DEFINING THERMODYNAMIC PARAMETERS FOR TEXTS FROM WORD RANK-FREQUENCY DISTRIBUTIONS

Andrij Rovenchak[1], Solomija Buk[2]
[1] *Department for Theoretical Physics, Ivan Franko National University of Lviv,*
12 Drahomanov St., Lviv, UA–79005, Ukraine
[2] *Department for General Linguistics, Ivan Franko National University of Lviv,*
1 Universytetska St., Lviv, UA–79000, Ukraine

(Received October 29, 2010; received in final form January 24, 2011)

We report the results regarding the calculation of a new parameter set obtained from the rank–frequency distribution of texts. The parameters are defined using the analogy between the rank–frequency distribution and the quantum Bose-distribution. The calculations are made for the translations of *The Little Prince*, a novella by Antoine de Saint-Exupéry, into forty languages from different language families. The obtained data suggest the connection between the type of the language grammar and the defined parameters.

Key words: word frequency, text parameters, Bose-distribution.

PACS number(s): 01.90.+g, 02.60.Ed, 05.30.Ch

> Le langage est source de malentendus
> (Antoine de Saint-Exupéry,
> *Le Petit Prince*, 1943)

I. INTRODUCTION

The underlying principles of human behavior were used by Pareto [1], Estoup [2], Zipf [3, 4], and others to explain some empirical laws in social sciences and humanities, cf. in particular [5] for some historical details. The principle of least effort as described by Zipf [4] has clear allusions to the least-action principle in physics. The application of physical approaches, or to be more precise, the approaches of statistical physics, seems thus natural to study the regularities in many domains, including linguistics and social sciences [6–8].

The empirical laws similar to that of Zipf hold in particular for the distribution of nucleotides in genomes [9–11], the distribution of metropolitan areas [12], and even patterns in civil violence [13].

The presented analysis is based on the approach introduced by the authors in [14] and relies on the analogy between the rank–frequency distribution in texts and the Bose-distribution (in the grand canonical formulation). The word frequncy distributions are characterized by a set of parameters, one of which is a correspondent of the temperature.

Several approaches to the notion of "temperature of texts" are known in the literature. Mandelbrot [15] proposed the term "informational temperature of texts" to denote a parameter in what is now known as the Zipf–Mandelbrot law, see also [16]. Kosmidis *et al.* suggested that the "temperature" can be used to measure the communicative ability [17]. Recently, Miyazima and Yamamoto [18] defined the "temperature of texts" by modelling the frequencies of the most frequent vocabulary with the classical Boltzmann distribution. We proposed an approach [14], which focuses mainly on the behavior of low-frequency words. In this paper, this approach is applied to study the translations of *The Little Prince*, a novella by Antoine de Saint-Exupéry.

The paper is organized as follows. In Section II we recall what the rank–frequency distribution is. In Section III the physical analogy is explained. Section IV uncovers the choice of the material for the analysis. The results are given in Section V together with a brief discussion.

II. RANK–FREQUENCY DISTRIBUTION

In this work, we analyze the frequency of words in texts. Various possible definitions of a word are known, cf. [19], and we stick to the so called "orthographic word". The latter is defined as an alphanumeric sequence between two spaces or punctuation marks. Thus, different word forms, like 'hand' and 'hands', 'write' and 'wrote', 'go', 'went', 'gone', etc. are treated as different words. On the one hand, it is done for simplicity, while, on the other hand, any stemming (or lemmatization) procedure requires some agreements to be superimposed, and those can be quite different in different languages.

In a word frequency list, the most frequent word is given rank 1, the second most frequent is given rank 2 and so on. The words with equal frequencies are ranked arbitrarily within a consecutive range of ranks (typically, according to the alphabetic order). Note that here we can see for the first time some analogy with the Bose distribution, which will be clarified a bit later.

Some examples of rank–frequency distributions are shown on Fig. 1.

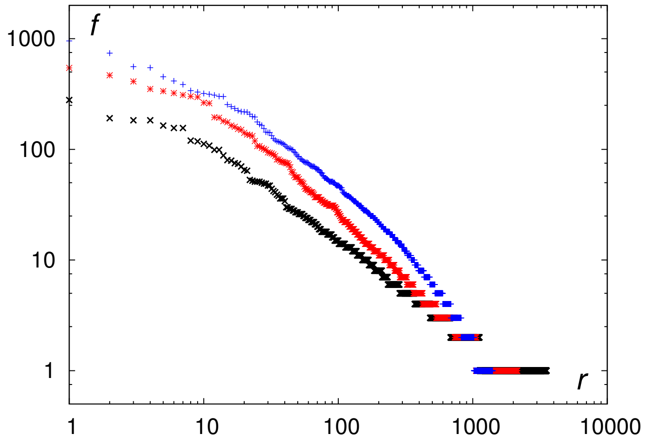

Fig. 1. (Color online) Examples of rank–frequency distributions of *The Little Prince* translations into Chinese, English, and Ukrainian. The data were compiled by the authors. Horizontal axis is rank (r), vertical axis is absolute frequency (f). **Legend:** + (blue) — Chinese data; $\ast$ (red) — English data; $\times$ (black) — Ukrainian data.

High ranks (and, respectively, low frequencies) are characteristic due to the long horizontal plateaus. These are caused by the fact that a large number of words have the same frequencies. The longest plateau, naturally, corresponds to the absolute frequency equal to unity. The words occurring only once in a given sample are known as *hapax legomena*, the term originating from the Bible studies. It is a Plural of the Classical Greek term *hapax legomenon* (ἅπαξ λεγόμενον) translated as '[something] said [only] once'. An interesting example from the Bible is in particular עֲצֵי־גֹפֶר 'gopher wood' (used to build Noah's Ark) [20].

About 40 to 60 per cent of the occurring words are hapaxes in large text samples [21, p. 72]. The proportion of *hapax legomena* slightly decreases as the text length N increases, presumably, according to the power law $N_{\rm hapax} = AN^b$.

III. A PHYSICAL ANALOGY

In order to proceed to the physical analogy [14] we invert the rank–frequency distribution in the sense of considering the relation between the **number of items** N_j having **absolute frequency** equal to j, see Fig. 2 for an example.

In this way, the similarities between the word distribution in the texts and the Bose-distribution in statistical physics become evident. We can identify the energy levels j with absolute word frequencies (the number of occurrences in a given text). Thus, the words with frequency 1 occupy the level $j = 1$, the words with frequency 2 occupy the level $j = 2$, etc.

We thus suggest [14] that the occupation of the jth level equals the number of different words with the frequency j. The use of the Bose-distribution is justified as a model since level occupations can have any value, in particular, significantly larger than unity for low frequencies. The lowest level corresponding to *hapax legomena* in this approach can be identified with the Bose-condensate.

The mathematical details of the above physical analogy are as follows [14]. In the Bose-statistics, the occupation of the jth level is given by

$$N_j = \frac{1}{z^{-1}e^{\varepsilon_j/T} - 1}, \tag{1}$$

where z is the fugacity, ε_j is the energy spectrum, and T is the temperature.

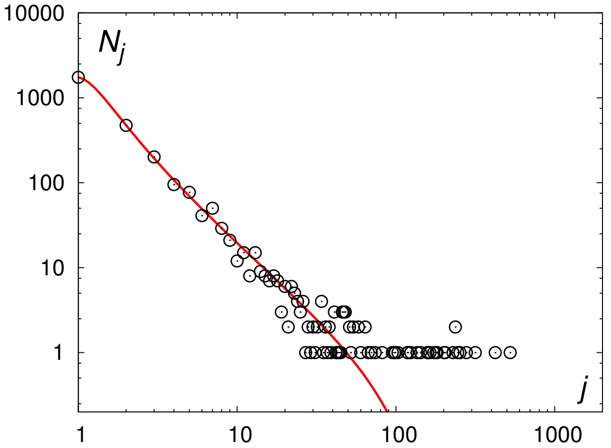

Fig. 2. (Color online) The fit of the power energy spectrum to the level occupations. The graph corresponds to the Greek translation of *The Little Prince*. Solid (red) line corresponds to the fitting function N_j given by Eq. (3), empirical data are denoted by empty circles.

Previously, we have shown that a power energy spectrum is a proper model to describe lower levels,

$$\varepsilon_j = (j-1)^\alpha. \tag{2}$$

The unity is subtracted for convenience in order to set the lowest level energy to zero.

In this work, we analyze only the levels with j larger than 10. For high levels, the energy spectrum must be corrected to have a weaker dependence on j. The respective analysis is yet to be done.

The parameters of the distribution function are defined as a result of fitting. Namely, the observed values $N_j^{\rm exp}$ are fitted using the function

$$N_j = \frac{1}{z^{-1}e^{(j-1)^\alpha/T} - 1} \tag{3}$$

with three parameters, z, α, T. They are defined in two steps. First, the "fugacity" z is calculated from the occupation number of the lowermost state, i. e., the number of *hapax legomena*:

$$N_{\rm hapax} = \frac{z}{1-z}. \tag{4}$$

Then, temperature T and exponent α in Eq. (3) are found from the fitting procedures for $j = 2 \div 10$.

It was discovered [14] that the ratio $\ln T/\ln N$ is a good variable for comparative linguistic studies, since the "temperature" parameter T scales as the size of a sample increases. We use this very ratio

$$\tau = \frac{\ln T}{\ln N} \tag{5}$$

for comparative analysis in this work as well.

IV. WHY *THE LITTLE PRINCE*?

The Little Prince (the French original title is *Le Petit Prince*) is a famous novella written by Antoine de Saint-Exupéry in 1943. Since its first publication, the novella was translated into over 190 languages (the site http://www.patoche.org/lepetitprince/gallima.htm listed 214 different languages, including some dialects and artificial languages, as of November 2010). In fact, not so many titles can be found in such a variety of translations. Moreover, a significant portion of these translations is available online. The Bible (being more precise, the New Testament) and the Universal Declaration of Human Rights (http://unicode.org/udhr/) certainly take the lead as to the number of translations, remaining though quite specific in language style. Still, those two texts are a good choice for some further comparative studies.

It should be noted that the novella under study – due to a high number of dialogues – approaches the colloquial genre in its frequency behavior. In particular, our calculations give the pronoun 'I' as the most frequent word in the Ukrainian translation and the second most frequent (after 'the') in the English one, which is a feature found only for colloquial texts, cf. [22].

For the analysis of *The Little Prince* we selected forty languages trying both to have several representatives of one family and to have different families covered. Our choice was also connected with the availability of the electronic texts of the translations. The following are the languages in the alphabetic order: Arabic, Armenian, Azerbaijani, Bamana, Basque (Euskara), Belarusian, Bulgarian, Catalan, Chinese, Croatian, Czech, English, Esperanto, Estonian, Farsi, French, Georgian, German, Greek, Hebrew, Hindi, Hungarian, Italian, Japanese, Korean, Latvian, Lithuanian, Lojban, Mauritian Creole (Morisyen), Mongolian, Polish, Portuguese, Romanian, Russian, Serbian, Spanish, Thai, Turkish, Ukrainian, Vietnamese.

The sources of the electronic texts were obtained from the following links:

- http://www.odaha.com/antoine-de-saint-exupery/maly-princ
- http://www.petit-prince.at/links.htm
- http://ukrlib.com.ua/books-zl/
- http://www.lib.ru/EKZUPERY/
- http://olddreamz.com/bookshelf/prince/littleprince.html
- http://www.xiaowangzi.org.
- and http://xorxes.110mb.com/LPP.html

V. RESULTS AND DISCUSSION

The results of calculations are summarized in Table and Fig. 3. So far we leave aside the analysis of the fugacity analogue z, its value is close to unity for all the analyzed texts due to the text size N being large enough.

Previous analysis [14] showed that the values of the temperature parameter and the exponent α correlate with the analyticity level of the language.

From Fig. 3 we can observe the grouping of languages into several domains on the $(\alpha;\tau)$ plane (the Thai data are not taken into consideration, see the caption of Table 1 for details). The domains are:

- Chinese, Vietnamese, Mauritian Creole, Bamana, Japanese, and Lojban;
- English, French, German, Hindi, Portuguese, and Italian;
- Czech, Croatian, Serbian, Farsi, Romanian, Spanish, and Mongolian;
- Belarusian, Polish, Russian, Ukrainian, Latvian, Lithuanian, Catalan, Arabic, Azerbaijani, Georgian, Hebrew, Hungarian, Korean, and Turkish (some internal subgrouping can be also seen).

Somewhat separately stand the following languages: Basque, Estonian, Armenian, Greek, Esperanto, and Bulgarian (in this Slavic language, the inflection was significantly reduced comparing to other related languages and this is probably the reason why it escaped from the relevant domains). Slavic languages split into two groups, one being composed of Belarusian, Polish, Russian, and Ukrainian, another one consisting of Croatian, Serbian, and – a bit unexpectedly – Czech. It is curious that Catalan parameters put this language in the domain, which would not be expected due to language relationships.

The obtained data can be another confirmation of the hypothesis on the connection between the values of the parameters α, τ and the analyticity of languages.

It is interesting to note that two artificial languages, namely Esperanto and Lojban, occupy quite different positions. This is due to the approach used while the respective languages were created. While Lojban is based on predicate logic and is highly analytical, even close to the "machine language", Esperanto is mostly an agglutinative language having common features with natural languages.

We have to emphasize that from close positions of languages in Fig. 3 one should not conclude in any case about a close relation between the languages in question. Rather, this means some similarities in the frequency structure of texts on the word level, which is linked to language analyticity / syntheticity.

In further works, we also plan to make similar studies with other texts, in particular, *The Universal Declaration of Human Rights*, for which even a higher variety of languages is available. This would also help to establish the correlation between parameter values and text genre.

Language	Original Title	N	N_{hapax}	α	T	$\tau = \ln T / \ln N$
Polish	Mały książe	11272	2025	1.58	814	0.718
Lithuanian	Mažasis princas	10899	2105	1.62	784	0.717
Russian	Маленький принц	12547	1915	1.55	831	0.712
Armenian	Փոքրիկ իշխանը	13794	1727	1.49	885	0.712
Azerbaijani	Balaca Şahzadə	12557	2667	1.58	820	0.711
Georgian	პატარა უფლისწული	10822	2519	1.60	741	0.711
Korean	어린 왕자의	11282	3048	1.55	748	0.709
Belarussian	Маленькі прынц	12391	1989	1.55	785	0.707
Latvian	Mazais princis	11527	1957	1.56	740	0.706
Hebrew	הנסיך הקטן	12105	3454	1.63	742	0.703
Catalan	El Petit Príncep	13954	1811	1.59	808	0.701
Hungarian	Kis herceg	12041	2476	1.56	725	0.701
Arabic	الأمير الصغير	1001	3367	1.62	607	0.696
Basque (Euskara)	Printze Txikia	11760	1957	1.50	683	0.696
Ukrainian	Маленький принц	11553	2205	1.55	673	0.696
Turkish	Küçük prens	11697	3173	1.65	673	0.695
Bulgarian	Малкия принц	12066	1609	1.39	625	0.685
Spanish	El Principito	13735	1479	1.47	669	0.683
Farsi	شازده كوچول	14214	2127	1.45	679	0.682
Czech	Malý princ	11398	2158	1.48	581	0.681
Serbian	Mali Princ	12217	1945	1.52	599	0.680
Estonian	Väike prints	11901	2056	1.41	587	0.679
Croatian	Mali Princ	12095	1958	1.49	587	0.678
Greek	Ο μικρός πρίγκιπας	14447	1742	1.58	653	0.677
Romanian	Micul prinţ	13188	1708	1.50	603	0.675
Mongolian	Бяцхан хун тайж	11819	2029	1.46	585	0.669
Italian	Il Piccolo Principe	12429	1734	1.45	528	0.665
Portuguese	O Pequeno Príncipe	12646	1591	1.47	531	0.664
Esperanto	La Eta Princo	11808	1555	1.55	505	0.664
French	Le Petit Prince	13926	1684	1.44	546	0.661
German	Der Kleine Prinz	14077	1554	1.49	544	0.659
Hindi	नन्हा राजकुमार	14014	1129	1.38	518	0.655
English	The Little Prince	16905	1030	1.43	579	0.653
Mauritian Creole	Zistoir Ti-Prins	12553	747	1.24	375	0.628
Japanese*	あのときの王子くん	19923	938	1.36	496	0.627
Vietnamese	Hoàng Tử Bé	17694	535	1.15	410	0.615
Chinese**	小王子	22806	420	1.24	458	0.611
Lojban	le cmalu noltru	17482	581	1.18	360	0.603
Bamana	Masadennin	16269	852	1.32	314	0.593
Thai***	เจ้าชายน้อย	2656	2241	1.47	50	0.497

* As Japanese texts do not have an explicit word-division, we used special software (UniDic, MeCab, and ChaSen) to obtain the word statistics.

** The data for the Chinese translation correspond to the frequency distribution of characters, not words, as Chinese texts do not contain word division.

*** In Thai texts, spaces separate not words but rather sentences or parts of sentences, thus the respective data occupy a separate position. This language is given mainly for future references.

Table 1. The parameters of energy spectrum and temperature of texts. N is the text size (total number of words), N_{hapax} is the number of hapaxes in the text, α and T are the fitting parameters in Eq. (3).

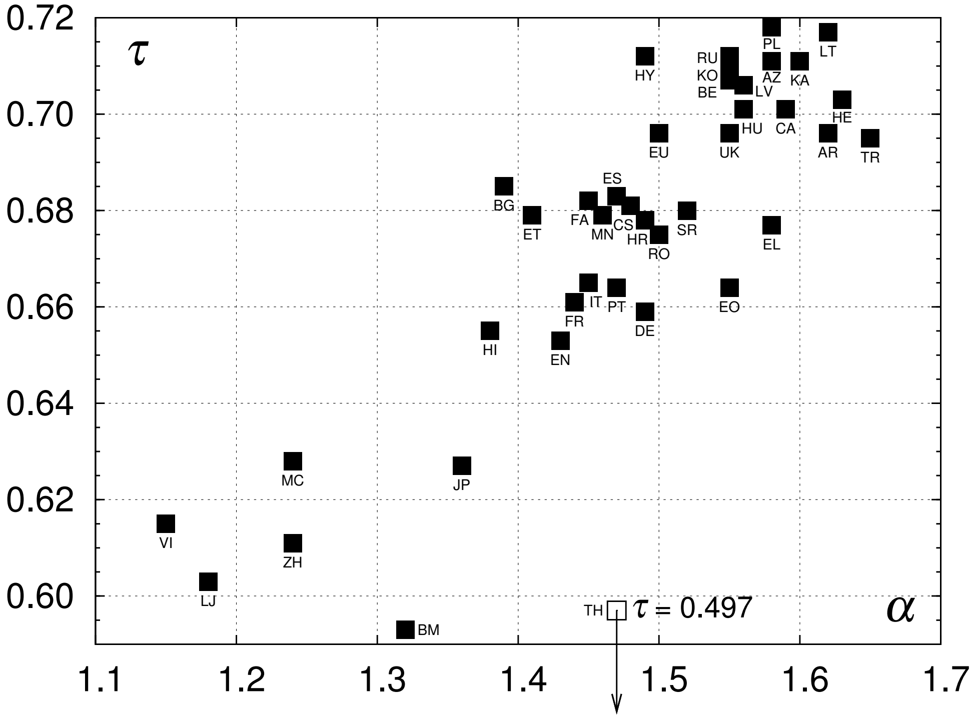

Fig. 3. The positions of different languages on the $(\alpha;\tau)$ plane. The languages are denoted by the ISO codes (where a two-letter code is not assigned, we used the closest two-letter notation): AR — Arabic, AZ — Azerbaijani, BM — Bamana, BE — Belarusian, BG — Bulgarian, CA — Catalan, CS — Czech, DE — German, EN — English, ES — Spanish, EU — Basque (Euskara), FA — Farsi, FR — French, EL — Greek, EO — Esperanto, ET — Estonian, HE — Hebrew, HI — Hindi, HR — Croatian, HU — Hungarian, HY — Armenian, IT — Italian, JP — Japanese, KA — Georgian, KO — Korean, LJ — Lojban, LV — Latvian, LT — Lithuanian, MC — Mauritian Creole (Morisyen), MN — Mongolian, PL — Polish, PT — Portuguese, RO — Romanian, RU — Russian, SR — Serbian, TH — Thai, TR — Turkish, UK — Ukrainian, VI — Vietnamese, ZH — Chinese.

Acknowledgements. We are grateful to Haruko Sanada for advices regarding the automatic word division in Japanese and to Valentin Vydrin for providing a copy of the Bamana translation of the novella.

[1] V. Pareto, *Cours d'economie politique* (Rouge, Lausanne, 1896–1897).

[2] J. B. Estoup, *Gammes stenographiques* (Institut Stenographique de France, Paris, 1916).

[3] G. K. Zipf, *The psychobiology of languages* (Houghton-Mifflin, Boston, Mass., 1935).

[4] G. K. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley, Cambridge, Mass., 1949).

[5] M. Petruszewycz, Math. Sci. Hum. **44**, 41 (1973).

[6] I. Kanter, D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[7] R. Ferrer i Cancho, Physica A **345**, 275 (2005).

[8] S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen, New J. Phys. **11**, 123015 (2009); Physica A **389** 330 (2010).

[9] M. L. Bender, Pr. Gill, Curr. Anthropol. **27**, 280 (1986).

[10] Neng-zhi Jin, Zi-xian Liu, Wen-yuan Qiu, Chin. J. Chem. Phys. **22**, 27 (2009).

[11] M. R. Dudek, S. Cebrat, M. Kowalczuk, P. Mackiewicz, A. Nowicka, D. Mackiewicz, M. Dudkiewicz, Computat. Meth. Sci. Technol. **13**, 5 (2007).

[12] C. M. Urzúa, Econ. Lett., **66**, 257 (2000).

[13] T. R. Gulten, Polit. Life Sci. **21**, 26 (2002).

[14] A. Rovenchak, S. Buk, Physica A **390**, 1326 (2011); arXiv:1011.5076 (2010).

[15] B. Mandelbrot, in: *Communication Theory*, ed. by W. Jackson (Academic, New York, 1953), p. 486.

[16] H. de Campos, J. M. Tolman, Poetics Today **3**, 177 (1982).

[17] K. Kosmidis, A. Kalampokis, P. Argyrakis, Physica A **366**, 495 (2006).

[18] S. Miyazima, K. Yamamoto, Fractals, **16**, 25 (2008).

[19] I.-I. Popescu *et al.*, *Word Frequency Studies* (Mouton de Gruyter, Berlin–New York, 2009).

[20] E. G. Hirsch, I. M. Casanowicz, J. Jacobs, M. Schloessinger, in: *The Jewish Encyclopedia* (Funk and Wagnalls, New York, 1901–1906), pp. 226–229; available online at `http://www.jewishencyclopedia.com`.

[21] A. Kornai, *Mathematical Linguistics* (Springer, 2008).

[22] S. Buk, *3 000 najchastotnishykh sliv rozmovno-pobutovoho stylju suchasnoji ukrajins'koji movy [3 000 most frequent words of the colloquial genre of the modern Ukrainian language]* (Lviv University Press, Lviv, 2006).

ВИЗНАЧЕННЯ ТЕРМОДИНАМІЧНИХ ПАРАМЕТРІВ ТЕКСТІВ НА ПІДСТАВІ РАНҐОВО-ЧАСТОТНИХ РОЗПОДІЛІВ ДЛЯ СЛІВ

Андрій Ровенчак[1], Соломія Бук[2]
[1] *Кафедра теоретичної фізики, Львівський національний університет імені Івана Франка, вул. Драгоманова, 12, Львів, 79005, Україна*
[2] *Кафедра загального мовознавства, Львівський національний університет імені Івана Франка, вул. Університетська, 1, Львів, 79000, Україна*

У роботі наведено результати розрахунків нового набору параметрів, отриманого з ранґово-частотних розподілів для слів у текстах. Ці параметри визначено на підставі аналогії між ранґово-частотним розподілом і квантовим розподілом Бозе. Обчислення виконано для перекладів казки Антуана де Сент-Екзюпері "Маленький принц" чотирма десятками мов із різних мовних родин. Отримані дані вказують на існування зв'язку між типом граматики певної мови і запропонованими параметрами.