Pour la segmentation automatique de l'arabe parlé : l'exemple de l'arabe tunisien

Fatma Ben Barka Messaoudi, Rayan Ziane, Meriem Beraik

Pour la segmentation automatique de l'arabe parlé : l'exemple de l'arabe tunisien

Fatma Ben Barka-Messaoudi[1], Rayan Ziane[2], Meriem Beraik[2]

(1) Laboratoire EMA, CY Cergy Paris Université INSPE

(2) CRISCO EA4255, Université Caen Normandie

Le traitement des données issues de l'oral est une tâche complexe, surtout quand il s'agit d'une langue privée d'orthographe usuelle. En effet, malgré le nombre croissant des travaux s'intéressant à l'étude du parler tunisien (Zribi & al., 2015 ; Ben Ahmed & al., 2018 ; Ben Barka 2018, Goncalves Teixeira & al., 2023), les tâches de transcription et d'annotation des corpus collectés sur cette langue ne cessent de soulever plusieurs interrogations, ce qui rend difficile leur exploitation linguistique.

Si plusieurs réponses ont été avancées concernant le choix du système, du mode ou bien des conventions de transcription, la question de la segmentation, indispensable à la description morphosyntaxique, n'a pas été suffisamment explorée.

Tel est le cas de la majorité des parlers arabes, l'arabe tunisien, langue fortement agglutinante, (Hamdi, 2015; Zribi 2016), se caractérise par le rattachement des clitiques aux mots simples, ce qui permet d'obtenir des formes plus complexes dites agglutinées. Séparer les clitiques de la forme fléchie est une tâche nécessaire dans une perspective de traitement des données par des outils de Traitement Automatique des Langues (TAL).

Conçues pour l'analyse de l'arabe standard moderne, les applications comme Arabic Part-of-speech Tagger (Khoja, 2001), Sakher (Chalabi, 2004) ou Aramorph (Linguistic Data Consortium, 2004) ne sont pas adaptées au traitement de nos données orales en arabe tunisien, transcrites en graphie latine sur TRANSCRIBER ou translittérées automatiquement grâce à l'API de l'outil Google Input[1] et le translittérateur ATAR (Talafha & al., 2021). Plus récemment, on compte parmi les ressources existantes le pipeline d'annotation multi-couche (Gugliotta & al., 2020) et le corpus TARC (Gugliotta & Dinarelli, 2020), qui est propre à la modalité écrite de l'arabe parlé au Maghreb annoté en partie du discours.

Soucieux de faire avancer le débat sur les opérations d'exploration lexicale et d'étiquetage morphosyntaxique de l'arabe tunisien, nous avons choisi de segmenter un échantillon de 876 mots, soit environ 5min d'enregistrements, en 1392 morphs (Ben Barka Messaoudi & al., 2022). Ce jeu de données nous servira de base d'inférence de règles et de corpus d'apprentissage pour entraîner un système de segmentation automatique (Harrat & al., 2019) pouvant traiter notre corpus comportant 109705 mots (Ben Barka Messaoudi & al., 2023). Cette étape est déterminante et préalable aux opérations d'étiquetage grammatical et d'analyse syntaxique étant donné que les clitiques possèdent leurs propres parties du discours et fonctions syntaxiques.

En amont du traitement des unités minimales d'analyse du corpus, il convient également, pour l'analyse syntaxique en dépendance, d'effectuer une autre segmentation au niveau des unités maximales traitées. Ces séquences doivent manifester une complétude sémantique, une

[1]Cet outil a été développé par Google pour faciliter le passage d'un système alphabétique à un autre.

autonomie syntaxique et une clôture intonative. Nous montrerons, à travers l'exemple de l'arabe tunisien, comment les récentes technologies d'apprentissage machine et en particulier la diarization automatique[2] (Plaquet & Bredin, 2023) peuvent aider au prétraitement de corpus arabes oraux.

Bibliographie:

BEN AHMED, Yossra, BADIN, Flora, et HRIBA, Linda. Constitution d'un corpus oral de l'arabe tunisien: une ressource essentielle pour l'étiquetage morphosyntaxique. In : *TALAF 2018: Traitement automatique des langues africaines (écrit et parole)*. 2018.

BEN BARKA MESSAOUDI, Fatma. Sur les enjeux méthodologiques de la construction d'un corpus d'arabe tunisien. In : *TALAF 2018 : Traitement automatique des langues africaines (écrit et parole),* Grenoble, France. 2018. ⟨halshs-04451827⟩

BEN BARKA MESSAOUDI, Fatma, KHOUDRI, Mustapha, et ZIANE, Rayan. Le treebank comme outil de description pour les langues orales. In : *CEDIL22 Sciences du langage : Enjeux théoriques et pratiques méthodologiques*. 2022. (hal-04095252)

BEN BARKA MESSAOUDI, Fatma, ZIANE, Rayan, et AISSANI, Anissa. L'apport des ESLO pour la documentation du continuum linguistique dans le petit Maghreb. In : *Actes des 11èmes Journées Internationales de la Linguistique de Corpus*. 2023. (hal-04356978)

GONCALVES TEIXEIRA Daphne, VANCAEYZEELE Charle, BAHRI Mohamed Malek. Transcription Automatique de l'Arabe Parlé à Tunis : Un Pont vers l'Analyse Linguistique. In Proceddings: Actes des 5èmes journées du Groupement de Recherche CNRS « Linguistique Informatique, Formelle et de Terrain ». pp.135, 2023.

GUGLIOTTA, Elisa, DINARELLI, Marco, et KRAIF, Olivier. Multi-task sequence prediction for Tunisian Arabizi multi-level annotation. *arXiv preprint arXiv:2011.05152*, 2020.

HAMDI, Ahmed. *Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard: application à l'étiquetage morphosyntaxique*. Thèse de doctorat. Aix-Marseille. 2015.

HARRAT, Salima, MEFTOUH, Karima, et SMAÏLI, Kamel. Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation. *Computación y sistemas*, 2019, vol. 23, no 3, p. 979-989.

PLAQUET, Alexis et BREDIN, Hervé. Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*, 2023.

ZRIBI, Ines. *Traitement automatique du dialecte tunisien: construction de ressources linguistiques*. 2016. Thèse de doctorat. Université de Sfax (Tunisie).

ZRIBI, Inès, ELLOUZE, Mariem, BELGUITH, Lamia Hadrich, *et al.* Spoken Tunisian Arabic corpus "STAC": transcription and annotation. *Research in computing science*, 2015, vol. 90, p. 123-135.

[2] Tâche de partition d'un flux audio en segments propres aux différents locuteurs