De la transcription à la translittération d'un corpus d'arabe tunisien : enjeux, problèmes et choix

Fatma Ben Barka Messaoudi, Rayan Ziane, Mustapha Khoudri

De la transcription à la translittération d'un corpus d'arabe tunisien : enjeux, problèmes et choix

Comme toutes les « langues à peine sorties de l'oralité, [et] pour lesquelles il existe une concurrence entre plusieurs alphabets » (Chaker & al. 2002), l'arabe tunisien souffre, malgré quelques tentatives récentes (OTTA[1] Zribi & al. 2013a, CODA[2] Zribi & al. 2014), du manque de ressources et d'outils conçus pour son traitement.

La nécessité de mettre en place un système de notation stable qui tienne compte des spécificités de cette langue a émergé dès notre première confrontation aux données que nous avons collectées à Orléans et en Tunisie, dans le cadre d'une étude doctorale intitulée *Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien.*

Durée (min)	772
Nombre de mots	109705
Nombre de locuteurs	19
Lieux d'enquête	Orléans - Tunisie
Situations de communication	entretiens - repas - cours universitaire

Tableau 1: le corpus d'arabe tunisien en chiffres

Afin de faire face à la rareté des travaux disponibles sur le parler tunisien, nous avons dû opérer certains choix méthodologiques et techniques (concernant l'outil, le système, le mode et les conventions de transcription).

Dans cette communication, nous exposerons, dans un premier temps, toutes les étapes effectuées lors de la transcription usuelle d'inspiration morphophonologique de nos enregistrements sur le logiciel TRANSCRIBER[3]. Nous expliquerons également les motifs des choix de la graphie latine comme système de notation et des conventions de l'Institut National des Langues et Civilisations Orientales pour codifier les caractères spéciaux, ainsi que celles des Enquêtes sociolinguistiques à Orléans pour noter les particularités de l'oral.

Ensuite, nous présenterons, dans un second temps, la démarche suivie pour convertir nos données brutes initialement traitées sur TRANSCRIBER vers le format propre au logiciel ELAN[4] propice à l'annotation multicouche (Parisse & al. 2020), ce qui nous a permis d'entamer par la suite la phase de translittération automatique du corpus vers la graphie arabe, grâce à l'API de l'outil Google Input[5] et le translittérateur ATAR[6] (Talafha & al. 2021).

---

[1]Orthographic Transcription of Tunisian Arabic
[2]Conventional Orthography for Dialectal Arabic
[3]C'est un logiciel, qui a été développé par Claude Barras et Edouard Geoffroy de la Direction Générale de l'Armement (DGA), pour transcrire plusieurs langues (européennes et/ou non européennes).
[4]Il s'agit d'un outil d'annotation pour les enregistrements audio et vidéo.
[5]Cet outil a été développé par Google pour faciliter le passage d'un système alphabétique à un autre.
[6]Attention-based LSTM for Arabizi transliteration

Enfin, nous tâcherons de montrer l'apport du traitement automatique des langues dans la manipulation des données issues des langues peu dotées, à partir de la démonstration des processus de segmentation morphologique et d'étiquetage syntaxique de notre corpus.

Nous souhaiterions, par ce travail, faire avancer le débat sur les enjeux théoriques et pratiques de la description d'une langue sous-documentée, en encourageant la communauté scientifique à reproduire, critiquer et améliorer la démarche entreprise sur l'arabe tunisien et à réfléchir aux « bonne pratiques » permettant d'harmoniser nos différentes approches.

Références bibliographiques

BEN BARKA MESSAOUDI, Fatma. *Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien.* PhD Thesis. Université d'Orléans. 2022.

CHAKER, Salem, et al. Les langues de France et leur codification-Ecrits ouverts, écrits divers. 2002.

PARISSE, Christophe, et al. A conversion tool for spoken language transcription with a pivot file in TEI. *Journal of the Text Encoding Initiative*, 2020, no 13.

TALAFHA, Bashar, et al. Attention-based LSTM for Arabizi transliteration. *International Journal of Electrical and Computer Engineering*, 2021, vol. 11, no 3, p. 2327.

ZRIBI, Inès, et al. Orthographic transcription for spoken Tunisian Arabic. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, Berlin, Heidelberg, 2013. p. 153-163.

ZRIBI, Inès, et al. A conventional orthography for Tunisian Arabic. In : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014. p. 2355-2361.