Travaux neuchâtelois de linguistique, 2024, 79, 1-4

Theoretical issues in the light of the interaction between quantitative and qualitative data: recent approaches and tendencies. An introduction

Francesca DELL'ORO & Corinne ROSSARI (eds.)
Institut des sciences du langage (ISLa), University of Neuchâtel
francesca.delloro@unine.ch & corinne.rossari@unine.ch

1. Qualitative and quantitative approaches to linguistic questions

The thematic thread of this special issue centres on how corpus data and quantitative methods can support, challenge and (re)shape linguistic theory. Within this broad topic, the papers in this thematic issue delve into the interaction of qualitative and quantitative methods in linguistic analysis.

The subject matter and the approach are particularly timely. The use of quantitative methods in the humanities has not only grown exponentially in the last decade, but it has also evolved very rapidly. As highlighted, for instance, in a recent opinion paper by Kortmann (2021), theoretical considerations are not absent from linguistic works focusing mainly on quantitative aspects. Nevertheless, the impact on linguistic theory needs to be reassessed, to emphasise the apparent–or expected future–impact of quantitative methods on theoretical issues in linguistics. The main question that this issue addresses is to what extent old and new theoretical assumptions can be nuanced, rejected or improved in the light of the interaction between qualitative and quantitative approaches.

2. From corpus data to massive datasets

The trajectory of this special issue spans from employing oral corpus data to challenge and refine our previous knowledge (Mithun) to utilising corpora containing an enormous quantity of textual data, leveraged through computational tools, to find new answers to both longstanding and emerging questions (Wälchli). In-between, the other papers showcase the potentialities of new tools, such as those enabling deconvolution (Beghini), along with the renewed investigation, thanks to statistical estimations and new visualisation techniques, of subjects that remained marginal (Marongiu).

In *Modality and Reality: The value of speech in context*, Marianne Mithun provides the first fine-grained outline of the functions of the prefix *a:-* in the Iroquoian languages with a focus on Mohawk, showcasing its diachronic

development. The prefix *a:-* works as an irrealis marker, corresponding to both types of irrealis categories identified by Cristofaro (2012), such as wishes and obligations along with unfulfilled obligations and desires, for instance. Moreover, the analysis of its contextualised uses reveals that this prefix can work as a modal marker, conveying basic modal types (cf. Nuyts 2016). Contrary to the first impressions of some native speakers, oral corpus data (sometimes from those same native speakers) shows that simple sentences just featuring *a:-* verbs abound in speech. Based on corpus data showing the frequent use of *a:-* on complements of modal matrix verbs to mark Irrealis, Mithun suggests that the high frequency of this pairing brought about contamination, lending the prefix a modal flavour. The prefix is also developing new functions as a purpose clause marker and as a marker of the purpose of referents. The contribution of frequency analysis coupled with the contextualisation of the corpus data provides new insights in the study of the prefix *a:-*.

An example of the integration of qualitative analysis with multiple quantitative techniques is provided by Federica Beghini in *Stylometry and Deep Learning: A case study on Milan Kundera's* Le Livre du rire et de l'oubli. Pursuing the goal of identifying and analysing the distinctiveness of Kundera's prose in the novel *Le Livre du rire et de l'oubli* with respect to contemporary French prose, she has designed a complex pipeline based on a corpus-driven approach. Alongside the (more conventional) convolution mechanism of deep learning, Beghini incorporates the use of the deconvolution mechanism (Vanni et al. 2018), a process capable of revealing the linguistic components upon which the algorithm has based its classification choices by outlining so-called 'activation zones'. The analysis is then conducted on multiple linguistic levels, including lexical forms, lemmas, lexical and morphosyntactic patterns. Such a methodology, combining quantitative data automatically retrieved by the computer algorithms with qualitative feedback, enables the identification of senses specific to the novel as well as, on a more general level, some patterns of Kundera's stylistic sensibility. For instance, Beghini highlights Kundera's preference for the use of common nouns instead of proper names. With such a methodology the paper shows an opening on the stylistic side, and thus supports a dialogue between linguistic and literary investigations.

Paola Marongiu's contribution, *Modal markers in co-occurrence: a study on Terence's comedies*, is a corpus-based investigation into the co-occurrence of modal markers–a theme that has been explored mainly from a qualitative perspective, and not at all for the Latin language. Moreover, Marongiu broadens her focus to comprehend complex syntactic structures within the same sentence. Her paper not only provides new insights on this topic, showing for instance that some markers tend to co-occur with themselves, but also detailed information about the procedure behind the quantitative analysis. It addresses corpus-related problems linked to the crucial question of the interpretation of statistical data, like for instance data sparsity. In the qualitative section, the

paper examines the interactions between modal markers and semantic types of modality. Drawing on the results and the visualisations obtained from the quantitative data, it identifies, for instance, which markers or types of modality are most likely to co-occur with themselves.

The paper by Bernhard Wälchli, *We need world-wide corpus-based typology: A parallel corpus study of restrictives ('only')*, is programmatic in advocating for the necessity of building an approach that combines corpus linguistics and typology. This is illustrated with the typological investigation of the–understudied–coding of the restrictive meaning 'only' as it is generally expressed worldwide (vs the specific uses of English *only*). To achieve this goal, he employs a massively parallel corpus, specifically that of the New Testament translations (Mayer & Cysouw 2014). The paper presents new findings, such as the universality of restrictives as well as a distribution of uses that clearly distinguish the Pacific linguistic area from the Afro-Eurasian one. The paper also highlights that the meaning of English *only* considerably differs from the generally expressed meaning 'only', pointing to a relevant bias in linguistic research when English meanings and uses are taken as the point of reference to investigate other languages. The methodology outlined in the paper enables to avoid such bias. It is also powerful in highlighting covert coding in the relevant domain, as this can be overt in other languages of the parallel corpus. Wälchli shows that, while quantitative tools play a crucial role in identifying major trends within a substantial amount of fuzzy data, qualitative methodology is equally essential to validate the accuracy of the obtained results. His approach further enables him to argue that the differences are deeply rooted in discourse (*parole*) rather than in grammar and lexicon (*langue*), and that they are so pervasive in discourse that they carry over to written Bible translations.

3. Concluding remarks: an invitation to readership

To sum up, each paper deals with a very specific theoretical issue by exploring the interrelations between qualitative and quantitative aspects. The volume as a whole is intended to encompass recent tendencies in the field, in particular with regard to the exploitation of corpora and of the genres that these corpora represent, including oral corpora (Mithun), literary corpora (Beghini, Marongiu), and massive parallel corpora of Bible translations (Wälchli). The contributions present various ways of applying quantitative methods, encompassing more traditional approaches based on the frequency of items and their surroundings (Mithun) to the use of more sophisticated tools based on statistical methods (Beghini, Marongiu, Wälchli). Finally, the authors take into account not only Indo-European languages (Beghini, Marongiu), but also non-Indo-European ones (Mithun) as well as a cross-linguistic perspective (Wälchli). The variety of the domains within language sciences (linguistic change, typology of languages, semantics and textometric approach to genre) as well as the various linguistic and discursive phenomena investigated (ancient ('dead') languages, linguistic

categories in non-Indo-European languages, modality and stylistic patterns of literary texts) show the extent to which the question of the interaction between qualitative and quantitative analysis is at the heart of the research in any field of linguistics.

The papers stem from presentations held at the University of Neuchâtel and organised within the framework of the activities led by the research pool on 'modality and corpora'.

REFERENCES

Cristofaro, S. (2012): Descriptive notions vs grammatical categories: Unrealized states of affairs and "Irrealis"? Language Sciences, 34, 131-146.

Kortmann, B. (2021): Reflecting on the quantitative turn in linguistics. Linguistics, 59(5), 1207-1226.

Mayer, T. & Cysouw, M. (2014): Creating a massively parallel Bible corpus. Proceedings of the International Conference on Language Resources and Evaluation (LREC), Reykjavik, 3158-3163.

Nuyts, J. (2016): Analyses of the modal meanings. In J. Nuyts, & J. Auwera, van der (eds.), The Oxford handbook of modality and mood. Oxford: Oxford University Press, 31-49.

Vanni L., Ducoffe M., Mayaffre, D., Precioso, F., Longrée, D. *et al*. (2018): Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne. https://hal.science/hal-01804310 (14.08.2023).