Annotation de segments textuels. Guide de découpage.

Lucence Ing, Matthias Gille Levenson

Les listes qui suivent présentent les tokens à retenir pour la tokénisation. En fonction de l'état du texte (présence de coquilles issues de l'OCR ou d'erreurs sur le témoin original conservées), il faudra également retenir les tokens qui correspondent au mot indiqué (exemple, corpus italien : lnsi pour Insi). Les listes ne présentent que les formes sans majuscule initiale.

1 Segments sans délimiteur

Certains segments ne comportent pas d'élément à retenir pour la tokénisation, par exemple : *fut filz au plus recreant homme. et au plus failli de cueur.*

Les tokéniseurs retenus le sont à des fins de découpage de segments textuels automatiques. Ils sont donc basés sur des éléments grammaticaux récurrents. Par peur de créer du bruit dans les données, ne seront pas étiquetés les éléments qui débutent des phrases, lorsque leur usage est également très courant à l'intérieur des phrases. Ne seront donc pas étiquetés, pour le moment, même si cela est peu satisfaisant d'un point de vue intellectuel :

- les pronoms personnels, même quand ils débutent des phrases. Ex. : —*Si li anoia moult li cheualliers Il oste son* : « Si » n'est pas étiqueté
- les prépositions qui débutent les phrases. Ex. : *ne seuent nulle noueles de lancelot. Au mains —sil seussent celes* : *Au* n'est pas étiqueté
- les appellatifs qui débute les prises de parole. Ex. : *Sire fet elle iou men priserai de miex —se iou* : *Sire* n'est pas étiqueté

N.B. : Les listes qui suivent présentent certaines des variantes graphiques du français et du castillan médiéval, mais ne sont pas exhaustives à ce sujet.

2 Pronoms relatifs

2.0.1 Français médiéval

- *que*, *ke*, *qe*, *qu*
  ex. : *s'en va tos les esclos —qu'il trueve*
- *qui*, *ki*
  ex. : *dit a celi —qui a lui parloit*

— *cui*
ex. : *ciax —cui messires .G.*
— *dont*, *dunt*
ex. : *plus de .xl. —dont nos somes si hontex et si dolent*
— *ou*, *o*
ex. : *erroit par la —ou il cuidoit trouer gent*
— *(le)quel*, *(la)quelle*
ex. : *en ce lit —qui est emprez uous —lequel est assez beau*
quel est étiqueté tokéniseur quand il est relatif, exclamatif ou interrogatif, pas quand il est simple déterminant indéfini (*fait mander la nouvele dou tornoiement et a quel terme* : pas de tokénisation), ni quand il est coupé de *que* (*en quel lieu —que il fust*)
quand *quel* est répété, seule la première occurrence est étiquetée : *—quel besoins et quel auenture lauoit la amene*

— *quantque*, *quamque*
ex. : *—Et elle crioit —quamque elle pooit*

→ Quand les pronoms relatifs sont précédés d'une préposition, c'est cette dernière qui est étiquetée : ex. : *—de quel tornoiement li chevalier de ceanz parloient anuit apres vespres* (cf. *infra*, section 2.4).

→ *que*, *qui*, *dont*, *ou*, *(le)quel* peuvent être aussi des pronoms interrogatifs. Ils servent dans ces cas-là également à al tokénisation. ex. : *—ou est il donques*

2.0.2 Castillan médiéval

— « como » quand il introduit une proposition
— « si »
— « cuando »
— groupes « las cuales », « los cuales »
— « cuanto » (« *Dueña diz Miles, bien me nienbra / quanto me avedes dicho,* »)
— « en que »

2.0.3 Italien médiéval

— *che*, *ch*, *que*
— *chi*, *qui*
— *cui*
ex. : *e domanda a cui est quello arnese*
— *donde*
ex. : *donde questo gentile huomo est morto*
— *ove*
ex. : *della grande servitudine —ove noi eravano*

→ les pronoms relatifs mentionnés peuvent également être pronoms interrogatifs. Lorsqu'ils occupent cette fonction, ils servent toujours de tokéniseurs. Ex. : *Santa Maria, —chi mi gittera di qui ?*

2.1 Conjonctions

2.1.1 Castillan médiéval

Conjonctions de coordinations Les conjonctions sont considérées comme des délimiteurs sauf cas d'énumérations : « *É fuemos a surgir sobre Çepta, / e deçendimos en tierra ; / e luego tomaron una caravela / e escrivieron a Caliz a fazerlo saber,* » mais « *pero tiene buen puerto e muchas tierras, e frutas, e aguas.* »

— « e », « et », « y », « & »
— négations : « ni »
— « o », « u »

Autres conjonctions

— « pues »
— « ca », « que » quand valeur causale
— « pero », « sino », « mas » si valeur adversative

2.1.2 Français médiéval

Conjonctions de coordination Les conjonctions de coordination sont retenues lorsqu'elles articulent deux propositions avec verbe exprimé.

— *et*
ex. : *—si salue les chevaliers —et il dient*
et ne doit pas être retenue dans le cadre des énumérations (*se prenent a flatir et a debatre. —et il commence* : seul le deuxième *et* est à retenir pour la tokénisation), ni dans les propositions dans lesquelles le verbe est absent (*—si la salue moult courtoisement et il lie* : seul *si* est étiqueté)
— *ou*, *o*
ex . : *—et pense —qu'il fera, —ou s'il l'ocirra*
— *mais*
ex. : *—Mais il nel puet trouer*
— *car*, *kar*, *quar*
ex. : *trop bien le faisoient/ —car ilz dennoient hardement aux leurs*
— *ainz*, *ains*
ex. : *ni ot an palais un tout seul —qui mot deist. —ains sagenoillent*
Aussi quand il précède *que*. Ex. :*—ainz que Lanceloz fust levez*
— *ne*
ex. : *de ce —que je vos ai dit, —ne que vos*
ne est étiqueté seulement lorsqu'il est conjonction de coordination de deux propositions (*que encques ne se remua —ne ne dit mot* : seul le deuxième *ne* est à retenir pour la tokénisation), pas lorsqu'il coordonne deux autres éléments, ni lorsqu'il est adverbe négatif

Conjonctions de subordination

— *que* ex. : *il dit —que*
ex. : *l'empainst si durement —qu'il* (consécutive)

— *se* lorsqu'introduit une conditionnelle avec verbe exprimé
ex. : *—Et dautre part —se il li otroie samor*
On n'étiquettera en revanche pas *se* dans *se par moi non.*

2.1.3 Italien médiéval

Conjonctions de coordination

— *e, ed*
ex. : *egli ebbe conquiso Marigart il Rosso e elli ebbe la damigella diliverata*
— *o*
ex. : *—e monsignor Galvano apelrenda —o elli morra*
— *né*
ex. : *non vo guarentisce, —né ella non vo puote guarire*
— *ma*
ex. : *elli il potesse vincere, —ma cio no gli*
— *ché*
ex. : *si potea, —ché v'avea una altra entrata*

Conjonctions de subordination

— *che*
ex. : *ben sapiate —che colui est*
— *se*
ex. : *saggio —se vo' l credete*
— *sicché*
ex. : *si l fiede sicché gli trinca la testa*

2.2 Éléments marquant le début d'un discours direct

2.2.1 Castillan médiéval

Verbes de parole

— Decir est systématiquement délimiteur, qu'il introduise du discours direct ou indirect (« *A la postremeria dixo :/"Yo hire a Iherusalem la çibdat.* » et « *E despues que esto ovo fecho, dixo / que se queria tornar para su tierra* »). À l'inverse : « *razon dulz e sabrosa,/ lo que dixo don Christo,* ». Quand le sujet est postposé, on coupe après celui-ci : « *E dixo el Rey :/ ¿Commo fue eso ?* »
— Le fonctionnement est le même pour les autres marqueurs de discours (« preguntar » par exemple).
— Quand le verbe de parole est en incise, on le conserve dans le token « *señor dixo bores / yo no puedo agora alla tornar* »

2.2.2 Français médiéval

— *haa, ha*
ex. : *—Haa sire fit elle —que auez uous*

— *hee*, *he*

Mots débutant le discours direct Certains mots débutent de manière récurrente les prises de parole des personnages au sein du récit, et peuvent ainsi servir de délimiteur, sans prêter à confusion (cf. *infra*, section « tokens rejetés »)

— *certes*
  ex. : *—mais il nen est pas encor temps. —Certes fait elle*
— *oil*, *oy*
  ex. : *Sire entendes vous —que ces lettres dient. —oil fait*
— *voire*
— *naie*
  ex. *—Naie fait il —se uous nel me dites.*
— la préposition *par* (cf. *infra*, 2.4)

2.2.3 Italien médiéval

Mots débutant le discours direct

— *vere*
  ex. : *Vere, disse quelli, per mio*
— *certo*
  ex. : *Certo, ladro, mal me venisti a fare onta*
— la préposition *Per* (cf. *infra*, 2.4)

2.3 Adverbes

2.3.1 Castillan médiéval

— « ahora » quand il débute une nouvelle proposition (« *et pues que vos havemos fablado algo de la Cavalleria / agora queremos vos decir alguna cosa de la lealtat.* »)
— « cuando », « entonces »
— « luego »
— « otrosí »
— « entonces »
— « como » quand il introduit une proposition (« *ca era de mala vida / e facie mal su facienda / como dixemos.* », mais « *e me acomiendo en la vuestra gracia como a padre e a sennor* »)
— « tanto » quand il est suivi de « que »
— « según que »

2.3.2 Français médiéval

— *si*
  ex. : *Lancelot passe oultre —si encontre le duc en son uenir*

L'adverbe *si* est très employé en français médiéval, et très polysémique. Il est retenu en tant que tokéniseur lorsqu'il débute une proposition (*—si furent uestues et atornees*) mais pas lorsqu'il est employé comme intensifieur (*—si l'empainst si durement —qu'il* : seul le premier *si* sert de tokéniseur)

— *or*, *ores*
ex. : *—Or uenes aueques moy*
Seulement lorsqu'il se trouve en début de proposition

— *lors*
ex. : *auecques luy/ —lors le fait mettre en prison auec les*

— *alors*
—Alors dist aux escuiers/ et cheualiers
Seulement lorsqu'il se trouve en début de proposition

— *puis*
ex. : *le laissa malade durement —puis luy compta —comment*

— *après*
ex. : *la lumiere de uostre proece. —Apres quant li*
Quand il se trouve en début de proposition ou dans *après ce* : *—Aprez ce ne demoura gaires que .iiii. escuiers*

— *ainsi*, *ansi*, *ensi*
ex. : (*AInsi crioient tous contre lancelot grans et petis*)
Seulement quand il est employé seul en début de proposition

— *ançois*
ex. : *si vilains —que a ceuls del pais sen prenge. —ancois sen viegne*
Également quand il précède *que* : *—ancois que je puisse retorner*

— *devant*
Seulement quand il est employé avant *que*.
ex. : *ne seray aise deuant que ie en saiche la uerite*

— *quant*, *quand*, *qant*
ex. : *—Quant bohors oy —que la damoisele li prioit*

— *comment*, *coment*
ex. : *—et cil les lui dist telles —comment il les scauoit*
comment peut également être adverbe interrogatif : *—comment fait lancelot. ne scait il pas encore*

— *come*, *comme*, *com*
ex. : *oncques si grant ioye —comme ie auroye de ceste chose*
Seulement quand *come* introduit une proposition. *trop bien se defrendoient, com a tel meschief* : on n'étiquette pas

— *tandis*
ex. : *—Tandis quilz parloient ainsi regarda monseigneur gauuain*

— *tant*
tant est retenu lorsqu'il précède *que* (*la selle uuider —tant quil*) ou *comme* (*Si se cueure de son escu —tant comme il*) ou lorsqu'il est employé seul (*—tant ont alé*, mais pas lorsqu'il sert d'intensifieur à un adjectif ou un adverbe, ni lorsqu'il se trouve en incise entre un verbe et son participe (*ont tant alé*)

— *atant*
ex. : *compter deuant ung aultre —ATant sen part lan. de*
— *maitenant*
ex. : *luy escrie si hault —que bien le peult ouir —maintenant luy uient*
Seulement quand il est employé seul en début de proposition
— *orendroit*
ex. :
Seulement quand il est employé seul en début de proposition (*—si fet orendroit* : on étiquettte *si*, pas *orendroit*)
— *totesfois*, *totesvois*
Seulement quand il est employé seul en début de proposition
— *car* adverbe (précède un ordre)
ex. : *Damoisele, fet il, —kar me dites*

2.3.3 Italien médiéval

— *si*
ex. : *vide Boordo —si-llo isgrida*
— *anzi*, *insi*
ex. : *—anzi me n'andro la dond'io vegno*
— *altresi*
ex. : *—e si rimise in suo camino, —altresi come el giorno davanti*
— *dunque*
ex. :
Seulement quand débute une proposition.
— *poscia*
ex. : *—Poscia comanda a quelli.*
Seulement quand débute une proposition (on n'étiquettera par exemple pas l'adverbe dans *—come l'avete voi poscia fatto*).
— *quand*, *quando*
ex. : *volli celare, —quand'ella mi disse*
— *allor*
ex. : *—Allor prende Lancelotto l'anello*
— *atanto*
ex. : *—Atanto sono venuti al castello*
— *tanto*
ex. : *giurato —che giamai non finiremo d'andare —tanto che*
Quant est en tête de proposition, avec ou sans *que*
— *or*
ex. : *—Or vo ne potrete andare*
— *inanzi*
ex. : *nol facesse soppellire nella Gioisosla Guardia —inanzi ch'e' vi fosse venuto*
— *immantanente*
ex. :
Seulement lorsqu'il est en tête de proposition.

— *mantanente*
ex. :
Seulement lorsqu'il est en tête de proposition.
— *perché*, adverbe interrogatif
ex. —*perché l demandate voi*
— *come*
ex. : *Come l'avete vo poscia fatto*

2.4 Prépositions

2.4.1 Castillan médiéval

— « por » quand il débute une proposition infinitive : « *con los Cavalleros de la vanda / por probar Cavalleria* »
— item pour « para », qui est délimiteur quand il est préposition introduisant une infinitive (« *E pecan que do ay buenos abogados / para tirar a otros ganançias toman ellos menos preçio* », mais « nin son menester para aquel pleito, »)
— hasta

2.4.2 Ancien français

— *pour*, *por*
La préposition est retenue lorsqu'elle débute une proposition infinitive (—*pour uous faire compaignie*) mais pas dans les autres cas (—*que pour ce luy feussent cheuz* ; —*et pour ce se seuffre* ; *por ce* —*que nee en* ; —*et por la perte* —*que je i ai*)
— *par*
par peut servir de tokéniseur lorsqu'il débute un discours direct (*Par foi*), ce qui est très fréquent (correspond aux mots débutant un discours direct, cf. *supra*, 2.2) et lorsqu'il introduit un pronom relatif ou un pronom démonstratif (ex. : *moi* —*par cui ses freres auoit este ocis* ; cf. *infra*, éléments multiples, 3.3)
— *de*, *a*, *par*, *en* lorsqu'elles précèdent un pronom relatif
ex. : —*a qui li paueillons estoit.*
ex. : *bracquet ta tollu* —*de quoy tu te plains*
— *jusque*
Seulement quand introduit l'antécédent d'une relative (cf. infra : *iucquez a ce que ayez parle a moy* : on étiquette) ; mais pas quand il fonctionne avec un substantif (*et por lespee que chascuns daus a sentie iussqua sanc* : on n'étiquette pas)

2.4.3 Italien médiéval

— *per* est retenu lorsqu'il sert d'introduction à une proposition infinitive
ex. : *cavalieri erano mossi* —*per chiedere Lancellotto*
Il est également retenu lorsqu'il permet de débuter un discours direct (cf.

les éléments qui débutent un discours direct)
ex. : *—Per santa croce, sire cavaliere, vo non mi scamperete*
— *de*, *a*, *per*, *in* quand précèdent un pronom relatif
ex. : *quella —per cui era la entro rimaso*
Également quand ils précèdent des adverbes en début de discours direct : *domenica a otto giorni. —a certo che*
—

3 Cas particuliers

3.1 Agglutinations

En fonction de la base du texte sur lequel on travail et même des principes d'édition, on pourra trouver des formes agglutinées telles que « conjonction de subordination + pronom personnel ». On retiendra ces formes pour tokénisation :

— *quil*, *quilz*
— *quelle*, *quelles*
— *sil*, *silz*
— *selle*, *selles*

3.2 Pratique

Impossible détermination Puisque les extraits à étiqueter sont produits de manière randomisée, ils sont souvent coupés (i.e. ils ne correspondent pas à des phrases entières), et l'on ne peut savoir s'il s'agit d'un élément à étiqueter ou non, tant en début qu'en fin de segment.

ex. : *nos couvenoit aler dusqu'a outrance, car il est preuz et vistes et* : le dernier *et* peut être un coordonnant d'un autre adjectif (à ne pas étiqueter) ou un coordonnant de proposition (à étiqueter). Dans le doute, on n'étiquette pas.

Si, au contraire, le segment manquant peut être complété de manière certaine, on étiquette (ou pas).

ex. : *qui ilz ont tant de maulx souffers* : *qui* est précédé d'une préposition et ne doit donc pas être sélectionné comme tokéniseur (ici non plus, on n'étiquette pas).

Tokénisation initiale des mots dans le texte source La tokénisation initiale des mots a une influence sur les éléments à sélectionner. Ainsi, dans *A tans sen taist le*, *A tans*, écrit en deux tokens, correspond à l'adverbe *Atant* : on étiquette *A*.

3.3 Éléments multiples

Lorsque deux éléments (ou plus) servant à la tokénisation se suivent, seul le premier est identifié. Une série d'exemples suit :

— *et quant il voit cele par cui* : *et*
— *par cui ses freres auoit este ocis dist il* : *par*
— *Tandis quilz parloient ainsi regarda monseigneur gauuain* : *Tandis*
— *uous lauez fait puis que ie ne uous uiz* : *puis*
— *Mais que sil pouoit par quelque moyen* : *Mais*
— *et qui trop bien se defrendoient* : *et*
— *et si tost com uos fussiez desarmez* : *et*
— *uoye a senestre entreray ie par ce que les lyens* : *par* (locution)
— *sera moult ioyeuse si tost comment elle laura ueu* : *si* (locution)
— *qui lors fust ne qui puis nasquist* : [*qui-1*], *ne* :
— *paumes et des ienos. Si quil recueure un autre cop si quil* : *Si*, *si*
— *li conte que, quant mesire Gauvain se fu partis de sa compaignie,* : *que*
— *par l'eve qui ert envenimee si que a poi qu'il* : *si*
— lorsqu'une proposition relative se trouve dans une proposition indépendante qui vient d'être introduite par *et*, on n'étiquette pas le pronom relatif : *—e l'acqua per dove...*

3.4 Les « que » à ne pas étiqueter

que pronom relatif, pronom interrogatif, pronom exclamatif, conjonction sert de tokéniseur. Il existe cependant un nombre de cas où il ne sera pas retenu :

— lorsque *que* est précédé de *si*, *devant*, *ainz*, *ançois*, *ainsi*.
ex. : *ne seray aise —deuant que ie en saiche la uerite*
— lorsque *que* a valeur de coordonnant
ex. : *—quil ert las et traueillies. que del combatre que del cheualchier*
— lorsque *que* est précédé de *que a peu* (*que* redondant)
ex. : *les dens —que a pou qu il ne les luy a brisiees*
— lorsque *que* est restrictif : *ne se relieue —que a grans peine et si estoit il*
— lorsque *que* est employé dans une comparative, sans verbe exprimé, il n'est pas étiqueté.
ex. : *cheualiers et preus —et plus pris sa cheualerie que la monseignor Gauuain* (on étiquetera en revanche : *a censeiller mieulx —que ung aultre ne fera*)
Cela vaut aussi pour *comme* (et *comment*).
ex. : *comme li lievres devant les chiens.*
— dans certains des "cas difficiles" évoqués ci-après

3.5 Cas difficiles

— On peut parfois hésiter à prendre un mot ou un autre pour tokéniseur, particulièrement dans les phrases du type : *Si li conte tot ensi com il li estoit*. Dans cet exemple, on étiquettera *com* comme tokéniseur, en prenant en considération *ainsi* comme COD du verbe. En revanche, lorsque l'adverbe débute la proposition, c'est lui qui sera pris comme tokéniseur : *quar il sera par tamps garis. —si comme*

— De la même manière, lorsque *que* fonctionne avec *tant*, il peut être difficile de savoir quel token étiqueter. On retiendra *tant* lorsqu'il a débute la proposition (*sen uait tous les escloz —tant quil aconceut messire gauuain a lentree* ; *frapa si durement —quil lui fist la selle uuider —tant quil le* ; *gli va cavalcando per di suso il corpo —tanto che di tutto*) mais pas dans cas où il fonctionne comme complément (complément de verbe : *Si len dist tant que la royne set bien que ce fu* ; *et vont tant —qu'il vienent a unes praeries qui estoient en mi*, même si, dans ce dernier cas, il reste toujours difficile de déterminer si le sens est "vont tant" "qu'il vienent" ou "vont" "tant qu'il vienent" ; fonctionnement avec une préposition : *sai ie rien fors tant —que se ceroit trop grans damages*).
— Un même questionnement peut se poser pour l'identification du tokéniseur sur *si* ou sur *que* lorsque les deux mots sont accolés. Lorsqu'il ne fonctionne pas comme complément, on étiquette *si*.
ex. : *la sapine —si comme li contes a deuise* ; *ancienne. —si que li mur en estoient fendu et creue. tout ensi*
Cas encore plus difficile, il peut arriver que *si* et *que* ne soient pas accolés, mais fonctionnent néanmoins ensemble : *—si à tanto andato che venne al poggio.* Dans ce cas, on étiquettera seulement le premier élément, *si*
— Les doubles éléments ne sont pas systématiquement accolés, mais peuvent être séparés par des éléments. Seul le premier élément reste à étiqueter (on espère que le modèle apprendra).
ex. : *Et la biautez que me vaut > Et que me vaut la biautez* : *Et* seulement est étiqueté.
— Lorsqu'un pronom relatif est précédé d'une préposition, on étiquette la préposition (*trueue .i. pont de fust —par ou en passoit au chastel* et cf. *supra*, 2.4), mais pas quand *ce* est placé entre la préposition et le pronom relatif : *iamais ior de uostre vie ne parleres de ce —que iou vous. Idem* quand un pronom démonstratif est présent : *il se met tantost es rens encontre cels —que mesire Gauvain aidoit ;*. Une exception concerne l'emploi de *jusque* précédent *(a) ce que* : *—iucquez a ce que ayez parle a moy*
— Lorsque *pour* introduit un mot qui est suivi d'une relative, on étiquette seulement le pronom relatif.
ex. : *pour le chault —qui ia estoit encommance.*
— lorsqu'une comparative contient une relative, c'est le pronom relatif qui est étiqueté : *si difende al meglio —ch'egli puote come colui —ch'assai* (on n'étiquette pas *come*)
— si l'on considère *si tost que* comme une locution, impliquant une tokénisation au niveau de *si*, on étiquettera *que* dans tous les autres cas (*si durement que*, *si bien que*...).
ex. : *Et mesires Gauuain recueure si bien —que*
— les conjonctions permettant l'introduction de propositions après un verbe de prise de parole et assimilés (*dire que*, mais aussi *savoir que*, *mander que prier que*) sont étiquetées.
ex. : *dist quil est pres de faire tout ce —que il volront.*. Même s'il s'agit

quil et *que* sont deux tokens très proches, on étiquette les deux.

— des exceptions à cette règle peuvent facilement être trouvées, lorsqu'on trouve *que* dans une suite de tokens potentiellement tokéniseurs.
ex. : *mes pour ce quil cuydoit quil feust mort nen* : *mes*, *quil-2* et pas *quil-1* (exclu car dans la suite *mes pour ce quil*).

— les mots débutant une phrase (pronoms personnels, subtantifs, prépositionns, etc.) ayant un usage massif en-dehors de ces débuts de phrase ne sont pas étiquetés. En revanche, les démonstratifs de début de phrase le sont.
ex. : : —*que vous dites* —*Che fu fait il a lassamblee del Roy* ; —*Celle nuit. . .*
italien : —*Quella notte feciono quelli di la entro molto*