Manuel du corpus PALAFRAFRO-V2

novembre 2017

Table des matières

1 Choix des textes

Le corpus PALAFRAFRO-V2 vise à documenter la période la plus ancienne de l'histoire du français pour étudier le passage du latin au français. Ce corpus intègre des ressources existantes, développées dans le cadre du projet *Corpus représentatif des premiers textes français*[1], auxquelles s'ajoutent de nouveaux textes, genres et annotations linguistiques. Le corpus constitue une sous partie de la *Base de français médiéval* (BFM)[2] et les textes sont accessibles dans les mêmes conditions que tous les autres textes de la BFM, sous licence CC BY-NC-SA 3.0 France.

Le corpus peut être interrogé sur le portail de la Base de français médiéval (http://txm.bfm-corpus.org) grâce à la plateforme TXM. Il peut également être téléchargé et exploité avec le logiciel TXM (version bureau, http://textometrie.ens-lyon.fr).

La sélection des textes s'appuie sur le jeu de métadonnées décrit dans le *Manuel de description des textes de la BFM* (http://bfm.ens-lyon.fr/spip.php?article301) et la *Présentation des descripteurs du projet CORPTEF* (http://corptef.ens-lyon.fr/spip.php?rubrique60). Les critères qui ont prévalu sont la date de composition des textes, la forme (vers/prose), le domaine et le genre discursif, la date du manuscrit et la qualité de l'édition de référence.

La partie latine du projet (corpus Palafralat) étant constituée de textes en prose (vies de saints, chartes, recueils de lois, formulaires, lettres et chroniques historiques), ces mêmes textes ont été privilégiés dans le corpus français. Le corpus PALAFRAFRO-V2 comporte par conséquent un nombre important de textes hagiographiques, de chartes et de textes historiques. Il contient une majorité d'œuvres en prose et du domaine religieux. Des textes en vers et littéraires ou didactiques ont été ajoutés lorsqu'ils étaient anciens. Les œuvres datent pour la plupart des 12ème et 13ème siècles mais quelques chartes et vies de saints du 14ème siècle ont également été intégrées afin de permettre l'étude diachronique de ces genres, de la latinité tardive à la fin du Moyen Âge.

1 http://corptef.ens-lyon.fr.
2 http://bfm.ens-lyon.fr.

2 Étiquetage morphosyntaxique

2.1 Principes de l'étiquetage morphosyntaxique

Tous les textes du corpus PALAFRAFRO-V2 ont été étiquetés automatiquement avec le TreeTagger et le modèle linguistique de la BFM. Dans la majorité des textes, les étiquettes ont été vérifiées par des médiévistes. La propriété *etiquetage*, attachée à chaque mot, permet de savoir pour chaque occurrence si l'étiquette morphosyntaxique a été vérifiée (etiquetage="vérifié" ou etiquetage="non vérifié").

Deux jeux d'étiquettes morphosyntaxiques ont été utilisés : le jeu spécialisé pour le français Cattex2009-min et le jeu multilingue Universal Dependencies (UD). Une étiquette CATTEX2009-min (*pos-cattex*) et une étiquette UD (*pos-ud*) ont été affectées à chaque mot. Lorsque son emploi en contexte le requiert, le mot a une étiquette *pos-cattex* pour sa catégorie morphologique et une étiquette *pos_syn-cattex* différente pour sa catégorie syntaxique.

Les formes contractées ont été dupliquées pour permettre l'utilisation de deux étiquettes distinctes :

« N'en parlez mais, se jo nel <pos-ud="ADV">*nel <pos-ud="PRON"»> vos cumant ! » (*Roland* v. 273)

« N'en parlez mais, se jo nel <pos-cattex="ADVneg">*nel <pos-cattex="PROper"»> vos cumant ! »

Les étiquettes doubles de Cattex2009 (comme ADVneg.PROper) ont été remplacées par deux étiquettes simples affectées à deux formes identiques. La seconde occurrence de la forme dupliquée est précédée d'un astérisque.

La liste des étiquettes CATTEX2009-min, les *Principes d'annotation Cattex09* et le *Manuel de référence Cattex09* sont accessibles sur le site de la *Base de français médiéval* : http://bfm.ens-lyon.fr/spip.php?article176. La liste des étiquettes et les principes d'étiquetage du jeu *Universal Dependencies* sont accessibles en ligne : http://universaldependencies.org.

2.2 Liste des étiquettes (pos) et traits (feats)

L'organisation des jeux CATTEX2009-min et UD et la répartition entre catégories (*pos*) et traits morphologiques (*feats*) sont est assez différentes. Le tableau qui suit établit la correspondance entre les étiquettes CATTEX2009-min (propriété *pos-cattex*) et les étiquettes UD (propriété *pos-ud*) avec les traits associés (*feats*). Chaque trait est une propriété (au même titre que *pos-ud*) dont les valeurs possibles sont détaillées dans le tableau. Par exemple, la propriété *numtype* peut avoir les valeurs "Card" et "Ord" (numtype="Card" ou numtype="Ord" ou *numtype* est vide).

Dans quelques cas, l'étiquette CATTEX2009-min n'a pas d'équivalent direct UD et l'étiquette UD correspondante est moins précise :

- ADJqua (pos-cattex) correspond à ADJ (pos-ud)
- ADVing (pos-cattex) correspond à ADV (pos-ud)
- ADVsub (pos-cattex) correspond à ADV (pos-ud)
- PROadv (pos-cattex) correspond à ADV (pos-ud)

- PROimp (pos-cattex) correspond à PRON (pos-ud)
- PONfbl (pos-cattex) correspond à PUNCT (pos-ud)
- PONfrt (pos-cattex) correspond à PUNCT (pos-ud)
- PONpdr (pos-cattex) correspond à PUNCT (pos-ud)
- PONpga (pos-cattex) correspond à PUNCT (pos-ud)
- PONpxx (pos-cattex) correspond à PUNCT (pos-ud)

Le trait *definite* du jeu UD a été utilisé dans quelques cas très limités. Cela concerne :
- l'article défini, qui combine les propriétés *pos* (pos-ud="Det") et *prontype* (prontype="Art") avec la propriété *definite* (definite="Def")
- l'article indéfini, qui combine les propriétés *pos* (pos-ud="Det") et *prontype* (prontype="Art") avec la propriété *definite* (definite="Ind")
- le déterminant et le pronom *ledit*, qui combinent la propriété *pos* (pos-ud="Det" ou pos-ud="Pron") avec la propriété *definite* (definite="Com") ; on utilise ici de manière détournée la valeur « Com » (pour *complex*) prévue dans le jeu UD.

pos-cattex	pos-ud	propriétés supplémentaires
ABR	X	abbr="Yes"
ADJcar	ADJ	numtype="Card"
ADJind	ADJ	prontype="Ind"
ADJord	ADJ	numtype="Ord"
ADJpos	ADJ	poss="Yes"
ADJqua	ADJ	
ADVgen	ADV	
ADVing	ADV	prontype="Int"
ADVint	ADV	prontype="Int"
ADVneg	ADV	prontype="Neg"
ADVsub	ADV	
CONcoo	CCONJ	
CONsub	SCONJ	
DETcar	DET	numtype="Card"
DETcom	DET	definite="Com"
DETdef	DET	prontype="Art" & definite="Def"
DETdem	DET	prontype="Dem"
DETind	DET	prontype="Ind"
DETint	DET	prontype="Int"
DETndf	DET	prontype="Art" & definite="Ind"
DETord	DET	numtype="Ord"
DETpos	DET	poss="Yes"
DETrel	DET	prontype="Rel"

ETR	X	foreign="Yes"
INJ	INTJ	
NOMcom	NOUN	
NOMpro	PROPN	
OUT	X	
PONfbl	PUNCT	
PONfrt	PUNCT	
PONpdr	PUNCT	
PONpga	PUNCT	
PONpxx	PUNCT	
PRE	ADP	
PROadv	ADV	
PROcar	PRON	numtype="Card"
PROcom	PRON	definite="Com"
PROdem	PRON	prontype="Dem"
PROimp	PRON	prontype="Prs"
PROind	PRON	prontype="Ind"
PROint	PRON	prontype="Int"
PROord	PRON	numtype="Ord"
PROper	PRON	prontype="Prs"
PROpos	PRON	poss="Yes"
PROrel	PRON	prontype="Rel"
RED	X	
RES	X	
VERcjg	VERB	verbform="Fin"
VERinf	VERB	verbform="Inf"
VERppa	VERB	verbform="Part" & tense="Pres"
VERppe	VERB	verbform="Part" & tense="Past"

3 Lemmatisation

Tous les textes du corpus PALAFRAFRO-V2 ont été lemmatisés automatiquement avec le logiciel TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger) et, pour certains textes, avec l'outil LGeRM (lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques, ATILF - CNRS & Université de Lorraine. http://www.atilf.fr/LGeRM). La lemmatisation s'est appuyée sur le lexique Frolex mis en ligne sur la plate-forme GITHUB sous licence CC-BY-NC-SA 3.0 FR.

Les lemmes utilisés sont ceux du *Dictionnaire de moyen français*, du Tobler-Lommatzsch, du dictionnaire de Godefroy, du *Dictionnaire électronique de Chrétien de Troyes*, du *Dictionnaire Anglo-Normand* (AND), du *Dictionnaire étymologique français* (FEW) et, dans quelques cas, ce

sont des créations de la BFM. La propriété *lemma_src* indique la source du lemme.

lemma_src[3]	Source du lemme
AND	*Anglo-Norman Dictionary*, http://www.anglo-norman.net/
BFM	Base de français médiéval (création)
DECT	*Dictionnaire Électronique de Chrétien de Troyes*, http://www.atilf.fr/dect/
DMF	*Dictionnaire du Moyen Français*, http://www.atilf.fr/dmf/
FEW	*Französisches Etymologisches Wörterbuch* de Walther von Wartburg, http://www.atilf.fr/few
GDF	*Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle* de Frédéric Godefroy
TL	*Tobler-Lommatzsch Altfranzösisches Wörterbuch*

Dans quelques textes les lemmes ont été intégralement vérifiés. La propriété lemmatisation, attachée à chaque mot, permet de savoir pour chaque occurrence si le lemme a été vérifié (lemmatisation="vérifiée" ou lemmatisation="non vérifiée").

Dans les textes vérifiés, la forme choisie comme lemme de référence est l'entrée du DMF[4]. À défaut, un lemme tiré de l'un des autres dictionnaires cités a été retenu. Dans les textes non vérifiés, les lemmes ne sont pas désambiguïsés et il est fréquent qu'un mot ait plusieurs lemmes, soit parce qu'il y a ambiguïté, soit parce que les lemmes des dictionnaires sont différents.

Il est donc nécessaire d'être attentifs à trois choses :

- Les indices numériques utilisés par les dictionnaires pour les homonymes (par exemple, *par1*, *par2* et *par3*) ont été supprimés et la source du lemme est également considérée comme non vérifiée. Pour les entrées du DMF par1, par2 et par3, le lemme indiqué sera toujours le même, de la forme « par ».

- Lorsqu'une même forme graphique a deux lemmes possibles, les deux lemmes sont encadrés et séparés par des barres droites. Par exemple, la forme *est* peut correspondre aux lemmes ÊTRE et AIDER. Le dernier cas est rare, mais attesté dans le lexique du dictionnaire du moyen français. Par conséquent, dans les textes où les lemmes ne sont pas vérifiés, les deux lemmes possibles sont proposés sous la forme suivante : « |être|aider| » (dans le lexique par exemple).

Un opérateur spécial *contains* du langage CQL permet d'interroger ces lemmes multiples d'une manière relativement simple. La requête suivante :

[lemma contains "être"]

signifie : « tous les mots pour lesquels au moins l'un des lemmes proposés est « être ».

- L'étiquette morphosyntaxique peut être utile pour réduire l'ambigüité de la lemmatisation, car la procédure automatique sélectionne toujours une étiquette (considérée comme la plus probable). Par exemple, la requête suivante permet de sélectionner les formes du verbe *avoir* tout en excluant les occurrences de la préposition *à* :

[lemma contains "avoir" & pos-ud="VERB"].

3 Source du lemme. Voir la section 4.

4 Sauf dans *Yvain* de Chrétien de Troyes (YvainKu), texte provenant du corpus du Dictionnaire Électronique de Chrétien de Troyes où les lemmes sont ceux du DÉCT.

4 Interrogation des textes et liste des propriétés de mots

Le corpus PALAFRAFRO-V2 est intégré au portail de la Base de français médiéval, les outils d'analyse et de lecture, ainsi que le langage de requêtes CQL sont les mêmes que ceux de la BFM. Le Tutoriel BFM (http://txm.ish-lyon.cnrs.fr/bfm/files/Tutoriel_TXM_BFM_V1.pdf) peut servir d'introduction à l'usage du corpus. D'autres ressources documentaires destinées aux utilisateurs de la BFM sont présentées sur le portail http://txm.bfm-corpus.org.

Le corpus PALAFRAFRO-V2 diffère du corpus principal de la BFM par son annotation linguistique et philologique au niveau des mots. Ces annotations peuvent être visualisées dans les index et les concordances grâce au bouton « Réglages » et peuvent être interrogées grâce au langage CQL.

Par exemple, pour chercher les occurrences de la forme *a* dont l'étiquette Cattex « VERcjg » (verbe conjugué) a été vérifiée, il faut écrire la requête suivante :

```
[word="a" & pos-cattex="VERcjg" & etiquetage="vérifié"]
```

Si la requête concerne un lemme, il faut utiliser l'opérateur contains :

```
[lemma contains "par" & pos-ud="ADV"]
```

La liste complète des annotations (propriétés de mots) disponibles dans le corpus PALAFRAFRO-V2 est présentée dans le tableau ci-dessous :

Nom de la propriété	Explication	Exemple
abbr	Étiquette de trait UD, indique si la forme est une abréviation	[abbr="Yes"]
definite	Étiquette de trait UD, indique la définitiude pour les déterminants et pronoms	[definite="Ind"]
etiquetage	Indique si l'étiquetage morphosyntaxique a été vérifié pour une occurrence donnée	[etiquetage="vérifié"]
foreign	Étiquette de trait UD, indique si l'occurrence est un mot étranger	[foreign="Yes"]
id	Identifiant du mot dans le corpus (peu utile pour les requêtes)	
lang	Langue du mot étranger	[lang="la.*"]
lemma	Lemme de l'occurrence	[lemma contains "avoir"]
lemma_src	Source du lemme, voir la section 3	[lemma_src="BFM"]
lemmatisation	Indique si le lemme est vérifié ou non	[lemmatisation="vérifiée"]
numtype	Étiquette de trait UD, indique si le mot est un numéral cardinal ou ordinal	[numtype="Card"]
orig	Indique si le mot est balisé <orig> (graphie non régularisée) dans le document XML-TEI, utilisée pour certaines ponctuations dans la *Queste del saint Graal*	[orig!="NA"]
pos-cattex	Étiquette morphologique Cattex 2009	[pos-cattex="ADJ.*"]
pos-ud	Étiquette « partie du discours » UD	[pos-ud="ADJ"]

pos_syn-cattex	Étiquette morphosyntaxique Cattex 2009, identique à pos-cattex, sauf dans les cas de substantivation, etc., identifiés dans certains textes vérifiés	[pos_syn-cattex="NOMcom" & pos-cattex="VER.*"]
poss	Étiquette de trait UD, indique si le mot est un possessif	[poss="Yes"]
prontype	Étiquette de trait UD, indique le type de pronom	[prontype="Pers"]
q	Niveau d'imbrication de l'occurrence dans un discours direct (« 0 » si hors du discours direct, « 1 » si dans un discours direct non imbriqué, « 2 » ou « 3 » si dans un discours direct imbriqué dans un ou deux autres discours directs)	[q="[123]"]
ref	Référence par défaut pour la concordance (peu utile pour les requêtes)	
sic	Indique si le mot est balisé <sic> (graphie visiblement erronée) dans le document XML-TEI,	[sic!="NA"]
sp	Indique si le mot se trouve dans une prise de parole d'un texte dramatique ou d'un dialogue (balise TEI <sp>), valeurs possibles : « t » (vrai) ou « f » (faux)	[sp="t"] ou [sp="f"]
supplied	Indique si le mot se trouve dans un passage balisé <supplied> (conjecture éditoriale) dans le document XML-TEI, valeurs possibles : « t » (vrai) ou « f » (faux)	[supplied="f"]
tense	Étiquette de trait UD, indique le temps verbal, uniquement utilisée pour les participes dans le corpus	[tense="Pres"]
verbform	Étiquette de trait UD, indique le type de forme verbale : conjugué (Fin), participe (Part) ou infinitif (Inf)	[verbform="Fin"]
word	La forme du mot, propriété par défaut. Une syntaxe raccourcie peut être utilisée si elle n'est pas combinée avec d'autres propriétés	a, "a" *ou* [word="a"] [word="a" & pos-cattex="VERcjg"]