

Modelling of growth of syntactic relations network in English and Russian

Bochkarev V., Shevlyakova A., Yu Lerner E.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2018 Institute of Physics Publishing. All rights reserved. Creation of the Google Books Ngram corpus opened up new opportunities for studying language evolution. This corpus consists of a large amount of digitized books written in 8 languages and contains information on frequencies of words, word combinations and syntactic relations of the last 500 years. In this paper, we present data on changes in the key characteristics of syntactic relations in English and Russian and propose a model which allows us to explain the observed changes. We used Google Books Ngram data (1800-2008) and performed modelling of network growth. Then, we compared the characteristics of the obtained model networks with the characteristics of the network of syntactic relations of the English and Russian languages. It was shown that selection of two parameters of the model allows us to obtain a very good correspondence between the changes in the clustering coefficient, the assortativity coefficient, and word distribution by the number of relations in the model network and in real networks of syntactic relations.

<http://dx.doi.org/10.1088/1742-6596/1141/1/012008>

References

- [1] Solé R V 2005 *Nature* 434 289
- [2] Ferrer R and Sole R V 2001 Santa Fe Institute working papers 01-03-016
- [3] Dorogovtsev S N and Mendes J F F 2001 *Proceedings of the Royal Society of London* 268 2603-2606
- [4] Masucci A P and Rodgers G J 2006 *Physical Review E* 74 026102
- [5] Kulig A, Drozd S, Kwapien J and Oswiecimka P 2015 *Phys Rev E* 91 032810
- [6] Job B, Blanchard P, Lücking A, Mehler A and Banisch S (ed) 2016 *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (Berlin: Springer)
- [7] Kleinberg J M, Kumar R, Raghavan P, Rajagopalan S and Tomkins A 1999 *Lecture Notes in Computer Science* 1627 1-17
- [8] Michel J-B et al 2010 *Science* 331 176-182
- [9] Newman M E J 2003 *SIAM Review* 45 167
- [10] Rosch E H 1973 *Cognitive Psychology* 4 328-350
- [11] Padó S and Lapata M 2007 *Computational Linguistics* 33 161