In Gabriel Altmann, Radek Čech, Ján Mačutek, Ludmila Uhlířová (eds.) Empirical Approaches to Text and Language Analysis. RAM-Verlag 2014 Lüdenscheid, ISBN: 978-3-942303-24-8, pp. 181-189.

Distribution of the Menzerath's law on the syllable level in Greek texts

Georgios Mikros, Jiří Milička

1. Introduction

Since 1980, when Gabriel Altmann published his famous article on the Menzerath's Law (Altmann 1980), the law has been corroborated on many linguistic levels and even non-linguistic material inspiring generations of linguists. Both Menzerath's Law and Altmann's equation assume that the relation between construct length and the constituent length is a monotonic decreasing function like the function depicted in Fig 1.

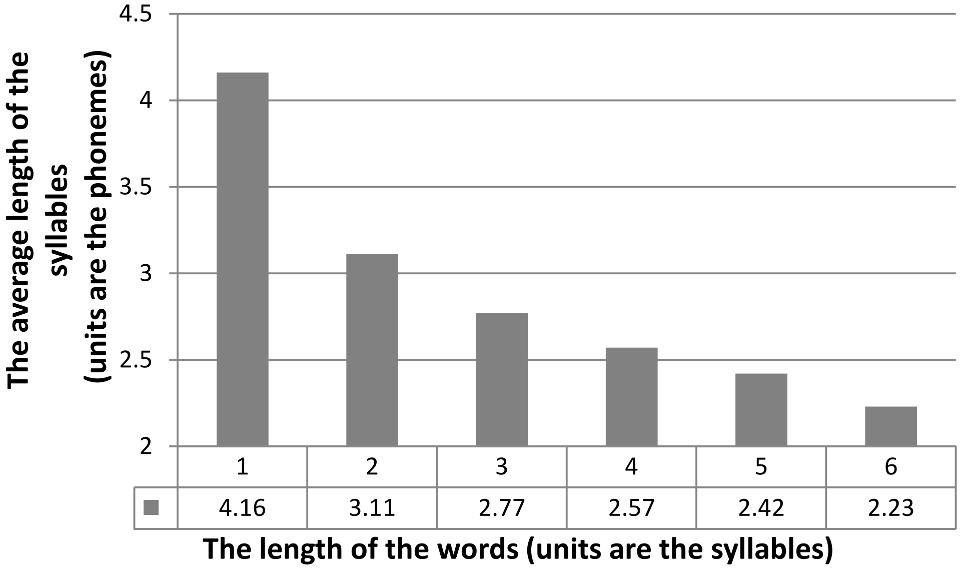

Fig 1. Menzerath's Law on the word-syllable-phoneme level in a German text (Source: Altmann et al. 1989, Table 5.1a, p. 38).

Counterexamples to this law can be found, for example in the relation between the length of a Greek word and its corresponding syllables in this text (Fig 2). But as the single counterexample could be only a random fluctuation from a trend (especially here, the text contains only 4 word tokens with 10 syllables), the following hypothesis needs to be tested:

H1: *On average, the function measured on a sample of texts is monotonly decreasing.*

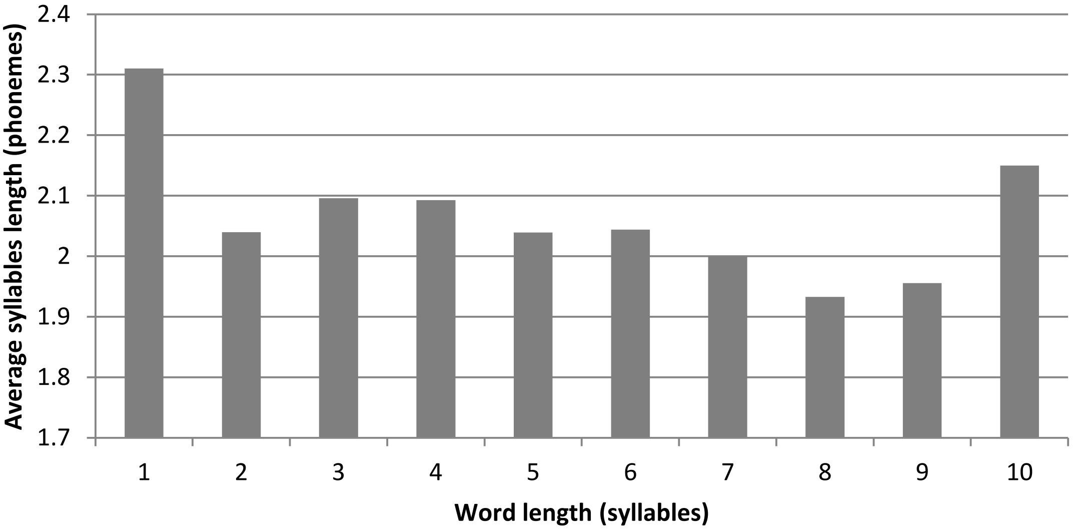

Figure 2. Menzerath's Law on the word-syllable-phoneme level in a Greek text[1].

2. Material and methods

Modern Greek is one of the least quantitatively studied modern European languages. Zipf's law and some basic quantitative characteristics in various linguistic elements have been measured in Hatzigeorgiu, Mikros & Carayannis, (2001) and Mikros, Hatzigeorgiu & Carayannis (2005). However, we don't have any published research focusing on Menzerath's law and its application to Modern Greek data[2]. This is the focus of the present research and in the following sections, we will describe the corpus, tools and methods used in this study.

2.1 The HNC corpus

The corpus used is the Hellenic National Corpus (HNC). HNC was developed by the Institute for Language and Speech Processing (ILSP) (Hatzigeorgiu et al. 2000) and is an ongoing effort. It currently contains more than 50,000 written MG texts, published from 1976 on, totaling 47 million words.

Texts in HNC are classified according to PAROLE standards (PAROLE, 1995), which follow the TEI (Sperberg-McQueen & Burnard, 1994) and

[1] The specific text is a scientific monograph belonging to the broader discipline of Epistemology (Title: "The peel of apricot", 2000, Ed. Ellinika Grammata, pp. 526).

[2] The only relevant research that we are aware of is the unpublished MA thesis of Zenetzi & Papachristos (2013). However, in this study Menzerath-Altmann's law has been studied in each text of a Modern Greek corpus separately using a very different methodology from the methodology that most quantitative linguistic studies use. A regression model was applied to the corpus data (mean word length in syllables per text and mean syllable length in letters per text) yielding a fit which confirmed a weak correlation of the investigated relationship.

EAGLES (EAGLES, 1994) guidelines. Texts are classified as regards to Medium, Genre, Topic, Detailed Genre, Detailed Topic and bibliographical information. As far as Medium is concerned, texts are classified into four categories, according to their source. The current percentage of words for each one of the four categories can be seen in Table 1.

Table1
Percentage of words according to text medium

Text Medium	Percentage
Newspaper	61.29%
Miscellaneous	23.08%
Book	9.41%
Magazine	5.89%
Internet	0.32%

We used a "clear" version of HNC texts, with all their metadata removed from the text and stored in an excel file with pointers to the related text files.

3. Tools

In order to test the application of the Menzerath law on Modern Greek (MG) we had to develop a number of tools that would permit the computational processing of the HNC. The first of these tools is the MGphontranscriber, a specialized PERL script for converting texts written using standard MG spelling to broad phonemic transcriptions. The script is based on 99 letter to phoneme rules implemented using appropriate regular expressions. The regular expressions apply sequentially to the text input and are ordered for producing the appropriate MG phonemic representations. The output of the tool was validated against human MG phonemic transcriptions and after some fine tuning produced 100% correct outputs. In the last preprocessing step, the tool tokenized the input removing all punctuation marks and putting each token in a separate line (vertical text representation).

The output of MGphontranscriber was used as input to a separate PERL script that calculated the basic variables of Menzerath's law, i.e. the length of words (measured in phonemes) and their number of syllables (measured as the number of vowels in each word).

The resulting data have been processed by Menzerath.exe software[3] package that was designed to measure Menzerath's Law for many kinds of pre-processed data. The bootstrap resampling has been done by the Bootstrapper software[4].

[3] Available at http://milicka.cz/en/menzerath

[4] Available at http://milicka.cz/en/bootstrapper

A simple bootstrap method was used i.e. taking a sample of *N* values from the original data and resampling from them to form a new sample that is also of size *N*. The bootstrap sample is taken from the original one using sampling with replacement so it is not identical with the original one. This process is repeated in our case 100,000 to 1,000,000 times, and for each of these bootstrap samples its mean is computed. Confidence intervals for the sample mean are estimated from the histogram of the bootstrap means.

4. Results

The Menzerath’s law for the phoneme-syllable-word level has been measured on a sample of 45,691 Greek texts. The distribution of the results is depicted in Fig. 3.

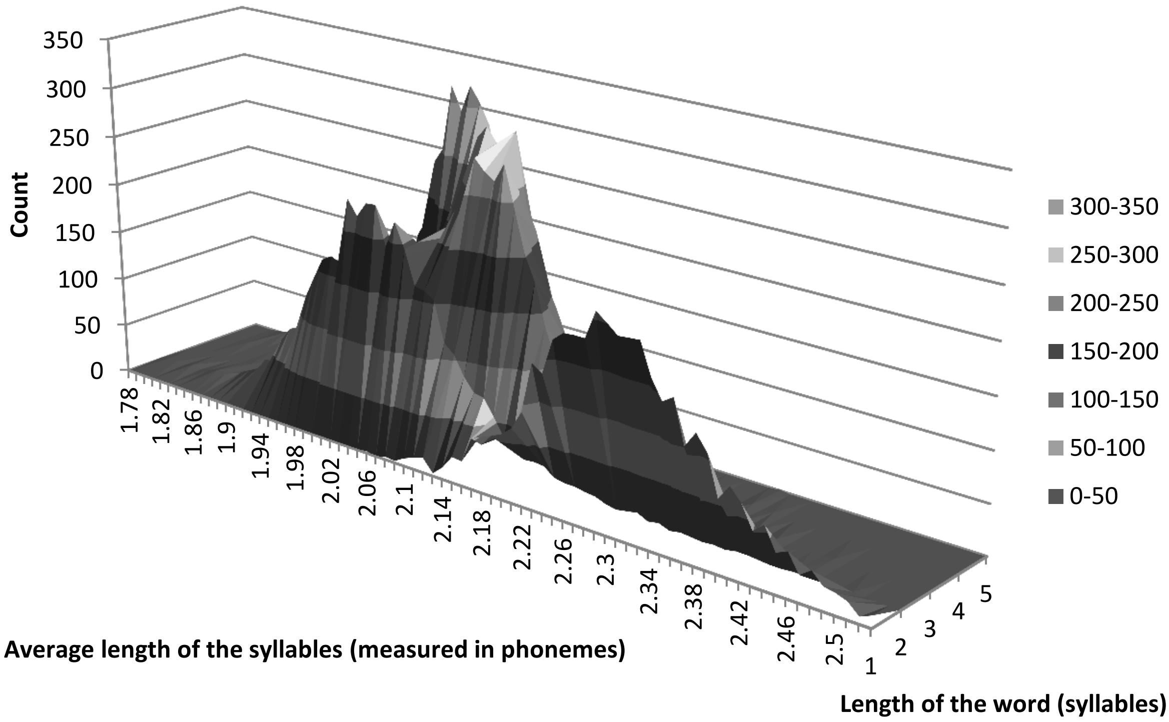

Figure 3. The distribution of results for words that contains 1–5 syllables.

From the distribution we bootstrapped confidence intervals[5] for the average values. Fig. 4. displays the resulting distribution produced by 100,000 bootstrap samples.

[5] For an introduction to the method see Efron (1979).

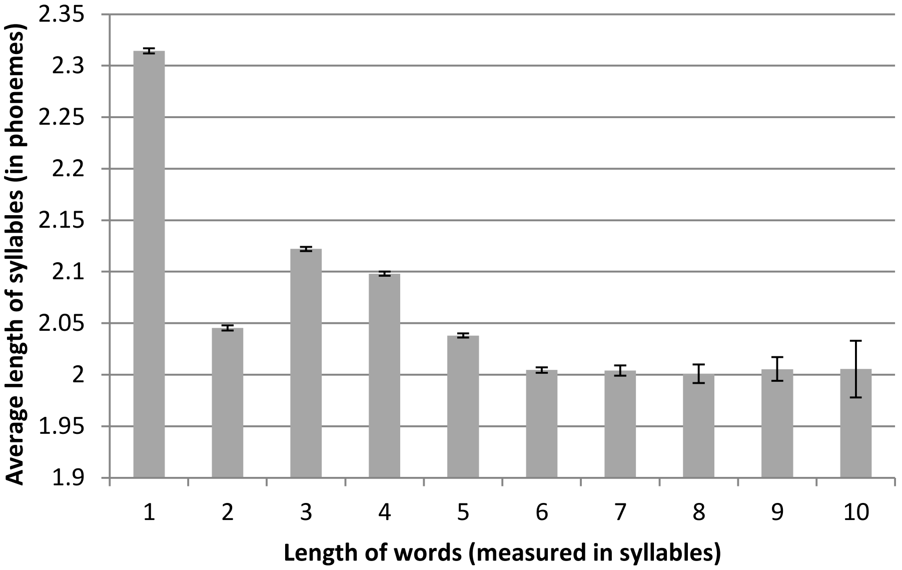

Figure 4. The Menzerath's law on the sample of texts. The error bars stands for the 95% confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11 – 20) were omitted.

Fig. 4 shows a break in the Menzerath-Altman (MA) law in the area of the 2-syllable words. The average length of syllables in words with 2 syllables is shorter than the average length of syllables in words with 1, 3 and 4 syllables making the function non-monotonic.

In order to investigate this phenomenon in detail on the single-text level we selected randomly 2 texts (#07439 & #35855) from the corpus and produced their word length histograms displayed below (Fig. 5). In Fig. 5 we see that the 2-syllable words exhibit a left-skewed distribution lowering their mean length and consequently producing bad fits to the MA law.

Another way to investigate further the differences between the various word lengths is to examine their relative ratios. More specifically, for each text, the ratios between values of neighbouring columns were calculated, i.e. average length of the syllables in words with n-syllables were divided by the average length of the syllables in words with n+1 syllables for each text. The distribution of the ratios is displayed in Fig 6.

Again, from the distribution we bootstrapped confidence intervals (confidence level 95%) for the average values, see Fig 7. (the confidence intervals are marked by the error bars). The number of bootstrap samples was 100,000. This measurement brings results equivalent to the results that are depicted in Fig. 4.

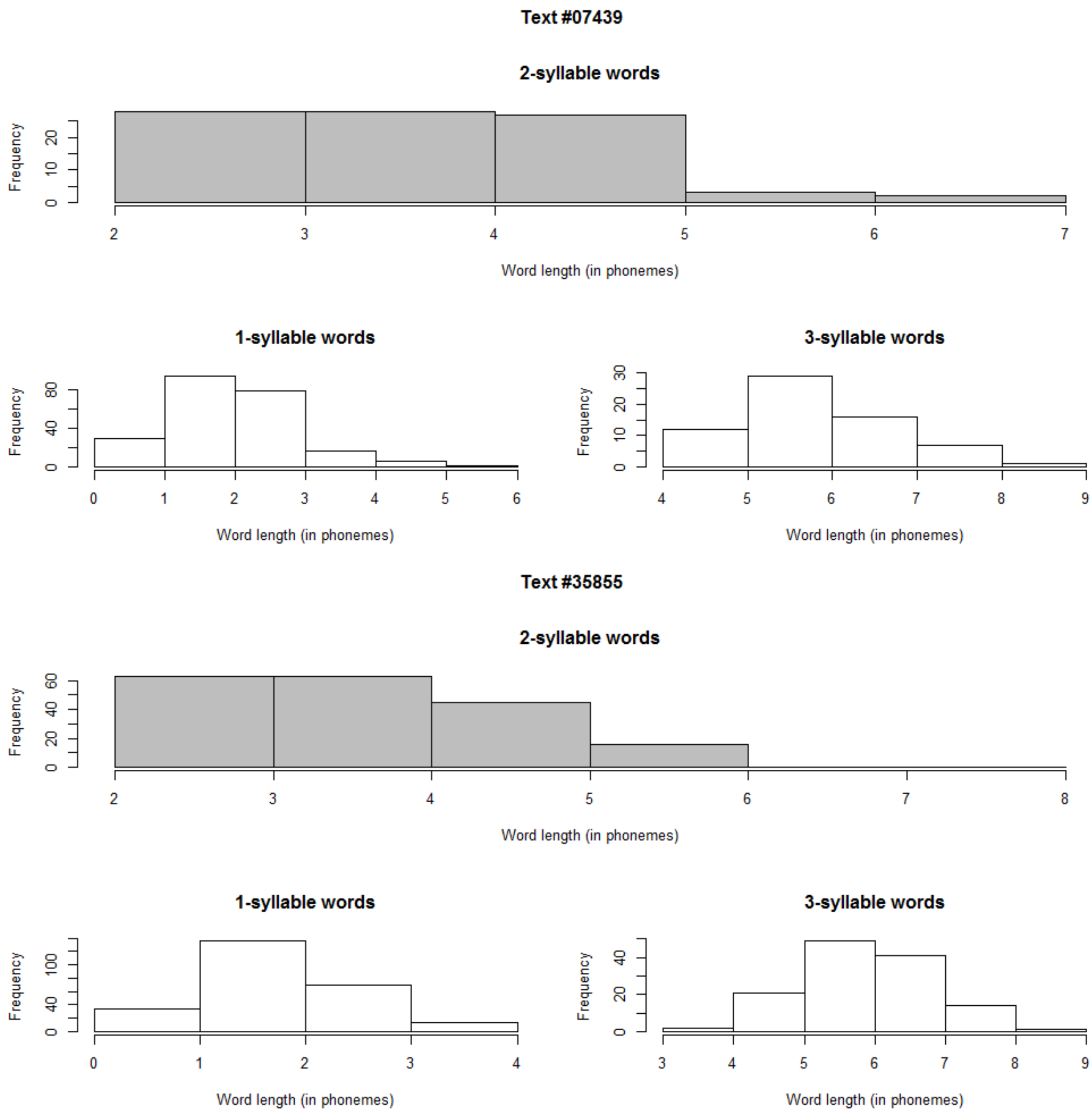

Figure 5. Distribution of word length (measured in phonemes) in 1, 2 and 3-syllable words in two random selected texts (#07439 & #35855).

On average, short words are the most frequent ones (for references to this phenomena see Straus, Altmann 2006). Simple theoretical consideration leads us to the suspicion that the abnormality observed in the short words can be caused or accompanied by some non-trivial abnormality in the most frequent words. Let us measure the Menzerath's law for the sets of types (or dictionaries) instead of the real texts. This should neutralize the impact of the most frequent words (which are somewhat special because the set of the most frequent words contains synsemantic words and many proper names).

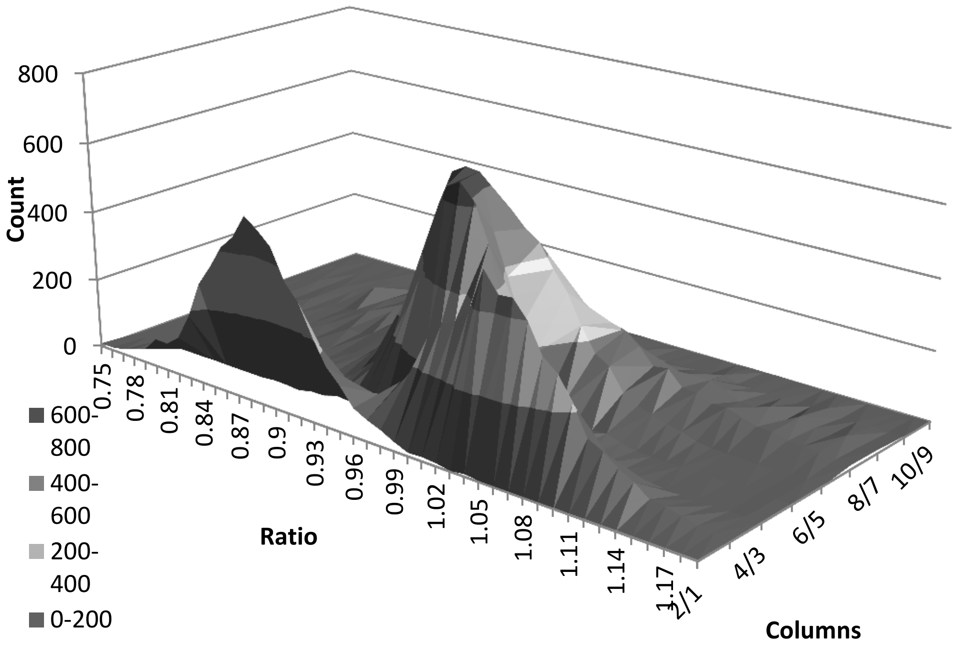

Figure 6. The distribution of ratios between neighbouring columns for columns 1–6.

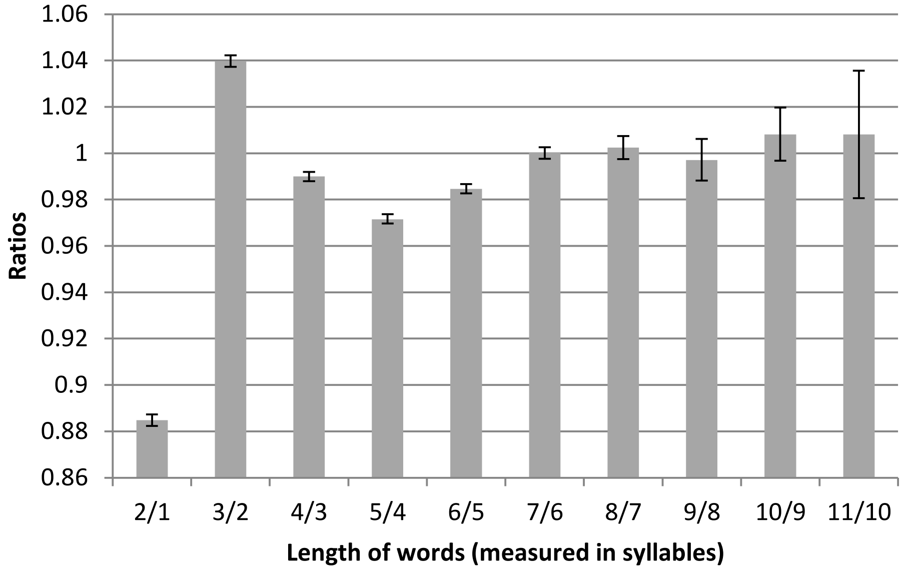

Figure 7. The Menzerath's law on the sample of texts. The error bars stands for the 95 % confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11 – 20 syllables) were omitted.

The Fig. 8 compares the values from Fig. 4 that were measured on word tokens with the values that were measured on the lists of types (for each text, a list of word types was extracted).

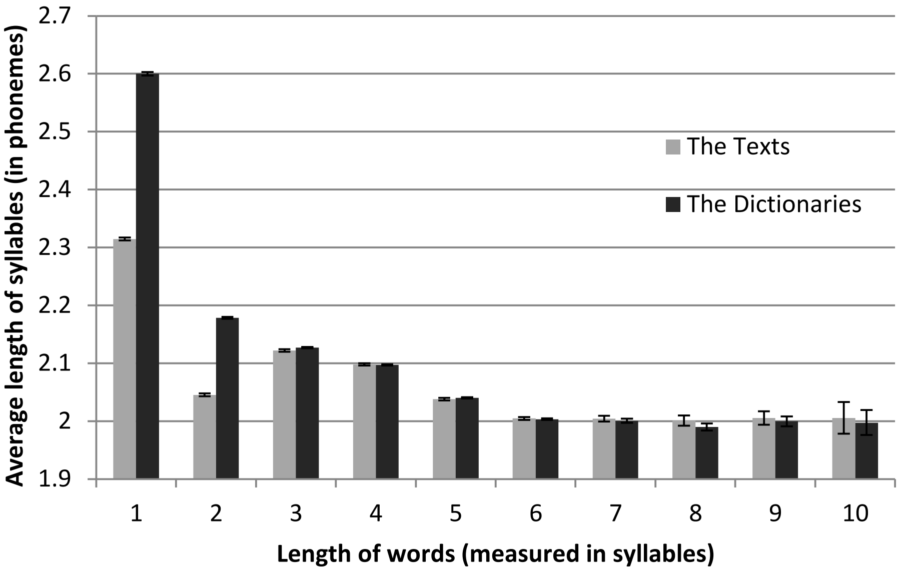

Figure 8. The Menzerath's law measured on the sample of texts and on the sample of lists of types (dictionaries) that were extracted from these texts. The error bars stands for the 95 % confidence interval that was bootstrapped from 100,000 samples. The lengths of words that were contained in less than 100 texts (11 – 20 syllables) were omitted.

As we can see, the Menzerath's law on the phoneme-syllable-word level measured on the list of types that were extracted from the Greek texts is in accordance with the theory (monotonically decreasing). At the same time we can observe that the values for mono-, di- and tri- syllabic words that were measured on the lists of types are significantly different from the values that were measured on the word tokens (original texts).

5. Conclusion

Examining a large corpus of Greek texts we found that the average length of syllables in the disyllabic words is lower than the average length of the syllable in monosyllabic words and lower than the average length of syllables in tri-syllabic words. This peculiar phenomenon seems to be ubiquitous in Greek texts and our future research is directed to a wider explanatory framework including both the phonological properties of the Greek syllables contained in two-syllable words, their frequency distribution and possible historical explanations.

Word frequency seems to be a factor that can partially explain the deviant behaviour of the two-syllable words. Data that were measured on lists of types (extracted from each text separately) are significantly different from the data that were measured on the raw texts, and the type data follow Menzerath's law.

We are aware that some well-taken objections can be raised, e.g.:

1. The sample was not really randomly chosen from the population of Greek texts (even if the population is limited for the texts written in some time period).
2. The validity of the Menzerath's Law on the phoneme-syllable-word level is sometimes considered to be a side effect of the validity of the law on the phoneme-morpheme-word level and thus less interesting than the latter one.

This is the first attempt to study systematically Menzerath's law in a large Modern Greek corpus. Despite the limitations of this study outlined above, we hope that this study will to draw attention to this problem and to foster the discussion about the proposed methodology.

Acknowledgments

This work was supported by the project "Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology", grant no. CZ.1.07/2.3.00/20.0161 which is financed by the European Social Fund and the National Budget of the Czech Republic.

References

Altmann, G. (1980). Prolegomena to Menzerath's law. In: Grotjahn, R. (ed.), *Glottometrika 2*, Bochum: Brockmeyer, 1–10.

Altmann, G., Schwibbe, H. (eds.) (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Georg Olms Verlag.

Andres, J. (2010). On a Conjecture about the fractal structure of language. *Journal of Quantitative Linguistics 17*, 101–122.

EAGLES (1994). Corpus encoding: Draft. Technical report, EAGLES. Document EAG-CSG/IR-T21.

Geršič, S., Altmann. G. (1980). Laut-Silbe-Wort und das Menzerathsche Gesetz. Frankfurter Phonetische Beitrage 3, 115-128.

Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A., Labropoulou, P., Mantzari E., Papageorgiou, H., Demiros, I. (2000). Design and implementation of the online ILSP Greek Corpus. In: Gavrilidou, M. et al. (eds.), *Proceedings of the LREC 2000 Conference*. Athens, 1737-1742.

Hatzigeorgiu, N., Mikros, G., Carayannis, G. (2001). Word length, word frequencies, and Zipf's law in the Greek language. *Journal of Quantitative Linguistics 8*, 175–185.

Hřebíček, L. (1990). The constants of Menzerath-Altmann's Law. In: Hammerl, R. (ed.), *Glottometrika 12*, Bochum: Brockmeyer, 61–71.

Hřebíček, L. (1992). *Text in comunication: Supra-sentence structures*. Bochum: Brockmeyer.

Hřebíček, L. (1994). Fractals in language. *Journal of Quantitative Linguistics 1*. 82–86.

Hřebíček, L. (1995). *Text Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Trier: Wissenschaftlicher Verlag .

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Ann. Statist. 71–26.

Menzerath, P. (1928). Über einige phonetische probleme. In: *Actes du premier Congres International de Linguistes*. Leiden: Sijthoff

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümler.

Mikros, G., Hatzigeorgiu, N., Carayannis, G. (2005). Basic quantitative characteristics of the Modern Greek language using the Hellenic National Corpus. *Journal of Quantitative Linguistics 12*. 167–184.

PAROLE (1995). *Design and composition of reusable har-monised written language reference corpora for euro-pean languages. Technical report*. PAROLE Consortium. MLAP, WP 4 - Task 1.1, 63–386.

Sperberg-McQueen C. M., Burnard, L. (1994). *Guide-lines for electronic text encoding and interchange: Tei-p3. Technical report*. Chicago and Oxford: ACH-ACL-ALLC Text Coding Initiative.

Strauss, U., Altmann, G. Word Length and Frequency. In: *Laws in Quantitative Linguistics*. [Available at http://lql.uni-trier.de/index.php/Word_length_and_frequency] [cit. 2013/11/25].

Zenetzi, K. Papachristos, D. (2013). *Mathematical – empirical methods in Greek language. A study of the Menzerath-Altmann Law*. Unpublished MSc. Thesis, Postgraduate Program "Technoglossia VI", National and Kapodistrian University of Athens and National Technical University of Athens, Athens.