EXTENDED ABSTRACT

Lexical analysis of scientific publications for nano-level scientometrics

Wolfgang Glänzel[*,**], Sarah Heeffer[*], Bart Thijs[*]

[*] *KU Leuven, ECOOM and Dept. MSI, Leuven, (Belgium)*
[**] *Library of the Hungarian Academy of Sciences, Dept. Science Policy and Scientometrics, Budapest (Hungary)*

Introduction

Components of text analysis are frequently applied in scientometrics, preferably in combination with link-based techniques. The objective is usually studying the structure of medium-sized or large document sets or monitoring the evolution of research fields/topics at the global and local level. Taking up the objectives of evaluative scientometrics, we would like to link the textual analysis of individual scientific papers (or smaller sets of those) to evaluative aspects of bibliometrics. Since we now apply techniques developed for large-scale analysis to the content of individual papers, we call this kind of studies nano-level analysis. The object is quite similar: Analysing structures, detecting (dis-)similarities and monitoring evolution. We proceed from earlier approaches used in quantitative linguistics but now applied to bibliometrics.

Methods and results

The statistical analysis of the text is based on the determination of the vocabulary. We have applied this method to 18 papers by András Schubert published in three different periods with six papers each: 1983–1985, 1993–1998 and 2010–2013. We have created two sets, all words (AW) and nouns only (NN). For the latter solution we have implemented a procedure based on NLP where we used the Stanford Parser (Klein & Manning, 2003) to extract the nouns. Subsets of the vocabulary of stemmed words/nouns are treated in a different manner depending on the actual application. In this context of this nano-level exercise we have to point to the particular function of stop-words, homonyms, synonyms, and various types of names. While common words, which are frequently used, form an essential part of a vocabulary and the use of certain common words can even be considered characteristic for a person's style, these high-frequency common terms are, on the other hand, considered noise in calculating document similarity since, from the cognitive viewpoint, those do not bear relevant information. Homonyms are always problematic and need to be resolved manually. Synonyms are regarded as enrichment and their use as typical of an individual's style. Consequently, we left synonyms unresolved.

In contrast to the traditional approach, where rank frequencies are used (e.g., in quantitative linguistics), we follow the approach by Telcs et al. (1985), who analysed four subsamples of an essay by Thomas Babington on Francis Bacon and the word frequency in Alexander Pushkin's story *The Captain's Daughter* using the newly developed statistical test on the basis of a Waring model. The word-distribution from the Bacon essay was based on nouns only, while the Pushkin sample referred to the complete text. They concluded that this might be one source of the observed considerable deviation in the parameters of both samples and they had to reject the model for the Pushkin text. That is the reason why we will also apply the same approach. We have to mention that vocabularies are limited and increasing the length of the text will not result in proportional growth of the vocabulary (cf. Kornai, 2002). The average use of words will grow to infinity with the length of the text. This also implies that the α parameter of the Waring distribution is expected to be close to the value 1, i.e., the word use has no specific (finite) expectation.

Before we proceed, still some words about the underlying model need to be said. Since the frequency of unused words is not known, we can make advantage of an important property of the Waring model, namely that a truncated distribution remains of Waring type with just a slight change in one of its parameters:

$$P(X=k|k>0)=\frac{P(X=k)}{1-P(X=0)}=\frac{\alpha}{N+\alpha+1}\cdot\frac{(N+1)\ldots(N+k-1)}{(N+\alpha+2)\ldots(N+\alpha+k)},\text{ for all } k\geq 1,$$

In order to estimate the two parameters N and α, we applied a hybrid MLE:

$$\alpha = (1-f_0)\cdot\left[\sum_{j=1}^{\infty}\frac{1}{\frac{\alpha}{f_0}+j}\sum_{i=j}^{\infty} f_i\right]^{-1} \quad \text{and} \quad N = \alpha\cdot\left(\frac{1}{f_0}-1\right),$$

where f_i denotes the observed relative word frequencies.

For the complete document set we obtain $\alpha = 0.820$ and $N = 0.515$ for all words (AW, $n = 2346$) and $\alpha = 0.915$ and $N = 0.214$ for the nouns (NN, $n = 1261$). The two distributions have similar parameters, both with $\alpha < 1$, that is, we could not observe the same effect as reported by Telcs et al. (1985) for the literary texts. The fit of the estimated distribution is in both cases more than satisfactory (Figure 1).

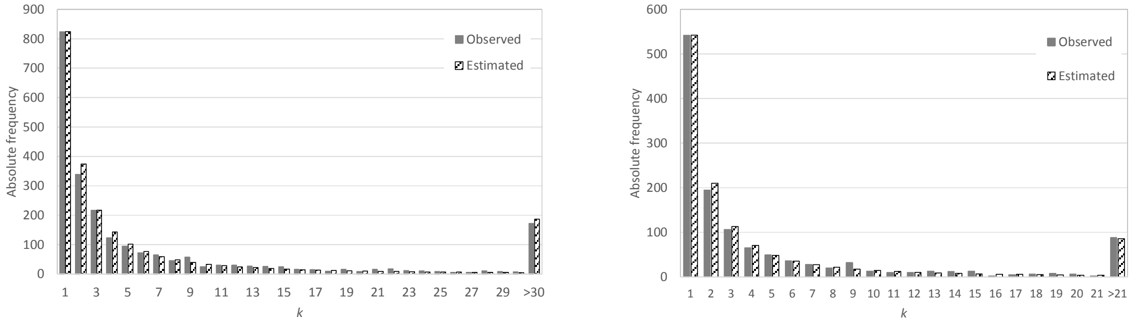

Figure 1: Observed and estimated word frequencies in 18 papers by Schubert using full text (left) and nouns only (right)

In a second step we looked at the evolution of Schubert's vocabulary. In order to do so, we have split up the complete corpus into two 'samples' of six papers each, one from the 1980's and one from the period 2010–2013. For the two blocks we have similar α parameters the values of which lie now slightly above 1. In order to determine the top frequently used words in the NN samples, we have applied the rule according to the method of *Characteristic Scores and Scales* (see Glänzel et al., 2014) providing the percentiles 70%, 21%, 6.5% and 2.5% (from low to high). The highest class comprises 17 nouns in the first (16 in the second) sample with a certain overlap. However, while the first period is more the time of creating and applying models (e.g., distribution models) and building indicators, the second one rather refers to network analysis, although scientometric indicators, notably Hirsch-type indices are still used in papers of the second sample. The further analysis of overlap (intersection) and difference (complement) of vocabularies will provide further insight in the dynamics of an author's academic writing.

Discussion and future perspective

The above approach just provides an introduction of the potential of mathematical models in analysing scientific text at the micro level. The options are manifold. Here we point to basic characteristics of an author's vocabulary and its change in time, the comparison of style and vocabulary of different authors and the detection of new research topics in an author's work. Another question to be answered is what the influence of co-authors on an author's style is, or, in turn, what the individual co-authors' contributions to a paper are.

More at the nano level, the vocabulary of different parts (sections) of the same text could be compared, provided the underlying text is long enough. But this task is subject of future research.

References

Klein, D. & Manning, Ch. D. (2003), *Accurate Unlexicalized Parsing*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 423–430.

Glänzel, W., Thijs, B. & Debackere, K. (2014), The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, 101 (2), 939–952.

Kornai, A (2002), How many words are there? *Glottometrics,* 4, 61–86.

Telcs, A., Glänzel, W. & Schubert, A. (1985), Characterization and statistical test using truncated expectations for a class of skew distributions. *Mathematical Social Sciences,* 10, 169–178.