Étiquetage morpho-syntaxique
Annotation de données

Guillaume Wisniewski
guillaume.wisniewski@linguist.univ-paris-diderot.fr

novembre 2019

L'objectif de ce TP est de vous montrer la difficulté et le coût de l'annotation manuelle de données pour le TAL. Ce TP se compose de 3 parties — pensez à recharger le sujet au début de chaque séance :

1. annotation des données ;
2. évaluation de la qualité des annotations (accord inter-annotateur) ;
3. développement d'un étiqueteur morpho-syntaxique.

1 Annotation manuelle de données

L'objectif de la première partie est d'annoter manuellement une partie du corpus JW300`https://www.aclweb.org/anthology/P19-1310/` avec des étiquettes morpho-syntaxique.

Les données peuvent être annoter en ligne à l'url `https://drive.google.com/drive/folders/19` (il faut ajouter `r-zl-b7JXrTQudnUUrxeo9PtrIZ8Gh5?usp=sharing` à la fin de l'url pour accéder au répertoire).

① dans le fichier `Attribution_fichiers`, indiquez votre nom à côté du numéro de fichier que vous allez annoter (en évitant d'annoter deux fois le même fichier !)

② vous pouvez ensuite modifier directement le fichier `corpus_XX.xlsx` correspondant au numéro choisi.

Chaque sous-corpus contient 100 phrases segmentées en mots. L'annotation consiste à :

— attribuer à chaque mot une des 17 étiquettes morpho-syntaxiques définie par le projet UD (dans la colonne à côté du mot) ;

— indiquer d'éventuelles erreurs de segmentation en mots ;

— indiquer d'éventuelles erreurs de segmentation en phrase.

La figure 1 donne un exemple de phrase annotée. Les erreurs de segmentation en phrases sont à indiquer en modifiant la zone 1 ; les erreurs de segmentation en mots en modifiant la zone 2.

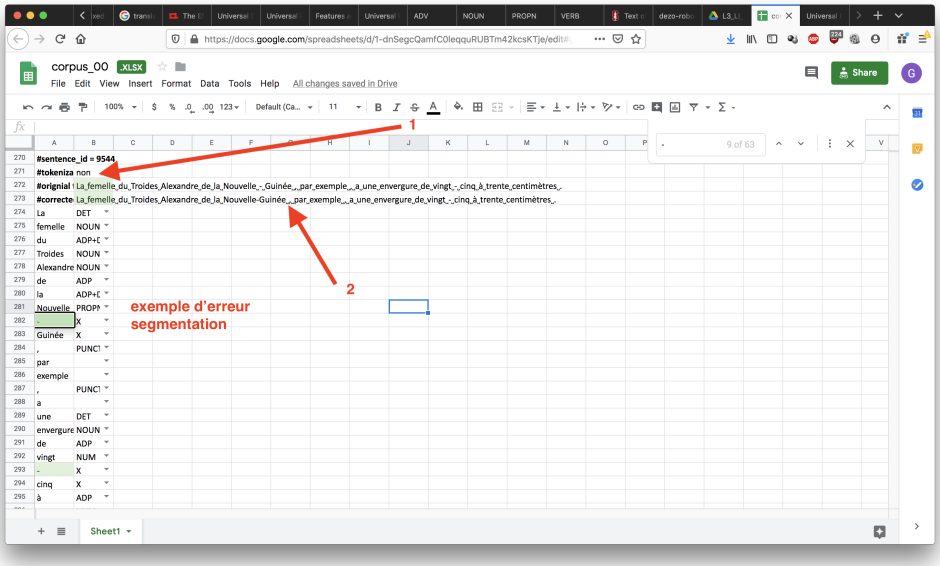

FIGURE 1 – Exemple d'annotation comportant une erreur de segmentation en mot.q

La définition des étiquettes est accessible à l'url `https://universaldependencies.org/u/pos/index.html`. Il est conseillé, en cas de doute, de regarder les décisions qui ont été prises pour annoter les corpus existants (p. ex. en cherchant, dans un corpus déjà annoté [1], les mots ou groupes de mots qui vous posent problème à l'aide de la commande `grep`). Nous considérerons également une étiquette supplémentaire ADP+DET correspondant à la contraction d'une préposition et d'un déterminant (p.ex. « au » qui correspond à la contraction de « à » et « le »).

1. Il est conseillé d'utiliser le corpus `UD_French-GSD`