Lexicometric and Informational Measures in Historical and Literary Corpora

Abstract

Florentina Armaselu,
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg

How the frequency of words may be interpreted in the context of an informational analysis of textual corpora? To what extent the frequency values and distribution could be an indicator of the "amount of information", the degree of "certainty" or of "informativeness" conveyed by a text? Are other factors, such as language and genre, influencing these informational measures?

The paper will address these questions from a comparative perspective. Two types of multilingual corpora (in Romanian, French, and English) are considered:

- a selection of minutes of plenary sittings (2005 to 2012) from the Digital Corpus of the European Parliament (DCEP–PV);
- a selection of poems by three authors (Eminescu, 2011; Hugo, 2009; Rossetti, 2005) from the Project Gutenberg.

The methodology consists of:

- a lexicometric analysis (part of speech tagging and lemmatization, corpus partitioning, frequencies and lexical tables for each part of a partition) by means of the TXM-Textometry software (Heiden et al. 2010);
- lexical tables import into Microsoft Excel and computing of three informational measures, entropy (1) (Shannon, 1948), energy (2) (Onicescu, 1966; Marcus, 1970), and informativeness (3) (adaptation of Carnap and Bar-Hillel, 1952; Dretske, 1999; Floridi, 2017), according to the following formulas:

$$H = -\sum_{i=1}^{N} p_i log_2 p_i \quad (1)$$

$$\delta = \sum_{i=1}^{N} p_i^2 \quad (2)$$

$$INF = -\sum_{i=1}^{N} \log_2 p_i \quad (3)$$

where N represents the number of unique lemmas for each part, and p_i the probability of lemma *i* calculated as the relative frequency inside a part of the corpus partition.

The presentation will include a discussion of the methodology and results, and will conclude on the interpretative aspects of the experiments from a lexicometric and informational perspective.

Bibliographic references

Carnap, R. Bar-Hillel, Y. "An Outline of a Theory of Semantic Information", *Technical Report No. 247*, October 27, 1952, Research Laboratories of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Digital Corpus of the European Parliament (DCEP). https://ec.europa.eu/jrc/en/language-technologies/dcep, downloaded from: https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html, document date: 11 March 2015.

Dretske, F.I. *Knowledge and the Flow of Information*, The David Hume Series, Philosophy and Cognitive Science Reissues, CSLI Publications, 1999.

Eminescu, M. The Project Gutenberg eBook, *Poezii*, Release Date: February 18, 2011, http://www.gutenberg.org/ebooks/35323.

Floridi, L. "Semantic Conceptions of Information", *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2017/entries/information-semantic/.

Heiden, S., Magué, J-P., Pincemin, B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement ». In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010* (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. https://halshs.archives-ouvertes.fr/halshs-00549779/fr/.

Hugo, V. The Project Gutenberg EBook of *Les contemplations*, v 2-2, Release Date: August 29, 2009, http://www.gutenberg.org/ebooks/29844.

Marcus, S. *Poetica matematică (Mathematical Poetics)*, Editura Academiei Republicii Socialiste România, Bucureşti, 1970.

Onicescu, O. "Energie informationnelle", *Comptes Rendus Acad. Sci.* Paris, 263, (1966) 22, 841-842, cited in Marcus (1970).

Rossetti, C.G. The Project Gutenberg EBook of *Goblin Market, The Prince's Progress, and Other Poems*, Release Date: October 26, 2005, http://www.gutenberg.org/ebooks/16950.

Shannon, C.E. "A Mathematical Theory of Communication", Reprinted with corrections from *The Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656, July, October, 1948. http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.