Statistical Parameterization of Text Corpora

Gregory Y. Martynenko and Tatiana Y. Sherstinova

Department of Phonetics, St.Petersburg State University,
Universitetskaya nab. 11, Saint-Petersburg, Russia
e-mail:gymart@ts4306.spb.edu

Abstract. Statistical parameters, usually used for diagnostic procedures, in many cases cannot be considered to be consistent ones from statistical point of view, being strongly dependent on sample size. It leads to considerable devaluation of diagnostic results. The paper concerns the problem of consistency verification of parameters in the initial (pre-classification) stage of research. Complete list of parameters, which may be useful for description of text lexicostatistical structure, was determined. Each of these parameters was exposed to the justifiability test. In the result a number of consistent parameters have been selected, which represent a description tool for the system characteristics of any text and corpora. Having rapid speed of convergence to the limit values, they may effectively perform classification procedures on text data of the arbitrary size. The proposed model of approximation makes it possible as well to forecast the values of all parameters for any sample size.

1 Consistency of text statistic parameters

In the multitude of statistical-classificational linguistic tasks (e.g., taxonomic, typological, attributional, etc.) the central position is occupied by a diagnostic task, which may be solved in two ways. The first approach implies that a "collection" of parameters supposed to be useful for classification procedures is formed, and thereafter from the whole multitude of parameters the actually essential ones (from the point of view of concrete task and corpora) are selected by means of multidimensional analysis procedures [1]. Such method has proved its high efficiency, especially in case of successful selection of the teaching population.

The second approach is based on traditional philological conceptions about the language structure and genre/stylistic differentiation of texts. In this case the main systematic/structural mechanisms of language functioning and its genre/stylistic types (correlation between parataxis and hypotaxis, preposition and postposition, compactness and distantness, static and dynamics) are taken into account, and further the concrete symptomatic attributes associated with these system characteristics are being determined [2].

However, not depending on general method the parameters that are usually used for diagnostic procedures in many cases cannot be considered as consistent ones from statistical point of view (i.e. they strongly depend on sample size). It leads to considerable devaluation of the diagnostic results because they are inevitably obtained on texts and corpora of different sizes.

2 Research material

The problem of convergence of text lingua/stylistic parameters to some limit values (that is their consistency) is being investigated on material of the Computer Anthology of Russian Short Stories of the XX-th century, which is created by the Department of Mathematical and Computational Linguistics of St.Petersburg University and on material of other corpora created in Russia and European countries.

Anthology of Russian Short Stories represents a text database consisting of about 2500 text samples (approximately 10 millions of word units). The corpus is divided into a number of "chronological cuts" (chronological periods), which merit special microanthologies to be created for them. In chronological periods we tried to present the maximum quantity of writers, active in the given literary epoch. For the outstanding writers (such as Anton Chekhov, Igor Bunin, Andrey Kuprin, Maxim Gorky, Fiodor Sologub, Andrey Platonov, Michael Zoshchenko and some others) the author's anthologies are created. The system of frequency dictionaries is being formed for the whole corpora, for each chronological period, and for individual prominent writers. The principles of dictionary structurization depend in each case on a number of parameters (in particular, see [1]).

3 Methods of data order in frequency dictionaries

Any frequency dictionary represents a lexicographic composition, any article of which contains the name of lexical unit and accompanying it statistical data of different kind (e.g., absolute frequency of the lexeme in concern, its frequency rank, quantity of lexical units with the identical frequency, etc.). Analyzing information accumulated in frequency dictionaries, it is possible to build statistical distributions, whose concrete types are determined by the fact, what particular information is functioning as dependent and independent parameters. The main distributions are the following: polynomial, rank, and spectral ones. In polynomial distribution the role of independent parameter plays the varying name of lexical units, whereas its frequency actions as dependent parameter. In rank distribution independent parameter is represented by frequency rank of the lexical unit, while dependent one - by its frequency (in this distribution parameter "name" simply disappears). In spectral distribution the frequency of lexical unit serves as an independent parameter, and the quantity of lexical units with the identical frequency functions as a dependent one.

4 Types of scales and corresponding parameters

Mathematical statistics distinguishes different scales (quantitative, ordinal, nominal) and has developed a number of data processing methods, which are applicable just for the appropriate scale. The most promoted system of techniques has been elaborated for the quantitative parameters. Basing to the considerable extent on the theory of moments, it implies an advanced system of mean and variance values, characteristics of distribution forms, etc., and effectively uses as well a system of ordinal statistics (mode, median, quartile, etc.).

We discussed in previous paragraph that the central distributions used in processing of frequency dictionaries are rank, spectral, and polynomial ones. Though rank and spectral distributions have the outward appearance of quantitative scales, they are characterized by an extremely great variance of parameters in both rank and frequency scales. This fact induces some researchers to doubt the possibility to apply here the theory of moments (because of their prone to infinity), and therefore to suggest instead some other characteristics, not depending on the sample size. What concerns polynomial distribution, it cannot apply here the theory of moments theoretically as its variance is of quality nature.

5 Statistic parameters and consistency test

Analysis of the recent scientific works and the results of our own investigations allowed us to determine a rather complete list of parameters, which may be used for description of text lexicostatistical structure. Each of these parameters was exposed to the consistency test. Table 1 presents the list of parameters subdivided into three groups according to the type of scale.

Nominal scale	Quantitative (frequency) scale	Ordinal (rank) scale
Mode (Mo)	Mean frequency ($\overline{f}$)	Rank mean frequency ($\overline{r}$) (differentiation coefficient)
Dictionary size (N)	Geametric mean frequency ($\overline{f}_g$)	Rank variance coefficient (V_r)
Maximal frequency ($f_{\max}$)	Frequency variance coefficient (V_f)	Rank median (Me) (equilibrium measure)
Entropy (E)	Frequency median (Me_f)	Rank golden section (G_r)
Maximal entropy ($E_{\max}$)	Golden section (G_f)	Rank mean deviaton (d_r)
Order coefficient $\frac{E}{E_{\max}}$	Diversity coefficient (S)	Variation coefficient on d_r (V_r)
Analytics measurement (A)		Coefficient of concentration ($\frac{\overline{f}_r}{N}$)
		Logarithmic concentration coefficient ($k = \frac{\log \overline{r}}{\log N}$)

Methodology for consistency test has been elaborated using the method of least squares with a number of principle modifications, caused by the complicated character of parameters dependence on the sample size [2]. In our hypothesis test the null hypothesis was stated as "all parameters converge to their limit values" (alternative hypothesis - "all parameters increase or decrease without limits"). For approximation of experimental data the Weibull function has been used for increasing dependencies, and the inverse of the Weibull function - for decreasing ones.

6 Conclusion

Our main results are the following: 1) Theoretically all the parameters have either upper or lower limits. That means that in principle they are statistically consistent. However, for the most of parameters actual

consistency is achieved only in the very big sample sizes, which are hardly attainable in ordinary linguistic tasks.
2) The most consistent parameters turned out to be (in decreasing order): order coefficient ($\frac{E}{E_{\max}}$), equilibrium measure ((Me_r)), logarithmic concentration coefficient ($k = \frac{\log \overline{r}}{\log N}$), diversity coefficient (a number of words whose absolute frequency equals to "1" - S), differentiation coefficient ($\overline{r}$). These parameters along with some other ones represent a description tool for the system characteristics of any text and corpora. Moreover, their rapid speed of convergence to the limit values allows to effectively perform classification procedures on text data of the arbitrary size.
3) The proposed model of approximation makes it possible to forecast the values of all parameters for any sample size.

References

1. Marusenko M.A: Attribution of anonymous and pseudonymous literary works by means of pattern recognition theory. Leningrad. Leningrad State University. 1990.
2. Martynenko G.Y.: Fundamentals of Stylometrics. Leningrad. Leningrad State University. 1988.
3. Frequency Dictionary of Chekhov's Short Stories: Ed. Martynenko G.Y. Collected by A.O.Grebennikov. St.Petersburg, 1998.
4. Khajtun S.D.: Scientific measurement. Present Conditions and Perspectives. Moscow. "Nauka", 1983.
5. Shrejder Y.A., Sharov A.A.: Systems and Models. Moscow. "Radio i sviaz", 1982.