LEXICAL ANALYSIS OF SCIENTIFIC PUBLICATIONS FOR NANO-LEVEL SCIENTOMETRICS

W. GLÄNZEL, S. HEEFFER, B. THIJS

KU Leuven, ECOOM & Dept MSI, Leuven (Belgium)

ECOOM

Structure of presentation

1. introduction
2. the dataset
3. methodological aspects
4. results
5. discussion

Introduction

Text analysis in scientometrics

Components of text analysis are applied in scientometrics, mostly in combination with citation-based techniques.

The objective is structural analysis of medium-sized or large document sets or to monitor the evolution of research fields and to detect emerging topics at both the global and local level.

☞ Lexical analysis proved most efficient if based on full text (e.g., 🗎 Glenisson et al., *Scientometrics*, 2005).

In the present study we apply quantitative text analysis at the micro/nano level to characterise individual vocabularies and to monitor possible topic changes.

Introduction

Although the objective is quite similar (analysing structures, detecting similarities, monitoring evolution and changes), the methodology completely differs from the cluster and community-detection based techniques used at the meso/macro level.

We proceeded from earlier approaches originated in quantitative linguistics but partially applied in bibliometric contexts.

- Telcs et al., *Mathematical Social Sciences*, 1985
- Mullins et al., *The structural analysis of a scientific paper*, 1988

The dataset

DOI numbers of the 18 selected papers by András Schubert

1983–1985	1993–1998	2010–2013
10.1007/BF02017143	10.1016/S1385-8947(98)00074-6	10.1007/s11192-014-1281-z
10.1007/BF02017147	10.1007/BF02457417	10.1007/s11192-012-0889-0
10.1007/BF02016759	10.1007/BF02129597	10.1016/j.joi.2012.04.004
10.1007/BF02025830	10.1007/BF02018114	10.1007/s11192-011-0559-7
10.1007/BF02097178	10.1177/0306312793023003007	10.1007/s11192-009-0072-4
10.1007/BF02095627	10.1007/BF02016790	10.1007/s11192-010-0167-y

The dataset

We have applied two approaches to the text corpora: all words (AW) and nouns only (NN).

- For the extraction of noun-phrases we have implemented a procedure based on NLP, where we used the Stanford Parser to extract the nouns, particularly, nouns (singular, mass and plural) and proper nouns (singular and plural).
- We have combined all documents into a large file and we have used the documents of the first and last periods (block 1 and 3) for comparison.
- All words in both sets (AW and NN) were stemmed, manually cleaned and then their frequencies were counted.

Methodological aspects

Two main methods are applied to study the word use: *rank frequencies* and *frequency distributions*. The choice of the method depends on the purpose.

- In quantitative and computational linguistics the first method is popular, e.g., if the index size in full-text databases is to be predicted (🗎 Gelbukh & Sidorov, *LNCS*, 2001).
  - ☞ Zipf-Mandelbrod-type laws are here the classical models but more complex approaches are used as well (🗎 Piantadosi, *Psychonomic Bulletin & Review*, 2014).
- The other approach deals with the frequency of word occurrence. Both approaches are not contradicting but just representing two different perspectives.

We use the second approach as has already been done by Telcs et al. (1985).

The Waring model

We say that X has a Waring distribution with real parameters $\alpha > 0$ and $N > 0$, if

$$P(X = k) = \frac{\alpha}{N + \alpha} \prod_{i=1}^{k} \frac{N + i - 1}{N + \alpha + i}, \quad k \in \mathbb{Z}_0^+ .$$

Since the number of unused words is unknown the distribution is truncated at $k = 0$. For a Waring distribution we thus obtain

$$P(X = k | k > 0) = \frac{P(X = k)}{1 - P(X = 0)} = \frac{\alpha}{N + \alpha + 1} \prod_{i=2}^{k} \frac{N + i - 1}{N + \alpha + i}, \quad k \in \mathbb{N} .$$

If we shift this distribution back to $k = 0$, we again obtain a Waring distribution but with parameters $(N + 1)$ and α.

Methodological aspects

Some relevant properties

The Waring distribution asymptotically obeys a power law since $\sum_{i=k} P(X = i) \approx c \cdot (N + \alpha)^{-k}$, if k is large enough.
While α is as characteristic parameter responsible for the heaviness of tail and the existence of finite positive moments, N reflects skewness.

Vocabularies are limited and increasing the length of the text will not result in proportional growth of the vocabulary (cf. Kornai, *Glottometrics*, 2002).

- Words will be increasingly repeated and the average use of a word will grow to infinity with the length of the text.
- The shape of the distribution depends on the length of the corpora.
- α is expected to be close to the value 1, that is, the word use has no specific (finite) expectation.

☛ Comparison of different texts by the same author should preferably be based on texts with comparable length.

Methodological aspects

A robust semi-ML parameter estimation

We say that X has a Waring distribution with real parameters $\alpha > 0$ and $N > 0$, if

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

$$N = \alpha\cdot \left(\frac{1}{f_0} - 1 \right) ,$$

where f_i denotes the observed relative word frequencies. We have applied the following extremely fast converging algorithm with arbitrary initial value α_1, which was stopped as soon as $|\alpha_{k+1} - \alpha_k| < 10^{-7}$,

$$\alpha_{k+1} = (1 - f_0)\cdot S(\alpha_k)^{-1} \text{ with } k = 1, 2, 3, \ldots .$$

Results

In scientific text technical terms are among the most frequent terms and fillers, embolalia can be assumed to be less common. The restriction to nouns seems to provide more significant results and of the similarity of estimated parameters did encourage us to restrict further analysis to nouns.

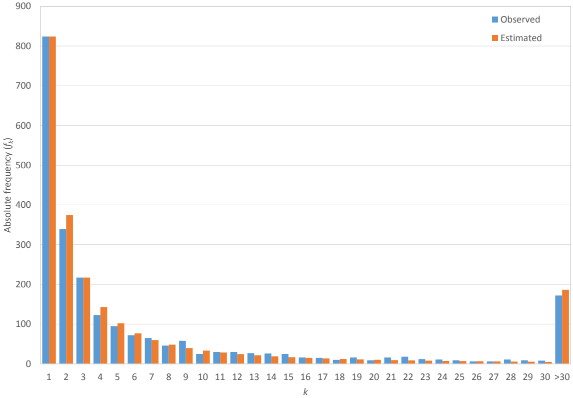
- As a “by-product” we also obtain that nearly 54% of all words were nouns.
- In total, the 18 documents contained 28,231 words, whereof 8,908 items were nouns and each noun has been used about 7 times on an average.

The two distributions have similar parameters, both with $\alpha < 1$, that is, we could not observe the effect reported by Telcs et al. (1985) for the literary text corpora.

☞ The fit of the estimated distribution is in both cases much more than satisfactory.

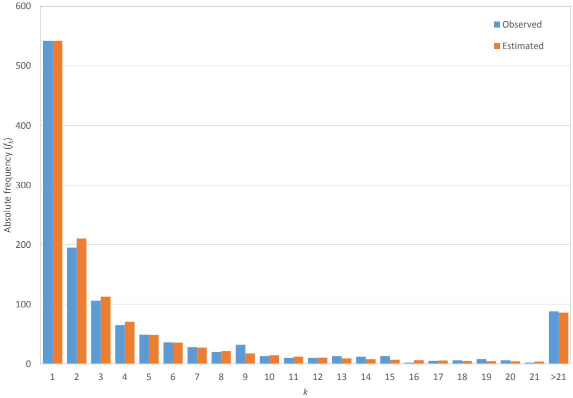
Results

Observed and estimated word frequencies in 18 papers by Schubert using all words ($\alpha = 0.820$; $N = 0.515$; $n = 2346$)



Results

Observed and estimated word frequencies in 18 papers by Schubert using nouns only ($\alpha = 0.915$; $N = 0.214$; $n = 1261$)



Results

For further analysis of style characteristics and possible changes of the vocabulary we have split up the complete document set into blocks.

We analysed and compared the first (1983–1985) and the last (2010-2013) block.

- The vocabulary is of comparable size and also their overall noun frequencies are similar.
- In both corpora a noun is used 4.6 times on an average.

☞ This reflects the stability of the author’s style but also illustrates the dependence of the mean of the document length, which amounted to 7.0 for the complete set.

Results

Waring parameters for the word frequency distributions in block 1 and 3

Parameters	Block 1	Block 3	Vocabulary (n)	Noun count
α	1.126	1.365	651	3,098
N	0.317	0.767	608	2,778

Results

The field of scientometrics is evolving fast and has changed much during three decades. On the other hand, Schubert's research topics might also have changed in time.

- In order to detect such changes in his vocabulary, we analysed the most frequently nouns.
    - We applied the percentiles derived from the method of *Characteristic Scores and Scales* originally developed for defining citation-based performance classes. (🗎 Glänzel et al., *Scientometrics*, 2014)

Waring parameters for the word frequency distributions in block 1 and 3

Rank	Block 1		Block 3	
	Noun	Frequency	Noun	Frequency
1	distribut	100	journal	139
2	valu	69	similar	77
3	paper	68	index	61
4	countri	64	citat	57
5	number	58	categori	53
6	journal	56	indic	53
7	review	52	distribut	52
8	book	51	field	50
9	citat	48	valu	47
10	public	48	cluster	46
11	author	47	refer	39
12	rate	45	measur	36
13	impact	44	case	31
14	sampl	37	paper	30
15	meet	35	partnership	28
16	price	34	scienc	28
17	factor	33		

Results

The content of CSS performance classes defined on heavy-tailed citation distributions follow 70–21–6.5–2.5 per-cent rule (from low to high end).

☛ The question arises of whether the same distribution is followed by word frequencies. Notwithstanding the above, the choice of 2.5% for the class of most frequently used nouns seems to be reasonable and close to other common approaches.

Despite a certain overlap between the two blocks there is also some change.

- The first period is characterised by model creation and application (e.g., distribution models – ‘Price’ is here used eponymically) and indicator building.
- The last block rather refers to network analysis, although scientometric indicators, notably Hirsch-type indices are still used.

☞ A further analysis of overlap (*intersection*) and difference (*complement*) of vocabularies could provide further insight in the dynamics of an author’s academic writing.

Discussion

We have sketched the potential of mathematical models in analysing scientific text at the micro level. The options are manifold and leave room for future research.

- *Micro level*
  - The analysis of 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.
  - Detection of divergence of an author's word use with respect to that by colleagues or other members of a team.
  - The influence of co-authors on an author's style.
  - Detection of new topics in individual research profiles. (Future research task)
  - Models for sentence length of scientific text (e.g., negative binomial or compound Poisson distribution used for classical prose. – cf. Sichel, *JR Statist Soc A*, 1974; Kelih & Grzybek, *Glottometrics*, 2004) (Future research task)

Discussion

- *Nano level*
    - Comparison of the vocabulary of different parts (sections) of the same document, provided the underlying text is long enough. (Future research task)
    - Identification of individual co-authors contribution to writing parts of the documents. (Future research task)
- *Further general remarks*
    - Finally we have to draw the attention to deviations in model parameters we found for scientific and literary text. Among some causes, which might also apply to sentence length, we mention:
        - Limited space in periodicals, proceedings and edited books (print media), but possibly less relevant for e-publications
        - Further possible deviation of articles and book chapters from monographic literature
        - Frequent use of subject-specific terms, phrases and acronyms
        - Possible correlation between "hardness" of science and conciseness of text (sentence length)

Thank you very much for your attention.

Köszönöm szépen a figyelmüket!