Constitution d'une collection de corpus oraux:
Une méthodologie pour l'étude des langues parlées en Afrique de l'Ouest

Abbie Hantgan
abbie.hantgan-sonko@cnrs.fr

Christian Chanard
christian.chanard@cnrs.fr

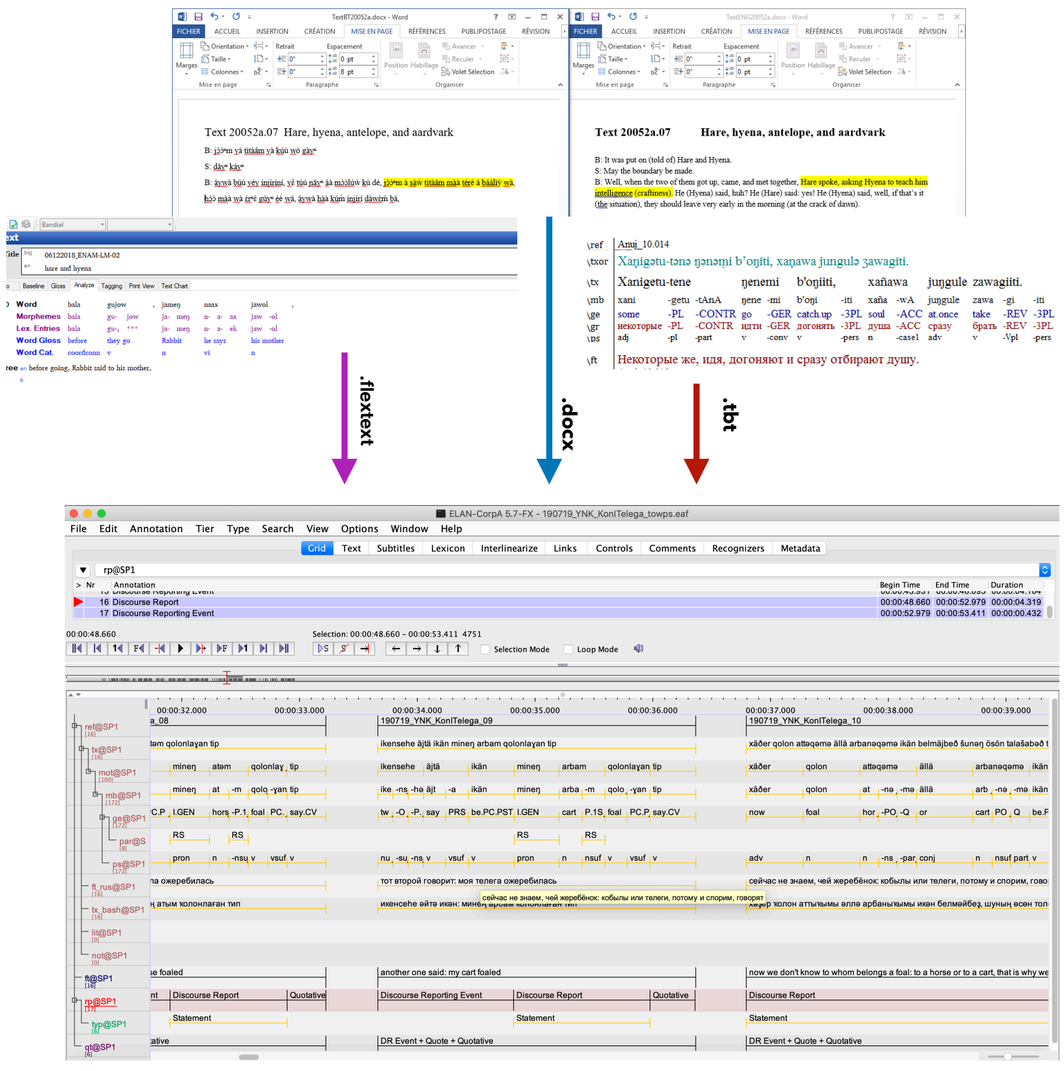
Katya Aplonova
aplooon@gmail.com

Introduction

Cette étude présente un cadre méthodologique pour la constitution d'un corpus multilingue comparatif avec traductions juxtalinéaires, composé des douze langues parlées en Afrique de l'Ouest. Les enregistrements audio et vidéo inclus dans le corpus ont été recueillis par plus d'une dizaine de chercheurs au Burkina Faso, Cameroun, Côte d'Ivoire, Mali, Nigéria, Sénégal et Togo.

Objectifs

Le but spécifique de cette recherche est d'examiner quelles sont les similitudes et les différences structurelles dans le cadre de contes traditionnels. Les données dans différents formats n'étant pas exploitables dans l'état, nous les avons fait converger vers un format commun. Nous avons choisi le logiciel ELAN pour sa capacité à importer des donnés provenant d'autres logiciels (Toolbox, Flex, Praat, texte tabulé...) et à maintenir un lien entre les transcriptions et les ressources media (audio, video) Plus spécifiquement, la version ELAN-CorpA (Chanard, 2018) qui permet d'annoter des textes à partir d'un lexique.

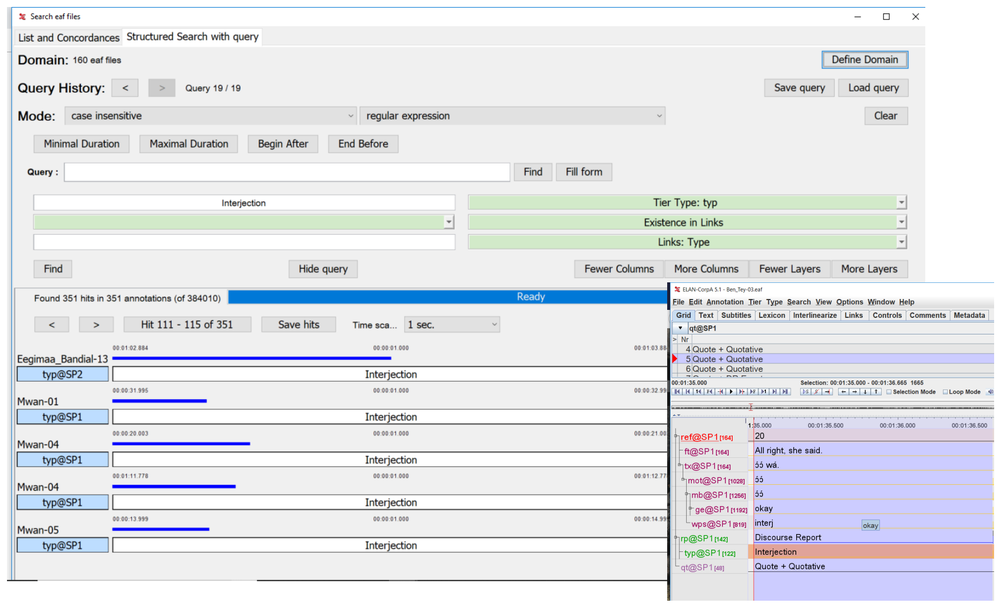


Annotation

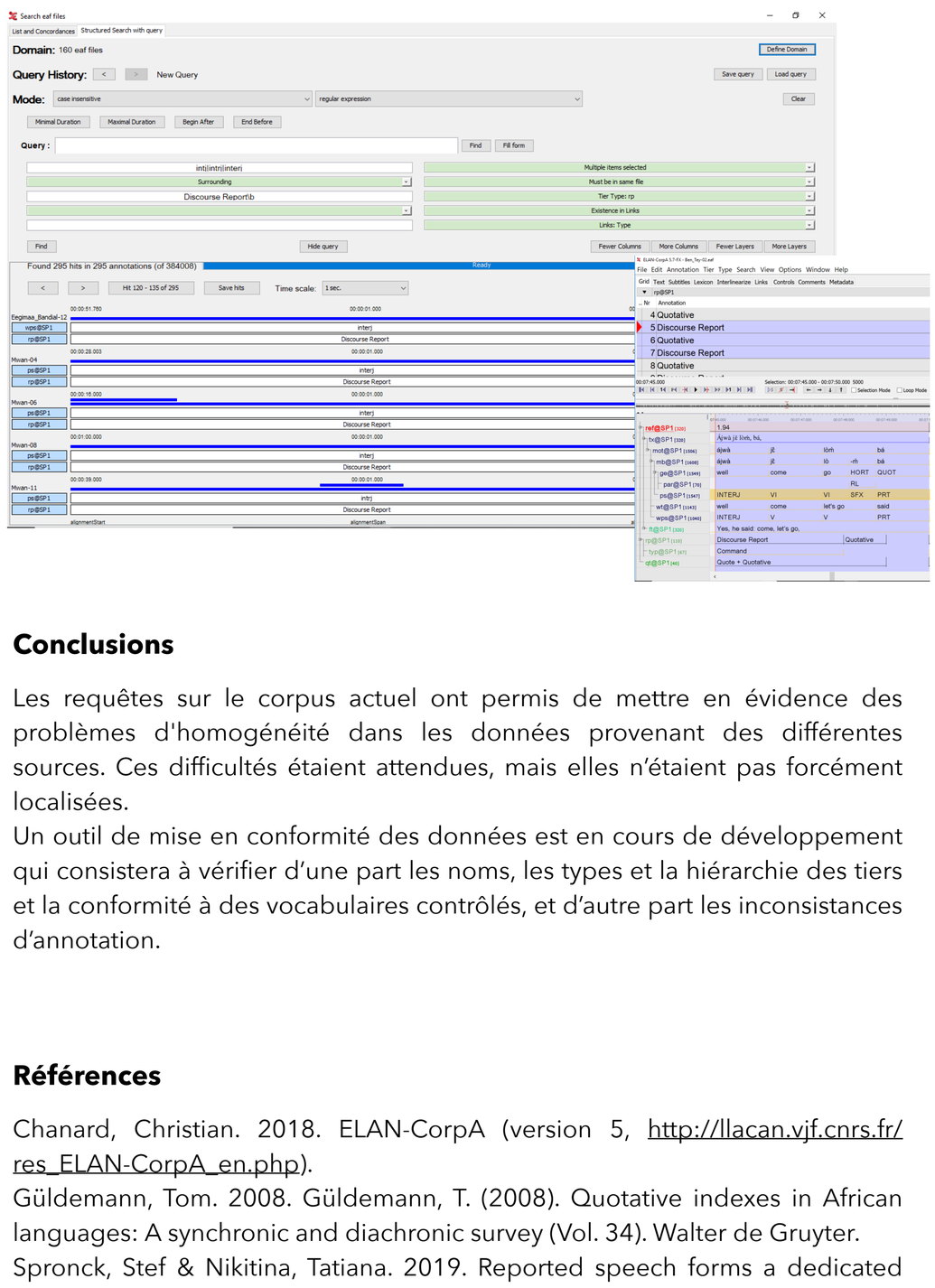
Les annotations morphosyntaxiques étaient nécessaires mais pas suffisantes pour l'étude du discours rapporté (cf. Spronck et Nikitina (2019): le discours rapporté constitue-t-il une catégorie syntaxique?). Par conséquent, les lignes d'annotation suivantes ont été ajoutées: RP (reported phrase), TYP (type of the reported phrase), QT (quoted type), PAR (participant) (Nikitina et al. 2019).

Example d'une requête : interjections

Les interjections sont souvent des éléments initiaux du discours rapporté (Güldemann 2008: 40, 220). En utilisant la fonction Multiple Files Search, il est possible de trouver toutes les occurrences des interjections qui apparaissent dans le discours rapporté. Il existe plusieurs variantes d'une telle requête. La requête la plus simple est de chercher l'annotation "Interjection" dans la tier *type de discours rapporté*.



Cette requête nous affiche les occurrences de discours rapportés où la parole est entièrement exprimée par l'interjection (e.g. ***Alright*** – *she said*). Afin de trouver les cas où les interjections ne font qu'une partie de la parole rapportée (e.g. ***Yes***, *he said, let's go*), il nous faut une requête plus compliquée où on cherche l'étiquette *interj* (ou *intrj*) dans la tier des parties de discours qui chevauche l'annotation "Discourse Report" dans la tier *segmentation de discours*



Conclusions

Les requêtes sur le corpus actuel ont permis de mettre en évidence des problèmes d'homogénéité dans les données provenant des différentes sources. Ces difficultés étaient attendues, mais elles n'étaient pas forcément localisées.

Un outil de mise en conformité des données est en cours de développement qui consistera à vérifier d'une part les noms, les types et la hiérarchie des tiers et la conformité à des vocabulaires contrôlés, et d'autre part les inconsistances d'annotation.

Références

Chanard, Christian. 2018. ELAN-CorpA (version 5, http://llacan.vjf.cnrs.fr/res_ELAN-CorpA_en.php).

Güldemann, Tom. 2008. Güldemann, T. (2008). Quotative indexes in African languages: A synchronic and diachronic survey (Vol. 34). Walter de Gruyter.

Spronck, Stef & Nikitina, Tatiana. 2019. Reported speech forms a dedicated syntactic domain. Linguistic Typology 119–159.

Nikitina, Tatiana & Hantgan, Abbie & Chanard, Christian. 2019. Reported speech annotation template for ELAN. (https://github.com/KatyaAplonova/ReportedDiscourseProject/blob/master/examples_of_annotated_files/SR-Template.etf)