Manuel d'annotation pour les corpus du projet PERCEO

Note : Ce document a été élaboré en nous inspirant fortement du manuel d'annotation d'Abeillé & Clément (2006)[1]. Nous nous sommes conformés à leurs propositions dès que cela s'avérait compatible avec notre propre démarche, ce qui a très souvent été le cas, en simplifiant souvent les analyses (notamment nous n'utilisons pas de traits morphologiques). Nous nous sommes contentés de changer quelque peu la rédaction, de mentionner le nom de nos propres étiquettes, etc., quand cela ne nous semblait pas conforme à ce que nous souhaitions. La plupart des exemples ont d'ailleurs été conservés. En revanche, nous avons effectué un important travail de lissage de la rédaction et de refonte du plan initial.

Toute erreur est évidemment de notre fait et ne saurait être imputable aux auteurs du document dont nous nous sommes inspirés.

Table des matières

[1] www.llf.cnrs.fr/Gens/Abeille/guide-morpho-synt.06.pdf

1. Présentation du projet

Le manque de corpus en français, écrits et surtout oraux, qui soient diffusables, normalisés (en TEI par exemple), échantillonnés et étiquetés en morphosyntaxe (parties du discours et lemmes) est un problème récurrent pour le TAL et la linguistique de corpus francophone. Nous nous sommes concentrés plus particulièrement sur ce dernier aspect ainsi que sur la diversification des sources. En effet, l'étiquetage morphosyntaxique est indispensable ne serait-ce que pour effectuer le dénombrement global des formes différentes ou pour retrouver plus aisément certaines formes ambiguës. Les corpus oraux étiquetés, quant à eux, peuvent aussi être utiles aux systèmes de transcriptions automatiques (Huet, Gravier, Sébillot 2006).

Pour l'écrit, la récente initiative de Ferraresi et al. (2008) et Baroni et al. (2009) a probablement amélioré cet état de fait en mettant à disposition le corpus FrWac (corpus aspiré sur le Web selon une méthodologie reproductible, d'une taille de 1,8 milliards d'occurrences). Néanmoins, comme le soulignent ses concepteurs, la taille et le mode de constitution du corpus FrWac n'ont pour le moment pas permis de statuer précisément sur son contenu et donc son échantillonnage, aspect indispensable pour la description linguistique sur corpus. De la même manière, la mise à disposition du corpus de l'Est Républicain sur le site du CNRTL représente un progrès mais ce corpus n'est ni échantillonné, ni étiqueté dans sa forme téléchargeable. Par ailleurs, disposer d'un nouveau système automatisé d'étiquetage pour la base textuelle Frantext représenterait une amélioration notable de celle-ci au vu des réflexions menées par le groupe Frantext2 animé par Véronique Montémont. Le projet TCOF, enfin, a contribué avec d'autres à la mise à disposition de corpus oraux, alignés texte-son et normalisés en TEI, mais ces données ne sont pour l'instant ni échantillonnées, ni étiquetées en morphosyntaxe.

Face à ces lacunes, l'objectif premier du projet que nous avons mené est de rendre disponible pour la communauté scientifique un étiqueteur en morphosyntaxe entraîné sur les données actuellement présentes sur le site du CNRTL, à savoir l'Est Républicain, les transcriptions du projet TCOF et les textes libres de droit de Frantext. Le travail réalisé pour « Frantext libre de droits » a pour objectif d'être directement utilisable pour étiqueter la future base textuelle Frantext2.

L'originalité principale du projet réside dans le fait d'aborder aussi bien les données écrites que les données orales, ce qui n'est pas le cas pour les autres systèmes d'étiquetage

automatique. Pour ce faire, nous avons entraîné l'étiqueteur TreeTagger[2] sur différents types de données et proposé autant de fichiers de paramètres que de types de données distinctes, ce qui représente l'autre versant original de notre approche. Notre choix s'est porté sur ce logiciel car il est libre de droits, multi-plateformes, supporte les deux encodages les plus courants, ISO et UTF-8 et permet l'élaboration de fichiers de paramètres spécifiques à chaque corpus d'apprentissage utilisé. Ainsi, nous fournissons plusieurs fichiers distincts pour Frantext en fonction des regroupements par type de texte, ainsi qu'un autre fichier pour l'Est Républicain et encore un autre pour l'oral.

Notre projet consiste donc principalement à développer divers fichiers de paramètres pour le logiciel TreeTagger basé uniquement sur l'apprentissage de corpus distincts : français parlé, littérature, etc. Le principe est de faire un apprentissage à l'aide du module TrainTreeTagger à partir de corpus que nous avons étiquetés automatiquement puis corrigés manuellement de manière systématique. Nos corpus bruts, pour la partie orale, sont issus du projet TCOF et librement téléchargeables à l'adresse : http://www.cnrtl.fr/corpus/tcof/. Il s'agit uniquement des corpus adultes. Pour l'écrit, les corpus sont issus d'une banque de données essentiellement littéraires annotées manuellement (voir supra) et de l'Est Républicain mis en ligne sur le CNRTL (PQN). Ils pourraient être complétés par l'utilisation du corpus scientifique SCIENTEXT mis à disposition et géré par Agnès Tutin dans le cadre du projet ANR du même nom.

Dans la version finale, chaque unité considérée comme étant un « mot » (unique ou multiple) est accompagnée d'une étiquette morphosyntaxique (traditionnellement appelée partie du discours) et du lemme correspondant. Nos étiquettes ne comportent pas de traits morphologiques contrairement à d'autres projets du même type. Cela provient de la méthodologie que nous avons suivie, à savoir celle de corriger et modifier directement la sortie du logiciel TreeTagger, celui-ci ne donnant pas directement les traits morphologiques. De plus, la correction des seuls lemmes et parties du discours requérait déjà un temps conséquent non extensible étant donné le faible nombre de personnes impliquées dans ce projet. La simplification extrême des étiquettes et la généralisation de certaines analyses sont également parfois imputables à ce même paramètre de limitation du nombre de personnes impliquées. Pour cette même raison, nous avons décidé de limiter le temps consacré à résoudre des questions ne portant que sur un nombre infime de cas, même si cela devait se faire au détriment de la cohérence de l'ensemble. Ces questions étant abordées depuis des siècles et d'une complexité difficile à surmonter, nous avons préféré procéder ainsi pour ne pas rester bloqués dans l'attente d'une solution ayant de toute façon des implications difficilement maîtrisables.

2. Présentation de la méthodologie suivie

La démarche suivie pour mener à bien l'apprentissage sur corpus et obtenir un outil d'annotation automatique entraîné de manière différenciée sur le français parlé non planifié (TCOF), le français du quotidien régional l'Est Républicain (PQR) et un échantillon de textes littéraires représentatifs de la base Frantext est la suivante. Le principe général consiste à

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

faire un premier passage avec TreeTagger (version étiqueteur) puis à corriger manuellement les étiquettes, en scrutant systématiquement l'intégralité des résultats, genre par genre (oral, PQR, échantillons littéraires). Dans un premier temps, une série d'heuristiques nous a permis d'accélérer le travail de correction (un programme en *python* injectant les étiquettes que l'on peut affecter automatiquement sans risque). Par ailleurs, nous sommes partis du jeu d'étiquettes par défaut proposé par TreeTagger, et avons essentiellement ajouté quelques parties du discours supplémentaires, quand cela s'avère nécessaire (pour les amorces de mots, les liaisons, les multi-transcriptions, etc.), et fait quelques modifications le cas échéant afin de rendre le jeu d'étiquettes plus cohérent.

Chaque fois que le volume de données annotées et vérifiées atteint 10.000 mots supplémentaires par tranche, nous relançons l'apprentissage à l'aide du module TrainTreetagger. Et nous repartons de cette base pour annoter automatiquement les textes suivants, et ainsi de suite. Les tests sont faits sur un échantillon appartenant au corpus d'entraînement afin de voir si les résultats sont particulièrement bons ainsi que sur un échantillon externe (le corpus de validation). L'objectif que nous poursuivons à plus long terme est d'obtenir des résultats comparables à l'état de l'art du domaine en renouvelant la phase d'entraînement sur des volumes de plus en plus importants mais aussi de diffuser un vaste corpus annoté de bonne qualité.

2.1. Corpus écrits

Le corpus écrit PQR ne présente pas de spécificités particulières du fait que la version actuellement diffusée de TreeTagger a été vraisemblablement entraînée sur le corpus du journal Le Monde, c'est-à-dire de la presse quotidienne nationale (PQN). Le corpus que nous avons étiqueté et le corpus d'entraînement sont donc similaires en genre. Cependant, des progrès importants restent à faire par rapport à l'état actuel de l'étiqueteur pour l'annotation des titres, des listes, que l'on trouve en grand nombre, et des dates et horaires (pharmacie de garde, programmes loisirs du week-end, séances de cinéma, etc.).

En revanche, les corpus écrits littéraires représentent une différence générique par rapport au corpus d'apprentissage à partir duquel a été entraînée la version actuelle de l'étiqueteur. En conséquence, nous avons utilisé le corpus annoté à la main par Josette Lecomte et Nabil Hatout en 1998 avec le jeu d'étiquettes de l'étiqueteur Brill (413.000 occurrences, puisées parmi 20 ouvrages essentiellement tirés de Frantext). Dans un premier temps, nous avons procédé à la traduction des étiquettes de Brill vers un jeu d'étiquettes le plus minimal possible compatible avec Treetagger et les choix opérés dans la partie orale (les catégories majeures, les principales catégories de mots grammaticaux et les catégories ambiguës lorsque cela est nécessaire, notamment NOM/ADJ, ADJ/VERBE, DET/PRO). Dans cette étape, nous avons rencontré les difficultés courantes de la mise en correspondance de deux jeux d'étiquettes dont les recouvrements ne sont pas nécessairement synthétisables. Ce cas de figure a en particulier été rencontré lorsqu'une étiquette de Brill, par exemple, PRO, correspondait à plusieurs étiquettes de TreeTagger. Dans ce cas précis, nous avons procédé à la traduction automatique correcte mais minimale, choix du subsumant PRO, puis à un

raffinement de cette étiquette minimale via un étiquetage manuel. Le corpus étiqueté manuellement et traduit dans un jeu d'étiquettes compatible avec TreeTagger a constitué le corpus et le lexique sur lequel nous nous sommes appuyés pour l'apprentissage de la tranche littéraire.

Après avoir mené à bien cette première étape, un échantillonnage du corpus étiqueté manuellement a été réalisé avec Véronique Montémont afin d'identifier les différents genres littéraires représentés dans ce corpus (sur une période allant de 1825 à 1929, 66% de romans, 18% d'essai et 14% d'autobiographie, les deux derniers genres représentés, mémoire et théâtre, l'étant de manière anecdotique). L'entraînement du module TrainTreetagger a ensuite été lancé pour chaque genre et les résultats analysés comme décrit ci-dessus.

A l'issue de ces travaux, des fichiers de paramètres propres à chaque genre littéraire ont été fournis afin d'étiqueter l'ensemble de la base Frantext et le corpus de l'Est Républicain a été mis à disposition sous sa forme étiquetée et accompagné de son fichier de paramètres propre.

2.2. Corpus oraux

Les corpus oraux présentent des spécificités qui ne sont pas prises en compte par les étiqueteurs morphosyntaxiques, élaborés à partir de données écrites et pour des données écrites. Même si l'étiquetage de corpus oraux ne représente pas un problème spécifique (Benzitoun, 2004), force est de constater que l'application directe d'un logiciel d'étiquetage ne donne pas des résultats pleinement satisfaisants. La solution envisagée est généralement d'adapter les outils existants (Dister, 2007 ; Blanc et al., 2008) ou de masquer certains phénomènes tels que les « disfluences » (Valli & Véronis, 1999).

Pour notre part, nous avons décidé de travailler directement à partir des transcriptions brutes avec un minimum d'aménagement pour ne pas dénaturer les données. Nous avons uniquement supprimés les commentaires et les pauses. Ainsi, nous avons décidé de nous attaquer au problème en profondeur, non en adaptant les outils existants, mais plutôt en développant une méthode spécifique, à l'image de ce que proposent Eshkol et al. (2010).

Etant donné que le module d'apprentissage nécessite également un lexique contenant la forme, le lemme et la partie du discours, en plus du corpus d'entraînement, nous avons pris l'option d'élaborer nous-même ce lexique à partir des annotations que nous produisons. La couverture sera évidemment beaucoup plus restreinte que celle que proposent des ressources telles que le *Lefff*[3] ou *Morphalou*[4], mais pour des raisons de compatibilité des étiquettes et de maîtrise optimale d'un maximum de paramètres, cette solution s'est imposée.

[3] http://www.labri.fr/perso/clement/lefff/telechargement.html

[4] http://www.cnrtl.fr/lexiques/morphalou/

Ce travail est mené, provisoirement, sur 25 transcriptions d'adultes extraites du projet TCOF[5]. Cet ensemble d'environ 50.000 mots nous a servi de premier échantillon test. Au terme de notre travail, nous avons diffusé librement un lexique syntaxique du français parlé, un étiqueteur morphosyntaxique (fichier de paramétrage spécifique aux corpus de français parlé) ainsi que le corpus étiqueté.

3. Principes généraux

3.1. Corrections apportées aux textes

On a corrigé les coquilles manifestes (aboutissant à un mot inexistant), les fautes d'orthographe et les erreurs de transcription.

c'est à dire => c'est-à-dire
à priori => a priori
rapelais => rappelais

Nous n'avons pas transformé les abréviations, ni les mots n'existant pas dans les dictionnaires mais utilisés par les locuteurs et respectant les conventions de transcription initiales. Certaines formes sont donc non standards.

3.2. Les catégories employées

Une étiquette comporte au maximum les deux parties suivantes séparées par deux points. Une première information, en majuscules, mentionnant la partie du discours proprement dite et une seconde partie en minuscules contenant la sous-catégorie ou le temps lorsqu'il s'agit d'un auxiliaire ou d'un verbe. Une étiquette doit obligatoirement comporter au moins la première partie.

Voici la liste intégrale des étiquettes utilisées dans notre projet :

ADJ	adjectif
ADV	adverbe
AUX:*tps*	auxiliaire (pour les temps, cf. ci-dessous les étiquettes verbales)
DET:def	déterminant défini
DET:dem	déterminant démonstratif (ce, cette, ces)
DET:ind	déterminant indéfini (chaque, quelque, un, des)
DET:int	déterminant interrogatif (quel)
DET:par	déterminant partitif (du)
DET:pos	déterminant possessif (ma, ta, etc.)
DET:pre	pré-déterminant (tout (le), toute (la), toutes (les))
EPE	épenthétique
ETR	mots étrangers
FNO	forme noyau (oui, non, d'accord, ...)
INT	interjection & particules discursives
KON	conjonction
LOC	locuteur

[5] TCOF est une base de transcriptions de français parlé constituée à l'ATILF et téléchargeable à l'adresse : http://www.cnrtl.fr/corpus/tcof/

MLT	multi-transcription (/x,y/, (n'))
NAM	nom propre
NOM	nom commun
NOM|NAM:sig	sigle
NUM	numéral
PRO:clo	clitique objet
PRO:cls	clitique sujet
PRO:clsi	clitique sujet impersonnel
PRO:dem	pronom démonstratif
PRO:ind	pronom indéfini
PRO:int	pronom interrogatif (comment, où, quand, quoi, etc.)
PRO:pos	pronom possessif (mien, tien...)
PRO:rel	pronom relatif
PRO:ton	pronom tonique
PRP	préposition
PRP:det	préposition+déterminant (au, du, aux, des)
PRT:int	particule interrogative (est-ce que, est-ce qui)
SYM	symbole
TRC	troncation – amorce de mot
étiq:trc	abréviation
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impératif
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitif
VER:pper	verbe au participe passé
VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passé simple
VER:subi	verbe au subjonctif imparfait
VER:subp	verbe au subjonctif présent
PUN	ponctuation
PUN:cit	guillemets
SENT	fin de phrase

3.3. La segmentation

Pour segmenter les textes initiaux, nous avons utilisé le tokenizer fourni avec Treetagger. Celui-ci utilise principalement une liste de caractères séparateurs, à l'image de la plupart des tokenizers, accompagné de quelques règles supplémentaires. Cependant, nous lui avons apporté des modifications, dont la plus importante est la constitution d'un fichier spécifique listant les mots multiples que nous estimons pertinents.

Les mots composés sont notés par la suite de leur composants séparés par des blancs, des tirets ou des apostrophes : *c'est-à-dire, aujourd'hui, pomme de terre*. Cependant, la démarche imposée par le logiciel ne nous permettait pas de tout segmenter à notre guise. Ainsi, de rares interventions manuelles ont été tout de même nécessaires.

3.4. Les critères pour déterminer les mots composés

Ces critères ne sont ni nécessaires ni suffisants. Ils permettent simplement d'établir un faisceau de propriétés liées aux mots composés.

Nous avons choisi d'être extrêmement restrictifs sur les mots composés en général car il faut nécessairement qu'ils soient contigus dans tous les contextes. Aucune insertion ne doit être possible. Ci-dessous, nous explicitons les critères qui ont présidé à la constitution de la liste des mots composés, en plus du critère de contiguïté qui est le plus important. Mais pour commencer, voici quelques mots composés en guise d'illustration. La liste exhaustive est fournie dans un fichier à part.

peut-être[ADV]
c'est-à-dire[KON]
rendez-vous[NOM]
garde-côtes[NOM]

3.4.1. Critères morphologiques et de contiguïté

La principale règle que nous avons suivie consiste à segmenter à chaque fois que l'on peut effectuer une insertion, quelle qu'elle soit. C'est la raison pour laquelle nous avons séparé l'auxiliaire et le participe passé, par exemple. Cela est lié au refus de faire des regroupements d'unités non contiguës.

Il arrive parfois que l'un des composants n'apparaisse que dans l'expression complexe. Si tel est le cas, nous ne segmentons pas :

au fur et à mesure
a priori
aujourd'hui

Lorsque les règles d'accord ne sont pas respectées, il faut également regrouper :

Une grand-mère
**Une grande-mère*

3.4.2. Critères sémantiques

Lorsque le sens global du mot composé n'est pas la composition sémantique du sens de ses composants, nous regroupons :

une sage-femme n'est pas une femme qui est sage.
carte bleue n'est pas une carte qui est bleue.

Nous en faisons de même lorsque l'on ne peut pas remplacer un des constituants par un synonyme ou un antonyme :

*à bas / *à haut*
*à la va vite / *à la va rapidement*

3.4.3. Autres cas de figure

Pour les prépositions complexes, nous segmentons généralement le *de* ou le *à* étant donné qu'un élément peut venir s'insérer avant et que nous ne souhaitons pas faire d'unités discontinues et multiplier les mots multiples du type *en dépit de, en dépit des, en dépit du* :

à l'insu[ADV] de[PRP]

Pour ces raisons, nous avons donc décidé de limiter très fortement le nombre de locutions conjonctives et prépositionnelles. Du coup, à chaque fois qu'il est possible d'insérer un élément, comme un adverbe par exemple, nous procédons à la segmentation.

Voici quelques exemples :

de part et d'autre[ADV]
de part et d'autre[ADV] de[PRP]
en face[ADV]
en face[ADV] de[PRP]
compte tenu[ADV] de[PRP]

A noter également que *il y a, y compris* et *d'ici* doivent être regroupés lorsqu'ils fonctionnent comme une préposition :

ils sont partis il y a[PRP] trois ans
y compris[PRP] à New York [=adv ?]
d'ici[PRP] trois ans

Attention : les autres instances de *il y a* doivent être analysées mot à mot.

En plus des critères énumérés ci-dessus, les propriétés syntaxiques du NOM composant (pas de déterminant ou pas d'adjectif insérable) doivent également être prises en compte :

*à cause[ADV] de[PRP] *à cette cause, *à la cause de, *à une grande cause de...*
*à côté[ADV] de[PRP] *à ce coté, *au côté de*

Ainsi, nous notons les "locutions conjonctives" comme des adverbes suivis d'une conjonction, sauf pour les unités qui ne peuvent pas être adverbe, que nous notons comme préposition (ou toute autre catégorie jugée pertinente) :

pour[ADV] que[KON], depuis[ADV] que[KON], pendant[PRP] que[KON] etc.

Remarque : *pour* et *depuis* sont considérés comme des adverbes car ils peuvent fonctionner sans régime, ce qui n'est pas le cas de *pendant*. Ici, *pendant* est donc considéré comme une préposition car il ne peut fonctionner que comme ça :

*il a fait ça pendant[PRP] le dîner / *il a fait ça pendant*

Nous ne regroupons que les unités qui n'ont aucune autonomie, comme *parce que* par exemple.

Ainsi que et *c'est-à-dire* sont également considérés comme des conjonctions complexes. On a normalement corrigé les mots composés non pertinents en contexte, par exemple :

C'est ainsi[ADV] que[KON] la Russie devint... (il ne s'agit pas de la conjonction *ainsi que*)

Les amalgames *des, du, au, au(x)quel(le)(s)* et *duquel, desquel(le)(s)* ne sont jamais segmentés en deux unités distinctes. Cependant, l'étiquette donnée tient compte du fait qu'il s'agit de deux catégories amalgamées quand l'on a affaire par exemple à la préposition (*à* ou *de*) suivie par un déterminant ou un pronom (relatif).

L'amalgame déterminant+préposition *des* est noté de la manière suivante :

c'était le moment des[PRP:det] vendanges
et le lemme est alors du.

En revanche, le déterminant indéfini est noté de la manière suivante :

j'ai travaillé dans des[DET:ind] camions

et le lemme est alors *un*.

Les partitifs en deux “mots” sont considérés comme des déterminants composés :

tu peux rajouter |de la|[DET:par] tourbe

et le lemme est alors *du*.

Ceux-ci ont été regroupés manuellement afin de ne pas générer des erreurs par un regroupement automatique systématique. On pourra d’ailleurs se servir de notre travail manuel pour essayer d’automatiser cette tâche dans des travaux futurs.

Dans le cas où plusieurs mots composés étaient candidats (exemple le nom composé "face à face" ou la préposition composés "face à") c'est le plus long qui a été privilégié.

Par ailleurs, étant donné que nous ne distinguons pas les préfixes, les mots ci-dessous sont considérés comme des composés et donc une seule et même unité. Ainsi, *méta, anti, archi*, etc. suivis d'un tiret ou non doivent être regroupés avec ce qui les suit.

un peintre italo-belge[ADJ] *Un sous-marin[NOM]* *Un trop-plein[NOM]* *c’est archi-sûr[ADJ]* *les relations outre-Atlantique[ADJ]*

Ci en suffixe est ADV. sauf *celui-ci, celle-ci* etc. qui sont des PRO composés.

cet homme -ci[ADV] qui me regarde *celui-ci[PRO:ton] me regarde*

Au niveau des adverbes, les unités telles que les suivantes doivent être regroupées en vertu des critères vus ci-dessus :

en revanche[ADV] *au contraire[ADV]* *par contre[ADV]*

De même, les unités suivantes doivent être regroupées :

*en effet[ADV] (*en un/cet effet)* *à terre[ADV] (*à sol...)*

mais pas :

en[PRP] partie[NOM] car on a : *en grande partie, en totalité...* *en[PRP] avance[NOM]* car on a : *en retard* *par[PRP] hasard[NOM] (par un hasard extraordinaire, par pur hasard, par chance)*

Attention : l'absence de DET est normale après *en* et ne constitue pas à soi seul un critère de figement (*en France, en linguistique, en or, en argent*...).

3.4.4. Les verbes composés

On a choisi d'en retenir très peu dans le corpus, car la plupart sont discontinus, et suivent une syntaxe régulière.

On a retenu les expressions verbales qui mettent en jeu un composant n'existant pas par ailleurs (faire fi de) ou celles qui sont très figées (V N sans déterminant possible: faire partie de) et non compositionnelles.

avoir beau[VER:infi]

mais pas:

avoir peur (avoir une grande peur), avoir faim, avoir soif...

La PRP ne fait pas partie de l'expression, car le VER figé ne se comporte pas comme un VER transitif direct, mais comme un V prenant un complément prépositionnel (cf. possibilité d'un pronom fort comme complément).

Les verbes contenant *se* ou *s'* (*s'apercevoir*) sont considérés comme une suite du type PRO:clo VER donc comme deux unités distinctes. En revanche, pour des cas comme *en avoir assez* ou *s'en aller*, dans lesquels le *en* est totalement figé, on procède au regroupement *en avoir* et *en aller*. Le *s'* ne fait pas partie du verbe complexe car il peut varier (*je **m'**en vais*).

3.5. L'annotation des lemmes

Nous entendons par lemme une forme conventionnelle unique renvoyant à l'ensemble des formes fléchies ou élidées d'un même terme lexical (pour une catégorie donnée). Par exemple, pour l'ensemble des formes fléchies d'un verbe, le lemme sera la forme infinitive.

Nous notons les homonymes partageant les mêmes flexions par un même lemme. Par exemple, le lemme {fraise} correspond à la classe des formes fléchies *fraise, fraises*, qu'il s'agisse de l'outil ou du fruit. Mais la forme *cuisinière* peut correspondre à deux lemmes différents {cuisinier} ou {cuisinière} (=fourneau).

Le lemme est parfois, mais rarement, ambigu pour une même forme et une même catégorie. C'est pourquoi il est important de le noter dans le corpus. Cela permet de désambiguïser l'ensemble des formes. Le lemme de la forme *suis* est par exemple ambigu entre {être} et {suivre}, les lemmes {fil} et {fils} conviennent tout deux pour la forme *fils*.

Pour les pronoms personnels sujet, le lemme est identique à la forme du pronom. Pour les pronoms objet, le même traitement doit être appliqué, mis à part pour *le, la* et *les* qui ont pour lemme la forme unique {le}. Et en ce qui concerne les forme *lui* et *leur*, le lemme est systématiquement {lui}. Dans les deux cas précédents, il s'agit simplement d'une variation en nombre ou en genre. En revanche, nous distinguons les formes *ça* et *cela* qui possèdent des lemmes distincts.

Le lemme du mot composé est la forme morphologiquement non marquée du mot composé (les noms et adjectifs sont au masculin singulier, les verbes à l'infinitif).

Les lemmes des formes tronquées correspondent à la forme longue (télé => télévision).

4. Principes généraux pour l'annotation

Nous ne détaillons pas l'ensemble des catégories à annoter, notamment nous n'abordons pas celles qui ne posent pas de problème particulier. Nous ne passons en revue que les éléments les plus difficiles à annoter ou présentant une ambiguïté particulière.

Nous avons privilégié les critères distributionnels par rapport aux critères morphologiques. On a ainsi des verbes ou adjectifs invariables ainsi que des adverbes variables :

Voilà[VER] Paul (car on a *le voilà* et la position postclitique est réservée aux V) *La roue avant[ADJ] droite[ADJ]* (car *avant* est dans une position interdite aux PRP) *Une pomme toute[ADV] rouge[ADJ]* (car prémodifieur d'ADJ est une position réservée aux ADV)

Cependant, nous avons décidé de minimiser les cas de recatégorisation contextuelle car :

- on veut que le corpus annoté puisse servir à extraire un dictionnaire fiable et réutilisable ;
- on ne veut pas que l'annotation automatique qui en résultera soit altérée par un trop grand nombre d'ambiguïtés ;
- il n'est pas évident de savoir s'il s'agit d'une position fonctionnelle ou d'un véritable changement de nature.

Par exemple, un NOM peut être épithète ou attribut (c'est un usage productif et la recatégorisation ADJ serait arbitrairement limitée aux cas rencontrés dans le corpus) :

Paul est très famille[NOM] *une recette maison[NOM]*

Un ADJ peut être tête de groupe nominal (avec ellipse du nom) :

la petite[ADJ] rouge *les meilleurs[ADJ]*

Un ADJ peut être employé comme modifieur de V (usage productif et distribution différente des Adverbes correspondants) : *voter utile, acheter malin, refuser net*, etc.

Nous n'avons pas de catégorie résiduelle pour les éléments qui n'entreraient dans aucune classification. Tous les mots ont été étiquetés.

5. Parties du discours à annoter

5.1. Les interjections / particules discursives

La classe des interjections regroupent ce qui est traditionnellement considéré comme lui appartenant. Mais elle contient également ce que l'on appelle parfois les particules discursives ou "petits mots de l'oral", tels que *bon, ben, eh ben*, etc.

5.2. Les identifiants des locuteurs

Ils sont notés LOC.

5.3. Les formes noyaux

Ce sont des unités pouvant fonctionner de manière autonome comme énoncés indépendants et ne trouvant pas de correspondance évidente dans la nomenclature traditionnelle des parties du discours. Il en va ainsi de *oui, non, voilà*, etc.

5.4. Les amorces et formes abrégées

Les lemmes des formes réduites volontairement (abréviations) correspondent aux formes longues/standards équivalentes et leurs étiquettes sont les étiquettes attendues suivies de deux points et de **trc** (pour **troncation**). En ce qui concerne les amorces de mots, il faut utiliser la seule mention TRC et le lemme doit être identique à l'amorce. Nous aurions préféré utiliser la forme longue dans ce cas aussi, mais la manière dont procède TreeTagger rendait cela peu opératoire.

Par ailleurs, étant donné que les amorces et les abréviations sont distinguées par les conventions de transcription, les amorces de mots comportant un tiret final, il n'est pas utile de faire la distinction au niveau de l'étiquette morphosyntaxique. Ci-dessous des exemples complets de la manière dont il faut noter ces phénomènes :

am- TRC am-
ciné NOM:trc cinéma

Les **sigles** sont notés de la manière suivante : NOM:sig ou NAM:sig (cf. partie sur les noms propres pour savoir comment distinguer noms propres et noms communs).

EDF[NAM:sig]
HLM[NOM:sig]

5.5. Les multitranscriptions

Elles sont notées comme un seul ensemble, possèdent l'étiquette MLT et le lemme est un copier-coller de la forme.

5.6. Ponctuation

Dans les transcriptions d'oral que nous avons annotées, il n'y avait pas de ponctuation. Cette section concerne donc seulement l'écrit.

Pour l'écrit, TreeTagger propose 3 étiquettes de ponctuation PUN, PUN:cit et SENT. SENT est utilisé pour les ponctuations de fin phrase et est très utile si elle est affectée correctement pour la suite des traitements en TAL. PUN concerne tous les autres types de ponctuation sauf les guillemets (PUN:cit).

5.7. Pronoms clitiques & pronoms toniques

Les pronoms clitiques (cl) sont aussi dits atones par opposition aux pronoms toniques (ton) : *je* vs *moi* ; *tu* et *te* vs *toi*, etc.

On a affaire au pronom clitique (cl) quand il est sujet direct ou complément de verbe directement à gauche du verbe. On a affaire au pronom tonique (ton) s'il est coordonné, après préposition ou adverbe, modifié par une relative ou un groupe apposé ou bien tout simplement isolé :

Nous[PRO:ton] nous[PRO:cls] nous[PRO:clo] en[PRO:clo] moquons
Lui[PRO:ton] le diplômé de ...
Ecoute-moi[PRO:clo]

De plus, le pronom tonique est mobile, ce qui n'est pas le cas du clitique.

Pour *moi-même, lui-même*, etc., l'ensemble est [PRO:ton].

Les pronoms clitiques se distinguent en morphosyntaxe par une sous-catégorie indiquant la fonction. Nous indiquons la fonction des pronoms clitiques (sujet ou objet, cette seconde catégorie comprenant également les réfléchis) en neutralisant les différences qui dépendent de la sous-catégorisation des verbes. Ainsi nous notons, objet pour datif et accusatif, objet également pour les clitiques réfléchis et sujet pour les clitiques nominaux.

Nous utilisons la notation PRO:cls pour désigner les clitiques sujet, PRO:clsi pour les clitiques sujets impersonnels (*il* uniquement) et enfin PRO:clo pour les clitiques objet et réfléchis.

nous et *vous* sont a priori ambigus entre cls et clo :

nous[PRO:cls] sommes partis ensemble.
nous[PRO:cls] nous[PRO:clo] sommes rencontrés.

te, me sont toujours des PRO:clo. En revanche, *toi* et *moi* sont très majoritairement des PRO:ton, excepté après un verbe à l'impératif.

Dis-moi[PRO:clo] si nous en reparlerons

Regarde-toi[PRO:clo] dans l'eau.

5.8. Pronoms relatifs et interrogatifs

Les interrogatifs sont presque toujours des pronoms, parfois des particules (pour *est-ce que* et *est-ce qui* uniquement). Il n'y a pas d'adverbe interrogatif. Lorsque *qu'* précède *est-ce que/qui,* il s'agit d'un pronom interrogatif et son lemme est *quoi*.

Les pronoms relatifs introduisent des relatives. Les pronoms interrogatifs, quant à eux, ont des places variées dans les interrogatives directes. Etant donné la difficulté à distinguer entre relative sans tête et interrogative indirecte, nous utiliserons le critère de substitution avec *comment* ou *pourquoi*. Chaque fois que la substitution est possible, nous notons l'élément introducteur PRO:int.

Moi qui[PRO:rel] ai fait ..
Qui[PRO:int] a fait cela ?
Je sais à qui[PRO:int] tu penses (je sais pourquoi tu penses à lui)
c'est moi qui[PRO:rel] ai dit ça

Où est considéré comme un pronom interrogatif ou comme un relatif :

l'époque où[PRO:rel] les bêtes parlaient
où[PRO:int] vas-tu ?
je sais où[PRO:int] tu vas

Quand, lui, peut être un pronom interrogatif ou une Conjonction de subordination (KON) :

Quand[KON] il pleut, il mouille
je sais quand[PRO:int] il arrive

Quel est déterminant interrogatif (noté DET:int) dans les interrogatives devant NOM, DET:ind dans les exclamatives devant NOM, et interrogatif (PRO:int) dans les autres cas :

je sais quelle[DET:int] tasse il veut
quelle[DET:ind] audace !
Quel[PRO:int] est cette odeur ?
quel[PRO:int] que soit son talent

Les formes en *n'importe* sont analysées comme PRO ou DET composés. Ce sont des formes indéfinies, ni relatives ni interrogatives :

n'importe qui[PRO:ind]
n'importe quel[DET:ind]
n'importe lequel[PRO:ind]
n'importe quand[PRO:ind]

Lequel, quand à lui, peut être un PRO relatif ou interrogatif :

L'homme pour lequel[PRO:rel] je t'ai quitté.
Lequel[PRO:int] veux-tu ?

Quoi est rare comme PRO relatif (après Préposition), plus fréquent comme PRO interrogatif :

A quoi[PRO:int] penses-tu ?
je cherche un outil avec quoi[PRO:rel] ouvrir cette boîte
quoi[PRO:int] que tu fasses

Attention, à l'oral, à bien le distinguer de la particule discursive *quoi* notée INT.

Les pronoms introduisant des « relatives » sans antécédent doivent être étiquetés PRO:rel :

je pense à qui[PRO:rel] tu penses
j'aime qui[PRO:rel] tu sais
je vais où[PRO:rel] tu vas

5.9. Les pronoms démonstratifs et indéfinis

En ce qui concerne les pronoms indéfinis, il s'agit des unités telles que *plusieurs, certains, tout, etc.* employées sans nom, en position de sujet ou de complément. D'un point de vue syntaxique ou phonologique, il s'agit également de pronoms toniques. Du coup, nous faisons comme si le trait tonique était sous-entendu.

Pour ce qui est des pronoms démonstratifs, cela concerne uniquement *ce, celui, ceux, celle* et *celles* dans des contextes tels que les suivants :

ce[PRO:dem] qui m'énerve
celui[PRO:dem] que j'ai vu hier

De même, en combinaison avec *-ci* ou *-là*, l'ensemble est considéré comme un PRO:dem.

5.10. Les mots démonstratifs

Ils sont PRO, ADV ou DET et sont susceptibles de recevoir la sous-catégorie DEM dans certains cas. Nous avons choisi de privilégier les sous-catégories clitique et tonique au détriment de démonstratif (cf. ci-dessus pour *ce* et *celui*).

ça[PRO:cls] me dérange
ce[DET:dem] truc me dérange
ceci, cela: toujours [PRO:ton]
celui-ci, celle-là etc. : toujours [PRO:dem]

En revanche, les particules *-ci* et *-là* collées à un nom sont ADV :

cette fois-ci[ADV]

5.11. Les mots négatifs

Ils sont PRO, ADV, ou DET.

n' et *ne* sont analysés comme des ADV. *Non* est majoritairement une FNO mais peut également être KON lorsqu'il est collé à *pas* (*non pas*). Dans ce cas, *non pas* forme une locution. Les forclusifs sont soit des ADV (*pas, plus, guère, jamais*) soit des PRO:ind (*rien, personne, aucun, nul*) soit des DET (*nul, aucun*) :

je ne[ADV] vois personne[PRO:ind]
Personne[PRO:ind] n'[ADV] est venu
Jamais[ADV] Paul ne[ADV] viendra
je n'[ADV] en vois aucun[PRO:ind]
je ne[ADV] vois aucun[DET:ind] chien

Personne en emploi négatif est PRO:ind, sinon c'est un NOM (*une personne*).

Rien est PRO:ind.

tu n'[ADV] as rien[PRO:ind] vu à Hiroshima

Nul peut être PRO, DET ou ADJ :

une partie nulle[ADJ]
Nul[PRO:ind] n'[ADV] est censé ignorer la loi
sans nul[DET:ind] doute

Aucun(e) est DET ou PRO :

Aucun[PRO:ind] ne[ADV] viendra
sans aucun[DET:ind] doute

ne pas, ne plus, ne jamais, placés devant un infinitif, sont notés de la même façon, à savoir comme une suite de deux ADV.

5.12. Les mots indéfinis

Ils sont PRO, ADJ, ADV ou DET et ont la sous-catégorie IND.

tout(e)(s) : voir supra.

Adjectifs indéfinis : *seul, divers, différents* : voir infra.

Autre(s) est toujours un ADJ :

un autre[ADJ] homme
D'autres[ADJ] viendront
j'en vois un autre[ADJ]
Les uns les autres[ADJ]

Certain(e)(s) est soit PRO (employé comme tête de groupe nominal), soit ADJ (attribut ou épithète) soit DET :

Certains[PRO:ind] pensent que ...
Certaines[DET:ind] femmes
une certaine[ADJ] femme
je suis certaine[ADJ] que ...

Quelque(s) est soit DET soit ADJ. *Quelque* (au singulier) peut également être un ADV.

Quelques[DET:ind] restes
ces quelques[ADJ] fleurs
quelques-uns[PRO:ind]
quelque[ADV] trois millions de francs

Quelqu'un, quelque chose : toujours [PRO:ind] composé, sauf emploi particulier avec déterminant => NOM :

un petit quelque chose[NOM]

5.13. Les mots possessifs

On doit choisir entre les catégorie DET, PRO, ADJ. Ils ont la sous-catégorie possessif. Les formes *mon, ton, son* etc. sont toujours DET:pos.

Pour *leur* : voir supra.

Leurs est soit DET soit PRO:pos (après Déterminant défini).

On a regardé si *leurs* commute avec un autre Déterminant (*mon, ton*) ou un autre Pronom (*sien, tien*). Après un déterminant défini, il est Pronom.

Leurs[DET:pos] premières armes
Je me souviens des leurs[PRO:pos]

Les formes *mien, tien, siennes* sont PRO:pos (après DET défini)

le mien[PRO:pos], le nôtre[PRO:pos]

5.14. Les quantifieurs : *beaucoup, trop, peu, assez, bien, tant, tellement, moins*

Ils sont généralement analysés comme des adverbes (*dormir beaucoup, beaucoup de gens*) mais ils peuvent également être des pronoms :

Beaucoup[ADV] de gens
Beaucoup[ADV] de ces gens

j'en veux beaucoup[ADV] (de fleurs)
il aime beaucoup[ADV] la musique
ce que j'aime le moins[ADV]
beaucoup[PRO:ind] considèrent que...
Ils sont beaucoup[PRO:ind]

En ce qui concerne *peu*, nous avons choisi de le noter NOM lorsqu'il est précédé d'un déterminant car il peut prendre un adjectif (*un tout petit peu de*…). Il peut également être ADV ou PRO dans certains contextes.

le peu[NOM] de vin qu'il boit ne peut lui faire du mal.
Il a bu un peu[NOM] de bière.
Il a bu peu[ADV] de vin.
Il est un peu[NOM] malade.
très peu[ADV] d'affaires
très peu[PRO:ind] sont diffusés
il viendra sous peu[ADV]

5.15. Les mots étrangers

Les mots d'origine étrangère utilisés en français reçoivent une étiquette normale. En revanche, nous étiquetons « ETR » les mots étrangers qui ne sont pas entrés dans la langue française et qui sont employés dans un contexte syntaxique peu clair. A chaque fois que l'étiquette normale est transparente, ils doivent être notés de la même façon qu'un mot français.

Un match de football[NOM]
Un building[NOM]
Il m'a dit shut up[ETR] !

5.16. Les nombres

Les nombres (cardinaux) sont étiquetés à l'aide de NUM et leur lemme est @card@ (en suivant en cela l'étiquetage par défaut de Treetagger).

Les nombres sont écrits en lettres (vingt-sept) ou en chiffres (27). Ils ont normalement été regroupés (100 000, quatre-vingt-sept). Ils sont systématiquement considérés comme des NUM.

En revanche, *premier, deuxième*, etc. sont considérés comme des ADJ (excepté dans les dates, les différentes vitesses et les classes de lycée).

Attention : TreeTagger ne tolérant pas de chiffres dans son lexique, tous les chiffres ont été réécrits en toutes lettres.

5.17. Les dates

Les noms de jour et de mois sont NOM (même avec une majuscule), les nombres sont NUM. *Demain, aujourd'hui, hier* sont des ADV.

Un jour, *ce matin* ne sont pas regroupés en adverbe composé, de même que *cette fois*, *une fois* (en revanche *des fois*, synonyme de *parfois*, est regroupé et considéré comme un ADV) :

une fois[NOM] Jean arrivé
le 1er[NUM] juillet[NOM] 1924[NUM]
le deux[NUM] est un dimanche[NOM].
les années[NOM] 80[NUM]
courant[PRP] décembre[NOM]

5.18. Les heures

Les heures notées en toutes lettres sont décomposées. Les heures notées en chiffres sont regroupées (2h15, 3h15mn13s, 3h15mn13'20'').

Les noms *demi-heure* et *quart d'heure* sont regroupés.

Deux[NUM] heures[NOM] et[KON] demie[NOM] plus tard il revenait
Vers 2h30[NUM]
ça fait une demi-heure[NOM] que je l'attends

5.19. Les consonnes épenthétiques

Nous notons le *-t-* euphonique accompagné de l'éventuel tiret collé au pronom ainsi que le *l'* collé au clitique *on* comme des consonnes épenthétiques (EPE).

le livre que l'[EPE] on m'a donné
que mangera -t-[EPE] on en 2050

6. Les mots les plus difficiles

On indique ici les principes de choix entre catégories pour des mots grammaticaux très fréquents.

6.1. CE

C' est toujours un clitique sujet (PRO:cls) :

c'[PRO:cls] est la vie
c'[PRO:cls] était un brave

On doit choisir pour *ce* entre DET:dem, PRO:cls, PRO:clo, KON et PRO:rel et savoir s'il forme une entité à part entière ou s'il doit être regroupé avec un autre mot.

Il est DET:dem devant NOM quand il commute avec un autre article, sa forme devant voyelle est *cet* :

devant ce[DET:dem] spectacle
selon ce[DET:dem] dernier

Il est PRO:cls devant un autre clitique (*ne*) ou un verbe conjugué (*être*) et PRO:clo devant un participe présent (*disant, faisant*) :

ce[PRO:cls] serait bien
ce[PRO:cls] serait pas plus mal
ce[PRO:clo] faisant

Il est considéré comme un PRO:dem dans les relatives et comme une KON composée dans les complétives. En tant qu'articulateur après *et*, nous considérons qu'il s'agit également d'un PRO:dem :

ce[PRO:dem] qui serait bien
il cherche à ce que[KON] tu viennes
je suis parti, et ce[PRO:dem] pour ne pas m'énerver

6.2. COMME

Comme est toujours étiqueté KON. Même si nous sommes conscients que cette notation est fort réductrice, les distinctions proposées habituellement nous ont semblé trop complexes à mettre en oeuvre pour faire l'objet d'une annotation systématique.

Comme[KON] il faisait beau, je suis sortie

6.3. DE- D' – DU - DES

6.3.1. DE-D'

De (ou *d'*) peut être préposition ou déterminant. Il peut former un déterminant simple ou complexe (*de la*, avec un déterminant défini qui suit). Pour trancher, il faut voir si *de* commute avec un déterminant ou avec une préposition.

Le déterminant *de* (ou *d'* singulier) peut être indéfini (DET:ind) ou partitif (DET:par).

De est déterminant indéfini ou partitif après une préposition (sauf cas de préposition complexe), après un verbe transitif direct, comme introducteur d'un groupe nominal sujet ou encore dans certains cas où il est précédé par *pas/plus*. Le lemme est alors *un*.

vous n'avez plus de[DET:ind] soucis
il mange de[DET:ind] délicieux fruits exotiques

De fait partie d'un déterminant composé partitif *de l'* ou *de la* :

Gagner de l'[DET:par] argent
Je veux de la[DET:par] farine

De est préposition avant un verbe à l'infinitif, avant ou après un pronom, avant un déterminant, après un verbe non transitif direct, comme introducteur de complément de nom ou d'adjectif, comme introducteur d'un complément circonstanciel et après certaines prépositions complexes :

Je viens de[PRP] finir
Il essaie de[PRP] les encourager
J'en ai une de[PRP] cassée
Il travaille de[PRP] ses mains
Je rêve de[PRP] beaux enfants
La maison de[PRP] Paul
Il est fier de[PRP] lui
Il vient de[PRP] chez lui
Une pomme de[PRP] trop
J'ai un fils de[PRP] malade

De est également considéré comme une préposition simple dans des cas comme les suivants :

En dépit[ADV] de[PRP] sérieuses difficultés
Au-delà[ADV] de[PRP] la mer

6.3.2. DU- DES

Tous deux peuvent être agglutinés PRP:det(*du=de le, des=de les*). Le lemme est alors *du*.

Il tombe du[PRP:det] train (on peut dire : *de ce train*)

Ou bien on a le déterminant composé partitif *du* (lemme=*du*).

Jean mange du[DET:par] gâteau
Je veux du[DET:par] sel

Ou enfin on a le déterminant indéfini des (lemme=*un*).

Malgré des[DET:ind] difficultés énormes

6.4. EN

Pour *en*, on doit choisir entre PRP et PRO:clo. *En* est PRP quand il introduit un groupe prépositionnel circonstanciel, complément de V, de N... Il est PRO:CLO devant un V (sauf participe présent, mis à part cas exceptionnels) et PRP partout ailleurs :

En[PRP] le voyant
Une bague en[PRP] or
Paul en[PRO:clo] parlera demain

6.5. LE – LA – LES - L'

Ambigus entre DET:def et PRO:clo. Ils sont clitiques objet devant V, *ne* ou devant un autre pronom clitique, ou après un V impératif. Ils sont DET:def partout ailleurs (devant NOM ou ADJ...) :

J'essaie de les[PRO:clo] encourager
En les[PRO:clo] voyant
Les[DET:def] trois brigands

Attention : il faut penser aux déterminants partitifs complexes pour lesquels c'est l'ensemble qui constitue le déterminant et non *la* tout seul.

|De la|[DET:par] salade

6.6. LEUR

On doit choisir entre DET:pos, PRO:clo et PRO:pos. Pour trancher, on a regardé s'il commute avec un autre déterminant, un autre clitique (*lui*) ou un autre pronom possessif (*sien, son*). *Leur* est DET:pos devant ADJ ou NOM. Leur est PRO:clo devant V, *ne* ou un autre clitique (*y, en*) et après V impératif.

Leur[DET:pos] enfant est malade
Il faut tenir compte de leur[DET:pos] performance
Je leur[PRO:clo] parlerai
Je préfère le leur[PRO:pos]

6.7. LUI

On doit essentiellement choisir entre PRO:clo et PRO:ton. Pour savoir si c'est le clitique ou le tonique, essayer la mise au pluriel (PRO:clo => *leur*, PRO:ton => *eux*). *Lui* est clitique devant V, *ne* ou un autre clitique (*en, y*), clitique aussi après un V impératif, tonique dans les autres cas (après PRP, avant relative) :

Je lui[PRO:clo] en parlerai
Je travaille pour lui[PRO:ton]

6.8. MEME(S)

On doit choisir entre ADJ et ADV.

Même est ADJ comme épithète antéposé ou avec ellipse de nom :

le même[ADJ] homme
ce sont les mêmes[ADJ]

Même est ADV (invariable) devant ADJ, ADV, KON, PRP etc. et en prédéterminant :

même[ADV] mort
même[ADV] avec des augmentations
même[ADV] les chiens sont plus propres
même[ADV] pas de blé

Attention : pour *même si* nous avons décidé de procéder systématiquement au regroupement.

même si[KON] j'ai tort…

Même en suffixe ou en épithète postposé est ADV sauf *moi-même, lui-même* etc. qui sont des PRO composés :

aujourd'hui:ADV même[ADV]
cet homme[NOM] -même[ADV] qui ...
j'y vais moi-même[PRO:ton]

6.9. VOICI - VOILA

Les mots *voici* et *voilà* ne sont PRP que comme tête de groupe prépositionnel (circonstanciel). Comme tête de phrase (principale ou enchâssée), ils sont VER et à ce titre peuvent prendre un Clitique complément :

Nous voici[VER] !
je pense que voici[VER] une occasion unique
je suis parti en Egypte voilà[PRP] trois ans

Mais majoritairement, *voilà* se comporte comme une FNO à l'oral :

Il a fait ça voilà[FNO]

6.10. PLUS

On doit choisir entre adverbe et marque de l'addition (KON). On ne distingue pas cet usage de celui de la négation. L'étiquette est la même.

Il mange plus[ADV] que toi
Le plus[ADV] grand des trois
le succès le plus[ADV] grand
deux plus[KON] deux égalent quatre

6.11. QUE- QU'

Que est ambigu entre Adverbe, Pronom relatif (PRO:rel) ou interrogatif (PRO:int) et Conjonction de subordination (KON).

Que est pronom introduisant une relative (PRO:rel) ou une interrogative directe (PRO:int). Etant donné la difficulté inhérente à la notion d'interrogative indirecte, nous avons préféré considérer qu'il s'agissait systématiquement d'un PRO:rel. Il est également pronom relatif dans les clivées :

la fille que[PRO:rel] je vois
je demande ce que[PRO:int] tu vois
Que[PRO:int] vois-tu ?
C'est Jean que[PRO:rel] je vois
quoi[PRO:int] qu'[PRO:rel] en pensent les collègues

Que est ADV dans *(ne)...que* (tour restrictif) :

Je ne vois que[ADV] toi
Ils n'[ADV] ont plus[ADV] que[ADV] la haine

Que est ADV dans les exclamatives :

Que[ADV] de bonbons !

Que est conjonction de subordination après un verbe (ou un nom ou un adjectif) à complétive, après une PRP, dans les comparatives ou les corrélatives (mêmes réduites) et dans les impératives.

Je veux que[KON] tu viennes
Comme il pleuvait et que[KON] tu venais ...
plus grande que[KON] vous

6.12. S'

Forme élidée de *se* (PRO:clo) ou de *si* (KON).

Ils veulent s'[PRO:clo] amuser
Je me demande s'[KON] il a tort

6.13. SI

On doit choisir entre FNO (affirmation), Adverbe (intensif, pouvant introduire une consécutive) et Conjonction de subordination (introduisant une subordonnée hypothétique ou une interrogative indirecte).

Si est toujours KON lorsqu'il introduit une subordonnée. Il est ADV devant un ADJ ou un autre ADV susceptibles de degrés et FNO dans les autres cas.

il est si[ADV] gentil
Si[KON] tu viens
Je me demande si[KON] tu viendras
Il m'a répondu que si[FNO]

6.14. TEL(LE)(S)

On doit choisir entre DET, PRO et ADJ.

Tel est ADJ en épithète et en attribut (souvent avec inversion du sujet)

un tel[ADJ] homme
tel[ADJ] était cet homme
Il était tel[ADJ] que tu l'avais laissé

Il est KON devant un déterminant :

Tel[KON] un lion

Tel est DET devant un nom sans déterminant ou dans la suite *tel quel* :

Tel[DET:ind] père tel[DET:ind] fils

Tel est PRO comme tête de groupe nominal :

avec tel[PRO:ind] ou tel[PRO:ind] de vos partenaires

6.15. TOUT(E)(S) - TOUS

On doit choisir entre Déterminant, Pronom, Adverbe et Nom.

Tout(e) est Déterminant devant un NOM ou un ADJ en tête de groupe nominal :

Tout[DET:ind] homme a le droit de...
Toute[DET:ind] nouvelle étudiante se verra remettre...

Tou(t)(e)(s) est DET:pre en position de prédéterminant :

Tout[DET:pre] le monde
Toutes[DET:pre] les femmes
Tout[DET:pre] cela est faux

Il l'est également dans la suite *tout ça*.

Tou(t)(e)(s) est PRO:ind employé seul comme sujet ou complément et comme quantifieur flottant :

Tous[PRO:ind] viendront
Elles viendront toutes[PRO:ind]
J'ai tout[PRO:ind] lu
C'est tout[PRO:ind]

Tout(e)(s) est Adverbe devant un adjectif (ou un participe passé employé comme adjectif), un adverbe ou une préposition :

Il est tout[ADV] rouge
J'ai vu des filles toutes[ADV] rouges
Une toute[ADV] nouvelle étudiante

Tout peut être Nom commun précédé d'un déterminant (rare) :

Le tout[NOM] c'est de ne pas mentir

6.16. UN(E)(S)

DET, PRO ou NUM.

Quand *un(e)(s)* est employé seul après un déterminant, il est PRO :

Les uns[PRO] ont tort
Une[DET:ind] belle fille
le tome un[NUM]

Attention : la *une* (presse) est un NOM. Pour le cas des NUM, le problème est très complexe et nous n'avons pas encore trouvé de solution satisfaisante.

6.17. DIVERS-DIFFERENT

divers(e)(s) et *différents* peuvent être déterminant ou adjectif :

divers[DET:ind] conseils:NCmp
des aventures diverses[ADJ]
ces diverses[ADJ] péripéties

7. Les ambiguïtés les plus fréquentes

On note ici les cas les plus systématiques d'ambiguïtés potentielles entre catégories ou sous-catégories, qu'on a dû trancher en contexte.

7.1. Adjectif(ADJ) / Participe passé (VER:pper)

Après un auxiliaire de temps, au passif, avec un complément en *par* (ou un complément d'agent en *de*), on a un participe passé.

Quand on a les mêmes compléments que le verbe conjugué on a un participe passé.

Même en position épithète ou attribut, on conserve généralement la notation de participe passé quand la forme verbale existe.

Quand le préfixe est *in-* (*inconnu, insatisfait*...) c'est un adjectif.

Si l'on peut paraphraser avec une relative à l'actif, il s'agit du verbe au participe. Cependant, le sens doit être pris en compte en tant que paramètre désambiguïsateur.

Un jugement prononcé[VER:pper] (qu'on a prononcé)

*Un recul prononcé[ADJ] des prix (*qu'on a prononcé)*
Un parking privé[ADJ]
Un enfant privé[VER:pper] de dessert
Il a parlé[VER:pper]
Je me suis trompée[VER:pper]
une canalisation bouchée[VER:pper]

En cas de doute, utiliser l'étiquette de participe passé comme étiquette par défaut.

7.2. Adjectif (ADJ) / Participe présent(VER:ppre)

Le participe présent est invariable et garde les compléments du verbe.

L'adjectif s'accorde avec le nom et ne prend pas de complément d'objet direct.

En lisant[VER:ppre] le journal...
Les erreurs existantes[ADJ]
Un déséquilibre persistant[ADJ]

7.3. Adjectif (ADJ) / Nom commun (NOM)

On ne recatégorise pas les Adjectifs en cas d'ellipse du nom sauf changement de sens ou de morphologie (invariable en genre, seulement déterminant défini, etc.) :

une grande[ADJ] bleue[ADJ]
il a perdu le gauche[ADJ]
ce dernier[ADJ]
la gauche[NOM] a gagné
le rouge[NOM] est une belle couleur
je veux les deux[NUM]

Dans les tours superlatifs, on a l'étiquette ADJ :

l'homme le plus grand[ADJ]

le plus grand[ADJ] des hommes

On ne recatégorise pas les noms employés comme épithètes ou comme attributs :

une veste sport[NOM]
Il est très famille[NOM]

7.4. Adjectif (ADJ) / Adverbe (ADV)

On ne recatégorise pas les adjectifs en fonction adverbiale.

chanter faux, crier fort, tomber juste => ADJ.

haut, bas peuvent uniquement être ADJ ou NOM :

en haut[NOM] de
un meuble haut[ADJ]

fort et *juste* sont ADV seulement comme prémodifieurs d'un adjectif, d'un adverbe ou d'une préposition.

il était fort[ADV] triste
fort[ADV] justement
un homme fort[ADJ]
parler trop fort[ADJ]

bien, mal ne sont jamais adjectif, mais seulement ADV ou NOM :

Le bien[NOM]
Bien[ADV] des années plus tard

un homme bien[ADV]

sauf est adjectif ou préposition, mais jamais adverbe.

l'honneur est sauf[ADJ]
Tous les fruits sauf[PRP] les pommes

7.5. Préposition (PRP) / Adverbe (ADV)

Une préposition prend toujours un complément ; sans complément, elle est adverbe. Cette approche est critiquable et pas très économique, mais elle a l'avantage d'introduire une distinction supplémentaire et elle n'est pas sujette à ambiguïté.

je vote pour[ADV]
je vote pour[PRP] la gauche

durant est toujours préposition même quand son complément est antéposé :

durant[PRP] trois heures
trois heures durant[PRP]

Pour les autres Préposition de temps, on peut avoir un prémodifieur nominal compatible avec un complément après la Préposition. En l'absence de ce dernier, l'étiquette est ADV :

trois heures avant[PRP] la fin
trois heures avant[ADV]

Voici, voilà sont PRP seulement comme tête de complément circonstanciel, sinon Verbe ou FNO :

nous voici[VER] en pleine tragédie
son arrivée voici[PRP] trois ans
Voilà[VER] sans doute le seul sens donné à notre existence.

Il y a est PRP comme tête de complément circonstanciel. Dans les autres cas, il est nécessaire de le décomposer :

ils sont partis il y a[PRP] trois ans
il[PRO:clsi] y[PRO:clo] a[VER:pres] trois ans qu'ils sont partis

7.6. Préfixes / Adverbe (ADV)

Avec les mots du type *méta-, anti-, franco-, auto-, ex-, super-* ainsi qu'une liste ouverte de termes, il n'est pas toujours aisé de savoir si l'on a affaire à un adverbe ou à un préfixe, et parfois même à un adjectif (comme *super* par exemple). Le préfixe est toujours un composant d'un terme lexical.

l' |ex-Union soviétique|[NAM]
l'ex-mari[NOM] de son client
Cet appareil est auto-régulé[VER:pper]

quasi est ambigu entre préfixe et ADV indépendant. Il est préfixe devant un nom, adverbe devant un adjectif ou un adverbe.

La quasi-totalité[NOM]
Une situation quasi-[ADV] parfaite

7.7. Conjonctions [KON] / Adverbe [ADV]

Nous ne faisons pas de distinction entre conjonctions de subordination et de coordination. Les deux sont notées [KON]. Cependant, nous distinguons des unités non mobiles comme

puis qui sont notées KON et des unités mobiles comme *donc* qui sont notées ADV. *Donc* peut toutefois être considéré comme une KON dans certains cas très restreints :

Il est donc[ADV] venu *donc[ADV] il viendra* *il est honnête donc[KON] pauvre*

soit est aussi une KON.

Sinon est un ADV car il est mobile.

aide-moi, sinon va faire tes devoirs

7.8. Noms communs / Noms propres

Plusieurs critères coexistent: la graphie (majuscule aux noms propres) , la syntaxe (Déterminant devant les noms communs), la sémantique (noms propres comme désignateurs rigides = référent unique).

Nous avons privilégié la sémantique, c'est-à-dire que nous admettons des noms communs avec majuscules et des noms propres avec Déterminant.

Prénoms, Noms: NAM

Les Dupont[NAM] *Faire un dessin à la Picasso[NAM]*

Titres : NOM

Mr[NOM] le Président[NOM] Chirac[NAM] *le Pape[NOM]*

Noms de société, d'institution unique, de parti politique, de même que la plupart des sigles : NAM

I.B.M.[NAM]

Les sigles sont NOM ou NAM selon les mêmes critères :

la S.N.C.F.[NAM] *un H.L.M.[NOM]* *les P.M.E.[NOM]*

Noms de langue, de nationalité (sauf épithète ou attribut => ADJ) ou de devises : NOM

Il est français[ADJ] *Il parle français[NOM]* *Les français[NOM] votent à gauche*

7.9. Adjectifs qualificatifs ou indéfinis ?

Nous ne distinguons pas ces deux sous-catégories, comme illustré ci-dessous.

un homme seul[ADJ] *un seul[ADJ] homme* *seule[ADV] cette femme* *avec pour seul[ADJ] but de l'entendre*