Sophie Prévost (Lattice, CNRS/ENS-PSL/Université Sorbonne nouvelle)
Mathieu Dehouck (Lattice, CNRS/ENS-PSL/Université Sorbonne nouvelle)
Loïc Grobol (Modyco, Université Paris Nanterre)
Mathilde Regnault (Institut für Linguistik/Romanistik, Universität Stuttgart)

Traitement automatique du langage et analyse de la variation

Mots-clés : TAL, variations, syntaxe, morpho-syntaxe

La variation se décline selon différentes dimensions « externes » : temps, espace, registre, domaine, forme (vers/prose)… ainsi que selon les contextes linguistiques. Ainsi en est-il, par exemple, en français, de l'obligation de l'inversion du sujet après certains adverbes à valeur épistémique (*peut-être réussira-t-il* vs *il réussira peut-être*), des différentes valeurs sémantiques (haut degré ou degré excessif) que revêt l'adverbe *trop* selon l'adjectif sur lequel il porte (*c'est trop bien* vs *c'est trop cher*) ; un même contexte peut cependant autoriser deux variantes, ainsi de la variation libre, en français moderne, entre expression et omission du sujet dans les contextes de coordination verbale immédiate (*il a raté la marche et [il] est tombé*), que l'on peut mettre en regard du caractère contraint de cette même variation, en diachronie, dans d'autres contextes (*et quant li rois voit ces letres, si dist à lancelot*….(deb. 13^{e} s.) vs *et quand le roi voit ces lettres,* ***il*** *dit à Lancelot* (traduction)). Ces différents exemples illustrent par ailleurs le fait que la variation peut affecter le moyen de codage (la forme) ou le sens (ou plus généralement la fonction), ou bien les deux (voir Frajzyngier 2015).
Ces différentes dimensions, externes et internes au système linguistique, qui peuvent se combiner, rendent complexe, en soi, l'appréhension de la variation. Par exemple, une situation de variation en synchronie peut évoluer et être ainsi soumise à la variation diachronique, comme l'exemplifie l'évolution de l'ordre des mots dans l'histoire du français. La variation qui caractérisait l'agencement des constituants majeurs en ancien français (SVO, OVS, SOV, …) a « varié » dans le temps, conduisant à la réduction progressive des différentes combinaisons au profit de SVO. On peut aussi croiser d'autres axes dia-, comme les axes diatopique et diachronique dans le cas de l'abandon de la déclinaison bicasuelle en ancien français qui a progressé d'ouest en est.
La variation joue un rôle particulièrement décisif dans le changement linguistique, puisque tout changement (sauf certaines innovations lexicales) résulte d'une situation de variation (plus ou moins longue, et qui s'instancie à travers différentes étapes ; cf. les modèles proposés par Heine 2002 et par Diewald 2002); mais toute variation ne débouche pas nécessairement sur un changement : la variante nouvelle peut disparaître, ou bien la variation peut se maintenir, généralement toutefois dans des contextes différents, qu'il soient linguistiques ou extra-linguistiques (par ex. l'alternance de la négation simple en *pas* vs la négation double, largement corrélée au registre de langue).

L'accès à des données suffisamment massives, et leur quantification, afin de repérer le plus précisément possible l'émergence de nouvelles variantes (qu'il s'agisse de la forme ou de la fonction d'une construction), et le possible recul (voire la disparition) de certaines est un outil précieux (parfois indispensable)

pour l'étude des variations, quelle que soit leurs dimensions (diachronique, spatiale, …) et quel que soit le domaine considéré (syntaxique, morphologique, …).
L'émergence de vastes corpus a ainsi permis de renouveler l'étude de la variation, qu'il s'agisse de confirmer, affiner ou parfois infirmer, des résultats acquis sur des ensembles de données plus restreints. Le TAL a largement contribué à ce renouveau, par le développement d'outils d'enrichissement (analyseurs et étiqueteurs syntaxiques), et d'exploration de ces corpus.
En retour, l'analyse linguistique, quand elle ne permet pas directement d'améliorer les performances des outils, par l'analyse des erreurs d'annotation, peut expliquer une partie de ces erreurs (Brigada Villa et Giarda 2023, Manning 2011) et ainsi remettre de la profondeur là où les mesures de performance tendent à tout traiter de la même manière.

Les performances des modèles d'analyse automatique sont aujourd'hui excellentes, quand ils sont appliqués à des données similaires à celles utilisées pour leur développement, mais elles deviennent rapidement bien moins satisfaisantes quand ils sont appliqués à des données s'en éloignant (Dereza et al. 2023, Manjavacas et Fonteyn 2022). Cela pose encore un certain nombre de problèmes notamment pour l'utilisation de données automatiquement annotées comme support aux études linguistiques (Beck et Köllner 2023, Faria 2014, Säily et al. 2011).

Le présent workshop vise à explorer les apports réciproques entre traitement automatique du langage et analyse de la variation dans les domaines de la morphosyntaxe et de la syntaxe, du point de vue diachronique et diatopique, mais également du genre, du domaine et de la forme (prose/vers), sans restriction de langue. Il sera l'occasion d'un échange sur les erreurs produites par les outils d'annotation automatique dans ce contexte spécifique (quel que soit le volume de données concerné), mais également sur les manières d'améliorer ces derniers (volumes de données toujours plus grands, meilleure annotation desdites données le long des dimensions externes de la variation, meilleure modélisation de la variation...).

Sont particulièrement bienvenues les contributions portant sur les apports et enjeux de l'application du TAL à l'analyse de la variation :

- La quantification de la variation le long de ses différentes dimensions, internes et externes, et leur possible interaction ;
- L'impact des erreurs d'annotation sur l'étude des structures marginales (émergentes ou en voie de disparition) ;
- La variation syntaxique induite par des changements sémantiques.

Mais aussi celles portant sur les apports de l'analyse de la variation au TAL :

- La mitigation de la variation (standardisation de l'orthographe, de l'ordre des mots...) ;
- L'adaptation des modèles à des données hors du domaine d'entraînement (entendu ici au sens large : époque, lieu, genre, forme) ;
- L'interprétation des erreurs (en domaine et/ou hors domaine) au regard de phénomènes de variation connus, parmi lesquels les phénomènes de (dé-)grammaticalisation (Grobol et al. 2021);

- L’évolution des catégories grammaticales et leurs impacts sur les modèles d’analyse.
- La place de l’étude de la variation dans le TAL dans le contexte des très grands modèles de langue.

Ces sujets sont indicatifs et le workshop accueillera toute proposition contribuant de manière significative aux apports réciproques entre TAL et analyse de la variation dans les domaines indiqués.

Références

Beck, C., & Köllner, M. (2023). « GHisBERT – Training BERT from scratch for lexical semantic investigations across historical German language stages ». *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, 33-45.

Brigada Villa, L., & Giarda, M. (2023). « Using Modern Languages to Parse Ancient Ones: a Test on Old English ». *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP,* 30-41.

Dereza, O., Fransen, T., & Mccrae, J. (2023). « Temporal Domain Adaptation for Historical Irish ». In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, 55-66.

Diewald, G. (2002). « A model for relevant types of contexts in grammaticalization ». In I. Wischer et G. Diewald (éd.) *New Reflections on Grammaticalization*. Amsterdam : John Benjamins, 103-120.

Faria, P. (2014). « Using Dominance Chains to Detect Annotation Variants in Parsed Corpora ». In *2014 IEEE 10th International Conference on e-Science*, 2, 25-32.

Frajzyngier, Z. (2015). Functional syntax and language change. In C. Bowern et B. Evans (éd.) *The Routledge Handbook of Historical Linguistics*, Londres / New York : Routledge, 308-325.

Grobol, L., Prévost, S. et Crabbé, B. 2021. Is Old French tougher to parse? In Daniel Dakota, Kilian Evang, and Sandra Kübler (éd.) *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*. Association for Computational Linguistics, Sofia, Bulgaria, edition., 27-34. https://aclanthology.org/2021.tlt-1.0.pdf.

Heine, B. (2002). On the role of context in grammaticalization. In I. Wischer et G. Diewald (éd.) *New Reflections on Grammaticalization.* Amsterdam / Philadelphie : John Benjamins, 83-101.

Manjavacas, Enrique, Lauren Fonteyn (2022). « Adapting vs. pre-training language models for historical languages ». *Journal of Data Mining & Digital Humanities*, pages 1–19.

Manning, C.D. (2011). Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?. In: Gelbukh, A.F. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2011. Lecture Notes in Computer Science, vol 6608. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19400-9_14

Säily, T., Nevalainen, T., & Siirtola, H. (2011). « Variation in noun and pronoun frequencies in a sociohistorical corpus of English ». *Literary and Linguistic Computing*, 26(2), 167-188.