Principes d'annotation Cattex09

Céline Guillot (Celine.Guillot@ens-lyon.fr)
Sophie Prévost (sophie.prevost@ens.fr)
Alexei Lavrentiev (alexei.lavrentev@ens-lyon.fr)

Version 2.0 – 8 avril 2013

Ce document définit les principes généraux d'utilisation des étiquettes du jeu morphosyntaxique Cattex09 (section 1) et présente le tableau synthétique des étiquettes Cattex09 (section 2).
Le *Manuel de référence du jeu Cattex09* expose étiquette par étiquette le détail de leurs conditions d'utilisation.

1. Principes généraux

1.1. Cattex09

Le jeu d'étiquettes morphosyntaxiques Cattex09 se décline sous deux versions :

- une version complète Cattex09max (= maximal), qui comprend des informations morphologiques sur le nombre, le genre, le temps, la personne…
- une version Cattex09min (= minimale), qui comprend des informations sur les seules catégories et leurs types (une spécification des catégories). Actuellement, les textes de la BFM sont étiquetés avec le jeu cattex09min.

« Cattex09 » désigne par défaut Cattex09min.

1.2. Définition et forme des étiquettes

Les étiquettes sont structurées en différents champs.
Pour Cattex09min, les champs sont les suivants : <catégorie> et <type>, les valeurs étant composées de trois lettres, en majuscules pour la catégorie, en minuscules pour le type.
Les catégories correspondent pour la plupart aux classiques parties du discours : VER (verbe) ; NOM (nom) ; ADJ (adjectif) ; PRO (pronom) ; DET (déterminant) ; ADV (adverbe) ; PRE (préposition) ; CON (conjonction) ; INJ (interjection)
Les autres catégories sont les suivantes : PON (ponctuation) ; ETR (mot étranger) ; ABR (abréviation), RED (mot redondant), OUT (catégorie temporaire).
Les types correspondent à des sous-classes des catégories, s'il y a lieu.

Les étiquettes ont donc un nom en 3 lettres quand elles sont composées de la seule catégorie (par ex ‘PRE’ pour préposition) ou, plus fréquemment, de 6 lettres quand elles sont composées de la catégorie et du type. Par exemple :
‘chevalier’ [<catégorie> = ‘NOM’, <type> = ‘com’ pour ‘commun’] correspond dans notre étiquetage à : ‘chevalier’ NOMcom (nom commun*)*

La composition des étiquettes des formes contractées (enclises/proclises) et des formes élidées est abordée en 1.4.

Pour le jeu Cattex09max, non encore mis en œuvre, outre les champs <catégorie> et <type> on a défini les champs suivants : <mode> ; <temps> <personne> ; <nombre> ; <genre> ; <cas> ; <degré>. Tous ne sont pas pertinents pour toutes les catégories.

La suite du document est consacrée à Cattex09min.

1.3. Principes d’étiquetage

- Chaque unité linguistique a **une étiquette morphologique et une seule** (voir infra les exceptions en 1.4.), composée des champs <catégorie> et <type> (sauf les catégories PRE, ETR, ABR, RED et OUT, composées de la seule catégorie).
Les champs comportent une valeur et une seule : la possible ambiguïté n’est pas prise en charge au niveau de l’étiquette, elle est résolue de deux façons :
- on ajoute une seconde étiquette dans un autre champ avec une valeur morphosyntaxique, par exemple dans les cas de nominalisation ;
- il existe un champ « note » dans lequel l’annotateur peut indiquer ses doutes.

- **L’étiquetage est morphosyntaxique** au sens où les catégories et les types sont déterminés en contexte selon des principes avant tout morphologiques (une annotation strictement syntaxique se fait par ailleurs). Dans quelques cas cependant on fait appel à des critères distributionnels pour déterminer plus précisément la valeur des étiquettes. Par exemple pour les démonstratifs, on distingue les emplois comme déterminants ou comme pronoms, selon qu’ils sont suivis ou non d’un nom :

  - Et la damoisele torne **cele part** si tost come il sont pres (*Graal*) [*cele* = DETdem]
  - Et **cele** dit que onques deseritee n' en fu (*Graal*) [*cele* = PROdem]

Les cas de ce type ont une double caractéristique :
- la distinction entre les différentes valeurs s’appuie sur des critères morphosyntaxiques bien identifiés et elle est largement admise ;
- aucune des valeurs n’est perçue comme plus essentielle à la forme que l’autre (*cele*, hors contexte, n’est pas plus déterminant que pronom, ou l’inverse).

Il existe deux autres cas en partie analogues à celui-ci, mais plus complexes :

a) Quelques unités linguistiques ont aussi plusieurs valeurs, selon les contextes, mais la distinction entre ces valeurs ne s’appuie pas toujours sur des critères morphosyntaxiques bien identifiés et objectifs. Elle peut relever d’un choix plus arbitraire. Il est par ailleurs difficile, pour ces unités comme pour les précédentes, d’identifier laquelle des valeurs serait essentielle à la forme ou première. On décide donc d’associer à ces formes la catégorie qui dans le contexte syntaxique paraît la plus appropriée.

Il s'agit majoritairement d'unités que l'on peut interpréter comme adjectifs qualificatifs ou comme noms communs : *ami ; ennemi ; fel/felon ; vassal ; ber/baron* (cf. liste en annexe 1 du *Manuel de référence du jeu Cattex09*).
Plus rarement, il s'agit d'unités qui peuvent être analysées comme des NOMcom ou des ADVgen/ADJqua : *bien, mal, voir* (cf. *Manuel de référence du jeu Cattex09*).

b) Pour d'autres unités linguistiques, on observe aussi plusieurs valeurs possibles, mais qui ne sont pas toutes enregistrées dans le lexique ; il existe une valeur morphologique de référence et d'origine, mais dans certains contextes morphosyntaxiques, l'unité acquiert, ponctuellement, une autre valeur. Par exemple dans :

- Et cele qui estoit la plus **dame** le menoit par la main et ploroit mout tendrement (*graal*),

le nom commun *dame* est ici assimilable à un ADJqua (présence de l'adverbe superlatif *plus*[1]).
Nous avons choisi, dans des cas comme celui-ci, de restituer la double valeur de l'unité linguistique, NOMcom et ADJqua. Mais, dans la mesure où chaque unité n'a qu'une étiquette, et que chaque étiquette n'a qu'une valeur, **nous avons opté pour un double étiquetage, morphologique et morphosyntaxique** (c'est-à-dire s'appuyant sur des principes syntaxiques pour l'attribution des valeurs des étiquettes).
Le jeu d'étiquettes utilisé pour les deux étiquetages est le même.
On a une étiquette **M (morphologique)** et une étiquette **MS (morphosyntaxique**).
Il existe donc pour toutes les unités deux étiquetages (et donc 2 étiquettes) correspondant à deux propriétés différentes. Par défaut les deux étiquettes ont la même valeur, mais dans certains cas la valeur de l'étiquette morphosyntaxique est différente de celle de l'étiquette morphologique.
Les principaux changements de catégorie - unités avec des étiquettes morphologique et morphosyntaxique différentes - sont les suivants (ils seront décrits de façon plus détaillée dans les paragraphes consacrés aux catégories et types en question dans le *Manuel de référence du jeu Cattex09*) :
NOMcom >ADJqua
ADJqua > NOMcom
VERinf > NOMcom
ADJqua > ADVgen
VERppe/VERppa > ADJqua
VERppe/VERppa > NOMcom

En annexe 1 du *Manuel* est proposée une liste (non exhaustive) des formes qui possèdent une double valeur morphologique ou qui changent de valeur dans certains contextes morphosyntaxiques.

- **On n'adapte pas les règles d'étiquetage à chaque texte** : le jeu, les principes et les règles sont les mêmes pour tous les textes et à toutes les périodes. Mais une unité linguistique peut évidemment changer de valeur d'un texte à l'autre (en particulier s'il y a un écart diachronique entre les textes). C'est généralement la valeur de l'étiquette morphosyntaxique qui change, plus rarement ce peut être aussi la valeur morphologique : ainsi *plaisir* est d'abord

[1] l'exemple est différent d'un exemple tel que « il est médecin », où *médecin* est bien interprété comme un NOMcom.

verbe, puis passe à un moment donné dans la catégorie des noms, et n'est plus attesté en tant que verbe (remplacé par *plaire*).

- **Latin et langues étrangères**

Tous les mots en langue étrangère ont la valeur ETR de l'étiquette M (morphologique).
Lorsque ces mots sont insérés dans un énoncé français, on leur attribue également l'étiquette Cattex09 qui correspond à leur catégorie en MS (morphosyntaxe).
Dans le cas des énoncés en langue étrangère (par exemple, en latin) qui sont complets et forment une unité syntaxique autonome, on n'associe pas aux mots étrangers d'étiquettes Cattex09 en MS et l'on peut utiliser un jeu morphosyntaxique spécialisé pour la langue en question si l'on souhaite décrire plus précisément la catégorie des unités.

1.4. Principes de segmentation graphique et de segmentation des unités linguistiques

L'étiquetage et la segmentation en mots des textes de la BFM s'opèrent sur la forme de surface des textes telle qu'elle a été définie par l'éditeur scientifique. On a donc pris le parti de respecter les graphies et les segmentations choisies par l'éditeur, même si les principes et les pratiques d'édition sont parfois hétérogènes. En général, chaque unité graphique est caractérisée par une étiquette qui lui est propre, l'unité graphique correspondant à une unité linguistique.

Il est cependant quelques cas bien identifiés, moins rares en français médiéval qu'en français moderne, où la segmentation graphique ne reflète pas la segmentation linguistique des unités. Nous avons donc été amenés à dissocier dans ces cas précis les unités linguistiques des unités graphiques. Mais nous n'avons jamais modifié la forme graphique et la surface des textes. L'étiquetage opère alors sur des unités linguistiques qui ne correspondent pas aux formes graphiques apparentes. Autrement dit, il arrive qu'à deux unités graphiques soit attribuée une seule étiquette. Il arrive à l'inverse qu'à une unité graphique soient associées deux étiquettes. Il arrive enfin qu'une unité graphique corresponde à deux unités contractées (enclise/proclise). Elle est dans ce cas associée à une seule étiquette, mais une étiquette complexe qui associe deux valeurs. On distingue donc trois cas de figure :

a) **fusion** : les cas où l'on regroupe deux formes graphiques pour leur associer une seule étiquette.
Cela ne se produit que dans 4 cas :
- ***quel que*** *lieu que* et ***ambes II***: sont traités comme une seule unité linguistique (car difficiles à analyser séparément) : déterminant relatif (DETrel) pour le premier, déterminant ou adjectif ou pronom cardinal (DETcar/ADJcar/PROcar) pour le second.

On traite aussi comme une seule unité :
- ***le dit*** = *ledit* (DETcom),
- ***le quel*** (et ses composés) = lequel (DETrel / PROrel)

Il convient de noter que l'espace blanc et l'apostrophe sont considérés dans ce cas comme faisant partie de la forme graphique.

b) **segmentation** : les cas où l'on segmente l'unité graphique pour lui associer deux étiquettes morphologiques
Cela ne se produit que dans deux cas :
- *parce* : on segmente en *par ce* : PRE + PROdem

- *sil/quil/qui* : dans de rares cas, l'éditeur a graphié la construction *se + il* en une seule unité, sans apostrophe comme s'il s'agissait d'un cas d'enclise (si + le) ; lorsque l'erreur d'interprétation de l'éditeur est manifeste, on la corrige en segmentant la forme graphique *sil* en deux unités linguistiques : CONsub + PROper[2]. On rencontre parfois aussi *qui* ou *quil* qui correspondent à des élisions non respectées (*qu'i* /*qu'il*) et donc à deux unités linguistiques : on segmente pareillement.
Dans tous ces cas, la segmentation informatique (ou la *tokenisation*) effectuée en amont de l'étiquetage permet de dissocier les deux éléments pour l'étiquetage et pour les outils de requêtes tout en préservant au niveau de l'affichage l'aspect de l'édition source. Il est par ailleurs possible de formuler des requêtes spécifiques sur ces éléments « agglutinés » (voir le *Manuel d'encodage*, <http://bfm.ens-lyon.fr/article.php3?id_article=158> pour les détails techniques et la *Référence rapide CQL pour la BFM* pour les exemples de requêtes).

Rappel : indépendamment des évolutions qui ont pu se produire dans la langue (tendance à la fusion de formes graphiques : *par mi > parmi*, *par ce > parce*, *ja mes > jamais*...), les habitudes graphiques des scribes étaient très fluctuantes pour certaines unités de même que le sont les pratiques de transcription des éditeurs ; dans la mesure où notre étiquetage est associé à une édition référencée d'un texte, nous avons choisi de respecter les choix opérés par l'éditeur de ce texte (à l'exception des quelques cas mentionnés ci-dessus), même si certains sont de toute évidence discutables.
Par conséquent, la segmentation de formes telles que *par mi/parmi* ; *ja mes / jamais, sicom/si com.*…est variable dans les textes, et le nombre d'étiquettes associées (une ou deux) aussi. Par exemple : *ja…mes* : ADVgen … ADVgen / *jamais* : ADVgen
par…mi : PRE… NOMcom / *parmi* : PRE
ne por quant : ne <ADVneg> por <PRE> quant <PROind>/ *neporquant* <ADVgen>

c) formes contractées et étiquettes complexes :
Les formes contractées sont beaucoup plus nombreuses et diversifiées en français médiéval qu'elles ne le sont en français moderne ; elles correspondent à des phénomènes d'enclise et de proclise On leur attribue une étiquette complexe, qui associe les deux valeurs des deux unités linguistiques contractées en une seule unité graphique.
Cette étiquette rend compte du caractère compositionnel/complexe de la forme, en incluant la catégorie et le type (le cas échéant) des deux composantes.
Par exemple *nel* est la contraction de *ne +le* , c'est-à-dire d'un adverbe négatif et d'un pronom personnel. L'étiquette de cette unité complexe a la forme compositionnelle suivante: ADVneg.PROper. Le point sépare les deux parties de l'étiquette complexe.
Voici ci-dessous la liste des différentes formes contractées ; elles sont traitées sous la première catégorie qui les compose (ADVneg.PROper est traité sous ADVneg) dans le *Manuel de référence*.

Liste des formes contractées

PROper.PROper (pronom personnel + pronom personnel) : *jel*
PROrel.PROper (pronom relatif + pronom personnel) : *kil, quel, kis, ques, quil, quis*
PROrel.PROadv (pronom relatif + pronom adverbial) : *quin*
PROrel.ADVneg (pronom relatif + adverbe de négation) : *quin*
ADVgen.PROper (adverbe général + pronom personnel) : *sil*

[2] Dans le cas plus fréquent où *sil* correspond à une enclise du pronom personnel *le* avec l'adverbe *si*, on utilise une étiquette complexe comme dans tous les cas d'enclise (voir ci-dessous).

ADVneg.PROper (adverbe de négation + pronom personnel): *nel*
PRE.DETdef (préposition + déterminant défini) : *del, ou*
PRE.DETcom (préposition + déterminant composé) : *audit*
PRE.DETrel (préposition + déterminant relatif) : *auquel, duquel*
PRE.PROper (préposition + pronom personnel) : *del*
PRE.PROrel (préposition + pronom relatif): *auquel, duquel*
PRE.PROint (préposition + pronom interrogatif): *asquex*
CONsub.PROper (conjonction de subordination+ pronom personnel) : *sel*

Les élisions sont traitées dans le corpus comme elles le sont généralement en français moderne. On considère l'apostrophe comme une limite de mot : « l'espee » correspond ainsi à deux unités, « l' » DETdef et « espee » NOMcom. Même chose pour « s'il » (< se il) CONsub + PROper, etc. Notons que l'apostrophe fait partie de la forme élidée.

Pour indiquer que la segmentation graphique ne correspond pas à la segmentation linguistique du texte, on utilisera l'étiquette temporaire OUT. Elle permettra de repérer les différences et de procéder ensuite à la fusion et au dégroupage des formes pour faire coïncider l'étiquetage avec le découpage en unités linguistiques : le <OUT> quel <PROrel> ou <DETrel>.

2. Tableau synthétique des étiquettes morphosyntaxiques CATTEX2009

Sophie Prévost, Céline Guillot, Alexei Lavrentiev, Serge Heiden

Le jeu CATTEX2009-max est défini par le tableau.
Le jeu CATTEX2009-min est défini par les colonnes CATEG et TYPE.

Légende :

- les cellules grisées correspondent à des traits non pertinents pour une catégorie ou un type donné, ces informations ne figurent pas dans l'étiquette compilée
- on utilise le tiret quand l'information est non pertinente pour une occurrence particulière (par exemple, le cas pour l'infinitif lié à un verbe conjugué modal ou le genre pour certains adjectifs cardinaux ou encore pour les pronoms relatifs adverbiaux comme *dont*)
- on utilise le *x* quand l'information n'est pas renseignée (par exemple, quand on ne sait pas si un guillemet est ouvrant ou fermant)
- on utilise le deux-points pour coder l'alternative
- on utilise le point pour les contractions

CATEG	***TYPE***	***MODE***	***TEMPS***	***PERS.***	***NOMB.***	***GENRE***	***CAS***	***DEGRE***	***Contr.***	***Commentaires, Exemples***
VER	cjg [1]	ind/imp/con/ sub [2]	pst/ipf/ fut/psp [3]	0/1/2/3 [4]	s/p [5]					Quant il la **voit** venir
	inf				s/p		n/r/- [6]			Quant il la voit **venir**
	ppe				s/p	m/f [7]	n/r			ce que j'ai toz jorz **celé**
	ppa				s/p	m/f	n/r			Et aussi fist il en **veillant**
NOM	com				s/p	m/f	n/r			en ceste **nuit**
	pro				s/p	m/f	n/r			**Boort**
ADJ	qua				s/p	m/f/n	n/r	p/c/s [8]		en pechié **mortel**
	ind				s/p	m/f/n	n/r			une **autre** nef
	car				s/p	m/f/n/-	n/r			apres ces **.ii.** vertuz
	ord				s/p	m/f/n	n/r			li **quarz** jorz
	pos			1/2/3	s/p	m/f/n	n/r			contre .i. **suen** voisin

PRO	per			1/2/3	s/p	m/f/n/-	n/r/i			Et aussi fist **il** en veillant
	per			1/2/3	s/p	m/f/n/-	n/r/i		.PROper	**jel** (= je le)
	imp			0	S	N	n/r			**il** me semble
	adv									il n'**i** dormist ja mes
	pos			1/2/3	s/p	m/f/n/-	n/r			je i lesseré le **mien**
	dem				s/p	m/f/n/-	n/r			Si estrange leu come **cist** est
	ind				s/p	m/f/n/-	n/r			si resgardent li **uns** l'**autre**
	car				s/p	m/f/n/-	n/r			si pria a **.ii.** de ses neveuz
	ord				s/p	m/f/n/-	n/r			Et quant il estoit venuz au **nuevieme**
	rel				s/p/-	m/f/n/-	n/r/i/-			Ce fu li premiers rois crestiens **qui** maintint le roiaume d'Escoce
	rel				s/p/-	m/f/n/-	n/r/i/-		.PROper	N'est hom **kil** (=qui le) veit
	rel				s/p/-	m/f/n/-	n/r/i/-		.PROadv	ki qu'en plurt u **kin** (= qui en) riet
	int				s/p/-	m/f/n/-	n/r/-			**Qui** estes vos ?
	com				s/p/-	m/f/n/-	n/r/-			**ledit** de Clerieux avoit creü
DET	def				s/p	m/f	n/r			Ce fu **li** premiers rois
	ndf				s/p	m/f	n/r			et mistrent **une** bele tombe sus lui
	dem				s/p	m/f	n/r			en **ceste** nuit
	pos			1/2/3	s/p	m/f	n/r			et i firent **son** non escrire
	ind				s/p	m/f	n/r			il sont parfet de **toutes** vertuz
	car				s/p	m/f	n/r			et quant il i a esté **.x.** ou **.xx.** ans
	rel				s/p	m/f	n/r			Il ne set **quel** chose puisse avenir

	int				s/p	m/f	n/r			**Quele** aventure vos a ça amenez ?
	com				s/p	m/f	n/r			**ledit** roy
ADV	gen							p/c/s/-		Il set **bien**
	gen								.PROper	**sil** (= si le)
	gen								.PROadv	**sin** (= si en)
	neg									**ne, pas, mie, point**
	neg								.PROper	**nel** (= ne le)
	int									**Quant** venistes vos ci ?
	sub									Il ne set **coment** il i puisse estre venuz
PRE										Il se mist ou grant chemin **de** la forest
									.DETdef	Il se mist **ou** (= en le) grant chemin de la forest
									.DETcom	lesquelz apportèrent **audit** (= a ledit) duc les clefz
									.DETrel	**auquel** (= a lequel) meffet tuit li oir partirent
									.PROper	Et tant fu desirranz **del** (=de le) veoir
									.PROrel	Tu sez **douquel** (=de lequel)
CON	coo									**et, mes…**
	sub									Et de cel serpent est tele la vertu **que se** nus hons tient nule de ses costes
	sub								.PROper	**sil** (= se il)
INJ										**Ha** fet la damoisele
PON	fbl									, : ; -
	frt									. ! ? …

	pga									« (
	pdr									»)
	pxx									" '
ETR										
ABR										
RED										Or dit li contes que quant mes sires Gauvains se fu partiz de ses compaignons **que** il chevaucha…
OUT										AOI (*Roland*)

[1] – cjg/inf/ppe/ppa : conjugué / infinitif / participe passé / participe présent
[2] – ind/imp/con/sub : indicatif / impératif / conditionnel / subjonctif
[3] – pst/ipf/ fut/psp : présent / imparfait / futur / passé simple (cette catégorie est non pertinente pour les modes impératif et conditionnel, seuls le présent et l'imparfait sont pertinents pour le mode subjonctif)
[4] – 0/1/2/3/ : impersonnel / 1ère personne / 2ème personne / 3ème personne
[5] – s/p : singulier / pluriel
[6] – n/r/i : nominatif (cas sujet) / régime / régime indirect (certains pronoms)
[7] – m/f/n : masculin / féminin / neutre (certains pronoms et adjectifs)
[8] – p/c/s : positif / comparatif / superlatif