**Natural language processing and linguistic variation analysis**

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)

Linguistic variation unfolds along various "external" dimensions : time, space, register, domain, form (verse/prose)... as well as linguistic contexts. Thus, in French for example, subject inversion is compulsory after certain epistemic adverbs (*peut-être réussira-t-il* vs *il réussira peut-être*), the adverb *trop* takes various semantic values (high or excessive) depending on the adjective to which it attaches (*c'est trop bien (so good)* vs *c'est trop cher (too expensive)*); however the very same context can also accommodate two variants, hence the free variation of the overt expression and the omission of the subject in direct verbal conjunction (*il a raté la marche et [il] est tombé*) in Modern French, that we can oppose to the conditionned nature of this very variation, diachronically, in different contexts (*et quant li rois voit ces letres, si dist à lancelot*....(early 13$^{th}$ c.) vs *et quand le roi voit ces lettres,* ***il*** *dit à Lancelot* (translation)). These examples also show that variation can touch on the coding mean (form), the meaning (or more broadly the function) or both at the same time (voir Frajzyngier 2015).

These different dimensions, external and internal to the linguistic system, can mingle and blur the analysis of variation. For example, a state of synchronic variation can evolve and undergo diachronic variation, such as is the case of word order in the history of French. Variation, that was characteristic of the order of core constituents in Old French (SVO, OVS, SOV, ...) has changed through time, leading to the eventual reduction of the different combinations in favour of SVO. Several axis can also interact, the abandon of the French two case system in a progressive eastward motion is an example of both diachronic and diatopic variation.

Variation plays a particularly important role in linguistic change, since, save a few lexical innovations, every change stem from a state of variation (shorter or longer in time and through several stages; cf. the models of Heine 2002 and Diewald 2002); but each state of variation does not necessarily end up with a change : the new variant can disappear, or variation can linger but in different contexts, be they linguistic or extra-linguistic (for example, the alternation of the French double negation with the bare *pas* which correlates heavily with the speech register).

Access to sufficient amounts of data and their quantification, in order to detect the emergence of new variants as precisely as possible (be it a new form or a new function of an existing construction), and the recession or even disappearance of others, is a precious (sometimes necessary) tool for the study of variations, whatever their dimensions (diachronic, diatopic, ...) and in whatever field (syntax, morphology, ...).

The appearance of large corpora has thus renewed the study of variation, be it to confirm, refine, or infirm, results obtained on smaller datasets. NLP has contributed largely to this renewal, providing tools for the enrichment (morphological taggers and syntactic parsers) and the exploration of these corpora.

In return, when linguistic analysis cannot directly help improve the performances of these tools, via annotation error analysis for example, can help explain some of these errors (Brigada Villa et Giarda 2023, Manning 2011) and thus deepen the picture where performance metrics tend to flatten out everything under a single number.

NLP annotation tools, such as syntactic parsers and morphological taggers, reach great performances nowadays when they are applied on similar data to those seen during their development. However, they quickly drop as the target data diverges from those of the training scenario (Dereza et al. 2023, Manjavacas et Fonteyn 2022). This raises a number of issues when it comes to using automatically annotated data to perform linguistic studies (Beck et Köllner 2023, Faria 2014, Säily et al. 2011).

This workshop aims at exploring bilateral contributions between Natural Language Processing and variation analysis in the fields of morphosyntax and syntax, from diachronic and diatopic perspectives but also from genre, domain or form of writing, without any restriction on the languages of interest.

We warmly welcome submissions dealing with the issues and contributions of applying NLP to variation analysis :

- Quantification of variation along its different dimensions (both external and internal ones as well as in interaction with each other) ;
- Impact of annotation errors on the study of marginal structures (emergent or recessing) ;
- Syntactic variation when it is induced by semantic changes.

But also submissions dealing with the contributions of variation analysis to NLP :

- Variation mitigation (spelling standardisation…) ;
- Domain adaptation (domain referring here to any variation dimension) ;
- Error analysis (in and out of domain) in light of known variation phenomena, amongst which (de-)grammaticalisation (Grobol et al. 2021);
- The evolution of grammatical categories and its impact on prediction models.
- The place of variation studies in NLP in the large language model era.

These themes are only suggestions and the workshop will gladly host any submission that deals substantially with the reciprocal contributions between NLP and variation analysis in the mentioned fields.

References :


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.