Electronic Journal of Applied Statistical Analysis
EJASA (2013), Electron. J. App. Stat. Anal., Vol. 6, Issue 1, 84 – 96
e-ISSN 2070-5948, DOI 10.1285/i20705948v6n1p84

THE NEGATIVE BINOMIAL MODEL OF WORD USAGE

Nina P. Alexeyeva[*], Alexandre Sotov

Department of Mathematics, St. Petersburg State University, Russian Federation

Received 10 June 2012; Accepted 01 March 2013
Available online 26 April 2013

***Abstract**: How people make texts is a rarely asked but central question in linguistics. From the language engineering perspective texts are outcomes of stochastic processes. A cognitive approach in linguistics holds that the speaker's intention is the key to text formation. We propose a biologically inspired statistical formulation of word usage that brings together these views. We have observed that in several multilingual text collections[1] word frequency distributions in a majority of non-rare words (occurring over 10 times) fit the negative binomial distribution (NBD). However, word counts in artificially randomized (permuted) texts agree with the geometric distribution. We conclude that the parameters of NBD deal with linguistic features of words. Specifically, named entities tend to have similar parameter values. We suggest that the NDB in natural texts accounts for intentionality of the speaker*

***Keywords**: Word frequency, negative binomial distribution (NBD), semantics, text randomization.*

1. Introduction

Metaphors of languages and linguistic entities as organisms have been current since the romantic scholarship of the 19th century. The idea that language is a living matter was expressed by Schlegel; Humboldt coined the term 'Sprachorganismus' [20] and believed that language exists as *energiea*, a manifestation of the vital power of the human spirit (cf. [7] for a summary of such views). In literary studies Propp's [15] reduction of the Russian fairytale to a number of functions was inspired by Goethe's functional definition of the parts of the organism. Recently, eco- and biolinguistics aim at applying biological models to language as a cultural system on the one hand, and as a cognitive, on the other. Much of that research is focused on the evolution of the human language faculty and changes in the linguistic system over time. The latter raises

[*] Email: ninaalexeyeva@mail.ru

criticism, particularly in historical linguistics: '…Biological organisms are not a good model for the "behaviour" of … the elements of language' [5]. Similar arguments were voiced in the domain of literary studies [17]. But the idea that languages and genres undergo transformations differently from organisms does not mean that the metaphor of the 'living word' should be automatically discarded. Statistical natural language processing (NLP), a direction of research flourishing with the unprecedented availability of textual resources and computational power, is based on the analysis of empirically observable features of text corpora, dealing with what was termed by Humboldt as *ergon*, the materiality of language. In the NLP literature one of the major topics is the analysis of distributional properties of linguistic entities for practical purposes of text data mining. Yet, viewing lexical units as populations brings about a reference to ecology which is concerned with modelling population dynamics, the change in the number of individuals in populations over time and space [10]. The parallel is quite striking if you consider that most such techniques are based on word frequency, i.e. population, counts.

We argue that linguistics can benefit from a careful consideration of methods that are well established in quantitative ecology. In this sense the biological metaphor, phrased in the flexible language of mathematics, does not lead the linguist astray; on the contrary, it sheds light on the process of text formation. A central question of linguistics is: How people understand and create sentences? From the language engineering perspective these are word sequences defined as an outcome of stochastic processes [7]. That view is substantiated by the complexity of language and subtleties of its use, thus the specific nature of the process is considered altogether irrelevant. On the contrary, a cognitive approach holds that text is the basic unit of linguistic analysis, that it possesses the property of coherence, or totality, and that it is structured by the speaker's intention – a concept reminiscent of Humboldt's *energeia* [13], [16]. This leads to the problem of lexical choice and ultimately to the search of techniques universally used by humans to create texts, not simply their parts, – irrespective of style, poetics or ideology. But texts themselves do not occur in empty space. They are a part of discourse [18]: a chapter of a novel, a newspaper article or a fairytale is integrated into a larger meaningful entity – a book, a genre, - a 'natural' collection (as opposed to larger corpora). Unlike an artificially constructed linguistic corpus of many million words, that is a population where each text is on the right place defined by its mode of existence in culture. Such constellations of texts are not infrequently - but not always - open systems (i.e.. articles of an ongoing newspaper vs. chapters of a novel) and can be compared to ecosystems. Modelling word frequency distributions thus remains in the crux of linguistic investigations with the addition that one has to account for systematic intentionality, i.e. human cognitive performance, manifested in 'natural' collections of texts.

The locus of ecologically inspired text modelling is in the interrelation of the two kinds of linguistic individuals: the word and the text. Our thesis is that the type of word frequency distribution is determined at the level of the collection and that individual character of words is reflected in the parameters of their distribution. That is to say, in linguistic texts the speaker's intention reveals itself in such a way that certain 'text individuals' are predisposed to certain 'word individuals'. We propose that that is modelled by the Gamma-Poisson scheme of the Negative Binomial Distribution (NBD). To test this we fitted the model to word counts first from natural and then to simulated (permuted) texts in Sanskrit, Russian and English, which indicated distinctions between different parts of speech. We intend to show that semantic features of lexicon yield distributional differences.

2. The Bayes Estimators of the parameter θ

Let us consider word frequencies in text as a random variable and a collection of n texts as a sample. We shall assume that texts contain an approximately identical number of words. So, for example, in our collection of n texts the word "the" is attested in the first text 8 times, in the second text 7 times and so on. Our observations indicate that such frequency tables typically fit NBD. Unlike the so-called Zipf's law, NBD does not consider the word's rank in the frequency table as its parameter. In fact it is not known why the Zipf's law holds not only for natural linguistic data but arguably also for random texts [21], but using the NBD model it is possible to explain word frequencies by the means of the Gamma-Poisson construction, which was successfully used for the description of the host-parasite system in parasitology [3].

2.1 The Gamma-Poisson construction of the negative binomial distribution

A regulatory scheme of the host-parasite relationships of warble flies in farm animals is deduced from the Gamma-Poisson construction of the NBD [4], [6]. We shall apply that model to the analysis of linguistic data viewing the use of words in analogy with the parasite invasion process. The Poisson distribution occurs when the probability of an event is small and the number of trials is large. In our case the number of trials is the number of words in a collection of texts. The probability to find a word in a piece of text is low because a typical word occurs rarely. In the corpora presented in this research such probability is around 0,015 in running text of 1000 words. Therefore we assume that we have the Poisson distribution of word occurrence with the intensity parameter λ in the case when the process of word usage is not regulated as such. However, let us consider that the speaker's intention regulates the process allowing some words to occur (to be 'kept' in the text) and some words to pass out (to be 'lost') with the probability $p\cdot$. We can spread the intensity on two components $\lambda = \lambda_1 + \lambda_2 = \lambda q + \lambda p,$ where $q = 1 - p.$ Accordingly, the number of trials is expressed in the sum $j + k,$ where k and j show respectively how many words are lost, and how many are kept by the speaker. Thus we have the Poisson distribution $P(j \mid \lambda_1)$ of the residual words number and $P(j \mid \lambda_2)$ of the number of lost words. One can get the distribution of the residual words number with the assumption that the lost words number is equal to k. The intensity λ_2 agrees with the gamma distribution $\gamma(\lambda_2 / k)$ with the scale parameter equal to 1 and the shape parameter equal to k from the fiducial Poisson and gamma distributions. Thus the probability that the residual words number is equal to j with the assumption that the lost words number is equal to k can be calculated in the following way:

$$\begin{aligned}
&\int_0^\infty P(j\,|\,\lambda_1)\gamma(\lambda_2\,|\,k)d\lambda_2 = \\
&=\int_0^\infty P(j\,|\,\lambda q)\gamma(\lambda p\,|\,k)d(\lambda p) = \\
&=\int_0^\infty \frac{\lambda^j q^j}{\Gamma(j+1)}e^{-\lambda q}\,\frac{\lambda^{k-1}p^{k-1}}{\Gamma(k)}e^{-\lambda p}d(\lambda p) = \qquad (1)\\
&=\frac{p^k q^j}{\Gamma(j+1)\Gamma(k)}\int_0^\infty \lambda^{k+j-1}e^{-\lambda}d\lambda = \\
&=\frac{\Gamma(k+j)p^k q^j}{\Gamma(j+1)\Gamma(k)} = \beta_-(j\,|\,k,p)
\end{aligned}$$

Now, NBD of the residual words number is determined by means of two parameters: p indicates the probability of a word not being used ('lost') and defines the speaker’s intention, while k indicates the number of 'lost' words and defines 'the invasive dimension' of a word, which for the lack of a linguistic term one can call the word's 'usualness'.

2.2 The negative binomial distribution properties

The mathematical expectation and variation of the distribution (1) are:

$$\mu=\frac{kq}{p}, \qquad \sigma^2 = \frac{kq}{p^2}. \qquad (2)$$

The parameters *p* and *k* can be obtained from μ and σ:

$$p=\frac{\mu}{\sigma^2}, \qquad k = \frac{\mu^2}{\sigma^2-\mu}. \qquad (3)$$

If $\sigma^2 > \mu + \mu^2$, then $k < 1$,else $k > 1$. Word extensiveness (*E*) as the probability of a word occurring at least once, and word intensity (*I*) as the ratio of the mean to the word extensiveness are obtained as follows:

$$E = 1 - p^k, \qquad I = \frac{\mu}{E} \qquad (4)$$

The number of texts where the word is attested at least once is the document frequency (*DF=nE*). By dint of k increasing word extensiveness increases fast under small p. For example, when certain words (such as proper and place names, or the so-called named entities) have small k and p values, then σ^2 is large, and μ is small: these words appear in only a few texts of the collection but when they do, they appear a lot (Figure 1).

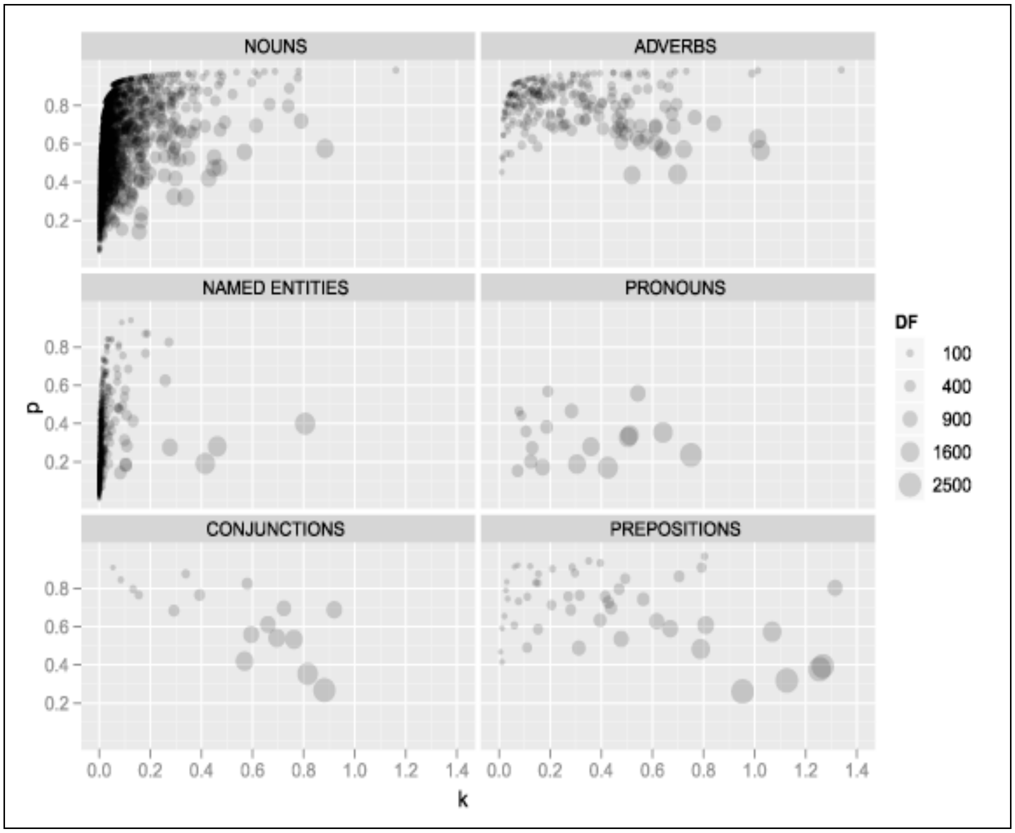

Figure 1. NBD parameters *k* and *p* in St. Petersburg Times collection: differences between some parts of speech. Unlike the majority of nouns, named entities tend to the lower left corner of the plot. DF is document frequency.

2.3 Confidence intervals and testing of fits

Let $x_1,\ldots,x_n$ be a sample of observations from the NBD with parameters p and k, $q = 1 - p$. Using the maximum likelihood method (MLE), the likelihood function is as follows:

$$\ln L(x_1,\ldots,x_n \mid k,p) = \sum_{i=1}^{n} \ln \beta\_(x_i \mid k,p), \quad \ln L(x_1,\ldots,x_n \mid k,p) = \sum_{i=1}^{n} \ln \beta\_(x_i \mid k,p), \text{ where}$$

$$\ln \beta\_(j \mid k,p) = \ln\Gamma(k+j) - \ln\Gamma(k) - \ln\Gamma(j+1) + k\ln p + j\ln(1-p) =$$

$$= \sum_{l=0}^{j-1} \ln(k+l) - \ln(\Gamma(j+1)) + k\ln p + j\ln(1-p).$$

By means of the parameter differentiation we obtain normal equations:

$$\frac{\partial}{\partial p} \ln L(x_1,\ldots,x_n \mid k,p) = \sum_{i=1}^{n} \left(\frac{k}{p} - \frac{x_i}{1-p} \right) = 0,$$

$$\frac{\partial}{\partial k} \ln L(x_1,\ldots,x_n \mid k,p) = \sum_{i=1}^{n} \left(\sum_{l=0}^{x_i-1} \frac{1}{k+l} + \ln p \right) = 0.$$

The maximum likelihood equations can be written as:

$$\frac{k(1-p)}{p} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{l=0}^{x_i-1}\frac{1}{k+l}\right) = \ln(\bar{x}+k) - \ln(k) \qquad (5)$$

The likelihood estimation $\hat{k}$ can be found by solving (5), $\hat{p} = \frac{\hat{k}}{\hat{k}+\bar{x}}$. Let us consider a vector

$g(x) = (g_1(x), g_2(x))^T = \left(\frac{\partial}{\partial p}\ln\beta_-(x\,|\,k,p), \frac{\partial}{\partial k}\ln\beta_-(x\,|\,k,p)\right)^T$,where x is from NBD.

$$g_1(x\,|\,k,p) = \frac{\partial}{\partial p}\ln\beta_-(x\,|\,k,p) = \frac{k}{p} - \frac{x}{1-p},$$

$g_2(x\,|\,k,p) = \frac{\partial}{\partial k}\ln\beta_-(x\,|\,k,p) = \psi(k+x) - \psi(k) + \ln p$, where $\psi(k) = (\ln\Gamma(k))'$.

Let us now construct the matrix of expectations $R = \mathrm{E}g(x)g(x)^T$ with elements:

$r_{22} = -\mathrm{E}\frac{\partial}{\partial p}\big(g_2(x\,|\,k,p)\big) = -\mathrm{E}(\varphi(k+x) - \varphi(k)) = \varphi(k) - \sum_{j=0}^{\infty}\varphi(k+j)\beta_-(j\,|\,k,p),$ where $\varphi(k) = \psi'(k)$. It is possible to take advantage of the approached calculation:

$$r_{22} \approx \varphi(k) - \varphi(k + \mathrm{E}x) = \varphi(k) - \varphi\big(kp^{-1}\big).$$

$$r_{11} = -\mathrm{E}\frac{\partial}{\partial p}\big(g_1(x\,|\,k,p)\big) = \frac{k}{p^2} + \frac{\mathrm{E}x}{(1-p)^2} = \frac{k}{p^2} + \frac{k}{pq} = \frac{k}{p^2q},$$

$$r_{12} = r_{21} = -\mathrm{E}\frac{\partial}{\partial k}\big(g_1(x\,|\,k,p)\big) = -p^{-1},$$

So we receive covariance matrix *(nR)*$^{-1}$ of $\hat{p}$ and $\hat{k}$.

$$S = R^{-1} = \frac{1}{|R|}\begin{bmatrix} r_{22} & p^{-1} \\ p^{-1} & kp^{-2}q^{-1}\end{bmatrix}, \qquad |R| = \frac{kr_{22} - q}{p^2q}.$$

The asymptotic variances $V_p = \frac{S_{11}}{n}$ and $V_k = \frac{S_{22}}{n}$ of $\hat{p}$ and $\hat{k}$ are expressed in terms of diagonal elements $S_{11} = S_{11}(k,p) = \left(\frac{k}{q} - \frac{1}{r_{22}(k,p)}\right)^{-1}$, $\quad S_{22} = S_{22}(k,p) = \left(r_{22}(k,p) - \frac{q}{k}\right)^{-1}$.

Thus we obtain asymptotic 95%-confidence intervals for parameters *p* and *k* of the form:

$$\left(\hat{p}-1.96\sqrt{\frac{S_{11}(\hat{k},\hat{p})}{n}};\hat{p}+1.96\sqrt{\frac{S_{11}(\hat{k},\hat{p})}{n}}\right) \text{ and } \left(\hat{k}-1.96\sqrt{\frac{S_{22}(\hat{k},\hat{p})}{n}};\hat{k}+1.96\sqrt{\frac{S_{22}(\hat{k},\hat{p})}{n}}\right).$$

In the experiments described below a series of goodness-of-fit tests was applied to test the fit of an observed to a theoretical distribution by means of the Kolmogorov-Smirnov (K-S) and the chi-squared tests. The K-S has certain advantages over the chi-squared test, but it tends to be more sensitive near the center of the distribution than at the tails. More importantly, the statistical model underlying the K-S method assumes a continuous distribution, and for that reason in our case it has a low power. Nevertheless, the test is sometimes applied to data from discrete distributions simply because its alternative is not perfect either. Indeed, such difficulties are known in biometry [3-4,6]: the chi-squared test looses power when NBD parameters are low. Problems also arise when the parameters are unusually high and with the decrease in the number of degrees of freedom. Overall, in our observations most non-rare words, occurring at least 20 times in the corpus, fit the NBD according to both, the K-S and the chi-squared tests. A truncated negative binomial distribution model, which we also applied, yielded an agreement of such words with a lesser limit of minimum probability in one class. Moreover, we argue that flaws in the testing of goodness of fit do not underestimate the applicability of NBD as a model of word usage.

3. Experiments and results

3.1 Text coherence

A natural, i.e. human, text is coherent because of the speakers' intention – a property which is absent in mechanically combined words and phrases. Under a null hypothesis of no intention-induced data structure, word frequencies in human-generated and randomly formed texts are drawn from the same distribution. In order to test this hypothesis we analysed distribution of 30 most frequent words in three grammatical categories in the Sanskrit source text of the *Ṛgveda* (SK, 16.5×10^4 word tokens), its Russian (RU, 24.5×10^4) and English (EN, 29.4×10^4) translations.

The *Ṛgveda* (N=1028, 10552 verses) is an ancient corpus of cultic poetry in Sanskrit [2],[8],[11]. We chose it because the texts in that collection are brought together by a tradition rather than a modern linguist, and since they contain relatively homogeneous material and are sizeable for corpus analysis. The translations were selected to generalize the results on grammatically different languages. Similarly to Sanskrit, but in a lesser extent, Russian has a synthetic structure, while English is an analytic language. The words were manually attributed to the following parts of speech, 10 items in each group, contrasted according to the semantic principle: prefixes (SK) and prepositions (RU, EN); personal pronouns (SK, RU, EN); and proper names (SK, RU, EN). Prepositions and prefixes are synsemantic words assumed to occur meaningfully only vis-à-vis other entities [14]. Similarly, pronouns are synsemantic relational variables that rely on neighbouring words for decoding them. Proper names are autosemantcs, i.e. independent, deictic entities. Examples of analysed Vedic prefixes are: ā, prá, ví, abhí; among pronouns were D., G., Acc. Sg. I person forms; N. Sg., Acc., D., G. Pl. and G., D. Du. II person; and V., Acc. Sg. Masc., N., Acc. Neut., N. Pl. Masc. III person. Proper names included V. N. Acc. forms of common deities (agní-, índra-, sóma-, and others). For the Russian translation prepositions included N. Acc. Sg. of all persons, N.Pl. of II and III persons, Acc. Pl. of I person. For the English text pronouns included archaic II person forms but did not include N. Sg. of I person.

To simulate the null hypothesis we performed (I) 100 random permutations of all verses, each roughly equivalent of a sentence, and (II) 100 random permutations of all words. This was done for the source texts and for the translations. In Experiment I the verses were randomly ordered to generate quasi-texts containing grammatical sentences; sentences are coherent, but resulting quasi-texts are not. In Experiment II words, rather than verses, were permuted, generating random ungrammatical sequences of variable length. In both experiments the number of quasi-texts was equal to the original size of the collection (N=1028).

In the natural texts K-S tests indicated a good fit of NBD to the counts of 29 of 30 word types in Sanskrit, and to all 30 in English and Russian. In simulated data the 0.75 level or higher was reached by 99 per cent of analysed words in Sanskrit and Russian, and by 98 per cent in English in all 100 randomizations in Experiment I, and by 99 per cent of analysed words in all languages in Experiment II.

For each word the two NBD parameters, k and p were estimated, using the MLE, prior to and after randomizations. A multivariate outlier detection method [1] indicated that the k parameter vector of all words in the natural data is an outlier for each set of the randomized data (all P<.0001). Hence, the hypothesis that the vector of k for the same words has an identical distribution in the natural and in the simulated data can be rejected. Indeed, meaningful texts and both types of random sequences are fundamentally different kinds of production.

With the increase in randomness k approaches to unity and p to a constant that depends on the word's absolute frequency in the collection, or its term count, $TC = \sum_{i=1}^{n} x_i$. In the case of $k = 1$ NBD converges to the geometric distribution, therefore in randomized data p approaches to $\frac{1}{\mu + 1}$. Mean k values in the three groups of words were worked out for all randomization runs; they are presented together with the mean values in the natural data in Table 1.

It can be seen that if in the natural data the word's k is less than unity randomization causes an increase in the value, otherwise it decreases towards unity. For example, in Experiment I the mean k value increased for autosemantic words: in Sanskrit texts the mean k value (± the standard error of mean) of autosemantics is 0.168±0.088 in the natural texts and 1.058±0.055 in the simulated data; in Russian 0.203±0.079 and 1.165±.039; in English 0.207±0.130 and 1.311±.057 respectively. The mean k decreased for non-deictic synsemantics: 0.939±0.088 and 0.991±0.055 in Sanskrit, 1.335±0.079 and 0.946±0.039 in Russian; 1.763±0.130 and 1.089±0.057 in English. In Experiment II MLE estimates further indicated that the values approximated to unity for all words. In comparison with the natural data this constituted a universal increase in the mean k values of deictic altogether, and a decrease in the mean k of non-deictic synsemantics.

The vanishing of differences between parts of speech in k values, arising from randomization, is especially vivid in deictic words – proper names and personal pronouns. In the Sanskrit collection in Experiment I in 87 of 100 verse permutation runs a series of Mann-Whitney U tests yielded a significant (P<0.05) difference between pronouns and proper names at the mean significance level P=0.032. In Experiment II the tests indicated a significant difference only in 10 of 100 word permutations, mean P-level insignificant (0.5). In the English collection the significance level of $P<0.05$ was reached in 95 verse permutation runs and only in 5 word permutations, mean $P=0.016$ and $P=0.44$ respectively. In the Russian collection the significance level was reached in 87 and 10 runs respectively, $P=0.022$ and $P=0.46$. That is to say, in *k* values name deictic differed from relational deictic in coherent sentences and were indistinguishable from each other in random word sequences. Non-rare words, regardless of their semantic profiles, were homogeneously distributed.

To explore the differences between the part of speech categories in the natural data we ran the Kruskal-Wallis analysis of variance by ranks test for each language. The tests indicated differences in group *k* values in all languages (all $P<0.0001$). Multiple Comparisons showed that the differences are significant ($P<0.05$) except between proper names and personal pronouns in Sanskrit texts ($P=0.17$). Overall, mean ranks for the categories are highest for proper names (SK: 7.3, RU: 5.9, EN: 5.7) and lowest for prepositions and prefixes (24.4, 25.1, 24.9; for personal pronouns: 14.8, 15.5, 15.9 respectively).

In the source texts and their translations, regardless of the language, proper names are ‘patchy’ distributed in natural data, which is characterized by lower *k* values. The use of such words is involved in text level cohesion; however, as deictic words, proper names and personal pronouns also play a part in sentence cohesion. Absence of structure in simulated texts, due to their non-intentional character, translated into a different model of word occurrence. The word frequency distributions universally converged to the geometric distribution which is an exponentially averaged mixture of Poisson. As is known, the exponential distribution has the maximum entropy among all distributions concentrated on the non-negative semiaxis, therefore word frequency counts obtained from meaningless simulated texts converge to an ‘averaged’ distribution with maximum entropy.

3.2 Semantics

Reflecting about vocabulary at large, one may expect that named entities occur in texts heterogeneously and hence have lower values of the NBD parameters than other kinds of words. To determine whether the means of the parameter estimates for different parts of speech are from the same population we analysed data from a collection of newspaper articles [19] containing every issue of St. Petersburg Times in 2010 (SPT: N=3438, 144×10^4 word tokens).

Word frequency counts were gathered for all word types attested at least 20 times in the corpus and fitted to NBD to obtain MLE estimates of the parameters. We consider such words 'non-rare'. As previously, K-S tests indicated a typically good fit to the model. Figure 2 presents an overview of the estimates.

Frequency list entries were parsed as to parts of speech using the CLAWS system [9], and then manually annotated for proper names, place names, and their derivatives, as well as the names of currencies. The following word groups were analysed: nouns, adverbs, adjectives, verbs, personal and/or possessive pronouns, conjunctions, prepositions, first names, and named entities. Some examples are presented on Figures 2 and 3.

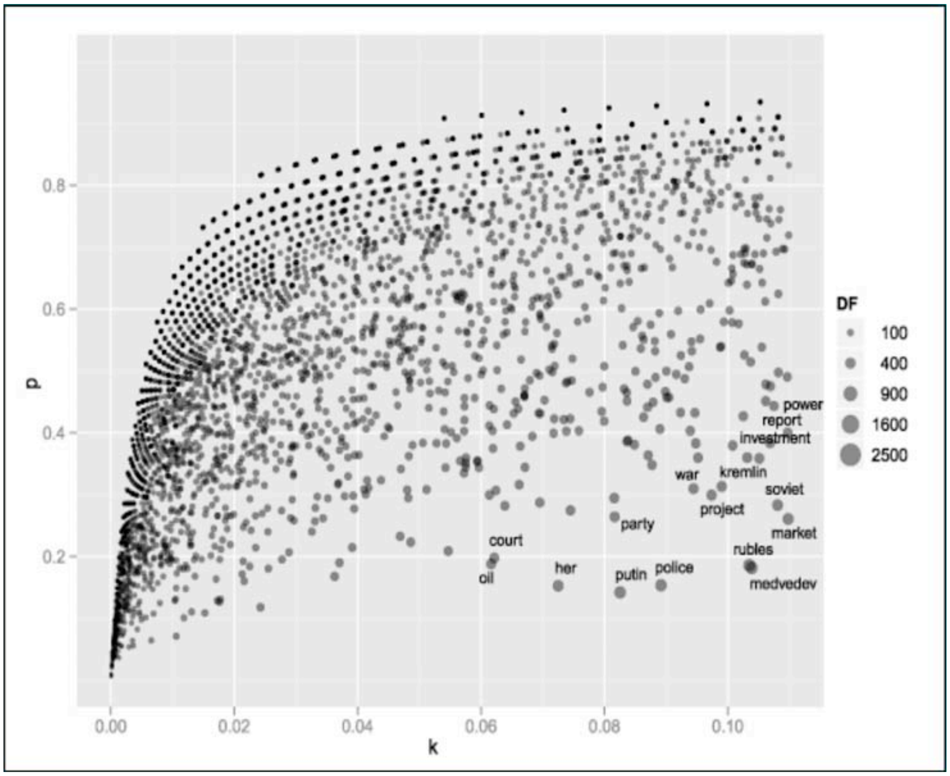

Figure 2. NBD parameters *k* and *p* in the St. Petersburg Times collection (a fragment). An orderly pattern on the edges of the figure is due to the dependence between the parameters for words with equal expected value: p=1-1/(k+1/μ). DF is document frequency.

A Kruskal-Wallis H test indicated differences between parts of speech in terms of *k* values (p<0.0001), as well as *p* values (p<0.0001).

Table 1. Mean±SE of different groups of words.

		Relational Entities		
		Prefixes, prepositions	Personal Pronouns	Proper Names
			Unnamed	Named
			Deictic Entities	
Data	Language	Synsemantics		Autosemantics
Natural	SK	0.939±.088	0.440±.088	0.168±.088
	RU	1.335±.079	0.655±.079	0.203±.079
	EN	1.763±.130	0.772±0.130	0.207±0.130
Ex. I	SK	0.991±.055	0.739±.055	1.058±.055
	RU	0.946±.039	0.843±.039	1.165±.039
	EN	1.089±.057	0.893±.057	1.311±.057
Ex. II	SK	1.026±.012	1.044±.012	1.077±.012
	RU	1.007±.010	1.057±.010	1.007±.010
	EN	1.011±.019	1.012±.019	1.105±.019

To determine which variables account for differences we applied a series of Kruskal-Wallis pair-wise post hoc tests in the differences of mean ranks. In both parameters named entities and nouns differed significantly from all analyzed parts of speech; first names differed only from nouns and named entities in both parameters, and from pronouns in *p*. Adjectives differed in both *k* and *p* from most categories except first names. Differences in *k* were observed for conjunctions, and in *p* for prepositions (Figure 1, Table 1).

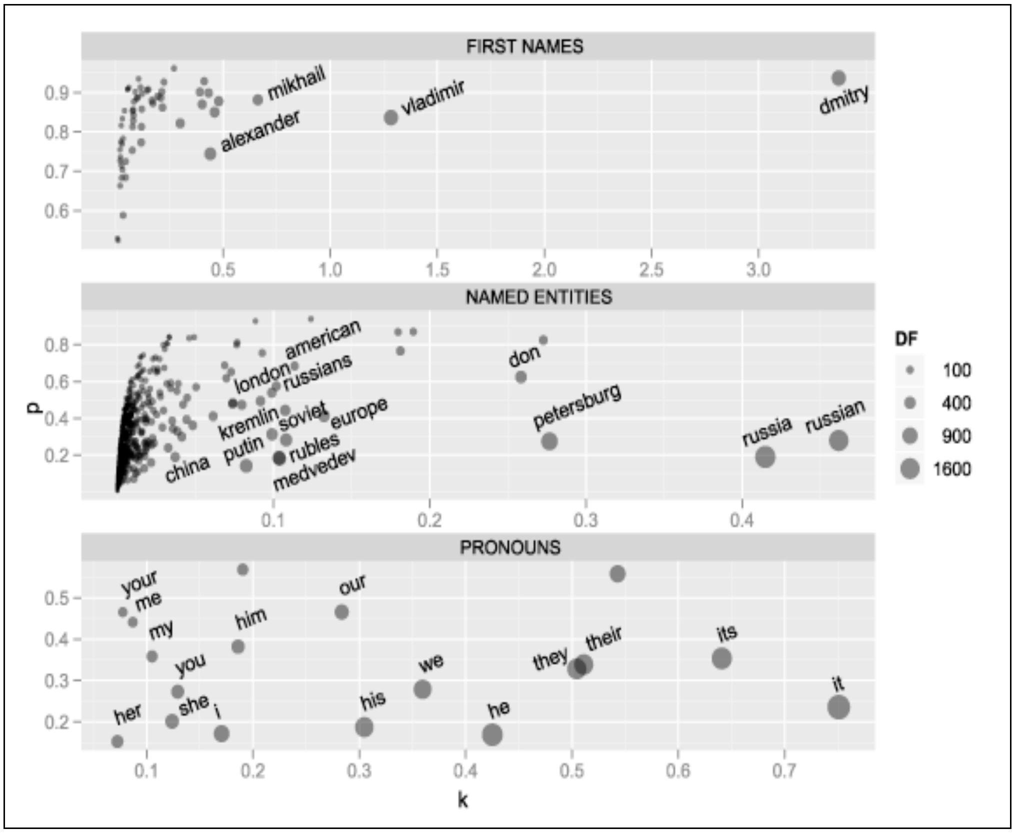

Figure 3. NBD parameters for some words in the St. Petersburg Times collection. First names are opposed to other named entities (proper and place names); pronouns are shifted on the k axis. DF is document frequency.

Table 2. Multiple Comparisons p values in k (above a diagonal) and p (below a diagonal) tested with a series of Kruskal-Wallis post hoc tests. Parts of speech: NP (named entities), NN (nouns), FN (first names), PNP (personal and possessive pronouns), AV (adverbs), AJ (adjectives), CJ (conjunctions), PR (prepositions), and V (verbs).

P values	NP	NN	FN	PNP	AV	AJ	CJ	PR	V
NP		0		0	0	0	0	0	0
NN	0			0	0	0	0	0	0
FN									
PNP		0.003				0.007			
AV	0	0		0		0			0
AJ	0	0		0	0.00004		0.0004	0.00002	0.0009
CJ	0.0002			0.01					0.011
PR	0	0.02		0.00001	0.02				0.007
V	0	0		0		0.03			

4. Conclusions

This paper introduced the Gamma-Poisson construction of NBD as a model of text formation. Words and texts are two fundamentally different units of linguistic analysis; using a biological example their interrelation is formally similar to a host-parasite system. We have analysed the occurrence of non-rare words in several collections of texts and concluded that a majority of them fits NBD. However, NBD parameters differ for grammatically different kinds of words. Named entities, specifically proper names, tend to occur frequently in just one or only a few texts, but are absent from the rest of the collection, i.e. they are heterogeneously distributed. We have described this phenomenon using NBD. Yet in permuted texts non-rare words follow the geometric distribution: since such texts lacked coherence of human-generated production, they were homogeneously distributed. We suggest that the distribution parameters capture semantic profiles of individual words because the way words are used by humans is inseparable from meaning [22]. Word frequency distributions are quantitative descriptions of the word's character.

References

[1]. Afifi, A.A., Azen, S.P. (1979). *Statistical analysis: a computer oriented approach*. New York: Academic Press.

[2]. Aufrecht, T. (1877). *Die Hymnen des Rigveda*. Bonn: A. Marcus.

[3]. Bart, A.G. (2003). *Analiz mediko-biologicheskikh sistem*. Sankt-Peterburg: Sankt-Peterburgskij Gos.Universitet.

[4]. Bart, A., Minář J. (1984). Basic regulatory parameters of the host-parasite system for warble flies of farm animals using Hypoderma bovis as an example. *Folia Parasitologica*, 31, 277-287.

[5]. Brian, J., Richard, J. (2003). *The handbook of historical linguistics*. Wiley-Blackwell.

[6]. Breev, K.A. (1968). O raspredelenii lichinok podkozhnykh ovodov v stadakh krupnogo rogatogo skota: Negativnoje binomialnoje raspredelenije kak model. *Parasitologija*, 2, 381-394.

[7]. Driem, G. (2008). *The Language Organism: Parasite or Mutualist? In Evidence and counter-evidence: essays in honour of Frederik Kortlandt, eds. F. Kortlandt, A. Lubotsky*. Amsterdam, New York: Podopi, 101-113.

[8]. Elizarenkova, T. (1999). Rigveda. Moskva: Nauka.

[9]. Garside, R., Smith N. (1997). *A hybrid grammatical tagger: CLAWS4. In Corpus Annotation: Linguistic Information from Computer Text Corpora*. Eds. R. Garside, G. Leech, A. McEnery. London: Longman, 102-121.

[10]. Gillman, M., Hails R. (1997). *An introduction to ecological modelling: putting practice into theory*. Wiley-Blackwell.

[11]. Griffith, R. (1963). *The Hymns of the Rigveda. 2 vols*. Varanasi: C.S.S.

[12]. Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2 (1), 15–59.

[13]. Leontiev, A.A. (1997). *Osnovy psikholingvistiki*. Moskva: Nauka.

[14]. Popescu, I., Mačutek J., Altmann G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

[15]. Propp, V. (1968). Morphology of the folktale. *Publications of the American Folklore Society*, v. 9. Austin: University of Texas Press.

[16]. Sakharnyj, L.V., Shtern A.S. (1988). *Nabor kl'uchevyh slov kak tip teksta. In Leksicheskije aspekty v sisteme professionalno-orientirovannogo obuchenija*. Perm: Permskij politekhnicheskij universitet, 34-51.

[17]. Steiner, P., Davydov S. (1977). The Biological Metaphor in Russian Formalism. *The Concept of Morphology*. Sub-Stance 6/7, 16 (July 1), 149-158. doi:10.2307/3684133.

[18]. Stubbs, M. (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

[19]. *The St. Petersburg Times Archive*. URL: http://times.spb.ru/index.php?action_id=5& i_number=1652year=2010.

[20]. Von Humboldt, W. (2010). *Ankündigung einer Schrift über die Vaskische Sprache und Nation, nebst Angabe des Geschichtspunctes und Inhalts derselben*. Berlin: Wissenschaftliche Buchgesellschaft.

[21]. Li, W. (1992). Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory*, 38 (6), 1842–1845. doi:10.1109/18.165464.

[22]. Wittgenstein, L. (1942). *The Blue and Brown Books*. London: HarperCollins.