André Valli - | José Deulofeu ORFEO/TALEP

Définition des POS.
Principes

- On essaie de dissocier le plus possible catégorie et fonction. Ce qui réduit les cas de double classement. Par défaut, une catégorie n'est pas attachée à une fonction. Par exemple, les adjectifs, canoniquement compléments d'adjectifs, peuvent aussi être compléments de verbes, de sorte que dans *il chante faux*, *faux* reste un adjectif comme dans un *faux billet* et ne devient pas adverbe (On met l'invariabilité au compte de la relation d'accord avec le gouverneur verbal : le genre est MASC par défaut, puisque le verbe ne varie pas en genre).

- On ne distingue pas deux catégories Pronoms interrogatifs et relatifs. On pose une seule catégorie PRQ avec des fonctionnements syntaxiques différents.

Limite du principe : On a parfois deux choix possibles du point de vue linguistique. Ce sont alors les critères TAL qui permettent de décider (voir catégorie PRO). On choisit l'analyse qui facilite le plus le travail de l'analyseur. Ces cas seront étudiés dans chaque rubrique.

1. Catégories variables :
1.1. Pour les catégories « ouvertes » : verbes, noms adjectifs, on reste fidèle au lexique du LEFF. Nous avons choisi de ne pas construire de sous catégories en fonction de caractéristiques morphologiques : masculin /féminin, pluriel / singulier.

VERBE
On distingue les POS : VRB, (verbe fini ou conjugué), VPP (verbe au participe passé), VPA (verbe au participe présent)
Certains lexèmes se voient attribuer deux POS. La répartition obéit à des critères syntaxiques.
Exemple : classement en VPP ou ADJ pour le lexème *serré* :

ADJ quand la distribution est celle d'un adjectif :
Je veux un café très serré

VPP si distribution verbale :
La vis trop serrée par le garagiste n'a pas tenu

NOM
On ajoute au critère morphologique de variabilité le critère syntactico-sémantique qu'un lexème, par ailleurs classé ADJ peut être classé comme nom dans des cas où il est tête de SN avec sens indépendant du contexte.

On analyse donc *bleu* comme NOM dans :
Les Bleus, un bleu (de travail)

Mais comme ADJECTIF dans :
La boule blanche et la bleue (COO DET ADJ)

Discussion de cas particuliers :
question, rapport, côté sont doublement catégorisés : PRE et N, car dans *question poisson, je préfère le thon,* la distribution de *question poisson* est celle d'un groupe prépositionnel.
On garde la catégorie NOM pour le modifieur dans *un paquet cadeau un exemple limite/ type, un livre témoignage*.

ADJ (adjectif)
Deux décisions méritent d'être signalées. Des lexèmes classés traditionnellement comme déterminants ont été reversés dans la catégorie ADJ (voir la section DET). Au contraire, pour les néologismes comme *il est super, trop*, notre dictionnaire conserve le double classement en adverbe et adjectif du LEFF.

1.2. Catégories variables fermées

DET (déterminant)
Cette catégorie regroupant traditionnellement articles et déterminants a été restructurée en fonction d'analyses linguistiques récentes et de considérations TAL.
1) Pour les objets « directs » partitifs et indéfinis *: je lis des livres, je veux de la tarte* , nous avons choisi l'analyse qui s'appuyant sur l'équivalence *je lis des livres / j'en lis* conduit à décomposer systématiquement *du, des* en préposition + déterminant (de + le/ les/la) dans tous les cas : ce qui signifie qu'il n'y a pas pour nous d'article ou de déterminant partitif, ni d'article indéfini pluriel en tant que tel.
2) Pour les autres cas
Dans les approches courantes, la fonction DET recouvre la catégorie DET. On a choisi de créer, suivant des approches récentes, une fonction SPE distincte de la catégorie DET. La fonction SPE est attribuée aux lexèmes qui permettent de faire fonctionner un NOM comme sujet. (Pour la définition détaillée de la fonction voir le Tagset fonctions.)

On classe alors dans DET les lexèmes simples qui font fonctionner un N comme sujet et qui ne sont pas combinables entre eux :
le, ce, mon, un, certains, chaque, plusieurs, quelque, zéro, aucun, tout, n'importe quel, quel.

On classe comme adjectifs faisant fonction de SPE ceux qui peuvent se combiner avec les précédents :
(ces) quelques, (les) différents, (un) autre, même, certain.
Ces éléments ont par ailleurs un fonctionnement d'adjectifs: *c'est certain / différent / tout autre*…

Cas particuliers : Pour *quel*, qui peut fonctionner à la fois comme SPE (*quel homme*) et comme PRQ (interrogatif) : *quel est-il ?* étendu à *tel quel*), on choisit un classement unique PRQ : puisqu' on admet que des ADJ peuvent fonctionner aussi comme SPE, on peut étendre cela aux PRQ.

NB. On ne considère pas comme DET les « déterminants complexes »**:**
Beaucoup de X, trop de, plein de, la plupart de sont décomposés. Le premier élément fonctionne comme tête avec sa catégorie habituelle : ADV ou PRO.

Conclusion : Syntaxiquement, la fonction SPE concerne tous les mots qui permettent à un GN de fonctionner comme sujet (qu'ils soient ADJ, PRQ ou DET.)

PROFORMES
Les proformes sont des lexèmes qui peuvent remplir à eux seuls les fonctions de groupes nominaux, adjectivaux ou prépositionnels. Nous avons suivi le LEFF pour ce qui concerne les pronoms (PRO) à distribution de SN, nous nous en sommes distingués en établissant les POS : CLS, CLN, PRQ, qui rassemblent des lexèmes variables et des invariables, des pro-SN (*il, le, qui*), et des pro PP *(y, en, où*).

PRO : sont classés PRO des items variables qui peuvent à eux seuls remplir les fonctions et occuper les positions d'un groupe nominal :
celui-ci, lui, ce, n'importe qui, qui que ce soit…

NB : Les items invariables qui jouent le rôle d'un groupe prépositionnel : *là, ici, alors, autant, ainsi, autrement, n'importe où*… sont comme dans le LEFF reversés à la catégorie ADV.

Tel, qui est en fait un pro-syntagme adjectival : *il est tel, tel il s'est présenté devant vous*, est classé PRO.

NB : L'équivalent de *tel* en langue parlée « *comme ça »* est classé adverbe.

CLI : nous avons crée une POS pour les proformes présentant la distribution particulière des clitiques, quelles soient variables (*il, ils, lui/ leur*) ou invariables (*y, en).* Avec une POS particulière CLS pour le clitique Sujet, qui joue un rôle important dans la syntaxe de la langue parlée.

PRQ : Cette POS rassemble les interrogatifs et relatifs, sur la base de leur distribution spécifique en tête de construction. Comme pour les clitiques, on trouve des PRQ variables : *qui/quoi* (pro SN) et des invariables *où, quand* (pro SP).
Qui et *que*, relatifs sujet et objet sont classés comme pronoms et non comme des conjonctions.

NUM
Compte tenu de la grande variété de fonctions syntaxiques que peuvent remplir les numéraux cardinaux (*deux, trois, cent-mille*), il a été décidé de réduire les erreurs de classement en POS en créant une POS spécifique pour eux. Il s'agit d'une décision purement pratique.

2. Catégories invariables

2.1. Principes

Nous distinguons les invariables qui figurent systématiquement en tête de construction (PRE, CSU, COO) de ceux dont la distribution est plus diverse (ADV, INT.)
Pour les invariables « tête de construction », on a renoncé à la solution radicale de poser une seule POS avec des combinatoires syntaxiques différentes. Le classement s'écarte cependant de la tradition sur les points suivants :

La distinction entre PRE et CSU repose principalement sur la propriété que les PRE ne peuvent construire des VRB à la différence des CSU.

**(Pour / après) il aille à Toulon*
(qu'/ comme / si / quand) il va à Toulon

Comme dans des études récentes, on considère que les PRE peuvent construire des séquences phrastiques *que* VRB : *pour qu'il aille à Toulon*.
Les conjonctions composées de la tradition *pour que, après que…* sont donc décomposées en PRE + CSU.
Les COO sont distinguées des CSU par la propriété classique de ne pouvoir être reprises par la CSU *que*.
On admet des multiples catégorisations, par exemple pour *comme* (PRE, CSU) ou *ainsi que* (COO, CSU).

2.2 Détail des POS invariables

PRE (préposition)

Caractéristiques :

- Ne gouverne pas un verbe fini VRB à la différence des CSU : **après il est parti*

mais peut gouverner une CSU + VRB : *après qu'il est parti*

- Construit canoniquement des SN ou des VIF ou d'autres PRE.
- construit des CSU à la fois dans des principales et dans des dépendantes :

avant que je revienne / fais-le avant que je revienne.

Cette propriété distingue la PRE d'adverbes comme *heureusement, peut-être,* qui ne construisent des CSU que dans des phrases racines (principales) :

*Heureusement qu'il est venu / *je pense que heureusement qu'il est venu*

NB 1 : La PRE peut être morphologiquement complexe comme les CSU

A l'instar de, à la différence de etc. au travers de

NB 2 : on a choisi de traiter des paires comme :

Au lieu de venir / au lieu qu'il soit venu

en distinguant *au lieu de* préposition complexe et *au lieu que* conjonction complexe.

Au lieu n'a pas été classé comme PRE, car on n'a pas : **au lieu le samedi venez le lundi*

ADV

La classe rassemble des éléments morphologiquement et distributionnellement hétérogènes comme dans le classement traditionnel du LEFF :

- Des items définis morphologiquement : adverbes terminés en *–ment*.
- Des items définis par une distribution particulière : entre auxiliaire et verbe sans rupture prosodique :

Aspectuels et modaux : *il a (encore / souvent / toujours/ bien/ peut-être) parlé à Marie*

Quantifieurs : *il a(trop / beaucoup / rien) mangé*

Négations : *pas, jamais*

- Des Pro PP : à la différence des items en 2., ils ont une distribution de PP :

ainsi, ailleurs, aujourd'hui, ça et là, comme ça, dehors, ensemble, ibidem, ici, jamais, là, partout, quelque (part,/fois).

On peut ajouter des PRO VRB comme *oui (il dit que oui)*

- Des PP figés : on cherche à en limiter le nombre. Mais on traite comme adverbes les groupes prépositionnels figés : *de fait, en fait, en somme, comme ça…*

NB1 : On s 'écarte du classement traditionnel sur 2 points :
-Les prépositions intransitives ou « orphelines » sont exclues de ADV et restent PRE *: avant, après, sans…*

-Des « adverbes » sont reclassés dans les POS CLI et PRQ (voir ces Pos)

NB2 : *jusque là, jusqu' (alors/ ici*) sont classés adverbes composés dans le LEFF, tout comme *jusque* malgré l'absence d'emploi isolé :* *j'attendrai jusque*.

En s'appuyant sur l'existence de séquences régulières PRE + ADV (*de là, par là)* et PRE + PRE.*(près de là par dessus le mur)*, on a décidé de classer *jusque* comme PRE. On a donc : *jusqu'à* PRE + PRE. Et *jusqu'ici* PRE + ADV. Ce classement permet en outre de prendre en compte les séquences non standard attestées : *jusque la prochaine fois, jusque le bout de la rue*.

CSU.
Caractéristiques :
A la différence des PRE, la CSU gouverne directement un verbe fini :
après **il est venu* /(*quand /si*) **il est venu**
A la différence des adverbes « connecteurs » (*donc, en effet*…). La CSU est toujours en tête de construction.
A la différence des COO, qui gouvernent aussi des VRB, la CSU peut être reprise par QUE.

NB : *car* et *or*, qui ne peuvent être reprises par *que*, sont classées COO
Peuvent être complexes (multilexicales) *bien que, alors que*, (voir rubrique composés)

COO
Voir CSU pour la propriété caractéristique.
On peut ajouter : construit un SN coordonné (même fonction) postposé au premier sur le modèle *et* et *mais* : *Pierre et Paul*, *Pierre ainsi que Paul*
ainsi que, excepté, sauf, hormis, sont donc classés COO par leur faculté de constituer des deuxièmes termes de coordinations, même si ils ont des emplois antéposés de prépositions.

INT
Contrairement à l'adverbe,

- Ne peut jamais être régi.
- Forme un tour de parole à interprétation autonome (peut initier un discours) :
  Hélas ! par opposition à *malheureusement* qui suppose un support discursif :
  Il ne viendra pas L2 malheureusement

INT regroupe de fait les interjections traditionnelles et les particules discursives hors catégories : *euh, bon, ben etc.…*

NB : les incises comme *je crois, je vois, tu sais* gardent leurs catégories d'origine (CLI+ VRB). Leur spécificité est traitée en syntaxe par relation DM (voir guide syntaxique)
C'est aussi le cas de : *pardon, merde, putain*

3. Traitements particuliers d'un domaine :
3.1. Les « connecteurs » Comparatifs et Consécutifs:
Autant que, plus que, moins que, ainsi que… Posent des problèmes particuliers. On peut hésiter entre catégories PRE, COO, CSU. On choisit de limiter le choix à COO et CSU.
On distingue :

- Les emplois où la séquence (*autant que, plus que, moins que, d'autant plus que)* est ou pourrait être « décomposée » :
  Il a travaillé autant qu'il s'est amusé
  Il a autant travaillé qu'il s'est amusé / que moi
  il a d'autant plus travaillé qu'on l'a bien payé

Dans ce cas, on décompose en ADV + CSU. Donc *que* est CSU dépendante du verbe du VPP ainsi que l'adverbe de quantité.

- Les séquences indécomposables homonymes des précédents :
  autant qu'on le fasse aujourd'hui
  **autant peut-être qu'on le fasse aujourd'hui*
  tant qu'il restera / tant alors qu'il restera*

Dans ce cas *autant que* et *tant que* sont CSU complexes

- Les autres séquences indécomposables: *ainsi que, plutôt que*

1. Dans le cas où la séquence gouverne une VRB, on aligne sur les CSU composées de type *alors que*
   Ainsi que je te l'avais dit, il est venu
   Il s'est comporté ainsi qu'on lui avait demandé
   Ainsi que les Grecs faisaient de la philo (de même) les Arabes faisaient

des maths

Plutôt qu'il aille à Paris, je préfère qu'il reste ici

2. Dans le cas où la locution gouverne un complément non phrastique équivalent d'une coordonnée :

Paul ainsi que Pierre viendront demain

Je prendrai un livre plutôt qu'un cahier

on a choisi de catégoriser COO même dans le cas d'antéposition :

Ainsi que Pierre, Jean fait du surf

Plutôt qu'un cahier, je prendrai un livre

3.2. Les exceptifs

Les items comme, *sauf, excepté, hormis,* sont classés comme COO sur la base d'exemples comme :

Les élèves sauf ceux de 3ème 2 auront piscine

L'analyse est étendue au cas de :

Sauf en cas d'urgence il ne faut pas utiliser ce téléphone.

4. Traitement des mots composés.

4.1. Composés fonctionnels versus composés pleins

Les « mots composés » selon la tradition ou les expressions « multimots » posent de nombreux problèmes de définition dont la solution ne fait pas l'unanimité parmi les linguistes. Cette question ne sera abordée ici que d'un point de vue pratique : choisir la solution qui minimise les erreurs d'analyse de l'analyseur.

4. 2. Objectif : réduire les erreurs d'analyse dues au figement du composé

Le problème majeur que pose à l'analyseur la décision de figer en un seul « token » une suite d'items qui fonctionnent par ailleurs comme des tokens séparés est que la séquence libre ne sera pas reconnue par l'analyseur. Ainsi si l'on décide de figer la suite *bien que* en un seul mot analysé comme conjonction de subordination dans le lexique, l'analyseur ne pourra pas donner une analyse correcte de la suite *bien que* dans la séquence : *je sais bien que ce n'est pas juste.*

Pour des raisons de temps, nous avons choisi de traiter le problème des composés « fonctionnels » appartenant aux catégories « fermées » de PRE, CSU, PRO, PRQ. Nous considérons en outre que les différences de tokénisation dans le domaine des catégories « ouvertes » (NOM, ADJ, VRB…) a moins d'incidence sur l'analyse en dépendances. Pour ces dernières catégories, nous avons suivi le LEFF.

4.3. Traitement des composés fonctionnels

Le lien de dépendance MORPH : Notre attitude pragmatique consiste à croiser

deux problèmes concernant les composés : définir les critères linguistiques selon lesquels une séquence de plusieurs tokens est considérée comme un seul mot dans le lexique. Adapter ces critères en fonction des conditions de fonctionnement de l'analyseur. Nous traitons les composés comme des séquences possédant un gouverneur qui est le seul à avoir un lien de dépendance avec le contexte. Les autres éléments de la séquence sont reliés au gouverneur par un lien spécifique MORPH représentant l'idée que ces éléments sont des extensions du morphème tête. Nous précisons ci-dessous les critères qui sont utilisés pour établir un lien MORPH.

4. 3.1. Critères linguistiques.

L'intuition de non compositionnalité sémantique de la séquence n'est pas un critère suffisant pour la déclarer composée. Elle peut fonctionner comme une heuristique, mais doit être accompagnée de critères formels. Nous prendrons l'exemple des suites comportant une CSU.

- La séquence multimots n'est pas une suite syntaxique régulière dans notre modèle.

Ex *: Pour que* ne sera pas composé parce que nous analysons la séquence comme la suite régulière : PRE + CSU, la CSU étant un dépendant régulier de PRE.

Dans le but que est une suite régulière, et sera donc décomposé.

Au contraire, *afin que*, *mis à part que* ne sont pas des suites régulières, car on ne peut attribuer une catégorie à *afin*, et *à part + que* ne sont pas des dépendants possibles pour *mis.* On a donc un seul mot dans le lexique.

Autres exemples de suites non régulières *: à condition que, à force que, toujours est-il que.*

- La séquence est régulière mais comporte des restrictions :

bien que peut-être analysé sur le modèle de *heureusement que*, mais sa distribution est celle d'une conjonction de subordination et non celle d'un adverbe (*bien* ne peut être root). Nous en faisons un candidat à la composition.

Les séquences candidates à la composition sont analysées au moyen du lien MORPH qui va d'un item choisi comme tête aux autres items du composé. Notre originalité tient à la volonté de motiver le plus possible le choix de la tête. Nous distinguons les composés endogènes où on peut considérer que l'ensemble a la distribution d'un des composants (*Bien que* composé a la distribution d'une conjonction de subordination comme *que)* et les composés exogènes où l'ensemble a une distribution différente de ses parties *: c'est pourquoi.* Nous avons constaté empiriquement que pour les CSU et les PRE, il y avait très peu d'exogènes. La situation est différente pour les COO : *ainsi que, plutôt que, au même titre que, c'est à dire,* qui sont classés COO, sont des composés exogènes dans ce fonctionnement.

4.3.1.2. Critères liés au fonctionnement de l’analyseur.

Notre intention initiale était d’annoter selon ces principes le corpus d’entraînement de l’analyseur et de lui laisser le soin de définir lui-même en fonction de son entraînement l’attribution du lien MORPH et donc des composés du corpus à annoter.

Cette technique devait en particulier permettre à l’analyseur de choisir entre l’analyse en composé et l’analyse standard dans le cas des séquences potentiellement ambigües : *bien que, alors que, pour autant que…*

Nous avons alors rencontré le problème suivant : la faible fréquence de certaines séquences pouvait rendre difficile l’apprentissage des différences d’analyse. Nous avons observé de fait que certaines séquences avaient un fonctionnement préférentiel : ainsi *tant que* est majoritairement décomposé (80% des occurrences sur notre échantillon), tandis que *alors que* est majoritairement un composé (88% des occurrences). Nous avons donc utilisé cette constatation pour simplifier l’analyse. Nous avons intégré directement dans le lexique des CSU les cas où l’ambiguïté était prévisiblement faible *: une fois que, à condition que, à part que*.

Pour le détail des classements, on se reportera au contenu des POS dans le lexique.