Principers de tokenisation du projet Presto

Principes de fusion et de segmentation

La liste des séquences ayant fait l'objet d'une fusion ou d'une segmentation, accompagnée de commentaires, est donnée dans le document « Liste_Fusions_Segmentations.ods »

1. Fusion : principes de décision

- **Principe 1** : Une séquence de 2 unités graphiques est fusionnée ssi 1 des unités qui la forme ne peut être traitée comme une unité linguistique car on ne dispose pas d'étiquette morphosyntaxique pour elle, quel que soit l'état de langue considéré.

Exemples :

La séquence *parce que* contient l'unité graphique *parce* à laquelle on ne peut assigner aucune étiquette morphosyntaxique. Diachroniquement, cette unité résulte d'une fusion graphique de la préposition *par* et du pronom *ce*. Face à ce type de cas, deux positions sont théoriquement possibles : soit on segmente *parce* [lemmes : PAR CE QUE[1] ; catégories : S Pd Cs], soit on fusionne *parce que* [lemme : PARCE QUE]. L'option choisie consiste à fusionner les deux unités graphiques.

La séquence *n'aguères* (*Essay sur l'histoire générale*,Voltaire, 1756) présente deux unités graphiques (*n'+ aguère*). Cette graphie est courante jusqu'au XVIII[e] s., *naguère* étant issu de la phrase *il n'y a guère*. L'UG *aguère*[2] ne peut se voir affecter une étiquette catégorielle. On procède à la fusion des deux unités graphiques.

- **Attention** : pour des états de langue anciens, il est possible que certaines UG aient un statut d'unité linguistique. Par ex. pour la séquence *au paravant* (*Lisandre et Caliste*, P. du Ryer, 1631), le *Complément du Dictionnaire de l'Académie* (dir. L. Barré, 1863) propose une entrée *paravant (adv).* Nous avons donc codé <Sp><Da><Rg> (lemmes À LE PARAVANT). De même pour *ce jourd'hui*, le même dictionnaire propose l'entrée *jourd'hui (nom)* ; nous avons donc codé <Dd> <Nc> (lemmes : CE JOURD'HUI)[3].

- **Principe 2** : On fusionne les séquences d'unités graphiques formant un nom propre dont l'orthographe moderne présente toujours une seule unité graphique.

Exemple :

Ménil-montant pour *Ménilmontant*

- **Principe 3** : Un choix d'analyse catégoriel nous a conduit à considérer une séquence spécifique d'unités graphiques comme relevant d'une classe d'unités linguistiques : les formes *de la*, *de l'* actualisant des noms massif ont été considérées comme variantes de l'article partitif DU.

2. Segmentation : principes de décision

Le choix dans Presto est d'opter pour une segmentation maximale.

- **Principe** : une unité graphique est segmentée en plusieurs unités graphiques / linguistiques ssi cette UG se présente *uniquement* sous forme segmentée en français moderne.

[1] Option adoptée par la BFM

[2] Ni d'ailleurs à *n'*, qui ne peut plus avoir le statut de morphème de négation dans ce contexte distributionnel (*ne* doit normalement porter sur un verbe) .

[3] Systématiquement, la prise de décision d'une fusion reposera sur l'absence d'entrée dictionnairique à l'un des dictionnaires proposés dans la base « Grand corpus des dictionnaires du 9e au 20e siècle » [GCDD9-20] des Classiques Garnier Numériques. Pour les cas de fusion, la date de consultation sera spécifiée dans les commentaires.

Exemple :

La forme *moymesme*, courante au XVII e s. notamment, est segmentée.

Rem : Les formes *ledit, notredit, votredit* etc. déterminant composé en a-fr est considéré comme appartenant au paradigme du déterminant composé défini « ledit » (étiquette : <Da>, lemme LEDIT).

3. Traitement des amalgames

Dans le cas des amalgames, étiquettes et lemmes sont séparés par + => Lemme1+lemme2; Etiquette1+étiquette2[4]

S+Da	amalgame préposition + art. défini : *du, des, au, au*
S+Pr	amalgame préposition + pronom relatif : *auquel, duquel, …*
S+Dr	amalgame préposition + déterminant relatif : *auquel, duquel, …*

Attention : d'autres phénomènes au XVI e s. à prévoir : *nel (ne le), ou, au (en le), …*
Pour la période IX e s. –XV e s., voir BFM.

[4] L'ordre des étiquettes est toujours lié à l'ordre des lemmes dans l'amalgame.