Offre de stage : Détection de la variation graphique dans une langue non standardisée (dialectes alsaciens)

Contexte

Les langues peu dotées présentent des défis spécifiques pour le Traitement Automatique des Langues (TAL) ; le manque de ressources textuelles volumineuses complique l'utilisation d'approches empiriques et, dans certains cas, comme celui de l'alsacien, l'absence d'une orthographe standardisée nécessite de gérer la variation graphique. Dans ce contexte, le projet MeThAL (Laboratoire LiLPa, Université de Strasbourg) est en train de créer un corpus large de théâtre en alsacien qui aidera à la création de ressources linguistiques pour les dialectes alsaciens ainsi qu'à une étude quantitative de la tradition dramatique alsacienne. Dans le cadre du projet, environ 4 000 pages de texte océrisé corrigé ont été produites, sur la base de numérisations en mode image créées par la Bibliothèque nationale et universitaire de Strasbourg. Une interface permet d'explorer les textes et métadonnées disponibles, et un sous-corpus de 300 000 tokens encodé selon les recommandations de la *Text Encoding Initiative* (TEI) a été publié. Pour pouvoir comparer le contenu des textes du corpus et effectuer des analyses thématiques ou textométriques, une représentation orthographique homogène du vocabulaire est nécessaire, et une neutralisation des variantes graphiques est incontournable. Elle serait aussi utile pour offrir une recherche en texte intégral sur le corpus.

Plusieurs approches ont été proposées pour la détection de la variation graphique. La plupart d'entre elles procèdent par normalisation, c'est-à-dire la transformation des variantes vers une forme standard. Une telle approche n'est pas applicable aux dialectes alsaciens, en raison de l'absence de standard orthographique stable. Millour & Fort (2019) ont utilisé le crowdsourcing pour collecter auprès des locuteurs de l'alsacien différentes graphies d'un mot donné. Les variantes alignées sont utilisées pour extraire automatiquement des règles de variation puis apparier automatiquement des graphies alternatives potentielles. Des méthodes non supervisées de clustering ont également été adoptées (Dasigi & Diab, 2011; Rafae et al., 2015). L'utilisation de ressources externes comme des lexiques bilingues ou des réseaux sémantiques multilingues a été proposée par Bernhard (2014). Il est également possible d'utiliser des méthodes supervisées, qui nécessitent toutefois des corpus annotés permettant d'identifier les variantes. Par exemple, Barteld et al. (2019) génèrent des variantes candidates qui sont ensuite filtrées à partir des n-grammes de caractères qu'elles contiennent et la similarité de leurs plongements de mots, ainsi que leurs contextes d'occurrence.

Activités du stage

Dans le cadre du stage, il s'agira dans un premier temps d'explorer les habitudes de scripturalisation (utilisation de certains caractères et n-grammes de caractères) en fonction des métadonnées disponibles (auteur, lieu de naissance, lieu de publication, maison d'édition, date, genre). La discriminativité des tendances dégagées pourra être éprouvée sur des tâches de classification en fonction des métadonnées. Le corpus pourra éventuellement être enrichi à l'aide d'un étiquetage morphosyntaxique automatique, dont la qualité sera à évaluer, compte tenu des spécificités du corpus (genre, période) : un intérêt particulier du

corpus est son caractère non-contemporain (1870-1940) ; il présente des divergences orthographiques par rapport aux pratiques actuelles qui demandent une adaptation des ressources existantes. Les activités suivantes sont prévues :

- Description approfondie du corpus (globale et par sous-corpus) : fréquence de caractères, de n-grammes, etc. (et, éventuellement, comparaison avec d'autres corpus de périodes plus récentes)
- Identification des propriétés discriminantes (p. ex. en proposant une représentation vectorielle des textes basée sur les différentes propriétés choisies)
- Induction de règles de variation et extraction automatique de paires de variantes au sein du corpus. Comparaison du résultat avec celui issu de l'application des règles obtenues par Millour & Fort, (2019) ; des différences sont attendues en raison des périodes des corpus respectifs
- Évaluation et proposition d'amélioration de la méthode

Profil recherché

Master en Informatique ou Traitement automatique des langues. Intérêt pour les problématiques du stage.

Conditions

Niveau : Master
Durée du stage : De 4 à 6 mois
Temps de travail : De préférence à temps plein
Date de début : Février ou mars 2022
Rémunération : Taux légal en vigueur (environ 575 € / mois)
Lieu : Télétravail ou hybride (au laboratoire LiLPa - Linguistique, Langues et Parole à Strasbourg)

Encadrement : Alice Millour (alice.millour@gmail.com), Delphine Bernhard (dbernhard@unistra.fr), Pablo Ruiz (ruizfabo@unistra.fr)

Comment candidater

Merci d'envoyer un relevé de notes récent et un CV aux encadrant.e.s avant le 7 janvier. Décrivez brièvement votre motivation pour le stage dans le corps du mail.

Références

Barteld, F., Biemann, C., & Zinsmeister, H. (2019). Token-based spelling variant detection in Middle Low German texts. *Language Resources and Evaluation*, *53*(4), 677–706. https://doi.org/10.1007/s10579-018-09441-5

Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources: The Example of Alsatian. *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, 23–29. https://hal.archives-ouvertes.fr/hal-00966820

Dasigi, P., & Diab, M. (2011). CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 318–326. https://aclanthology.org/I11-1036

Millour, A. & Fort, K. (2019). Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling*, RANLP,* 776–784. https://aclanthology.org/R19-1090.pdf

Rafae, A., Qayyum, A., Moeenuddin, M., Karim, A., Sajjad, H., & Kamiran, F. (2015). An Unsupervised Method for Discovering Lexical Variations in Roman Urdu Informal Text. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 823–828. https://aclanthology.org/D15-1097.pdf