Bootstrapper son corpus oral ou comment développer un corpus d'arabe parlé dans le petit Maghreb

Rayan Ziane, Fatma Ben Barka Messaoudi

Bootstrapper son corpus oral ou comment développer un corpus d'arabe parlé dans le petit Maghreb

Rayan Ziane[1,2] Fatma Ben Barka Messaoudi[3]
(1) Centre de Recherches Inter-langues sur la Signification en COntexte (CRISCO), Université de Caen
(2) Laboratoire Ligérien de Linguistique (LLL), Université d'Orléans
(3) Ecole, Mutation, Apprentissage (EMA), CY Cergy Paris Université INSPE
rayan.ziane@univ-orleans.fr, fatma.messaoudi1@cyu.fr

Les dernières années ont connu un essor des techniques d'analyse linguistique automatique basées sur l'adaptation (fine-tuning) de modèles pré-entraînés à partir de petits corpus annotés. Peng et al. (2022) ont notamment démontré l'efficacité d'une approche agile de développement de corpus arborés, consistant à enrichir progressivement un corpus d'apprentissage par des données issues d'analyses automatiques d'abord imprécises, corrigées manuellement par des experts. La transcription contribue au cycle de documentation des langues en se plaçant en interface des autres types d'analyses du plan phonétique à l'analyse du discours en passant par la morpho-syntaxe.

Dans cette communication, nous proposons de répliquer cette méthodologie pour le développement de corpus oraux. La démocratisation des techniques de reconnaissance automatique de la parole (ASR) ouvre de nouvelles perspectives en ce sens (Baevski et al., 2020; Pratap et al., 2023). Par affinage d'un modèle multilingue pré-entraîné sur des quantités colossales de données, quelques heures transcrites manuellement sont suffisantes afin d'entraîner un premier système imparfait pour faciliter la transcription manuelle de plus de données avant de parfaire le système et ses prédictions durant un processus itératif qui se nourrit des transcriptions validées par l'expert.

La détection automatique des tours de parole et la discrimination des locuteurs bénéficient également de ce bond technologique (Plaquet et Bredin, 2023). Là encore, le modèle pré-entraîné généraliste peut être adapté avec des segmentations éditées manuellement. Cette approche permettrait de gérer efficacement les spécificités du langage oral, telles que les chevauchements de paroles, les faux départs et les interruptions fréquentes. En identifiant précisément les tours de parole, la qualité de la transcription peut être améliorée et rendre compte fidèlement des dynamiques de l'interaction entre locuteurs.

En outre, Guillaume et al. (2022) montraient comment cette méthode ouvre des perspectives prometteuses pour le traitement et l'inclusion des langues peu dotées. Les langues minoritaires ou moins représentées dans les grands ensembles de données peuvent bénéficier de cette approche itérative.

Le contexte de notre contribution est celui du continuum linguistique de l'arabe parlé au petit Maghreb (Algérie, Maroc, Tunisie), dont les variétés orales sont dépourvues de système orthographique standardisé. Pendant longtemps, le paysage linguistique dans le petit Maghreb se caractérisait par la prédominance de la variété standard non seulement dans le milieu

scolaire mais aussi médiatique et scientifique pour des considérations politiques et idéologiques d'inspiration fortement religieuse. Néanmoins, depuis quelques années, nous assistons grâce à la multiplicité des outils informatiques et à l'avènement de l'algorithmique (apprentissage profond) à des tentatives de documentation des variétés parlées qui ont été jusqu'à présent largement mises à l'écart.

Pour remédier à la non-accessibilité des données existantes et à la non-adaptabilité des corpus disponibles pour nos besoins, nous avons choisi de collecter un corpus du parler maghrébin, dans la perspective de mise à disposition. Animés par la volonté d'alignement sur un ensemble de bonnes pratiques déjà bien établies (Abouda & Baude, 2006; Baude et al., 2006; Gadet & Guerin, 2016; Wilkinson et al., 2016), notre méthodologie de collecte s'inspire de celle des ESLO (Enquêtes Sociolinguistiques à Orléans) (Baude & Dugua, 2016), adaptée aux spécificités du contexte maghrébin. Nous évoquerons la phase déterminante de conception du corpus qui comprend la définition des objectifs de la recherche et des types de données nécessaires (entretiens, conversations spontanées, etc.), ainsi que la méthodologie d'échantillonnage de locuteurs en termes de lieu géographique, de tranche d'âge, de genre et catégorie socio-professionnelle.

Ce corpus inclut des données variées, situées, protégées, transcrites, translittérées et segmentées en tours de parole. Il répond ainsi à une diversité de besoins et espère poser les bases de la constitution d'un corpus de référence et comparable entre les trois pays du petit Maghreb. Partagé entre les idées selon lesquelles la transcription serait une voie à éviter (Bird, 2020) et la réalité des pratiques dictées par la tradition et l'implantation forte d'outils d'analyses, nous proposons ici de reconsidérer une phase du développement de corpus oraux longtemps tenue pour subsidiaire par une étude exploratoire sur l'arabe parlé au petit Maghreb. Face aux interrogations impossibles à résoudre quant à la convention de transcription adéquate, l'interaction entre la transcription manuelle et les prédictions du système automatique dans le processus itératif peut s'avérer bénéfique dans la prise de décision guidée par des contraintes pragmatiques.

Bibliographie:

Abouda, L., & Baude, O. (2006). *CONSTITUER ET EXPLOITER UN GRAND CORPUS ORAL : CHOIX ET ENJEUX THEORIQUES. LE CAS DES ESLO*. https://halshs.archives-ouvertes.fr/halshs-01162506

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations* (No. arXiv:2006.11477). arXiv. https://doi.org/10.48550/arXiv.2006.11477

Baude, O., Blanche-Benveniste, C., Calas, M.-F., Cappeau, P., Cordereix, P., Goury, L., & Jacobson, M. (2006). *Corpus oraux, guide des bonnes pratiques 2006*. 209.

Baude, O., & Dugua, C. (2016). Les ESLO, du portrait sonore au paysage digital. *Corpus*, *15*. https://doi.org/10.4000/corpus.2924

Bird, S. (2020). Sparse Transcription. *Computational Linguistics*, *46*(4), 713‑744. https://doi.org/10.1162/coli_a_00387

Gadet, F., & Guerin, E. (2016). Construire un corpus pour des façons de parler non standard : « Multicultural Paris French ». *Corpus*, *15*. https://doi.org/10.4000/corpus.3049

Peng, Z., Gerdes, K., & Guiller, K. (2022). Pull your treebank up by its own bootstraps. In L. Becerra, B. Favre, C. Gardent, & Y. Parmentier (Éds.), *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)* (p. 139‑153). CNRS. https://hal.science/hal-03846834

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *24th INTERSPEECH Conference (INTERSPEECH 2023)*, 3222‑3226. https://doi.org/10.21437/Interspeech.2023-205

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2023). *Scaling Speech Technology to 1,000+ Languages* (No. arXiv:2305.13516). arXiv. https://doi.org/10.48550/arXiv.2305.13516

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18