A sociolinguistic perspective on the 500 billion word corpus

Anna Zięba
Adam Mickiewicz University
azieba@amu.edu.pl

Being able to follow changes in the frequency of words or phrases over decades or even centuries seems a valuable opportunity to a socio-cultural researcher. Hopes for such a possibility grew with the introduction of Google Books Ngram Viewer (GBNV). The creators of this tool, i.e. the Cultural Observatory, Harvard University, Encyclopaedia Britannica, the American Heritage Dictionary, and Google maintain (Michel et al. 2010) that the corpus enables investigators to study cultural trends quantitatively, and that it has opened a new field of research, namely *culturomics*, a field drawing a connection between changes in word frequency and linguistic and cultural shifts. However, the tool has also received criticism. Its simplistic interface (Davies 2014) or inaccuracies in the scanned text and in the metadata (Nunberg 2010) were indicated as its weaknesses. The objective of this paper is to investigate if GBNV, a tool working on a database of 361 billion words in English, and enabling quick recovery of data on word frequency in a diachronic perspective, is indeed useful for studies concerning transformations of social and cultural phenomena.

In the paper we present examples of application of the tool paying special attention to a study performed by Greenfield (2013) who applies the program to her *Ecological Analysis.* Her research based on hypotheses on a theory of social change from the *gemeinschaft* environments into the *gesellschaft* environments encouraged us to perform a similar study with the use of GBNV. Thus, we followed the trends in changes in word frequency throughout the 19th and 20th centuries to observe if these changes correspond to one of the major socio-cultural transformations that took place in the studied period, i.e. mediatization (Hjarvard 2008, Lilleker 2008). Regrettably, the study reveals not only advantages but also many shortcomings of the tool. As these shortcomings have also been addressed by Davies (2014) we turn to his advanced interface for Google Books and reflect on whether its application helps in managing the encountered difficulties effectively. Hopefully, the conclusion will open a discussion on the usefulness of investigating word frequency in a diachronic perspective in the vast area of socio-cultural research.

References:

Davies, M. (2014). Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics 19* (3): 401-16.

Greenfield, P. M. (2013). The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science, 24*(9), 1722-1731. doi:10.1177/0956797613479387

Hjarvard, S. (2008). The Mediatization of Society. A Theory of the Media as Agents of Social and Cultural Change. *Nordicom Review, 29*(2), 105-134.

Lilleker, D. (2008). *Key Concepts in Political Communication*s. London: SAGE

Michel, J. B. et al. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science, 331*(6014), 176-182.

Nunberg, G. (2010). Counting on Google Books. *The Chronicle of Higher Education*. December 16, 2010. Available at: https://chronicle.com/article/Counting-on-Google-Books/125735/ (accessed February 2015)