George K. Mikros, Ján Mačutek

Word length distribution and text length: two important factors influencing properties of word length motifs

Abstract: Word length motifs and their properties in Modern Greek blogs are investigated. We observe that the parameter of the type-token relation decreases with the increasing text length, and that the parameter values in real and randomized texts are not significantly different. Relations between the number of motifs in a text and the text length is modelled by the linear function for Modern Greek and Ukrainian texts, with the parameter of the model being in both languages inversely proportional to the mean word length.

Keywords: word length motifs, text length, word length.

1 Introduction

In linguistics, motifs are understood as continuous sequences of values of some language units. In this paper we focus exclusively on word length motifs, i.e., on sequences of words with non-decreasing lengths, with word length measured in the number of syllables (for on overview of linguistic motifs in general cf. Köhler 2015).

Word length motifs (henceforward denoted only as motifs) were introduced by Köhler (2006), who was inspired by a study of a sequential structure of note duration in music by Boroda (1982). Motifs were later investigated from a theoretical point of view by Köhler (2008a, 2008b), Mačutek (2009, 2015), Sanada (2010), Mačutek & Mikros (2015), and Milička (2015). Hints at possible applications of motif properties (mainly) to an automatic text classification can be found in Köhler & Naumann (2008, 2010), and also in some of the abovementioned papers.

This paper, using data material which consists of 1000 blogs written in Modern Greek (cf. Section 2 for data description), tries to provide some insight into two related areas. First, we follow the line of research presented by Milička

1 University of Athens; gmikros@isll.uoa.gr
2 Comenius University; jmacutek@yahoo.com

(2015). In that paper, a question is asked whether the word length distribution observed in a text determines motif properties (or, as the author says, whether motif properties are inherited from the word length distribution). It was shown that motif frequencies in real texts differ from those in randomized texts (in which the word length distribution is preserved, but the order in which particular lengths occur is random). This finding indicates that motif properties carry some information which is not included in the word length distribution. On the other hand, type-token relations for motifs (cf. Mačutek 2015) in real and randomized texts are basically the same (statistically speaking, they are not significantly different, cf. Section 3).

Second, we concentrate on several interrelations among motif properties, namely the type-token relation, text length (measured in the number of words), and the number of motifs observed in the text. It seems that text length plays a very important role (cf. Section 4).

Computations were performed using the software environments R (www.r-project.org) and NLREG (www.nlreg.com).

2 Data description

During last decade Internet evolved from a static field of simple information provision into a digital carrier of language production characterized by interactivity and dynamic configuration of the online textual content.

Blogs are among the most known web tools that have transformed web communication and overcame the unidirectionality of standard online communication. Until 2011 worldwide have been created approximately 181 million blogs producing 900,000 posts every day which are being read by 77% of internet users[1]. Since many blogs are important information nodes and attract many more readers than most of the traditional printed media, they can exert influence in language usage and produce linguistic innovations accelerating linguistic change. For this reason, blog's language usage has started to attract attention and become a challenging research subject in the linguistic community.

Blogs represent a new text genre with interesting characteristics. They combine personal views, news and reporting on current events (Mishne, 2007). Their structure is a hybrid containing both monologue and dialogue features. At the same time, they are both log entries reflecting personal opinions and open

1 Source: NM Incite.

calls for public discussion (Nilsson, 2003). Mishne (2007, p. 34) studied in detail various properties of linguistic usage in English blogs and showed that they present increased usage of personal pronouns and words relating to personal surroundings emerging personal experience. Furthermore, he examined the linguistic complexity of the blogs using the perplexity measure (Brown et al., 1992) and the out-of-vocabulary rate (OOV) and found that their linguistic structure was more complex than most of the similar written genres (e.g. personal correspondence). Increased perplexity, according to Mishne (2007), equates with increased irregularity in linguistic usage (i.e. free-form sentences, decreased compliance with grammatical rules etc.). In addition, blogs presented increased OOV rates, meaning that blog texts exhibit a topical diffused vocabulary, with many neologisms, possible typographical errors and increased level of references to named entities from the blogger's personal environment.

Another interesting characteristic of the blog's linguistic structure is its equilibrium between spoken and written language. Sentence construction in blogs is highly variable using selectively structures from both spoken and written norms (Chafe & Danielewicz, 1987). Equally important effect in language usage in blogs is the age of the bloggers. Half of the them are aged 18-34[2] . For this reason, formality in language usage is decreased, with shorter that average sentence lengths and lower readability scores in the most famous readability formulas (like, e.g., Gunning-Fog, Flesch-Kincaid, SMOG; cf. Bailin & Grafstein 2016).

Since there was not any social media corpus available for Modern Greek, we decided to compile ours from scratch. For this reason, we harvested the Greek blogosphere from September 2010 till August 2011 and manually collected 100 Greek blogs, their authors being equally divided to 50 male and 50 female bloggers. Since topics can induce significant bias into stylometric measurements (Mikros & Argiri, 2007), we decided to explore only a part of the collected corpus, using blogs that share the same topic. In this study we used 20 blogs (10 male and 10 female authors) with the common topic (personal affairs), making the total of 1,000 blog posts counting 406,460 words. For each author we collected 50 most recent blog posts. The posts contain from 24 (the minimum) to 3652 (the maximum) words, the mean text length being 406.46 words.

All measurements related to syllable length have been performed using a customized PERL script we developed for this research. Syllable counting in Modern Greek is a complex task since its orthography is considered deep, with

2 Source: The Social Media Report, Q3, 2011, NM Incite, Nielsen.

complex correspondences between graphemes and phonemes. Moreover, a large number of inconsistent phoneme-grapheme correspondences have been created due to the diglossia existent in the Greek language for over a century.

Our syllable counter works using a two-stage algorithm. In the first pass, the software performs a broad phonemic transcription of the written text using the basic grapheme to phoneme rules for Modern Greek. In the second stage, there is a hand-made lexicon with phonetic transcriptions of the 20,000 most frequent words of Modern Greek. The software checks the phonetic transcriptions produced in the first stage and corrects them if it finds a difference between its list and the relative phonetic transcription produced by the rules. A series of reliability tests of our script showed that its final phonemic transcriptions have a 95-98% accuracy, a performance that guarantees that our motif length measurements are reliable as well.

3 Type-token relation for motifs revisited

The type-token relation is, when motifs are considered (cf. Wimmer 2005 for a comprehensive paper on type-token relations in linguistics), the relation between the number of all motifs (tokens) and different motifs (types) observed in a text. Mačutek (2015) analyzed 70 Ukrainian texts of seven different genres and showed that the development of the relation can be modelled by the function

$$f(x) = x^b, \tag{1}$$

with $f(x)$ being the number of different motifs among the first x motifs; b is a parameter.

Our results confirm that model (1) is appropriate also for the motif type-token relation in Modern Greek texts, which, given that motifs behave similarly to their basic units, i.e., words, and that the type-token relation for words is one of established language laws, is not surprising.

However, a much more interesting finding can be reported. Milička (2015) investigated motif frequencies in both real and randomized Czech and Arabic texts. He concluded that there were statistically significant differences in motif frequencies in these two types of texts. We randomized 100 times each of the first ten of the blog posts from our corpus. The process of randomization was performed in two different ways. First, sequences of word lengths were permuted and each permutation was considered to correspond to a new, randomized text, i.e., word length frequencies were preserved, with only the

order of the words changed. Second, we generated random numbers from the distributions of word length in the texts, i.e., the type of the word length distribution was preserved, with word length frequencies fluctuating randomly. Then motifs were constructed in the randomized texts. Both approaches to the randomization lead to word length sequences for which not only model (1) described the type-token relation for motifs sufficiently well, but they also yielded practically the same values of the parameter b (statistical tests did not reveal significant differences).

One could object that 100 randomization are not reliable enough; however, the values of the parameter b seem to be very stable, and, in addition, one is likely to reject (almost) any statistical hypothesis if the sample size is very large (cf. Mačutek & Wimmer 2013). The same result was obtained for a long Modern Greek text (novel *The mother of the dog* by Matesis, containing 47,852 words, cf. Mačutek & Mikros 2015); here, because of time consuming computations, randomizations were performed only ten times.

Hence, it seems that the parameter of the model for the type-token relation is influenced more by language and by text length (cf. Section 4) than by the author or by the genre (attempts to apply the parameter values to an automatic authorship attribution did not bring convincing results).

It is to be noted that not all motif properties in real texts coincide with those in randomized texts. In addition to already mentioned motif frequencies (Milička 2015), also the parameters of the Menzerath-Altmann law in such texts differ (Mačutek & Mikros 2015). Moreover, even in the cases where properties of real and randomized texts are the same, one cannot exclude the possibility that an apparent randomness is in fact a result of an interference of several competing language laws (which can „cancel“ each other).

4 Interrelations among some motif and text properties

Within the framework of quantitative linguistics, language units and their properties are not considered to be isolated entities, but, on the contrary, it is supposed that they influence each other. One of the aims of linguistic research is to build a network of established (i.e., mathematically modelled and statistically tested) interrelations between its particular elements (i.e., units and their properties). Such an approach to building a linguistic theory is known as synergetic linguistics (cf. Köhler 2005).

One of the first attempts to present interrelations among properties of linguistic motifs was presented by Mačutek (2015), based on an analysis of Ukrainian texts. We confirm these results on new data material[3] from another language.

Analogously to words, also for motifs it holds that the value of the parameter b in model (1) tends to decrease with the increasing text length (see Fig. 1). This behaviour is easy to explain – the maximum of word length observed in a text is usually relatively low and long words occur but seldom, hence motifs containing long words appear only rarely, and those in which there are several long words are possible, but the probability of their occurrence is close to zero. It means that new types of motifs appear, once the usual ones (consisting of relatively short words) are already observed, more and more seldom as the text gets longer. At the present we do not try to derive a mathematical model for this relation. The trend is quite obvious, but the values of b exhibit a relatively high variability. Therefore we suppose that text length is not the only important factor, and a more complicated mathematical model with more than one explanatory variable (e.g., also some parameters of the word length distribution) will probably be needed to fit the data sufficiently well.

3 Numerical values of properties discussed in this section can be sent upon request. Given the size of the data material (1000 texts), it is not possible to present them in this paper.

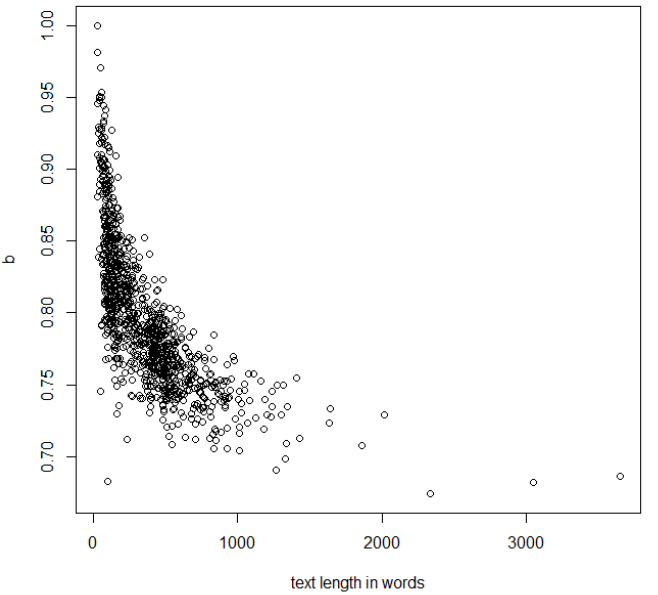

Fig. 1: Relation between text length in words and value of parameter b in type-token relation (Modern Greek blogs).

According to Mačutek (2015), there is also a relation between the number of all motifs in a text and the text length. The number of all motifs, seemingly, increases linearly with the increasing text length. He also formulated a hypothesis that, if the relation really is linear, the slope of the linear function would be inversely proportional to the mean word length in the text. The linear model without an intercept,

$$f(x) = ax, \tag{2}$$

gives a very good fit also for Modern Greek blogs, see Fig. 2. This model not only fits data very well, but it also has two additional advantages.

First, it is a special case of a very general model suggested by Wimmer & Altmann (2005), which is currently considered as one common mathematical expression of (nearly) all language laws. Admittedly, linguists prefer power laws (Naranan & Balasubrahmanyan 2005) over linear functions, but, again, a linear model is a special case of a power law for certain parameter values

(which is the case of our data – fitting the function $f(x) = ax^b$ yields values of b very close to 1).

Second, the only parameter of model (2) is directly interpretable in terms of word length. Denote MWL the mean word length (measured in the number of syllables) in the corpus and $NM(x)$ the number of motifs in a text consisting of x words. The fit (in terms of the determination coefficient R^2, cf. Mačutek & Wimmer 2013) remains very good if the parameter a is estimated not by the classical least square method, but as the inverse proportion of the MWL. We thus obtain the model

$$NM(x) = \frac{1}{MWL}x, \tag{3}$$

with x being the number of words in a text. The line in Fig. 2 represents model (3). For the corpus of Modern Greek blogs we have $MWL = 2.306$, which leads to the model $NM(x) = 0.434x$ with $R^2 = 0.996$.

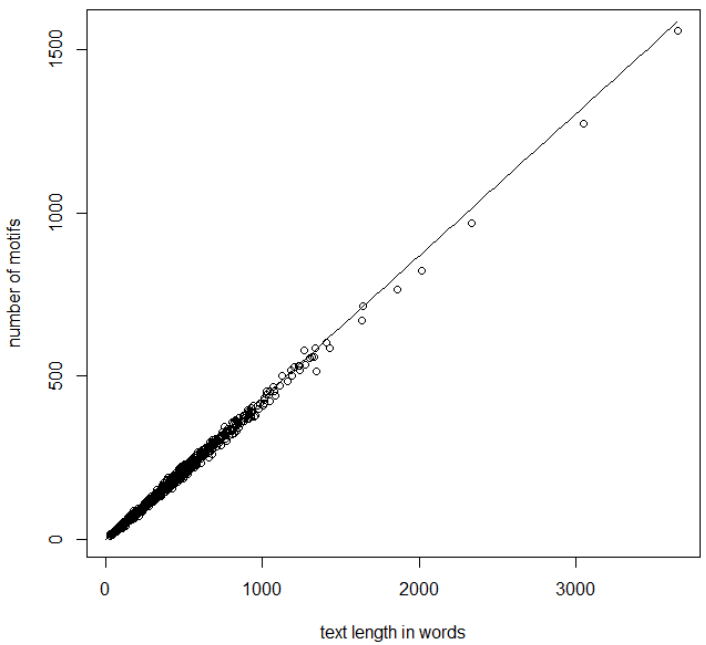

Fig. 2: Relation between text length in words and number of all motifs in text (Modern Greek blogs).

The same behaviour - the linear relation between the number of motifs in a text and the text length, with the inverse proportion of the *MWL* from the corpus substituted as the value of parameter a - can be observed also in 70 Ukrainian texts of seven different genres (cf. Mačutek 2015). The *MWL* in the Ukrainian corpus is 2.660 and after inserting the value into (3) we have $NM(x) = 0.376x$, with $R^2 = 0.9765$. The data and the line representing model (3) can be seen in Fig. 3.

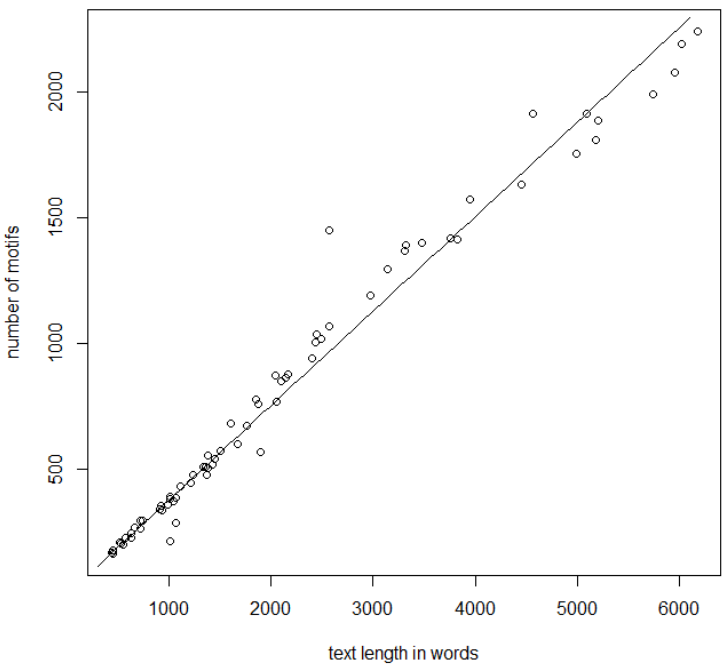

Fig. 3: Relation between text length in words and number of all motifs in text (70 Ukrainian texts, cf. Mačutek 2015).

For the sake of simplicity, the *MWL* was computed from the corpora of Modern Greek and Ukrainian texts, respectively. Similar results are obtained if one computes the MWL for each text separately, i.e., if the value of the parameter in model (3) is specific for each text. Hereby wo do not claim that the *MWL* does not differ among genres or authors. The good fit for the *MWL* value from a corpus can be caused by a relatively homogenous language material.

This approach, however, is not directly applicable to five very long Modern Greek texts (cf. data in Mačutek & Mikros 2015, text lengths range from 9,761 to 77,692). The linear function (2) fit data excellently, but not if we take the $\frac{1}{MWL} = 0.482$ from the corpus of the five novels as the parameter. On the other hand, the model

$$NM(x) = 0.482x^b \tag{4}$$

with $b = 0.985$ achieves $R^2 = 0.9988$.

With respect to this different pattern of behaviour of long texts we remind that word length seems to decrease with the increasing text length (cf. Kelih 2012 for an analysis of a Russian novel and its translation to Bulgarian). This aspect has not been sufficiently investigated yet, and it deserves a closer attention. The linear model may not be sufficient if a text is long, but it seems that even in such a case the inverse proportion remains an important player. Anyway, the question is whether we can consider a long novel to be one text, one whole, or whether it is more appropriate to see in it a collection of several shorter texts (e.g., chapters).

Mačutek (2015) speculated over the nature (linear vs. non-linear) of the relation between the number of all motifs and the number of different motifs in texts. We can conclude that the relation is non-linear (Fig. 4), and that the function $f(x) = ax^b$ yields a good fit ($a = 3.398$, $b = 0.52$, $R^2 = 0.9337$). We postpone attempts to interpret the parameters until data from more languages are available.

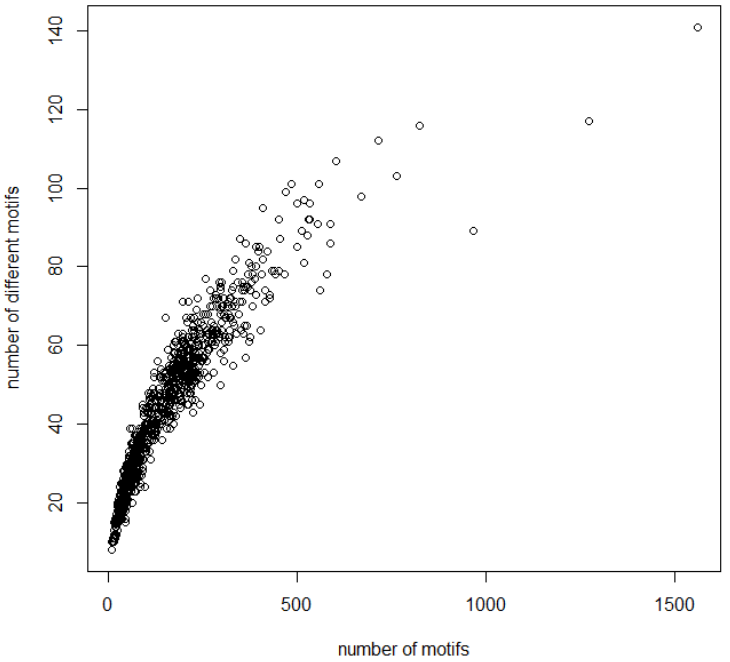

Fig. 4: Relation between number of all motifs and number of different motifs (Modern Greek blogs).

5 Conclusion

The type-token relation for word length motifs behaves analogously to that for words (its parameter decreases with the increasing text length), confirming thus once more the expectations that motifs display properties similar to the units of which they consist (words here). The parameters of the type-token relations for motifs in real and randomized text attain the same value for texts analyzed in this paper.

The number of motifs is determined by the mean word length (namely, by its inverse proportion) in relatively short texts. In long novels, the suggested mathematical model has two parameters, the inverse proportion of the mean word length being one of them.

Based on analyses of Modern Greek and Ukrainian texts we allow ourselves to state (at least tentatively) that the word length distribution is a crucial factor influencing the type-token relation for word length motifs. If the distribution is fixed, even randomized texts do not differ from real ones as far as the type-token relation is concerned. The text length, together with the mean word length in the text (which is given by the word length distribution), determine the number of all motifs which occur in the text.

Acknowledgement

J. Mačutek was supported the grant VEGA 2/0047/15.

References

Bailin, A., & Grafstein, A. (2016). *Readability: Text and Context*. London: Palgrave Macmillan.

Boroda, M.G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: J.K. Orlov, M.G. Boroda & I.Š. Nadarejšvili (Eds.), *Sprache, Text, Kunst. Quantitative Analysen* (pp. 231-262). Bochum: Brockmeyer.

Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J., & Lai, J.C. (1992). Class-based *n*-gram models of natural language. *Computational Linguistics, 18*(4), 467-479.

Chafe, W., & Danielewicz, J. (1987). Properties of spoken and written language. In R. Horowitz & J. S. Samuels (Eds.), *Comprehending Oral and Written Language* (pp. 83-113). New York: Academic Press.

Kelih, E. (2012). On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In: S. Naumann, P. Grzybek, R. Vulanović & G. Altmann (Eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems* (pp. 67-80). Wien: Praesens.

Köhler, R. (2005). Synergetic linguistics. In: R. Köhler, G. Altmann & R.G. Piotrowski (Eds.), *Quantitative Linguistics. An International Handbook* (pp. 760-774). Berlin, New York: de Gruyter.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: J. Genzor & M. Bucková (Eds.), *Favete Linguis. Studies in Honour of Viktor Krupa* (pp. 145-152). Bratislava: Slovak Academic Press.

Köhler, R. (2008a). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory, 1*(1), 115-119.

Köhler, R. (2008b). Word length in text. A study in the syntagmatic dimension. In S. Mislovičová (Ed.), *Jazyk a jazykoveda v pohybe* (pp. 416-421). Bratislava: Veda.

Köhler, R. (2015). Linguistic motifs. In: G.K. Mikros & J. Mačutek (Eds.), *Sequences in Language and Text* (pp. 89-108). Berlin, Boston: de Gruyter.

Köhler, R., & Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 635-646). Berlin, Heidelberg: Springer.

Köhler, R., & Naumann, S. (2010). A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In: P. Grzybek, E. Kelih & J. Mačutek (Eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives* (pp. 81-89). Wien: Praesens.

Mačutek, J. (2009). Motif richness. In: Köhler, R. (Ed.), *Issues in Quantitative Linguistics* (pp. 51-60). Lüdenscheid: RAM-Verlag.

Mačutek, J. (2015). Type-token relation for word length motifs in Ukrainian texts. In: A. Tuzzi, M. Benešová & J. Mačutek (Eds.), *Recent Contributions to Quantitative Linguistics* (pp. 63-73). Berlin, Boston: de Gruyter.

Mačutek, J., & Mikros, G.K. (2015). Menzerath-Altmann law for word length motifs.In: G.K. Mikros & J. Mačutek (Eds.), *Sequences in Language and Text* (pp. 125-131). Berlin, Boston: de Gruyter.

Mačutek, J., & Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics, 20*(3), 227-240.

Milička, J. (2015). Is the distribution of L-motifs inherited from the word length distribution? . In: G.K. Mikros & J. Mačutek (Eds.), *Sequences in Language and Text* (pp. 133-145). Berlin, Boston: de Gruyter.

Mikros, G.K., & Argiri, E.K. (2007). Investigating topic influence in authorship attribution. In: B. Stein, M. Koppel & E. Stamatatos (Eds.), *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (vol. 276, pp. 29-35). Amsterdam: CEUR.

Mishne, G. (2007). *Applied Text Analytics for Blogs*. Amsterdam: University of Amsterdam.

Naranan, S., & Balasubrahmanyan, V.K. (2005). Power laws in statistical linguistics and related systems. In: R. Köhler, G. Altmann & R.G. Piotrowski (Eds.), *Quantitative Linguistics. An International Handbook* (pp. 716-738). Berlin, New York: de Gruyter.

Nilsson, S. (2003). The Function of Language to Facilitate and Maintain Social Networks in Research Weblogs. Umeå: Umeå Universitet.

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: P. Grzybek, E. Kelih & J. Mačutek (Eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives* (pp. 181-193). Wien: Praesens.

Wimmer, G. (2005). The type-token relation. In: R. Köhler, G. Altmann & R.G. Piotrowski (Eds.), *Quantitative Linguistics. An International Handbook* (pp. 361-368). Berlin, New York: de Gruyter.

Wimmer, G., & Altmann, G. (2005). Unified derivation of some linguistic laws. In: R. Köhler, G. Altmann & R.G. Piotrowski (Eds.), *Quantitative Linguistics. An International Handbook* (pp. 791-807). Berlin, New York: de Gruyter.