L'oeuvre de l'outil, la part de l'humain. Pour une approche philologique de la constitution de corpus de données spontanées

Julie Glikman*[1], Nicolas Mazziotta*[2], and Christophe Benzitoun*[3]

[1]Analyse et Traitement Informatique de la Langue Française (ATILF) – Université de Lorraine, Centre National de la Recherche Scientifique – France

[2]Traverses [Université de Liège] – Belgique

[3]Analyse et Traitement Informatique de la Langue Française (ATILF) – Université de Lorraine, Centre National de la Recherche Scientifique : UMR7118, Centre National de la Recherche Scientifique – Université de Lorraine, 44 Av de la Libération, BP 30687 54063 Nancy Cedex, France

Résumé

D'immenses progrès ont été faits dans les outils automatiques pour traiter les corpus, tant dans la qualité des résultats qu'avec le développement d'outils libres et l'accessibilité, avec le développement d'interfaces permettant l'utilisation de ces outils sans avoir besoin de compétences techniques. Ces outils facilitent ainsi grandement le traitement des données et permet d'envisager la constitution de corpus enrichis de manière moins couteuse.
C'est dans cette optique que nous avons envisagé la constitution du corpus Les Vocaux, corpus de SMS vocaux originaux. Cette nouvelle pratique est intéressante pour les linguistes car il s'agit de nouvelles données écologiques, permettant potentiellement d'observer des phénomènes linguistiques émergents, peu présents dans les entretiens et les conversations présents dans les corpus de français parlé " classiques ". L'objectif du projet est de constituer un corpus distribué librement dans différents formats et avec différentes couches d'annotation. En premier lieu, le respect du RGPD a eu une incidence sur la procédure et une première phase manuelle d'anonymisation et de vérification du contenu des messages a été nécessaire. Nous avons ensuite utilisé des outils automatique à différents niveaux de notre chaine de traitement :

- transcription automatique des messages reçus

- détection automatique des pauses dans le signal sonore et alignement au phonème

- lemmatisation, étiquetage morphosyntaxique et analyse syntaxique

Cependant, si les résultats sont très bons, à tous les niveaux, une intervention humaine est nécessaire, que ce soit pour la préparation des données ou pour la correction / vérification des résultats. Cela est d'autant plus vrai dans la constitution d'un corpus de données hétérogènes, produites dans des contextes variés et donnant des matériaux de qualité différente, comme cela est le cas de notre corpus de SMS vocaux.

*Intervenant

Sur des grands corpus, on peut envisager de se passer de la vérification, en proposant que le nombre permettra de lisser les erreurs. Cependant, dans des données écologiques telles que nos SMS Vocaux, où ce sont justement les phénomènes rares qui peuvent se montrer les plus intéressants, il serait dommage de s'en passer. En fonction des étapes sur la chaine de traitement, la vérification est d'autant plus importante que la moindre erreur se répercute à tous les niveaux de l'analyse, comme pour la transcription. Il en est de même pour interroger les interfaces entre les différents plans d'analyse (phonétique / morphosyntaxe / syntaxe).

Dans cette contribution, nous proposerons un retour d'expérience sur le traitement de nos données, montrant à chaque étape l'oeuvre de l'outil et la part de l'humain, et en plaidant pour une approche philologique dans la constitution de corpus de données écologiques.