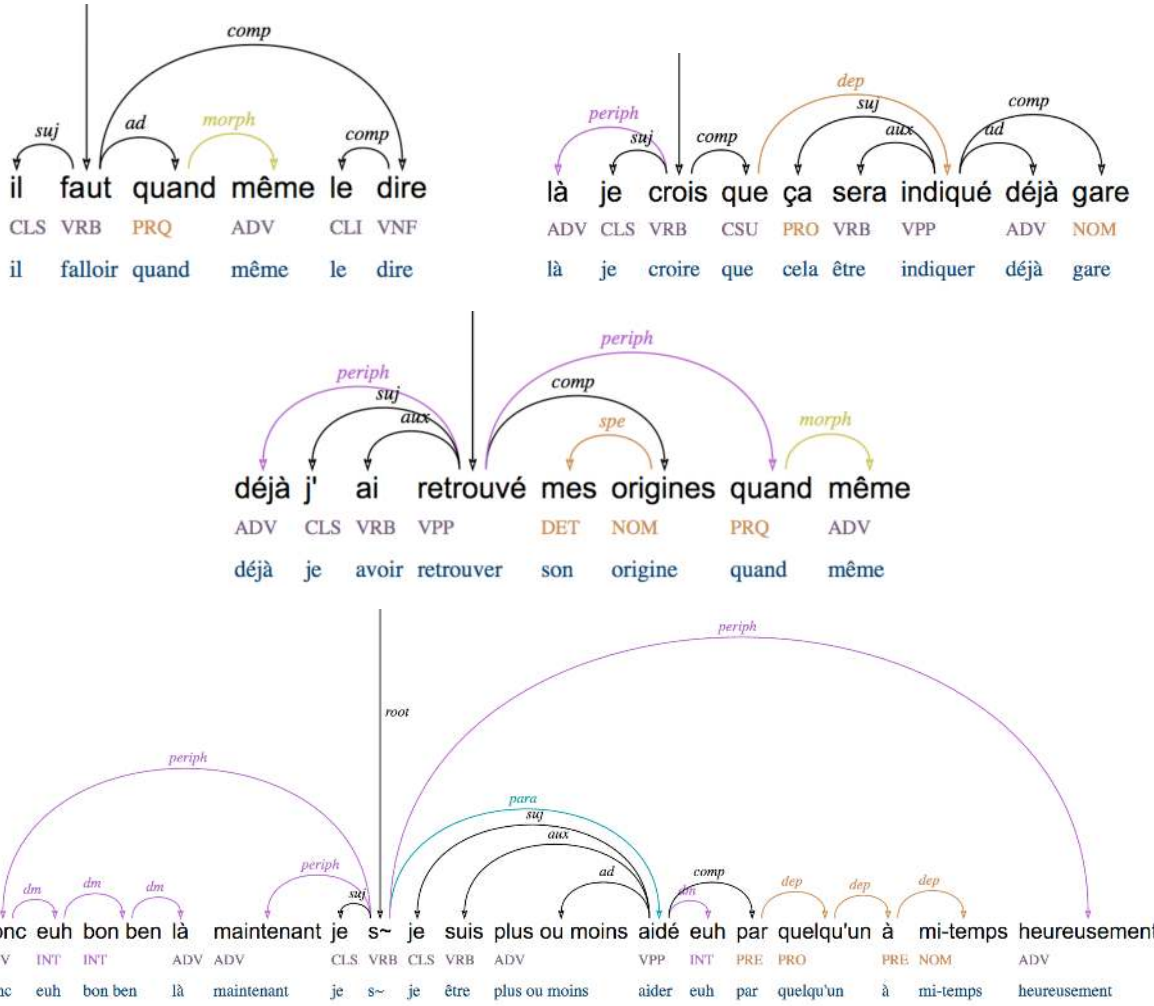


Figure 62. Adverbes *ad* vs *periph*

Ils sont également analysés (sauf les PRQ) comme *periph* quand ils sont devant le sujet d'un VRB dépendant (en subordonnée) :

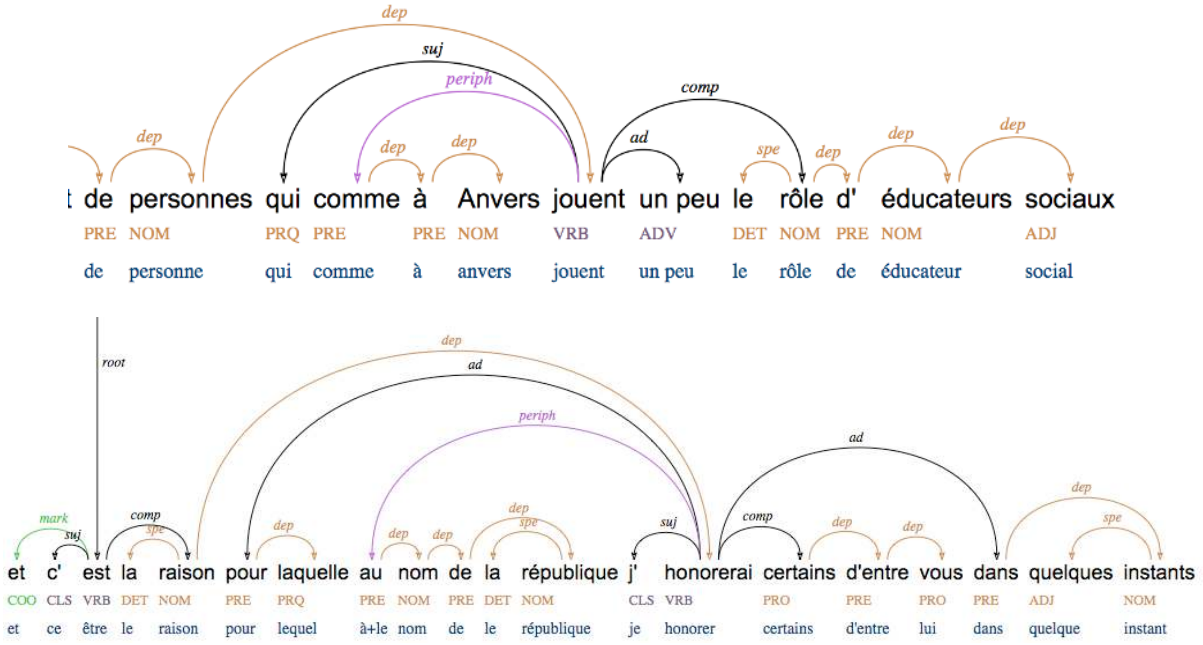


Figure 63. *periph* dans une « subordonnée »

Les adverbes qui apparaissent dans des listes paradigmatiques dépendent du conjoint (voir section 3.5.4). Certains adverbes sont traités comme des marqueurs de discours (voir section 5.2 qui suit). Les autres sont *dep*, même ceux qui sont détachés et pourraient être analysés comme *periph* (cf. *jusqu'à présent* ci-après) :

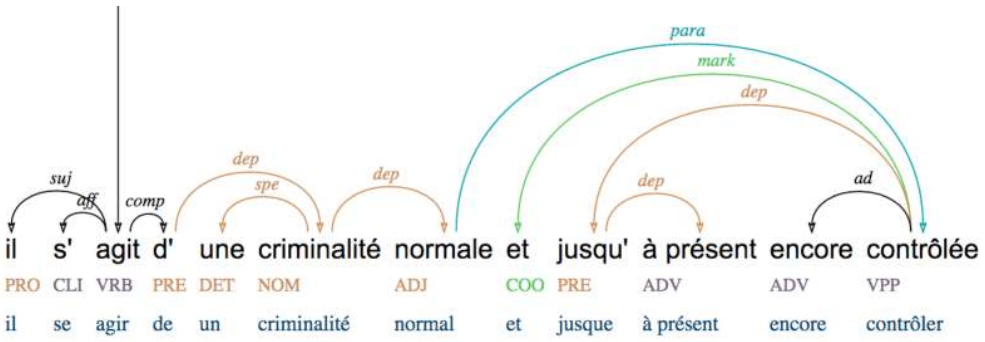


Figure 64. pas de *periph* dans un entassement

On évite les *periph* de *periph* même quand il y aurait de bonnes raisons de le faire. Par exemple, *moi mon vélo le guidon il est cassé* est analysé avec 3 *periph* dépendant du noyau. Idem dans le cas suivant :

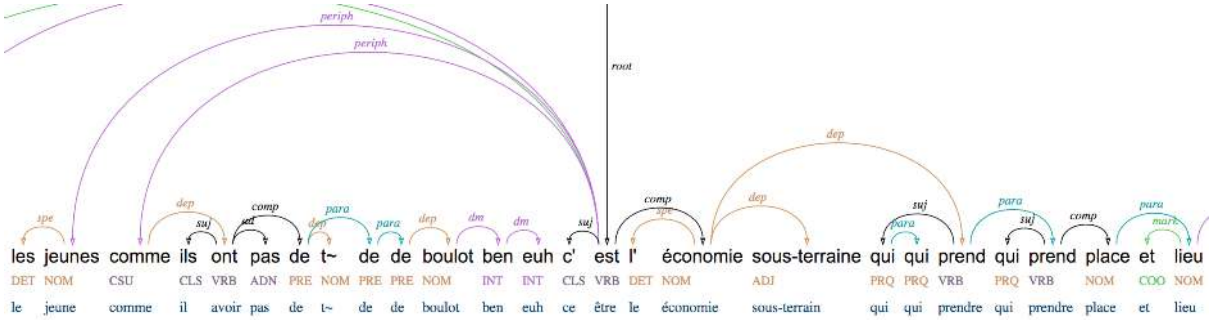


Figure 65. Double *periph*

5.2 Marqueurs de discours : *dm*

Les marqueurs de discours sont des éléments plus flottants que les *periph*. Un élément *dm* est rattaché à l'élément qui le précède directement, ou à la racine s'il est en position initiale.

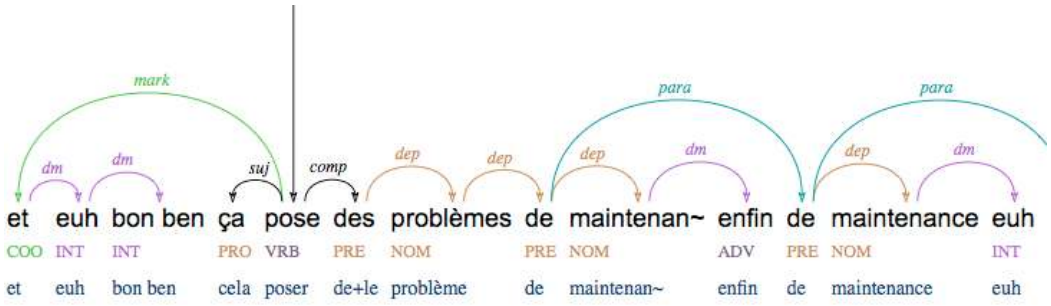


Figure 66. *dm*

Seule exception : lorsque le *dm* est en tête d'un discours direct, il est rattaché au verbe principal du discours direct et pas à l'élément qui précède.

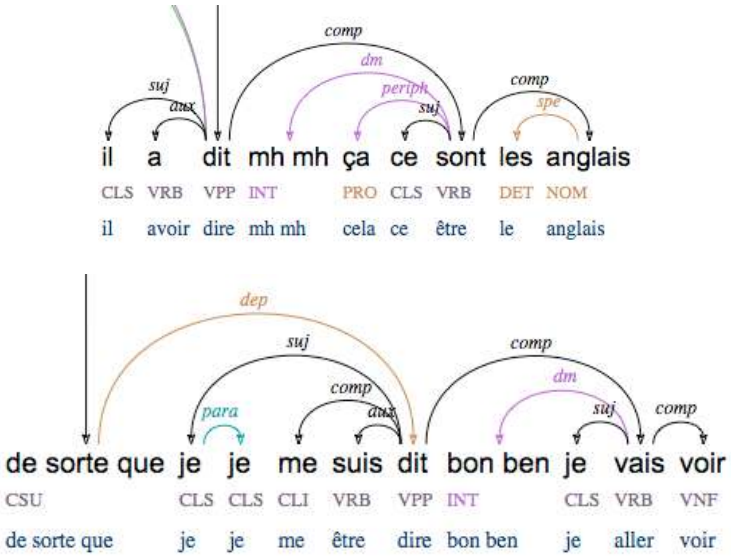


Figure 67. *dm* et discours direct

<u>Liste des DM</u>

La liste des DM extraite depuis les fichiers Rhapsodie :

Interjection*s : euh, , bon, hein, bah, enfin, mh mh, voilà , oui, non, ben, peuh, eh, ah , eh ben, eh bien, oh(là) oh la la, ouh, et oui, ouh la la la la,, waouh, eh oui, ah bon, ouais, bref, pff, quoi, non mais, fff, OK, en tout cas, attention,*

Incises verbales « sans complément *» : je dirais, je veux dire , on dit, on va dire, je dois dire, disons, je te dis , je me disais, on dirait, si je puis dire, c'est-à-dire, je cite, , je vous signale*

je sais, tu sais, je sais pas, vous savez,

je vois, vous voyez, voyez-vous, tu vois , tu as vu, voyez, tu verras , vous verrez, vous allez voir

il me semble, je crois , je pense, je trouve, j'imagine, tu imagines , je me souviens, si vous voulez, si tu veux, allez, remarque, remarquez, écoute, écoutez,, attends, attendez, ça y est, pardonnez-moi, pardon, n'est-ce pas, excusez-moi,

Certains adverbes comme *alors* ou *donc, en fait, enfin (fig 61)*, et *en tout cas* sont normalement des *dep* ou des *periph*, mais certains locuteurs en font des tics de langage et les utilisent comme des phatiques ponctuant la plupart de leurs énoncés. Dans ce cas, on peut être amenés à en faire des *dm*, mais la frontière est difficile à tracer.

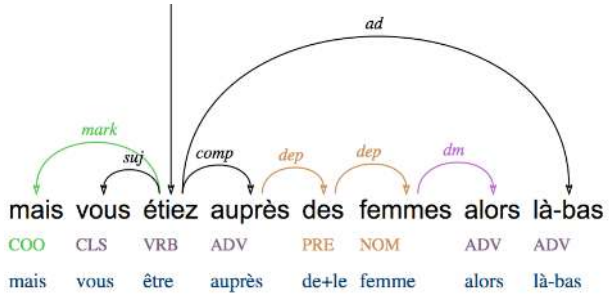


Figure 68. *alors* comme *dm*

Certains adverbes en *-ment* comme *finalement* ou *évidemment* pourrait être analysés comme des *dm* dans certains cas, mais pour éviter les ambiguïtés d'annotation nous les traiterons systématiquement comme des *dep* ou *periph*.

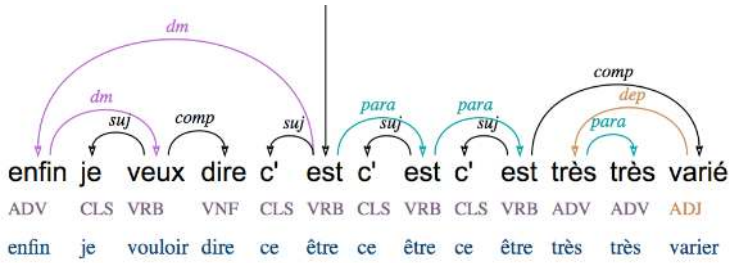


Figure 69. *enfin* comme *dm*

5.3 Incises

Les incises, comme *répéta-t-il* ou *dit le diable*, sont étiquetées *insert* (angl. pour incise) et attachées à la racine du segment d'énoncé qui précède (et pas au mot précédent comme les *dm* et les *parenth*). Les incises se distinguent :

- des *dm* car elles ne sont pas figées et acceptent des modifieurs ;
- des *parenth* (parenthétiques) car ne sont pas saturées et prennent l'énoncé sur lequel elle s'appuie comme « objet » du verbe de *dire*. De ce point de vue elles se situent entre micro et macrosyntaxe.
- des *dm* et des *parenth* car le sujet est inversé.

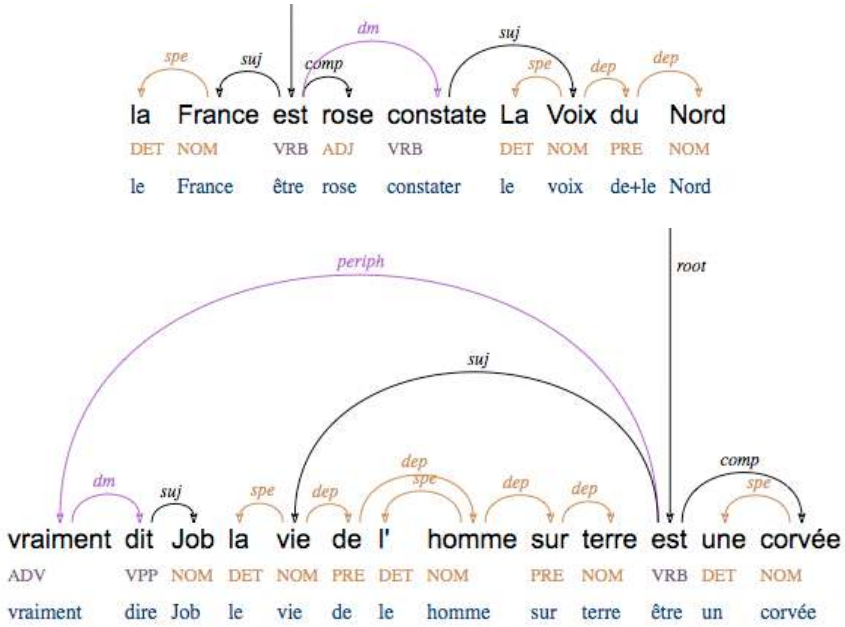


Figure 70. Incise de discours rapporté

Les constructions verbales saturées (et sans inversion du sujet) comme *vous l'avez vu* ou *il a dit* seront traitées comme parenthèses :

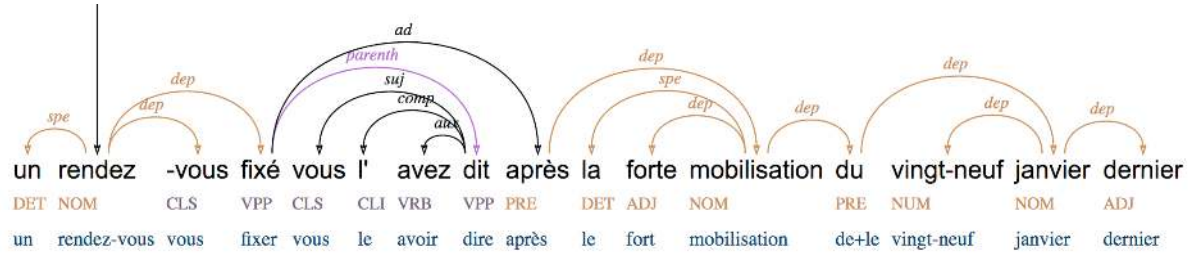
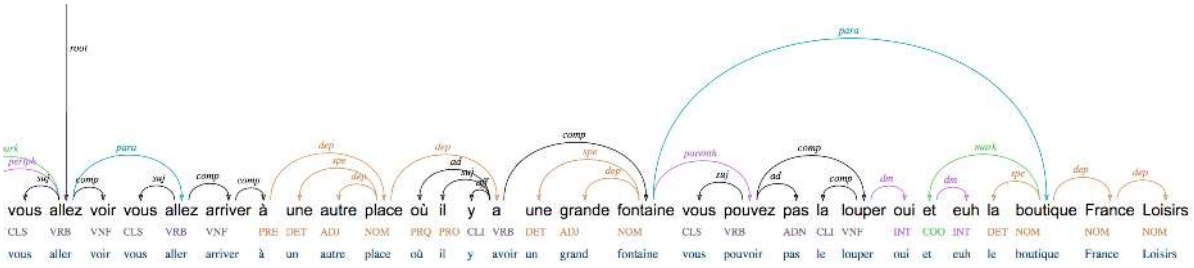


Figure 71. Incises *vs* parenthèses

5.4 Parenthèses

Les parenthèses sont explicitement identifiées et analysées *parenth*. Le gouverneur de la parenthétique est, comme pour les *dm*, le mot qui précède.



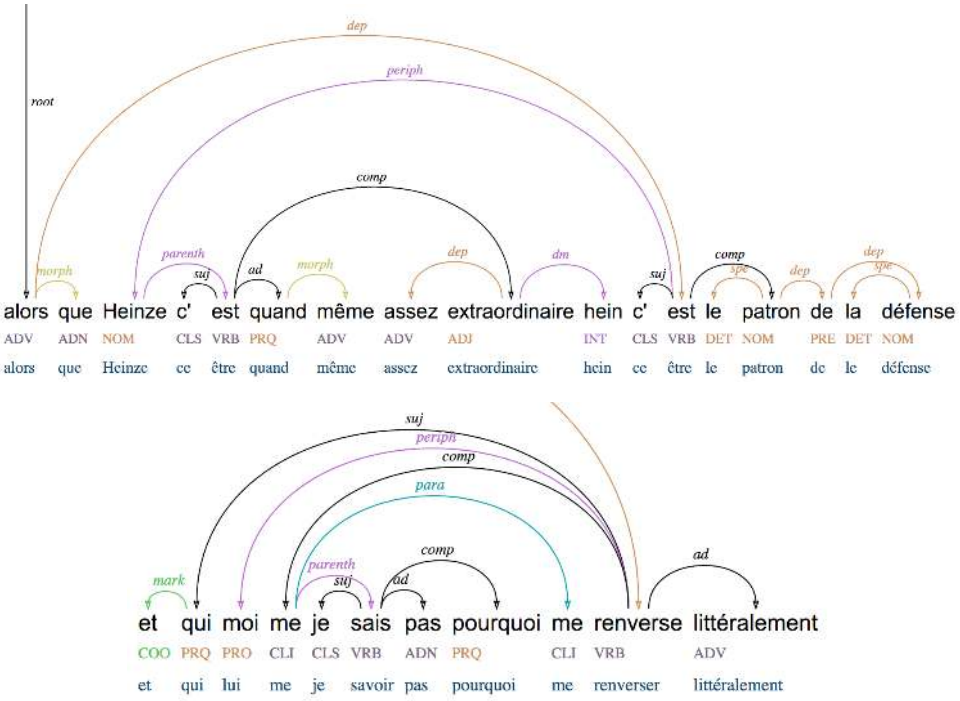
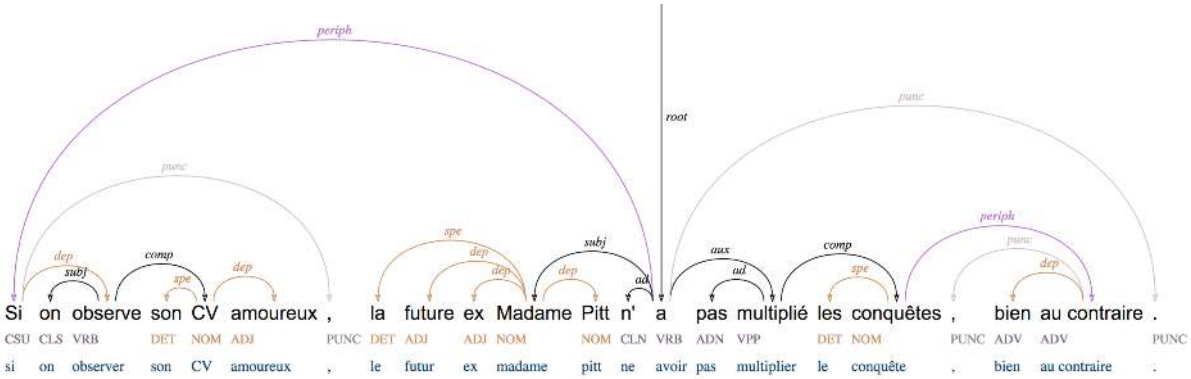


Figure 72. Parenthèses

6. Ponctuation

Les ponctuations à l'écrit forment des tokens séparés de catégorie PUNC et de fonction *punc*. Il faut distinguer deux types de ponctuation.

Ponctuation simple : il s'agit d'une ponctuation qui fonctionne seule et marque le début ou la fin d'un syntagme. Une telle ponctuation est toujours rattaché au dépendant de la relation qui la couvre. Dans l'exemple suivant, la prémière virgule est couverte par la dépendance *periph* entre *a* et *Si* : elle est dont rattaché au dépendant qui est *Si* et marque ainsi la limite droite du syntagme dont *Si* est la tête. La même chose s'observe avec la deuxième virgule : elle est également couverte par un lien *periph* et s'attache au dépendant *au contraire* de lien, marquant ainsi la limite gauche du syntagme dont au contraire est la tête. Enfin, la ponctuation finale qui n'est « couverte » que par le lien *root*, s'attache aussi au dépendant de ce lien, c'est-à-dire à la racine de l'arbre, ici *a*.



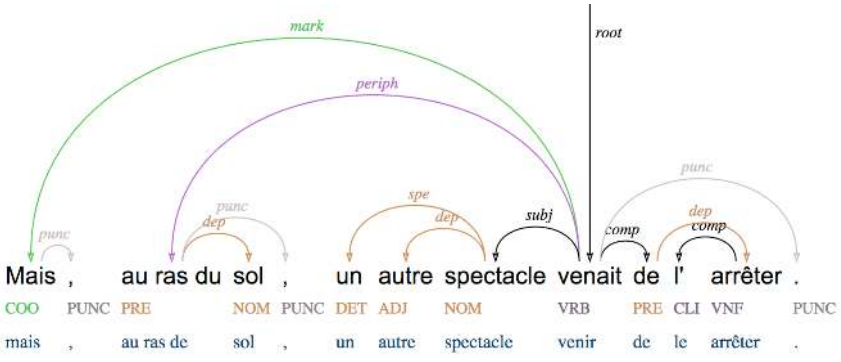


Figure 73. Ponctuation simple

Les virgules qui marque des listes paradigmatiques sont traitées de la même façon. Elles sont couvertes par un lien *para* et s'attache au dépendant de ce lien qui se trouve à leur droite ; elles marque ainsi la frontière gauche du conjoint qui les suit, comme le font aussi les COO.

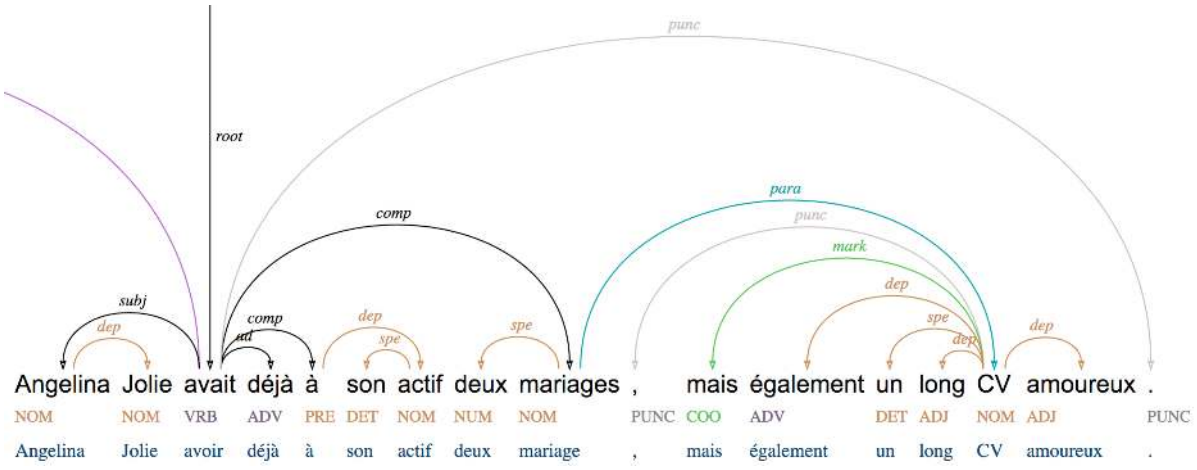
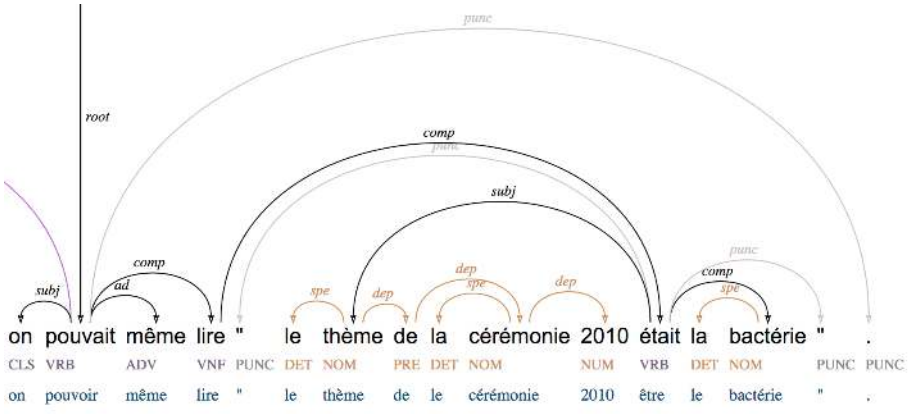


Figure 74. Ponctuation simple dans une liste paradigmatique

Ponctuation double : il s'agit d'une paire de ponctuations de même nature (deux virgules, deux guillemets, deux parenthèses, deux tirets) qui marque les frontières gauche et droite d'un même syntagme. Dans ce cas les deux ponctuations sont rattachées à la racine de ce syntagme



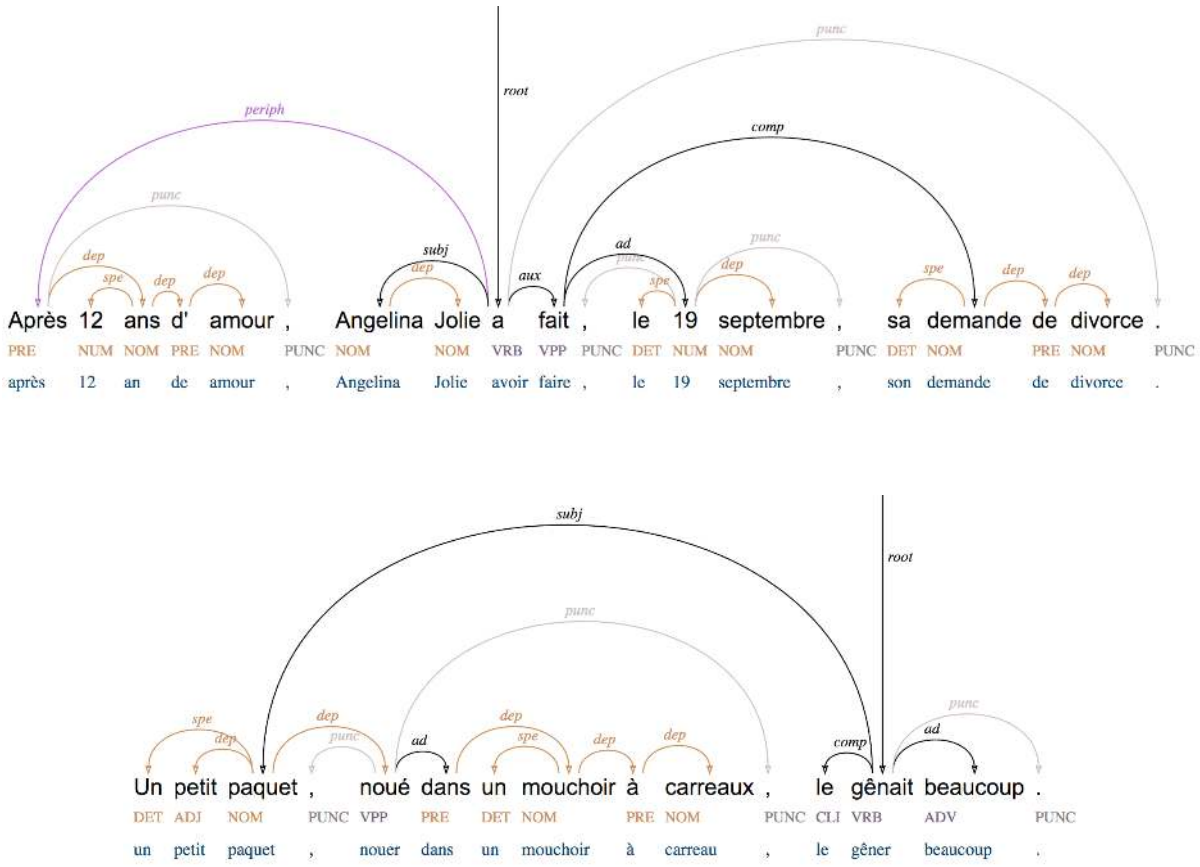


Figure 75. Ponctuations doubles