Le treebank comme outil de description pour les langues orales : le cas de l'arabe tunisien

Fatma Ben Barka Messaoudi, Mustapha Khoudri, Rayan Ziane

Le treebank comme outil de description pour les langues orales :

le cas de l'arabe tunisien

Fatma Ben Barka-Messaoudi[1], Mustapha Khoudri[1], Rayan Ziane[2]

(1) LLL UMR7270, Université d'Orléans

(2) CRISCO UR4255, Université Caen Normandie

Mots-clés: arabe tunisien, corpus oral, langue peu dotée, universal dependencies, treebank

Introduction

La constitution et l'enrichissement d'un corpus est une tâche délicate, surtout quand il s'agit d'une langue privée de ressources et d'outils qui en facilitent le traitement. Nous nous reposons dans cette communication sur l'exposition de la situation de la plupart des parlers maghrébins à travers l'exemple de la réalisation d'un petit treebank issu d'un corpus échantillonné et diversifié d'environ 13 heures d'enregistrements de l'arabe tunisien effectués, à Orléans et en Tunisie, dans le cadre d'une étude doctorale (Ben Barka Messaoudi, 2022). Il s'agira de présenter ce type de ressources et leur apport en tant qu'outils de description de l'oral et des phénomènes liés à l'oralité, notamment les chevauchements, les coénonciations et le codeswitching.

Durée	772 (12h58min)
Nombre de mots	108705
Nombre de locuteurs	19
Lieux d'enquête	Orléans - Tunisie
Situations de communication	entretiens - repas - cours universitaire

Figure 1. Le corpus d'AT en chiffres

De la transcription au treebank

Nous aborderons, dans un premier temps, les principaux choix méthodologiques et techniques opérés afin de répondre aux contraintes rencontrées lors de la constitution de ce corpus arboré.

Dans cette première partie, il sera question tout d'abord d'expliquer les étapes faites pour transcrire en graphie latine nos données sur le logiciel TRANSCRIBER en adoptant les conventions de l'INALCO[1] pour la notation des caractères spéciaux et celles d'ESLO[2] pour la codification des particularités de l'oral. Ensuite, nous présenterons la démarche que nous avons suivie pour convertir ces données initialement transcrites sur Transcriber vers Elan (Parisse et al. 2020), ce qui offre la possibilité d'une annotation multicouche effectuée au format EAF. Une fois que nos données ont été traitées, nous avons entamé la phase de translittération du corpus vers la graphie arabe que nous avons pu automatiser grâce à l'API de Google Input-tool[3] et à l'outil ATAR (Talafha et al. 2021). Cette phase était indispensable dans la mesure où elle nous a permis d'accomplir un étiquetage hybride en catégories syntaxiques effectué par les parsers Farasa (Abdelali et al. 2016, Darwish et al. 2020) et Spacy (Honnibal & Montani 2017). Grâce au format EAF, nous avons pu exploiter les outils développés, pour le treebank UD_Beja-NSC[4] dans le cadre de l'élaboration du premier treebank pour le Bedja, afin d'extraire les données et d'en faciliter le traitement.

[1] Institut national des langues et civilisations orientales

[2] Enquête Socio-Linguistique à Orléans

[3] https://www.google.com/inputtools/

[4] https://universaldependencies.org/treebanks/bej_nsc/index.html

Nous proposons pour ce type de données une analyse morphosyntaxique au format CONLLU. Ayant la volonté de standardiser l'enrichissement de nos données, nous avons opté pour une réadaptation du cadre de travail Universal Dependencies (Nivre et al. 2020) (UD) en introduisant une tokenisation morphologique (Park et al. 2021, Kahane et al., 2021a) et en exploitant le jeu de relations de dépendance ainsi que le type d'analyse proposés par Gerdes et al. (2018).

Un treebank pour l'oral

À l'heure actuelle, la majeure partie des recherches sur les banques d'arbres syntaxiques se fondent sur l'écrit. Avec la diffusion de framework comme Universal Dependencies, les langues dépourvues de système d'écriture sont désormais intégrées dans la démarche et on voit de plus en plus d'initiatives allant en ce sens. Dans ce cadre, nous proposons une standardisation de la chaîne de traitement pour le développement de treebank pour les langues orales et peu dotées à travers l'exemple de l'arabe tunisien. Cette chaîne de traitement revisite notamment l'unité minimale de recherche en syntaxe grâce à une tokenisation morphologique (Kahane et al., 2021a). De nombreux linguistes de terrain disposent de données déjà analysées en gloses interlinéaires et sont prêts à enrichir leur corpus avec une annotation syntaxique. Il est *de facto* nécessaire d'offrir la possibilité d'une annotation basée sur les morphèmes, qui leur permettra de conserver cette structure. Nous nous attarderons également sur l'unité maximale en abandonnant le concept de phrase pour celui d'unité énonciative.

Enfin, nous aspirons au dépassement de l'arbre syntaxique comme objet fermé pour mettre en valeur les dynamiques de l'oral. Nous exposerons une annotation des chevauchements et des coénonciations en bénéficiant du système de métadonnées libre d'UD et de relations de dépendance syntaxique entre les énoncés (Kahane et al. 2021b, Oloff 2008). Par ailleurs, nous mettrons en avant le traitement automatisé du phénomène de code switching par une détection basée sur le "Lexique des formes fléchies du français" (Sagot 2020).

L'objectif de notre travail est de constituer un corpus de référence d'arabe tunisien qui pourra faire l'objet de prochaines recherches sur cette langue peu dotée, tout en incitant la communauté scientifique à reproduire, critiquer et améliorer notre démarche en apportant son expérience sur d'autres langues.

Références

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa : A fast and furious segmenter for arabic. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, 11-16.

Darwish, K., Attia, M., Mubarak, H., Samih, Y., & Abdelali, A. (2020). Effective Multi Dialectal Arabic POS Tagging. *Natural Language Engineering*, *1*(1), 18.

Ben Barka Messaoudi, F. (2022). *Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien*. http://www.theses.fr/s264750

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018, novembre). SUD or Surface-Syntactic Universal Dependencies : An annotation scheme near-isomorphic to UD. *Universal Dependencies Workshop 2018*.

Honnibal, M., & Montani, I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, *7*(1), 411-420.

Kahane, S., Vanhove, M., Ziane, R. & Guillaume, B. (2021a). A morph-based and a word-based treebank for Beja. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria.

Kahane, S., Caron, B., Gerdes, K., & Strickland, E. (2021b). Annotation guidelines of UD and SUD treebanks for spoken corpora. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria.

Nivre, J., Marneffe, M.-C. de, Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). *Universal Dependencies v2 : An Evergrowing Multilingual Treebank Collection*.

Oloff, F. (2008). La complétude négociée des unités de construction de tour : Les complétions différées comme ressource en français parlé. *Congrès Mondial de Linguistique Française 2008*, 085.

Parisse, C., Etienne, C., & Liégeois, L. (2020). TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI. *Journal of the Text Encoding Initiative*. https://halshs.archives-ouvertes.fr/halshs-03043572

Park, H., Schwartz, L., & Tyers, F. (2021). Expanding Universal Dependencies for Polysynthetic Languages : A Case of St. Lawrence Island Yupik. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 131‑142. https://doi.org/10.18653/v1/2021.americasnlp-1.14

Sagot, B. (2010, mai). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. *7th international conference on Language Resources and Evaluation (LREC 2010)*.

Talafha, B., Abuammar, A., & Al-Ayyoub, M. (2021). ATAR: Attention-based LSTM for Arabizi transliteration. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(3), 2327‑2334.