L'apport des ESLO pour la documentation du continuum linguistique dans le petit Maghreb

Fatma Ben Barka Messaoudi, Rayan Ziane, Anissa Aissani

L'apport des ESLO pour la documentation du continuum linguistique dans le petit Maghreb

Ben Barka Messaoudi Fatma [1], Rayan Ziane [2] et Anissa Aissani [2]
[1] Laboratoire EMA, CY Cergy Paris Université INSPE [2] Laboratoire CRISCO, Université Caen
fatma.messaoudi1@cyu.fr, rayan.ziane@unicaen.fr, anissa-fella.aissani@etu.unicaen.fr

Introduction

Au cours de ces dernières années, les avancées informatiques ne cessent de renouveler les recherches linguistiques. Comme le note Baude (2007 : 85), les toutes nouvelles technologies de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils (transcriptions synchronisées sur le signal, annotation, etc.) ouvrent des perspectives prometteuses pour les études sur les corpus de langues parlées. Dans ce mouvement collectif de collecte et de diffusion de données orales, le Laboratoire Ligérien de Linguistique a joué un rôle important par la mise à disposition des enquêtes sociolinguistiques à Orléans (ESLO) à tout.e chercheur/euse intéressé.e. Ce corpus dispose de trois principaux atouts : 1) des données en masse (10 000 000 mots), 2) des données situées (métadonnées informant sur le profil du locuteur en termes d'âge, de sexe et de la catégorie socioprofessionnelle) et 3) des données micro-diachroniques (ESLO1 1968, ESLO2 à partir de 2008). Cette sélection de données tenant compte de plusieurs variétés langagières nous amotivés à construire un grand corpus comparable en arabe maghrébin.

Situation linguistique dans le petit Maghreb

Le paysage linguistique dans le petit maghreb se caractérise par la coexistence de l'arabe standard moderne (ASM) et l'arabe parlé. Bien que les parlers tunisien, algérien et marocain constituent les langues maternelles des locuteurs maghrébins, ils ont été largement déconsidérés et mis à l'écart. En effet, l'idéologie arabo-musulmane a longtemps occasionné la non reconnaissance de toute forme d'évolution linguistique en considérant que ces parlers sont le résultat de la décadence dues au contact. À l'exception du contexte médiatique et plus particulièrement du domaine publicitaire, ces variétés, privées de descriptions grammaticales et d'orthographes stables, sont exclues des programmes d'éducation, des écrits administratifs et parfois des émissions télévisées au profit de l'ASM.

Ne bénéficiant ni de statuts officiels, ni de ressources linguistiques [1], ces langues sous-documentées manquent de données et d'outils en libre accès facilitant leurs descriptions sans tomber dans les pièges de l'intuition.

Ayant la volonté de faire avancer le débat sur les enjeux théoriques et pratiques de la documentation et de la description des langues peu dotées, nous avons exploité la démarche de la constitution d'ESLO pour constituer notre grand corpus d'arabe maghrébin [2].

Méthodologie

Exploitation de la méthodologie d'ESLO

Le réservoir ESLO illustre un ensemble de bonnes pratiques à préconiser pour minimiser la complexité des procédures de collecte et de transformation des données primaires en données secondaires. Dans la mesure du possible, nous les avons suivies afin de construire un corpus partageable et interopérable.Nous avons favorisé comme mode de collecte des données l'entretien semi-dirigé en face-à-face, "situation certes très formelle, mais qui avait l'avantage d'être (...) contrôlable."(Abouda et Baude, 2006 : 4). Notre corpus compte aussi un certain nombre d'entretiens réalisés via des plateformes

1. Hormis quelques tentatives récentes de recueil des parlers tunisien et marocain (Graja & al. 2010, Zribi & al 2015, Moukrim 2010 ...) s'inscrivant dans le sillage du développement général de la linguistique de corpus et du traitement automatique du langage.

2. Deux premières enquêtes en parlers tunisien et algérien ont été déjà entamées 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 (2022) et d'un mémoire de recherche (en cours) portant sur les représentations langagières en Algérie. L'objectif suivant est de mener la troisième enquête en arabe marocain.

de visio-conférences (en particulier pour la partie algérienne du corpus). Le guide d'entretien a été puisé dans les principaux thèmes retenus par ESLO (logement, travail, loisirs, langues, ville d'habitation), tout en ajoutant d'autres thématiques (sur la révolution tunisienne et sur les représentations langagières en Algérie) susceptibles de faire parler les locuteurs tunisiens et algériens.

Afin de maintenir un certain équilibre au sein de notre corpus, deux autres genres interactionnels de « contrôle », i.e. les repas et les conférences universitaires sont et seront intégrés, à hauteur de 20%, tout en respectant les exigences de nos terrains. En ce sens, quelques aménagements ont dû être faits lors du recueil des conférences universitaires. En effet, nous nous sommes rendus compte que, dans ce contexte formel, l'arabe parlé cède sa place à l'ASM. Nous avons donc décidé de remplacer les conférences des ESLO par des cours universitaires. Néanmoins, cette décision n'a pas été suffisante pour résoudre ce problème étant donné que la plupart des cours universitaires privilégient la variété standard. Nous nous sommes donc retournés vers les cours artistiques (de musique, de danse et de dessin) où les professeurs ont plus de liberté pour s'exprimer dans la langue vernaculaire du pays.

D'autres choix méthodologiques et techniques ont été opérés afin de faire face aux contraintes rencontrées sur terrain. Nous pouvons citer à titre d'exemple, le changement du lieu de l'enquête (d'Orléans en Tunisie) ou le recours à des enregistrements en visio-conférence (données algériennes) pour certains enregistrements afin de gérer le manque de quelques profils sociologiques ou le déséquilibre intérieur en termes de variations diaphasique et diatopique.

Dans l'objectif de représenter la quasi totalité du continuum linguistique maghrébin, nous avons également opté pour un échantillonnage des locuteurs en nous basant sur les principes suivants :

- nous reposer sur trois catégories socioprofessionnelles (CSP) : cadres, employés, ouvriers ;
- classer les locuteurs sélectionnés de chaque CSP en trois tranches d'âge 15-35 ans, 35- 55 ans et plus de 55 ans ;
- choisir un homme et une femme de chaque catégorie d'âge.

Données situées protégées

Conscients des avantages offerts par les données situées dans le traitement des faits linguistiques, nous avons veillé à récolter tous les renseignements concernant nos enregistrements ainsi que leurs contextes de production en mettant en place deux formulaires :

- un formulaire Témoin comportant des informations sur l'âge, le sexe, le niveau scolaire, la profession, la CSP, les langues parlées et lieu de naissance.
- un formulaire Enregistrement informant sur la situation de communication enregistrée, sa date, son lieu, sa durée et le nombre de participants.

Afin de protéger nos locuteurs et nos données des éventuels problèmes de collecte, de transmission et de la propriété intellectuelle, nous avons demandé aux témoins la signature d'un texte de consentement écrit synthétisant le cadre et les finalités de nos enquêtes.

Traitement

L'absence de standardisation du code orthographique pour les parlers maghrébins nous a poussés à questionner les pratiques opérées de façon systématique dans les transcriptions du corpus ESLO. Là où une transcription phonétique serait trop coûteuse en termes de temps, nous avons adopté une transcription morphophonologique, proche des formes usuelles répandues sur les réseaux sociaux communément caractérisées par le terme Arabizi (Yaghan, 2008). C'est aussi par souci d'universalisation des données que nous avons eu recours à la graphie latine et aux caractères relatifs à l'API permettant de combler les lacunes du premier alphabet, en suivant les conventions de l'Institut National des Langues et Civilisations Orientales (INALCO). Néanmoins, pour codifier les spécificités de l'oral, nous nous sommes servis des propositions d'ESLO. Tenant compte des faveurs des outils conçus pour le traitement automatique de la langue arabe et des méthodes d'apprentissage automatique, nous avons ensuite réalisé une translittération automatique des données transcrites en caractères latins vers les

caractères arabes, grâce à l'outil API de Google Input et le translittérateur ATAR (Talafha et al. 2021).

De la diffusion à l'archivage

Notre projet s'inscrit dans une démarche de science ouverte et de diffusion des données de la recherche, ainsi nous prévoyons de mettre à disposition les transcriptions et toutes formes d'annotations ajoutées. Toutefois, les enregistrements audio constituent des données à caractères personnels sensibles, ce qui exclut leur diffusion publique. Par ailleurs, un archivage des données pérenne est planifié à l'issue de la réalisation du projet afin de conserver ce paysage sonore des parlers du petit Maghreb.

Conclusion

Face aux innombrables "corpus fantômes" cités dans la littérature sur les parlers du petit Maghreb, nous proposons le premier corpus "équitable" pour ce continuum linguistique. Équitable étant donné qu'il répond aux principes FAIR (D. Wilkinson et al, 2016) qui doivent êtres opérés comme une ligne directrice pour tout projet qui vise à partager et à rendre accessible les données de la recherche. Notre corpus sera aussi Findable/facile à trouver, cela implique qu'il se doit d'être stocké dans un site/plateforme scientifique et décrit par des métadonnées pour faciliter les recherches. Notre corpus sera également Accessible et ouvert, en accès libre, à tous les acteurs et actrices de la communauté scientifique. Findable et Accessible via la plateforme HUMA-NUM, nos données seront aussi Interopérables car elles seront sous format largement diffusé et faciles à lire et à traiter, au format XML dans notre cas. Notre corpus sera ensuite Reusable/réutilisable puisque nous permettrons son partage et sa réutilisation.

Enfin, ce corpus a pour vocation à être doté d'une annotation morphosyntaxique sous la forme de treebank en nous inspirant de la méthodologie proposée par Kahane et al. (2021) dans les formalismes Universal Dependencies (De Marneffe et al., 2021) et Surface Syntactic Universal Dependencies (Gerdes et al, 2018).

Références bibliographiques

Abouda, L., & Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO. *In Corpus en Lettres et Sciences sociales, Des documents numériques à l'interprétation.*

Baude, O. (2007). Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux. *Revue française de linguistique appliquée*, (1), 85-97.

Ben Barka Messaoudi, F. *Étude contrastive du subjonctif en français parlé à Orléans et de ses éventuels équivalents en arabe tunisien.* PhD Thesis. Université d'Orléans. 2022.

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

De Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational linguistics, 47*(2), 255-308.

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.*

Kahane, S., Vanhove, M., Ziane, R. & Guillaume, B. (2021). 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.

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

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9.

Yaghan, M. A. (2008). " Arabizi" : A Contemporary Style of Arabic Slang. *Design Issues, 24*(2), 39-52.

Index