Studies
in
Quantitative Linguistics
8

Ioan-Iovitz Popescu
Ján Mačutek
Emmerich Kelih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

Vectors and Codes of Text

RAM - Verlag

Vectors and codes of text

by

Ioan-Iovitz Popescu

in cooperation with

Ján Mačutek
Emmerich Kelih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

2010
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, II + 161 pp.

Gedruckt mit Unterstützung der Karl-Franzens-Universität Graz

ISBN: 978-3-942303-02-6

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

The present book is a continuation of our endeavour to introduce in textology new quantitative methods and evaluate some older ones (cf. Popescu et. al. 2009, Popescu, Mačutek, Altmann 2009; Tuzzi, Popescu, Altmann 2010). We illustrated the measurements and performed evaluations of texts from many languages. Needless to say, all results ensuing from the use of vectors, codes and chains must be tested on further languages and texts. Since this is ongoing empirical research, some modifications and adaptations of the methods presented may be necessary.

Nevertheless, the more we advanced the clearer we saw the abysmal playground hidden in texts. With some progress in using and elaborating quantitative methods in text analysis some new problems and formerly unrecognized phenomena appeared, thus we were confronted not only with methodological challenges, but also with new questions and problems in linguistic text theory in general.

We restricted ourselves to formal features (frequencies, codes and chains) accessible to the cooperating linguists and avoided sociolinguistic, psycholinguistic and other problems. We nevertheless hope that the methods presented could be made useful for other investigations, too.

The book consists of nine chapters. In Chapter 1 we introduce briefly the extensive domain of possible problems concerning comparisons and research strategies. In Chapters 2 to 6 we examine different vectors of texts, show their behaviour, compare texts and languages and take a step towards capturing the text dynamics looking at it from different points of view.

In Chapters 7 and 8 we goedelize the text in one special way, show breaks in the syntactic continuity in the text and ascribe to it its binary code which can be compared and tested.

The last chapter, Chapter 9, is devoted to chaining phenomena restricted here to Belza-Skorochoďko chains, revealing many new vistas touching behaviour, perseveration, psycholinguistic and other aspects. As a matter of fact, each topic could be developed infinitely but we strived for presenting simple methods, developed tests and showed a way of ternary plotting.

We hope that other scholars will adopt the methods for different purposes and for analyzing other languages, in order to get stronger corroboration of the procedures presented.

In this place we want to express our gratitude to Claudiu Vasilescu, who patiently wrote for us dilettantes all the Excel programmes and discretely concealed his amazement about our naivety. We had suffered much more without his kind help. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach).

I.-I. Popescu

Contents

1. Introduction

> "Not everything that can be counted counts, and not everything that counts can be counted."
> *A. Einstein*

The comparison of vocabularies of two texts in the same language can be performed in principle in two different ways:

(A) with regard to the identity of individual words,

(B) without regard to the identity of individual words.

In case (A) there are again two possibilities:

(A-1) The vocabularies of the two texts are considered sets and these sets are compared for similarity, with the frequencies of individual words ignored.

(A-2) The frequencies of individual words are taken into account as a kind of weight, and the weights of identical words are compared. This is the most common practice in quantitative lexicology (cf. Brunet 1988; Muller 1992; Labbé C., Labbé D. 2001, 2003, 2006; Labbé D. 2007; Merriam 2002; Rudman 1998; Tuldava 1971, 1998; Viprey, Ledoux 2006). The weights (= frequencies) are usually relativized because of different text lengths.

Needless to say, an analysis of type (A) can be practised only in texts of the same language. However, general textology is interested also in possible tendencies existing in all languages and must take into account some properties of the text for whose computation the identity of words is irrelevant.[1] Thus one must go beyond the level of lexicology and consider some abstract forms formed by the words of the text. There are several possibilities here, but two of them are quite conspicuous, namely the comparison of

(B-1) the rank-frequency sequence of words which can be considered either as a distribution or as a simple sequence, or of

(B-2) the frequency spectrum of words, where the random variable X is the occurrence number (= frequency) (x = 1,2,3,…) and $f(x)$ is the number of words having frequency x. This version can be attained by a simple transformation of (B-1).

In case (B-1) one takes into account the identity of ranks, in case (B-2) one takes into account the identity of occurrences. In all B-cases one can use for comparison some non-parametric tests, e.g. the chi-square, or one can reduce the data to some moments of the distributions and perform the comparison using Ord's scheme (Ord 1972) for which only the mean, the variance and the third central moment are necessary. Ord´s scheme is represented by the vector <I, S> = $<m_2/m_1´, m_3/m_2>$, where m_i are the individual moments. In this way one can transcend the material base of the text but still take into account some rather

[1] In case of comparison of vocabularies of a text and its translation in another language there is seldom a one-to-one correspondence of words.

abstract properties of (B-1) and (B-2). Here we shall present another vector which can easily be computed for any text and any variant of (B).

The above-said shows that there is not only *direct* text comparison based on word identity; as a matter of fact, texts have an infinite number of properties all of which can in principle be quantified and their numerical forms compared. Even psychological/psycholinguistic or aesthetic properties have already been quantified (cf. e.g. Paivio, Yuille, Madigan 1968; Paivio 1971). Hence, there are different aspects of research for which text comparisons are necessary. Let us mention only some of them:

(a) Text unfolding, i.e. observing the dynamics of a property in the course of text;

(b) properties of genres, i.e. observing the common features of different texts even in different languages;

(c) style identity, used also in forensic linguistics but especially in music, concerning similar technical means used in different texts of the same author;

(d) historical development of texts in a language, i.e. the change of a property in the history of written texts, beginning from simple forms up to modern novels;

(e) ontogenetic development of texts in children;

(f) the speech of individual persons in a stage play;

(g) general textology surpassing the boundaries of individual persons, languages and epochs and using rather abstract properties.

All these approaches can be combined and must lead to the establishment of a special aspect of text theory.

Our procedure is rather explorative; we bring some results but are not always able to unveil the secrets of the background mechanisms whose existence must be assumed. However, the way of their operation is far from being known or even hypothesized. We try to go new ways offering new methods important for the description of individual texts or groups of texts rather than results. The tiresome work with text processing for different evaluations must be left to interested researchers specialized in individual domains.

Methodologically, our way in the depth of the text can be described in four steps. First, we consider it a whole and process it as a whole. Only a complete text contains the complete information. In the second step we reduce it to distributions of various entities and try to model them. Here we search for the genesis of attractors without the existence of which no communication is possible. Self-regulation is an intrinsic principle of language stability and this is warranted by the existence of attractors. In the third step, we reduce the properties of a certain attractor to a vector consisting of three components, study its form and compare texts. At last, in the fourth step, we reduce a property to a single number, the binary code of text, and show its applicability to different properties. Graphically, the procedure can be presented as follows:

Text

↓

Distribution = a ranked set of numbers

↓

Vector = three numbers

↓

Binary code = a single number

The binary code, though it is only a number, can be partitioned in a sum of numbers which reveal the given special structure of the text. Its study is not very advanced but here at least the first steps are made.