HAL open science

Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles)

Simon Gabay, Jean-Baptiste Camps, Thibault Clérice

To cite this version:

Simon Gabay, Jean-Baptiste Camps, Thibault Clérice. Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles). 2020. hal-02571190v1

HAL Id: hal-02571190

https://hal.science/hal-02571190v1

Submitted on 12 May 2020 (v1), last revised 18 Apr 2022 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Manuel d'annotation linguistique pour le français moderne (XVIe-XVIIIe siècles)

Simon Gabay[1], Jean-Baptiste Camps[2] et Thibault Clérice[2]

[1]Universités de Neuchâtel et Genève
[2]Centre Jean Mabillon (EA 3624), École nationale des chartes, Université PSL

Version A (« Absidale Argélie »)
12 mai 2020

Sommaire

1 Segmentation en mots (*tokenisation*) 3

2 Lemmatisation 4

3 Étiquetage morpho-syntaxique (parties du discours) 7

4 Morphologie 11

5 Entités nommées 12

Annexes 14

A Exemple d'annotation 14

B Abréviations pour le tokeniseur (extrait) 15

C Jeu d'étiquette CATTEX 16

Références 18

Lassé de Mars j'aspire aux douceurs de la paix.
J'habite l'Austrasie aux bois les plus épais.
Là je consacre un temple à ce puissant Mercure,
Qui m'ouvre les clartez d'une science obscure :
Qui m'apprend loin du bruit les secrets curieux
Des enfers, de la mer, de la terre, et des cieux

DESMARETS, *Clovis*, 1657

La conversation finirait mal, ne l'entamons point, tirons nos chausses.

DANCOURT, *Les Vacances*, 1697

À plus !

Cri de guerre de la maison de Rohan

Introduction

Ce manuel a pour objectif d'accompagner la constitution de corpus d'entraînement pour la lemmatisation, l'étiquetage morpho-syntaxique et morphologique, et la reconnaissance d'entités nommées. Il s'intéresse principalement aux problème soulevés par les textes français modernes, de la Renaissance aux Lumières. Il propose des règles simples et s'appuie sur des standards nationaux (par ex. CATTEX ou *LGeRM*) et internationaux (par ex. le format BIO) pour permettre la production de données du type de celles présentées en table 1.

Token	Lemme	POS	Morphologie	Entités
Michel	Michel	NOMPro	NOMB.=s|GENRE=m	PER-B
annote	annoter	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s	0
les	le	DETdef	NOMB.=p|GENRE=m	0
textes	texte	NOMcom	NOMB.=p|GENRE=m	0

Table 1 – Exemple d'annotation

Son objectif est d'encoder des états de langues anciens sur la longue durée. Il doit donc tenir compte des particularités de ces états de langue, et assurer une interopérabilité minimale entre différentes époques tout en conservant une annotation assez riche pour permettre des requêtes fines.

1 Segmentation en mots (*tokenisation*)

La segmentation en mots obéit à quelques règles simples. Sauf erreur manifeste, qui doit être corrigée :

- Les tokens sont toutes les chaînes de caractères séparées par des espaces et des retours chariot en l'absence de tiret de fin de ligne.
- La non agglutination (*ce pendant* vs *cependant*) n'est pas corrigée, et les locutions ne sont pas regroupées (*bien que* =2 tokens).
- Les agglutinations ne sont pas non plus corrigées (*tresobeissant*, *parce* =1 token).
- Les signes de ponctuation (point, virgule...) et équivalents (tiret, cadratin...) constituent, sauf exception (apostrophe, cas spéciaux mentionnés *infra*), un token.

Ces choix radicaux ont pour objectif pratique de simplifier la tokenisation d'une part, et pour objectif linguistique de simplifier l'analyse diachronique en évitant un quelconque parti pris sur les phénomènes de figement et d'agglutination, qui sont difficiles à dater précisément. Revers de la médaille : l'analyse grammaticale n'est en revanche pas toujours aisée, notamment pour les locutions – dont nous avons multiplié les exemples dans ce document afin d'aider l'utilisateur.

- Les locutions adverbiales (*à gauche*) ne forment pas un token mais deux (`à PRE` + `gauche NOMcom`)
- Tant qu'une espace graphique est maintenu (loc. *ce pendant*) nous conservons plusieurs tokens (`ce PROdem` + `pendre VERppa`)

Concernant l'apostrophe, elle n'est pas considérée comme indépendante et se trouve rattachée au token élidé (*c'est* → `c'` + `est`, *quelqu'un* → `quelqu'` + `un`, *l'on* → `l'` +

on). Quelques rarissimes cas d'élisions internes au mot peuvent exister pour des raisons stylistiques, surtout en littérature contemporaine, comme *V'la c'te voiture* : ils doivent être maintenus en un seul token quand cela est possible.

Concernant le trait d'union, la situation est plus complexe. Il est considéré comme un token à part entière (*beau-frère* → `beau + - + frère`) sauf dans les cas où il est suivi :

— d'un pronom personnel (*-je*, *-tu*, *-il*, *-elle*, *-on*, *-nous*, *-vous*, *-ils*, *-elles*)
— d'un pronom (régime) tonique (*-moi*, *-toi*, *-lui*, *-eux*)
— d'un pronom (régime) atone (*-me*, *-m'*, *-te*, *-t'*, *le*, *la*, *-les*, *-leur*, *-leurs*)
— d'un pronom démonstratif (*-ce*)
— d'un pronom adverbial (*-en*, *-y*)
— d'un pronom indéfini (*-un*, *-uns*, *-une*, *-unes*)
— de certains adverbes (*-ci*, *-là*, *-aussi*)
— d'un adjectif indéfini (*-même(s)*)

Attention au cas particulier du *t* euphonique (*-t-*) uni avec le verbe qui précède : *a-t-il* est ainsi segmenté `a-t / -il`.

Nous considérons en effet que, dans les cas précédemment mentionnés, la présence du trait d'union est un phénomène syntaxique (par exp. la postposition) et non lexical (*grand-mère*). Il en va logiquement de même pour les noms propres, qui restent en plusieurs tokens (*Jean-Baptiste*, *Aulu-Gelle*, *Saint-Étienne-du-Mont*...).

Concernant les abréviations, il est impératif de distinguer les points qui terminent les phrases de ceux terminant une abréviation. Ainsi, dans la phrase *C'est D. Juan.*, le premier point indique que *D* est la forme abrégée de *Don*, tandis que le second marque la fermeture de la phrase. Une liste (évidemment non-exhaustive) des abréviations possibles a été définie en annexe (cf. annexe B), à laquelle il convient d'ajouter les lettres uniques suivies d'un point, qui peuvent indiquer le prénom (*D.* pour *Damien*), les points cardinaux (*N.* pour *nord*) ou des mots précis (*M.* pour *monsieur*, *S.* pour *saint*, *P.* pour *père*, *v.* pour *voir*). Il est à noter que dans certains cas des interférences sont inévitables, car *lit.* peut autant être la formée abrégée de *littérature* que le substantif *lit* en fin de phrase (*Il va au lit.*). Le point qui marque l'abréviation n'est pas considéré comme un token et reste avec le mot abrégé (*M. le prince* → `M. le prince` et non `M . le prince`).

2 Lemmatisation

Le choix du lemme n'est pas libre : il doit se trouver dans le référentiel qui dérive de *LGeRM MODE* (*i.e. Moderne étendu*) [1]. Ce choix permet de garantir une interopérabilité minimum en amont avec *LGeRM AF* (*i.e. Ancien Français* [2]) et en aval avec le *TLFi* [3]. La lemmatisation obéit à quelques grandes règles simples :

1. Sascha Diwersy, Achille Falaise, Marie-Hélène Lay et Gilles Souvay, « Ressources et méthodes pour l'analyse diachronique », *Langages*, N° 206–2 (août 2017), p. 21-44, URL : `https://www.cairn.info/revue-langages-2017-2-page-21.htm` (visité le 03/12/2018).

2. G. Souvay et Jean-Marie Pierrel, « LGeRM Lemmatisation des mots en Moyen Français », *Traitement Automatique des Langues*, 50–2 (2009), p. 149-172, URL : `https://halshs.archives-ouvertes.fr/halshs-00396452`.

3. J.M. Pierrel, Jacques Dendien et Pascale Bernard, « Le TLFi ou Trésor de la Langue Française informatisé », dans *Proceedings of the 11th EURALEX International Congress*, dir. Geoffrey Williams et Sandra Vessier, Lorient, France, 2004, p. 165-170.

— Le lemme est, autant que possible, la forme contemporaine du mot et non une ancienne forme (*avecque* → `avec`), le masculin singulier pour les adjectifs et les substantifs, l'infinitif pour les verbes.
— Dans le cas où il existe une forme masculine et féminine d'un même mot, le lemme est dans la très grande majorité des cas la forme masculine pour les substantifs (*comtesse* → `comte`) comme pour les adjectifs (*grande* → `grand`). Ce n'est en revanche pas le cas pour les noms propres (*Jeanne* → `Jeanne` et non `Jean`). Il existe quelques exceptions qui possèdent deux lemmes différents (*dame* vs ancien français *don*) malgré une racine commune (<*dominus*, *a*) pour conserver un interopérabilité dans le temps (*don* ayant disparu en français).
— L'existence d'une entrée dans un dictionnaire historique (préférablement celui de Furetière, mais aussi ceux de Richelet ou même de l'Académie) ou dans un dictionnaire scientifique (comme l'*Altfranzösisches Wörterbuch* de Tobler et Lommatzsch ou le *Dictionnaire du Moyen Français*) est le principal critère pour l'ajout d'un nouveau lemme dans le référentiel.
— Certains tokens posent problème :
  — Ils ne sont pas analysable hors contexte, comme *parce* (→ dans *parce que*), *afin* (dans *afin que*), ou *ledit*. Nous créons alors un lemme (`parce`, `afin`, `ledit`).
  — Ce sont des enclises (*du*, *des*, *au*, *dudit*, *auquel*) : les deux lemmes originels sont alors conservés et séparés par un tiret bas (`de_le` ou `à_le`).
  — Ce sont des mots-valises (*tresobeissant*...) : nous utilisons la même méthode que pour les enclises (`tres_obeissant`...).
— Comme pour les autres tokens, nous considérons que le lemme des noms propres est leur forme contemporaine. Nous privilégions des formes communes malgré des variations diachroniques (*Jehanne* → `Jeanne`) ou graphiques à la marge (*Denys* → `Denis`, *Remus* → `Rémus`...). Si la forme est clairement dans la langue étrangère, le lemme est alors dans cette langue (*Jan* → `Jan`...et non `Jean`, *Vespasianus* → `Vespasianus`...et non `Vespasien`, *Demosthenes* → `Demosthenes` et non `Démosthène`).
— Nous considérons comme nom propre un token qui commence par une majuscule : ainsi *la mer* ***Noire*** → `le mer Noir`, *madame de* ***Sévigné*** → `madame de Sévigné` (mais *madame* ***De Sévigné*** → `madame De Sévigné`). L'annotation des entités au format BIO permet de capturer des ensembles plus larges, contenant titres, prénoms, etc. (voir sect. 5).
— Le lemme retenu pour les pronoms personnels (sujet, réfléchi, objet direct, indirect ou disjoint) est la forme du pronom personnel sujet (*moi* → `je`), le cas échéant singulier (*eux* → `il`) masculin (*elles* → `il`). L'annotation morphologique (voir sect. 4) fournit le complément d'information nécessaire pour des requêtes fines.
— Nous ne connaissons pas l'adjectivation des participes, car le passage d'une catégorie à l'autre, surtout en diachronie, est beaucoup trop difficile à identifier précisément. Si un infinitif existe, nous considérons donc qu'il s'agit d'un participe et l'infinitif sert de lemme (*le retour éclatant* → `éclater`, *une âme affligée* → `affliger`). Notons que la présence d'une marque de flexion (*des traits charmants*) n'est pas considérée comme une raison suffisante pour en faire une forme autonome (*charmants* → `charmer` et non `charmant`).
— Le lemme d'un chiffre en toutes lettres est ce chiffre en toutes lettres (*quatre* →

`quatre` et non `4`). Le lemme d'un nombre en chiffres arabes ou romains écrit (par ex. *XII*, *12* → `12`) est la version arabe du chiffre.

— Pour des langues étrangères, en accord avec les principes éditoriaux que nous exposons ici, on privilégiera l'infinitif pour le verbe, le singulier masculin (nominatif) pour les substantifs ou adjectifs. On aura de préférence recours à un référentiel existant (par exemple, pour le latin, un dérivé de *Forcellini*[4]), ou à défaut d'un dictionnaire faisant autorité (comme le *DMF*[5] pour le moyen français).
— On respecte les ligatures pour les lemmes, peu importe si elles sont dans l'occurrence analysée ou pas (`nœud`, `vœu`...et non `noeud`, `voeu`).
— Les majuscules sont accentuées (`Etiopie` → `Éthiopie`).
— Les lemme des verbes pronominaux ne contiennent pas le pronom, car cela serait redondant avec le pronom qui est tokenisé à part (`s'enfuir` → `se` + `enfuir`), et que cela permet de ranger sous une même étiquette différents emplois d'un même verbe (emploi transitif vs pronominal d'*abaisser*).
— Dans le cas où il existe deux formes très proches dont l'une s'est imposée, notamment pour les états de langue les plus anciens (par exp. *finablement* vs *finalement*), on conserve bien deux lemmes distincts (en l'occurrence `finablement` et `finalement`).
— Pour les mots abrégés, nous ne développons le token que s'il s'agit d'un substantif (*M.* → `monsieur`), mais pas s'il s'agit d'un nom propre (*le Père* ***R.*** avec *R.* pour *Rapin* → `le père R.` et non `le père Rapin`).

Quelques exemples pour le débutant

Des exemples simples pour commencer :
— *je vais à Genève* → `je aller à Genève`
— *un sort assez propice* → `un sort assez propice`
— *il m'a sauvé toutefois des ravages du temps* → `il je avoir sauver toutefois de_le ravage de_le temps`
— *c'est que Vespasian me regardait pour lui* → `ce être que Vespasien je regarder pour il`
— *vous sentez-vous impropre au* ***matrimonium*** → `matrimonium`
— *Lettre 12. 16 septembre 1676. À M. R.* → `lettre 12 . 16 septembre 1676 . à monsieur R.`

Il est essentiel de faire attention aux homographes :
— Dans *soit…soit* ou *tu le veux?* ***soit*** le lemme est `soit`. En revanche, dans *ainsi* ***soit****-il* il s'agit évidemment du verbe `être`.
— Dans *il ne* ***leur*** *manquera rien* il s'agit du pronom `il` mais dans ***leur*** *vif éclat* il s'agit du déterminant possessif `leur`.
— Dans ***la*** *fin* il s'agit du déterminant `le` mais dans *je* ***la*** *vois* il s'agit du pronom `il`.

Attention aussi aux lemmes composés :
— *dudit* → `de_ledit`

4. Thibault Clérice, *Référentiel du Latin pour Pyrrha, d'après le dictionnaire et travaux du LASLA de D. Longrée et al*, mai 2020, DOI : `10.5281/zenodo.3822040`.
5. ATILF-CNRS et Université de Lorraine, *Dictionnaire du Moyen Français (1330-1500)*, 2015, URL : `http://www.atilf.fr/dmf`.

— *au*(*x*) → `à_le`
— *duquel* → `de_lequel`
— *auxquels* → `à_lequel`
— *du* → `de_le`
— *des* → `de_le` (ou `un`, comme dans `des beaux garçons sont arrivés` : attention au contexte !)

Les entités nommées sont complexes à gérer :
— *don Carlos* → `don Carlos`
— *la mer Rouge* → `le mer Rouge`
— *Monsieur de La Rochefoucault* → `monsieur de La Rochefoucauld`
— *François, duc d'Enghien* → `François , duc de Enghien`
— *M. Du Plessis* → `monsieur Du Plessis` (et non `de_le` pour *Du*, qui a d'ailleurs une majuscule en français contemporain).
— *Denys d'Halycarnasse* → `Denis de Halicarnasse`

3 Étiquetage morpho-syntaxique (parties du discours)

Nous reprenons le jeu d'étiquettes *CATTEX-max*[6] (qui inclut la morphologie, sur laquelle nous revenons *infra*) et dont nous rappelons les étiquettes en annexe (voir annexe C). Concernant les principes d'annotation, on se reportera au manuel détaillé conçu par ses créateurs[7]. Nous nous bornons ici à rappeler quelques grandes règles. Le choix de *CATTEX* plutôt que d'un autre système permet de garantir l'interopérabilité avec les corpus médiévaux comme la *Base de français médiéval*[8] ou ceux développés à l'École des chartes[9].

— L'annotation n'est pas morphologique mais morpho-syntaxique, c'est à dire que l'identification des catégories grammaticales est faite en contexte.
— Le principe général veut que :
  — un token qui précède un nom sans adjectif est un déterminant
  — un token précédé d'un déterminant et suivi d'un substantif est un adjectif
  — en l'absence d'un substantif, le token est un pronom
— Les adjectifs ou les adverbes peuvent être substantivés – ce que l'on reconnaît à la présence d'un déterminant. Ainsi *beau* est `ADJqua` mais peut être substantivé *un* ***beau*** → `NOMcom`. De même *paravant* est adverbe, mais *au paravant* →

6. Sophie Prévost, Céline Guillot, Alexei Lavrentiev et Serge Heiden, *Jeu d'étiquettes morphosyntaxiques CATTEX2009*, rapp. tech., version 2.0. `http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf`, Lyon, École normale supérieure de Lyon, 2013.

7. C. Guillot, S. Prévost et A. Lavrentiev, *Principes d'annotation Cattex09*, rapp. tech., version 2.0. `http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf`, Lyon, École normale supérieure de Lyon, 2013.

8. C. Guillot, S. Heiden et A. Lavrentiev, « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques : Revue de Linguistique française diachronique*–7 (déc. 2017), p. 168-184, URL : `https://halshs.archives-ouvertes.fr/halshs-01809581` (visité le 19/12/2019).

9. Jean-Baptiste Camps (éd.), *Geste : un corpus de chansons de geste, 2016-… (Version 02)*, Paris, 2016-2020, URL : `http://doi.org/10.5281/zenodo.2630574`, ainsi que les données provenant des thèses de doctorat, en préparation d'Ariane Pinche (*Édition nativement numérique du recueil hagiographique 'Li Seint Confessor' de Wauchier de Denain d'après le manuscrit 412 de la Bibliothèque nationale de France*, dir. Corinne Pierreville et Bruno Bureau, Univ. Lyon 3) et Lucence Ing (*Disparitions lexicales en diachronie : traitements automatiques sur le Lancelot en prose*, dir. Frédéric Duval, École nationale des Chartes).

PRE.DETdef NOMcom.

— Nous considérons comme nom propre un token qui commence par une majuscule : ainsi *la mer **Noire*** → DETdef NOMcom NOMpro, *madame de **Sévigné*** → NOMcom PRE NOMpro (mais *madame **de Sévigné*** → NOMcom PRE NOMpro). L'annotation des entités au format BIO permet de capturer des ensembles plus larges, contenant titres, prénoms...(voir sect. 5).
— La personnification (marquée par la majuscule, voir *supra*) entraîne automatiquement la requalification en nom propre : dans *la **Fortune*** → NOMpro et pas NOMcom).
— Les tokens *Dieu* et *dieu* peuvent être NOMpro si la détermination non pertinente (***Dieu** merci*), potentiellement avec un adjectif (*le bon **Dieu***). Dans les autres cas le token est NOMcom (*le **dieu** Jupiter*).
— Il n'y a pas d'adjectivation du participe : si un infinitif existe, nous considérons qu'il s'agit d'un participe (*un retour **éclatant*** → VERppa, *une âme **affligée*** → VERppe). Notons que la présence d'une marque de flexion (*des traits **charmants***) n'est pas considérée comme une raison suffisante pour en faire une forme autonome (*charmants* → VERppa).
— Pour les locutions conjonctives : *sauf* ou *sans* suivis de *que* sont toujours PRE, *puis* ou *outre* suivis de *que* sont toujours ADVgen.
— Les tokens qui sont le résultat d'une soudure (*parce* < *par ce*, *afin* < *à fin*) et n'existent que dans une locution sont analysés par analogie : ***parce** que* → ADVgen car *parce* ≈ *bien* dans *bien que*. Cela permet d'éviter la création de classes d'étiquette d'effectif trop faible du type *afin* → PRE.NOMcom ou *parce* → PRE.PROdem.
— La graphie joue un rôle important dans l'annotation. Ainsi, *d'avantage* (→ PRE NOMcom) n'est pas la même chose que *davantage* (→ ADVgen), *au par avant* (→ PRE.DETdef PRE ADVgen) n'est pas la même chose que *au paravant* (→ PRE.DETdef NOMcom) ni que *auparavant* (→ ADVgen). Dans des cas aberrants (*n'aguères*) il convient de retokeniser (*n'a guères* → ADVneg VERcjg ADVgen ou *naguère* → ADVgen.
— Les tokens *voici* et *voilà* sont considérés comme des verbes conjugués (VERcjg). *Idem* pour *vive* (*vive / vivent les vacances*) où *vive(nt)* est VERcjg.
— Pour les chiffres, rappelons que le traditionnel adjectif cardinal est annoté comme déterminant cardinal s'il suivi d'un substantif (***18** ans* → DETcar NOMcom) sauf s'il est précédé d'un déterminant (*ses **18** ans* → DETpos ADJcar NOMcom). Les cardinaux qui suivent une référence éditoriale (*lettre **1**, tome **2**...*) ou une référence temporelle (*l'an **1671***) sont considérés comme des adjectif cardinaux (ADJcar). En revanche, dans les dates données en style moderne, on considère qu'il y a ellipse des noms et que les cardinaux sont donc des pronoms : i.e., *le dixième jour de mars de l'an 1632* → DETdef ADJord NOMcom PRE NOMcom PRE DETdef NOMcom ADJcar,
mais en revanche, *le 10 mars 1632* → DETdef PROcar NOMcom PROcar.
— Les emprunts ou les passages en langue étrangère sont annotés avec ETR.
— Les mots abrégés sont étiquetés avec ABR seulement s'il est impossible de connaître leur partie du discours. Dans les cas où il es possible de savoir s'il s'agit d'un nom propre (*le père **R.***), d'un substantif (***M.** Dupuis*), d'un verbe (***voy.** page 12*), on utilise ces étiquettes.

Quelques exemples pour le débutant

Des exemples simples pour commencer :

— *je vais à Metz* → `PROper VERcjg PRE NOMpro`
— *un sort assez propice* → `DETndf NOMcom ADVgen ADJqua`
— *il m'a sauvé toutefois des ravages du temps* → `PROper PROper VERcjg VERppe ADVgen PRE.DETdef NOMcom PRE.DETdef NOMcom`
— *c'est que Vespasian me regardait pour lui* → `PRODem VERcjg CONsub NOMpro PROper VERcjg PRE PROper`
— *Lettre 12. 16 septembre 1676. À M. R.* → `NOMcom ADJcar PONfrt PROcar NOMcom PROcar PONfrt PRE NOMcom NOMpro`.

Il est essentiel de faire attention aux homographes :

— *soit…soit* → `CONcoo` mais *tu le veux? **Soit*** → `ADVgen` ou bien encore *ainsi **soit**-il* → `VERcjg`
— *il ne **leur** manquera rien* → `PROper` mais ***leur** vif éclat* → `DETpos`
— ***la** fin* → `DETdef` mais *je **la** vois* → `PROper`.
— ***il** mange* → `PROper` mais ***il** semble* → `PROimp`
— *je cours **même** sort* → `DETind` mais *faire de **même*** → `ADVgen` et *le **même** sort* → `ADJind`.
— *en **tel** désespoir* → `DETind` mais *un **tel** objet* → `ADJind`.
— ***un** jour* → `DETndf` mais *l'**un**, l'autre* → `PROind`.
— ***ce** jour* → `DETdem` mais ***c'** est* → `PRODem`.
— ***15** juillet* → `PROcar` mais ***15** jours* → `DETcar` et *chapitre **15*** → `ADJcar` .

Attention aussi aux lemmes composés :

— *dudit* → `PRE.DETcom`
— *au*(*x*) → `PRE.DETdef`
— *duquel* → `PRE.DETrel`, `PRE.PROrel` ou `PRE.PROint` en fonction du contexte.
— *auxquels* → `PRE.DETrel`, `PRE.PROrel` ou `PRE.PROint` en fonction du contexte.
— *du* → `PRE.DETdef`
— *des* → `PRE.DETdef` (ou `DETndf`, comme dans `il voit des arbres` : attention au contexte !)

Deux types de tokens sont particulièrement problématiques.

— Les noms propres et assimilés (titres...) :
  — *don Carlos* → `NOMcom NOMpro`
  — *la mer Rouge* → `DETdef NOMcom ADJqua`
  — *Monsieur de La Rochefoucauld* → `NOMcom PRE NOMpro NOMpro`
  — *François, duc d'Enghien* → `NOMpro PONfbl NOMcom PRE NOMpro`
  — *M. Du Plessis* → `NOMcom NOMpro NOMpro` (et non `PRE.DETdef` pour *Du*, qui a d'ailleurs une majuscule en français contemporain).
  — *Denys d'Halycarnasse* → `NOMpro PRE NOMpro`
  — *Mesnil montant* → `NOMpro VERppa`
— Les locutions
  — *afin de* → `ADVgen PRE`
  — *afin que* → `ADVgen CONsub`
  — *à fin que* → `PRE NOMcom CONsub`
  — *puis que* → `ADVGen CONsub`

— *puis donc que* → `ADVGen ADVgen CONsub`
— *de ce que* → `PRE PROdem CONsub`
— *lors que* → `ADVGen CONsub`
— *pourvu que* → `ADVgen CONsub`
— *vu que* → `VERppe CONsub`
— *tandis que* → `ADVgen CONsub`
— *selon que* → `ADVgen CONsub`
— *bien que* → `ADVgen CONsub`
— *pres que* → `ADVGen CONsub`
— *parce que* → `ADVgen CONsub`
— *par ce que* → `PRE PROdem CONsub`
— *pource que* → `ADVgen CONsub`
— *pour ce que* → `PRE PROdem CONsub`
— *quant à* → `ADVgen PRE`
— *par tout* → `PRE PROind`
— *à peine* → `PRE NOMcom`
— *tout à fait* → `ADVgen PRE NOMcom`
— *la plus part* → `DETdef ADVgen NOMcom`
— *au paravant* → `PRE.DETdef NOMcom`
— *tout à coup* → `ADVgen PRE NOMcom`
— *d'avantage* → `PRE NOMcom`
— *n'aguères* → `A_RETOKENISER`
— *auprès de* → `ADV PRE`
— *là-dessus* → `ADVgen PONfbl ADVgen`
— *en dessus* → `PRE ADVgen`
— *au dessus* → `PRE.DETdef NOMcom`
— *au moins* → `PRE.DETdef NOMcom`
— *du moins* → `PRE.DETdef NOMcom`
— *auprès de* → `PRE PRE`

Ajoutons quelques cas particuliers :

— *premier* est `ADJord`
— *dernier* est `ADJqua`
— *16 septembre 1676* → `PROcar NOMcom PROcar`

4 Morphologie

Le recours à *CATTEX-max* implique l'étiquetage morphologique précis des tokens en plus de la catégorie grammaticale.

Catégorie	Valeurs possibles
GENRE	*m*, *f*, *n*
NOMB	*s*, *p*
GENRE	*m*, *f*
MODE	*ind*, *imp*, *con*, *sub*
TEMPS	*pst*, *ipf*, *fut*, *psp*
PERS	*0*, *1*, *2*, *3*
CAS	*n*, *r*, *i*

TABLE 2 – Valeurs possibles pour la morphologie

Quelques remarques générales :

— Les différents emplois du pronom ayant tous un même lemme (*je*, *me*, *moi* → `je`) on utilise le cas pour les distinguer (respectivement `CAS=n` pour le nominatif, `CAS=r` pour le régime direct et `CAS=i` pour le régime indirect).

— La question du nombre des possessifs est complexe, car le choix de *CATTEX* de ne retenir que trois personnes (1, 2 et 3, pluriel ou singulier) pose problème : dans ces conditions, le nombre du possessif est-il celui de la personne ou de son référent ? Ainsi *mes* est-ce la première personne du singulier (`PERS.=1|NOMB.=s`) ou un déterminant possessif dont le référent est un pluriel (`PERS.=1|NOMB.=p`). Nous avons retenu la seconde option.

— Dans le cas où le contexte immédiat ne permet pas de désambiguiser l'information, on laisse la valeur `x` : ***vous** êtes odieux* → `GENRE=x` et `odieux NOMB.=x`, *je cherche un enfant qui **joue** dehors* → `MODE=x`. On considère que c'est toujours le cas pour les pronoms personnels sans marque morphologique de genre (*i.e.*, *je*, *tu*, *l'* → `GENRE=x`, mais *la* → `GENRE=f`).

— Il ne faut pas confondre `GENRE=x` et `GENRE=n`. Les adjectifs qualificatifs résultant de pronoms impersonnels (*il est **clair***) sont décrits comme de genre neutre (`NOMB.=s|GENRE=n`). Il en va de même pour le pronom impersonnel(***il** est clair*), auquel on ajoute qu'il n'a pas de personne (`PERS.=0|NOMB.=s|GENRE=n|CAS=n`). C'est aussi le cas du pronom démonstratif neutre *ce*.

— Il n'y a pas de temps pour les modes conditionnel (`MODE=con`) et impératif (`MODE=imp`) : *Viens ici* → `MODE=imp|PERS.=2|NOMB.=s`.

— Nous ne connaissons pas les temps composés : *J'ai mangé* est composé d'un indicatif présent (`VERcjg MODE=ind|TEMPS=pst`) et d'un participe passé (`VERppe`).

— Les tokens *voici* et *voilà* sont des indicatifs présent sans personne ni nombre (`MODE=ind|TEMPS=pst|PERS.=x|NOMB.=x`).

— Pour certains tokens (`PRE`, `PON`, `ADV`...) il n'existe pas d'information morphologique : on met la valeur `MORPH=empty`.

— Contrairement aux prénoms qui ont un genre, les noms de famille n'en ont pas (*Julien **Sorel*** → `NOMB.=s|GENRE=x`) sauf s'ils sont précédés d'un déterminant (*le*

Sorel → `NOMB.=s|GENRE=m`) ou qu'ils renvoient à une personnage précis (**Calvin**, **Cromwell** → `NOMB.=s|GENRE=m`). Pour les ville, même si un tendance générale porte au féminin, les usages sont fluctuants. On ne met un genre que lorsque le contexte immédiat permet de trancher :(*la **Rome** éternelle* → `NOMB.=s|GENRE=f`, mais *il revient de **Venise*** → `NOMB.=s|GENRE=x`).

Quelques exemples pour le débutant

— *Je suis là, **et** vous?* → `MORPH=empty`
— ***Je** suis là* → `PERS.=1|NOMB.=s|GENRE=x|CAS=n`
— ***mes** yeux* → `PERS.=1|NOMB.=p|GENRE=m`
— ***Venez** ici!* → `MODE=imp|PERS.=2|NOMB.=p`
— ***il** y a* → `PERS.=0|NOMB.=s|GENRE=n|CAS=n`
— *c'est **évident*** → `NOMB.=s|GENRE=n`.

5 Entités nommées

Une entité nommée est une expression linguistique référentielle, souvent associée aux noms propres et aux descriptions définies. Nous proposons d'en rester aux quatre étiquettes proposées pour la *ConLL shared task 2003* [10] :

— `PER` pour les personnes (*Jacques*)
— `LOC` pour les lieux (*Paris*)
— `ORG` pour les organisations (*L'Académie française*)
— `MISC` pour les autres noms (*L'Annonciation*)

Il conviendrait de réfléchir à l'utilisation d'autres étiquettes, notamment pour des valeurs (par ex. `DATE` pour les dates, `QUANTITY` pour les quantités ou `MONEY` pour les sommes d'argent) comme défini par d'autres projets [11]. Ce jeu d'étiquettes est utilisé avec le format *BIO - beginning inside outside* afin d'identifier les suites de tokens.

— Afin de compenser nos choix de lemmatisation et d'étiquetage morpho- syntaxique très minimalistes pour ce qui concerne les noms propres (uniquement les tokens commençant avec une majuscule), nous optons ici pour une approche très maximaliste. Tout ce qui se rapporte à l'entité est annoté : titre (*Monsieur*, *Président*, *Marquis*...), particules (*de*, *Le*) et divers noms (*Bussy-Rabutin*).
— Le premier token est étiqueté avec le suffixe `-B` (`LOC-B` ou `PER-B`) et tous les suivants avec le suffixe `-I` (`LOC-I` ou `PER-I`), y compris la ponctuation (*Bussy-Rabutin* → `PER-B PER-I PER-I`). Le suffixe `-B` est donc placé en fonction du contexte (*le comte de Bussy-Rabutin* → `O PER-B PER-I PER-I PER-I PER-I`)).

10. Erik F. Tjong Kim Sang et Fien De Meulder, « Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition », dans *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142-147, URL : `https://www.aclweb.org/anthology/W03-0419`.

11. Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, *et al.*, *OntoNotes Release 5.0 with OntoNotes DB Tool*, rapp. tech., v0.999 beta. `https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf`, 2012.

Quelques exemples pour le débutant

— *j'ai vu don Carlos* → `O O O PER-B PER-I`
— *je suis à la mer Rouge* → `O O O O LOC-B LOC-I`
— *Monsieur de La Rochefoucauld* → `PER-B PER-I PER-B PER-I`
— *François, duc d'Enghien* → `PER-B PER-I PER-I PER-I PER-I`
— *M. Du Plessis* → `PER-B PER-I PER-I PER-I.`
— *Denys d'Halycarnasse* → `PER-B PER-I PER-I`

Remerciements

Un grand nombre des cas problématiques commentés dans ce manuel proviennent de discussions avec Jean-Baptiste Tanguy, Marie Puren, Frédéric Duval, Lucence Ing, Suzanne Duval, et Maxime Cario : leur aide fut donc précieuse lors de la rédaction. Une pensée toute particulière va à Florian Cafiero, frère d'arme lors de la campagne *CornMol*.

Annexes

A Exemple d'annotation

Ami	ami	NOMcom	NOMB.=s|GENRE=m
,	,	PONfbl	MORPH=empty
j'	je	PROper	PERS.=1|NOMB.=s|GENRE=x|CAS=n
ai	avoir	VERcjg	MODE=ind|TEMPS=pst|PERS.=1|NOMB.=s
beau	beau	ADVgen	MORPH=empty
rêver	rêver	VERinf	MORPH=empty
,	,	PONfbl	MORPH=empty
toute	tout	DETind	NOMB.=s|GENRE=f
ma	mon	DETpos	PERS.=1|NOMB.=s|GENRE=f
rêverie	rêverie	NOMcom	NOMB.=s|GENRE=f
Ne	ne	ADVneg	MORPH=empty
me	je	PROper	PERS.=1|NOMB.=s|GENRE=x|CAS=r
fait	faire	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s
rien	rien	PROind	NOMB.=s|GENRE=x
comprendre	comprendre	VERinf	MORPH=empty
en	en	PRE	MORPH=empty
ta	ton	DETpos	PERS.=2|NOMB.=s|GENRE=f
galanterie	galanterie	NOMcom	NOMB.=s|GENRE=f
.	.	PONfrt	MORPH=empty
Auprès	auprès	ADVgen	MORPH=empty
de	de	PRE	MORPH=empty
ta	ton	DETpos	PERS.=2|NOMB.=s|GENRE=f
maîtresse	maîtresse	NOMcom	NOMB.=s|GENRE=f
engager	engager	VERinf	MORPH=empty
un	un	DETndf	NOMB.=s|GENRE=m
ami	ami	NOMcom	NOMB.=s|GENRE=m
,	,	PONfbl	MORPH=empty
C'	ce	PRODem	NOMB.=s|GENRE=n
est	être	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s
,	,	PONfbl	MORPH=empty
à	à	PRE	MORPH=empty
mon	mon	DETpos	PERS.=1|NOMB.=s|GENRE=m
jugement	jugement	NOMcom	NOMB.=s|GENRE=m
,	,	PONfbl	MORPH=empty
ne	ne	ADVneg	MORPH=empty
l'	il	PROper	PERS.=3|NOMB.=s|GENRE=x|CAS=r
aimer	aimer	VERinf	MORPH=empty
qu'	que	CONsub	MORPH=empty
à	à	PRE	MORPH=empty
demi	demi	NOMcom	NOMB.=s|GENRE=m
.	.	PONfrt	MORPH=empty

B Abréviations pour le tokeniseur (extrait)

Acad.	*académie*
Adj.	*adjectif*
Agricol.	*agricole*
Agricul.	*agriculture*
Apocal.	*Apocalypse*
anc.	*ancienne*
Bot.	*botanique*
Botan.	*botanique*
Botaniq.	*botanique*
ca.	*capitulum*
cap.	*capitulum*
capi.	*capitulum*
Cf.	*confer*
Cha.	*chapitre*
Chap.	*chapitre*
Col.	*colonne*
Dic.	*dictionnaire*
Diction.	*dictionnaire*
Dictionn.	*dictionnaire*
Eccl.	*ecclésiastique*
Écon.	*économie*
Élem.	*élément*
Fig.	*figure*
Fr.	*français(e)*
Geog.	*géographie*
Gram.	*grammaire*
Gramm.	*grammaire*
Hist.	*histoire*
Ibid.	*ibidem*
Ibid.	*ibidem*
Inst.	*institution*
Jard.	*jardinage*
Jurisprud.	*jurisprudence*
Latit.	*latitude*
Li.	*liber*
Lib.	*liber*
Libr.	*liber*
Lig.	*ligne*
Lit.	*littérature*
Littérat.	*littérature*
Liv.	*livre*
Long.	*longitude*
Mar.	*marin*
Mat.	*mathématiques*
Mathém.	*mathématiques*
Mech.	*méchanique*
Med.	*médecine*
Med.	*médical*
Mem.	*mémoire*
Menuis.	*menuiserie*
Milit.	*militaire*
Mod.	*moderne*
Mor.	*moral*
Mr.	*monsieur*
Monsr.	*monsieur*
nat.	*naturel*
natur.	*naturel*
N.b.	*nota bene*
Orat.	*oratoire*
Ornith.	*ornithologie*
Ornythol.	*ornithologie*
Ornitholog.	*ornithologie*
Part.	*partie*
Pag.	*page*
Pharm.	*pharmacie*
Phil.	*philosophie*
Philos.	*philosophie*
Pl.	*planche*
Pl.	*pluriel*
Politiq.	*politique*
P.S.	*post scriptum*
Phys.	*physique*
Physiq.	*physique*
Sr.	*sieur*
St.	*saint*
Subst.	*subtantif*
s.f.	*substantif féminin*
S.M.	*Sa Majesté*
s.m.	*substantif masculin*
Tab.	*tableau*
Théât.	*théâtre*
Trév.	*Trévoux*
Tom.	*Tome*
Vol.	*Volume*
V. n.	*Verbe neutre*
V. a.	*Verbe actif*
V. act.	*Verbe actif*
Zoo.	*zoologique*
Zoolog.	*zoologique*

C Jeu d'étiquette CATTEX

Type	Étiquette	Définition
Verbes	VERcjg	Verbe conjugué
	VERinf	Verbe infinitif
	VERppe	Verbe p.passé
	VERppa	Verbe p.présent
Noms	NOMcom	Nom commun
	NOMpro	Nom propre
Adjectifs	ADJqua	Adjectif qualificatif
	ADJind	Adjectif indéfini
	ADJpos	Adjectif possessif
	ADJcar	Adjectif cardinal
	ADJord	Adjectif ordinal
Pronoms	PROper	Pronom personnel
	PROper.PROper	Pronoms personnel composés (*jel/jol*)
	PROimp	Pronom impersonnel
	PROadv	Pronom adverbial
	PROpos	Pronom possessif
	PRODem	Pronom démonstratif
	PROind	Pronom indéfini
	PROcar	Pronom cardinal
	PROord	Pronom ordinal
	PRORel	Pronom relatif
	PROint	Pronom interrogatif
	PROcom	Pronom composé (*ledict*, *ladict*)
Déterminant	DETdef	Déterminant défini
	DETndf	Déterminant non défini
	DETdem	Déterminant démonstratif
	DETpos	Déterminant possessif
	DETind	Déterminant indéfini
	DETcar	Déterminant cardinal
	DETrel	Déterminant relatif
	DETint	Déterminant interrogatif
	DETcom	Déterminant composé
Adverbes	ADVgen	Adverbe général
	ADVgen.PROper	Adverbe général + pronom personnel (*sil*, *sel*)
	ADVgen.PROadv	Adverbe général + pronom adverbial (*sin*)
	ADVneg	Adverbe de négation
	ADVneg.PROper	Adverbe de négation + pronom personnel (*nel*)
	ADVneg.PROadv	Adverbe de négation + pronom adverbial (non = ne + en)
	ADVint	Adverbe interrogatif
	ADVing	Adverbe interrogatif négatif
	ADVsub	Adverbe subordonnant
Prépositions	PRE	Préposition (sauf, par, de, en, por, sans)
	PRE.DETdef	Enclise du déterminant défini après préposition
	PRE.DETcom	Enclise du déterminant composé après préposition
	PRE.DETrel	Enclise du déterminant relatif (ou interrogatif en interrogative indirecte) après préposition

Type	Étiquette	Définition
	PRE.PROper	Enclise du pronom personnel après préposition
	PRE.PROrel	Enclise du pronom relatif (ou interrogatif en interrogative indirecte) après préposition
Conjonctions	CONcoo	Conjonction de coordination
	CONsub	Conjonction de subordination
	CONsub.PROper	Enclise du pronom personnel après conjonction de subordination
Interjections	INT	Interjection
Ponctuations	PON	Ponctuation
	PONfrt	Ponctuation forte (délimite les phrases)
	PONfbl	Ponctuation faible (interne à une phrase)
	PONpga	Guillemet ou parenthèse ouvrants
	PONpdr	Guillemet ou parenthèse fermant
	PONpxx	Guillemet droit (quand on ne sait pas si c'est ouvrant ou fermant)
Redondance	RED	"que" redondant
OUT	OUT	Ce qui ne doit pas être pris en compte dans l'analyse linguistique

Références

ATILF-CNRS et Université de Lorraine, *Dictionnaire du Moyen Français (1330-1500)*, 2015, url : http://www.atilf.fr/dmf.

Jean-Baptiste Camps (éd.), *Geste : un corpus de chansons de geste, 2016-… (Version 02)*, Paris, 2016-2020, url : http://doi.org/10.5281/zenodo.2630574.

Clérice (Thibault), *Référentiel du Latin pour Pyrrha, d'après le dictionnaire et travaux du LASLA de D. Longrée et al*, mai 2020, doi : 10.5281/zenodo.3822040.

Diwersy (Sascha), Falaise (Achille), Lay (Marie-Hélène) et Souvay (Gilles), « Ressources et méthodes pour l'analyse diachronique », *Langages*, N° 206–2 (août 2017), p. 21-44, url : https://www.cairn.info/revue-langages-2017-2-page-21.htm (visité le 03/12/2018).

Guillot (Céline), Heiden (Serge) et Lavrentiev (Alexei), « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques : Revue de Linguistique française diachronique*–7 (déc. 2017), p. 168-184, url : https://halshs.archives-ouvertes.fr/halshs-01809581 (visité le 19/12/2019).

Guillot (Céline), Prévost (Sophie) et Lavrentiev (Alexei), *Principes d'annotation Cattex09*, rapp. tech., version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf, Lyon, École normale supérieure de Lyon, 2013.

Pierrel (Jean-Marie), Dendien (Jacques) et Bernard (Pascale), « Le TLFi ou Trésor de la Langue Française informatisé », dans *Proceedings of the 11th EURALEX International Congress*, dir. Geoffrey Williams et Sandra Vessier, Lorient, France, 2004, p. 165-170.

Prévost (Sophie), Guillot (Céline), Lavrentiev (Alexei) et Heiden (Serge), *Jeu d'étiquettes morphosyntaxiques CATTEX2009*, rapp. tech., version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf, Lyon, École normale supérieure de Lyon, 2013.

Souvay (Gilles) et Pierrel (Jean-Marie), « LGeRM Lemmatisation des mots en Moyen Français », *Traitement Automatique des Langues*, 50–2 (2009), p. 149-172, url : https://halshs.archives-ouvertes.fr/halshs-00396452.

Tjong Kim Sang (Erik F.) et De Meulder (Fien), « Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition », dans *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142-147, url : https://www.aclweb.org/anthology/W03-0419.

Weischedel (Ralph), Palmer (Martha), Marcus (Mitchell), Hovy (Eduard), Pradhan (Sameer), Ramshaw (Lance), Xue (Nianwen), Taylor (Ann), Kaufman (Jeff), Franchini (Michelle), *et al.*, *OntoNotes Release 5.0 with OntoNotes DB Tool*, rapp. tech., v0.999 beta. https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf, 2012.