A computational approach to detect discourse traditions and register differences: a case study on historical French

Historical sociolinguists have demonstrated the crucial role of register/genre in mediating the spread of innovations throughout language communities (Nevalainen and Raumolin-Brunberg 2017). However, the traditional conceptualization of genre has been challenged by the concept of Discourse Traditions (Kabatek 2005), henceforth DTs.

The core idea in the DTs framework is that language is not a monolithic object and one cannot dispense with the impact of textual traditionality to study the evolution of individual phenomena (Kabatek 2005). Additionally, the detection of DTs represents a challenge for quantitative corpus linguistics, as each texts can allow for global or internal classifications (Kabatek 2013: 19). Although previous research has discussed distinctive classification features for textual genre, descriptions might be biased by a researcher's particular interest or object language. It is therefore worthwhile to explore whether thorough philological analysis can be complemented by bottom-up generated classifications.

The first goal of our paper is to leverage on the popularity of computational models for the semantic representation of words and texts, so-called 'vector space models' (Boleda 2020), for the unsupervised, bottom-up identification of DTs. Document-based vector space models represent a document's content by means of a frequency profile (i.e., vector) of the terms occurring therein. Afterwards documents can be compared by calculating a similarity value based on those frequency vectors. The rationale is that the co-occurrence of certain terms in a document will be correlated with certain DTs. For this endeavor we explore a corpus of 1400+ historical French theater plays dated between 1600 and 1930 (Author 2023). This corpus is annotated in terms of sub-genres (e.g., comedy, tragedy, pastoral, etc.), which might correlate with different registers.

The second goal is to verify how the automatic classification of documents improves or complements a traditional genre classification provided by the corpus metadata. Building on previous work on Spanish (Author *in press*), we evaluate this comparison by checking the impact of both classifications on a case study of syntactic alternation in French, namely the distribution and change in inverted (1) and clefted (2) interrogatives.

(1) *Aimez-vous voyager?*
'Do you like to travel?'
(2) *Est-ce que vous aimez voyager?*
'Is it that you like to travel?'

By including these two different operationalizations of DTs in a logistic regression, we show how the bottom-up classification (a) improves the overall fit of the regression model, (b) reveals unattested differentiation within theater texts, and (c) functions as a principled approach to distinguishing 'change from above' and 'change from below'. Overall, the proposed approach evidences the relevance of computational-semantic methods for historical (socio)linguistic research.

References

Author 2023.

Authors in press.

Boleda, Gemma. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*. 6(1). 213–234.

GITHE. 2015. Codea+ 2015. *Corpus de documentos españoles anteriores a 1800*.

Kabatek, Johannes. 2005. Tradiciones discursivas y cambio lingüístico. *Lexis: Revista de lingüística y literatura* 29(2). 151–177.
Kabatek, Johannes. 2013. ¿Es posible una lingüística histórica basada en un corpus representativo? *Iberoromania* 77(1).
Nevalainen, Terttu and Helena Raumolin-Brunberg. 2017. *Historical Sociolinguistics : Language Change in Tudor and Stuart England*. London: Routledge.