Atelier : Préparer son corpus pour le rendre utilisable avec des outils d'analyse textuelle

Loïc LIEGEOIS [1,2] et Achille FALAISE [2]
[1] CLILLAC-ARP – Université Paris Cité
[2] LLF – CNRS – Université Paris Cité
loic.liegeois@u-paris.fr, achille.falaise@u-paris.fr

Introduction

L'atelier que nous proposons est adapté de sessions de formation assurées dans le cadre des activités du consortium CORLI (Consortium d'Huma-Num).

Notre atelier s'adresse à des personnes qui :

- disposent de données textuelles dans un format « brut » : texte brut, formats traitement de texte comme MSWord ou OpenDocumentText, PDF… Les données textuelles peuvent être issues de données primaires nativement écrites (web, articles de presse, œuvres littéraires, manuels scolaires...) ou obtenues suite à la transcription de données primaires orales (captations en situation naturelle d'interaction, entretiens, émissions de radio…)
- souhaitent analyser leur données textuelles avec un outil d'analyse de corpus oral ou écrit : AntConc (Antony, 2022), CLAN (MacWhinney, 2000), ELAN (MPIP, 2022), Hyperbase (BCL, 2023), Le Trameur, Lexico, TigerSearch, TXM (Heiden, Magué et Pincemin, 2010)…
- ne savent pas comment procéder au formatage de leur corpus afin de le rendre utilisable par les outils d'analyse de corpus oral ou écrit

L'atelier pourra se dérouler sur une journée entière, et privilégiera les personnes non spécialistes (non linguistes et/ou totalement novices en linguistique de corpus). En lien avec la thématique des JLC, nous encouragerons les didacticiens et didacticiennes (enseignants du premier et du second degré, enseignant-chercheur en didactique ou en sciences de l'éducation…) à s'inscrire à l'atelier.

Objectifs

Les objectifs de l'atelier sont les suivants :

- formation théorique rapide à la notion de données textuelles : format de document et encodage des caractères
- formation théorique rapide à la notion de corpus textuel : recueil, structuration, annotation, analyse outillée
- formation pratique à la préparation d'un corpus linguistique en vue de son exploitation outillée

L'objectif final est le suivant : nous souhaitons qu'à la fin de l'atelier chaque participant ou participante reparte avec une version de leur corpus exploitable par un outil d'analyse textuelle.

Organisation

Les participants et participantes à l'atelier seront contactées par les formateurs plusieurs jours en amont afin de leur communiquer leurs données (ou un échantillon), leurs besoins et leurs objectifs.

Après la brève formation théorique, chaque jeu de données fourni sera observé collectivement. Une discussion avec l'assemblée aura ensuite lieu dans le but de déterminer collectivement les actions à entreprendre afin de structurer les données textuelles dans un ou des formats permettant l'exploitation du corpus à l'aide d'un ou de plusieurs outils d'analyse textuelle.

Ainsi, en fonction des données dont nous disposerons, nous nous attendons à devoir aborder les problématiques techniques suivantes (liste non exhaustive) :

- encodage des caractères
- tokenisation, lemmatisation, étiquetage morphosyntaxique, étiquetage syntaxique en dépendances
- conversion des données à l'aide d'outils comme teiconvert (Liégeois et al., 2015 ; MoDyCo, 2016) par exemple pour convertir du texte brut vers le format TXM, du texte au format MSWord vers le format ELAN ou bien des données au format CLAN vers le format TXM
- nettoyage, conversion ou formatage « manuel » des données à l'aide des expressions régulières pour, par exemple, rendre des données au format texte brut analysables grâce à des outils comme AntConc, CLAN ou Lexico

Conclusion

Notre atelier s'adresse donc à un public varié et souhaite principalement aider les personnes néophytes à s'initier à la linguistique de corpus en se focalisant sur l'étape de la préparation des données. L'atelier n'a donc pas pour vocation de former les participants et les participantes à l'utilisation des outils d'analyse textuelle mais bien de se focaliser sur la méthodologie à mettre en œuvre afin de constituer un corpus linguistique exploitable par ces outils. Les formateurs sauront en revanche orienter et conseiller les personnes qui souhaiteraient se former à l'analyse textuelle outillée.

Références bibliographiques

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bases, Corpus, Langage (2023). Hyperbase (Version 10). Nice : Université Côte d'Azur et CNRS.

ELAN (Version 6.4) (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Heiden Serge, Magué Jean-Philippe, & Pincemin Bénédicte. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data (pp. 12 p.). Rome, Italie.

Loïc Liégeois, Carole Etienne, Christophe Parisse, Christophe Benzitoun, Christian Chanard. Using the TEI as a pivot format for oral and multimodal language corpora. Text Encoding Initiative Conference and Member's meeting 2015, Oct 2015, Lyon, France.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Modèles, Dynamiques, Corpus - UMR 7114 (MoDyCo) (2016). teicorpo [Outil]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, v1, https://hdl.handle.net/11403/teicorpo/v1.