Ján Mačutek and George K. Mikros

Menzerath-Altmann Law for Word Length Motifs

1 Introduction

Motifs are relatively new linguistic units which make possible an in-depth investigation of sequential properties of texts (for the general definition cf. Köhler, this volume, pp. 89-90). They were studied in a handful of papers (Köhler 2006, 2008a,b, this volume, pp. 89-108; Köhler and Naumann 2008, 2009, 2010, Mačutek 2009, Sanada 2010, Milička, this volume, pp. 133-145). Specifically, a word length motif is a continuous series of equal or increasing word lengths (measured here in the number of syllables, although there are also other options, like, e.g., morphemes).

In the papers cited above it is supposed that motifs should have properties similar to those of their basic units, i.e., words in our case. Indeed, word frequency and motif frequency, as well as word length (measured in the number of syllables) and motif length (measured in the number of words) can be modelled by the same distributions (power laws, like, e.g., the Zipf-Mandelbrot distribution, and Poisson-like distributions, respectively; cf. Wimmer and Altmann 1999). Also the type-token relations for words and motifs display similar behaviour, differing only in parameters values, but not in models.

We enlarge the list of analogous properties of motifs and their basic units, demonstrating (cf. Section 3.1) that for word length motifs also the Menzerath-Altmann law (cf. Cramer 2005; MA law henceforth) is valid. The MA law describes the relation between sizes of the construct, e.g., a word, and its constituents, e.g., syllables. It states that the larger the construct (the whole), the smaller its constituents (parts). In particular, for our data it holds the longer is the motif (in the number of words), the shorter the mean length of words (in the number of syllables) which constitute the motif. In addition, in Section 3.2 we show that for randomly generated texts the MA law is valid as well, but its parameters differ from those obtained from real texts.

2 Data

In order to study the MA law for word length motifs we compiled a Modern Greek literature corpus totaling 236,233 words. The corpus contains complete versions of literary texts from the same time period and has been strictly controlled for editorial normalization. It contains five novels from four widely known Modern Greek writers published from the same publishing house (Kastaniotis Publishing House). All the novels were best-sellers in the Greek market and belong to the "classics" of the Modern Greek literature. More specifically, the corpus consists of:

– *The mother of the dog*, 1990, by Matesis (47,852 words).
– *Murders*, 1991, by Michailidis (72,475 words).
– *From the other side of the time*, 1988, by Milliex [1] (77,692 words).
– *Dreams*, 1991, by Milliex [2] (9,761 words) - Test novel.
– *The dead liqueur*, 1992, by Xanthoulis (28,453 words).

The basic descriptive statistics of the corpus appear in table 1:

Table 1: Basic descriptive statistics of the data.

Authors	Matesis	Michailidis	Milliex [1]	Milleix [2]	Xanthoulis
number of words	47,852	72,475	77,692	9,761	28,453
number of motifs	19,283	29,144	32,034	4,022	11,236
different motifs	316	381	402	192	289
mean word length in syllables	2.09	2.03	2.10	2.13	2.07
mean motif length in words	2.48	2.49	2.43	2.43	2.53
mean motif length in syllables	5.20	5.05	5.09	5.17	5.23

3 Results

3.1 MA law for word length motifs in Modern Greek texts

The results obtained confirmed our expectation that the MA law should be valid also for word length motifs. The tendency of mean word length (measured in the number of syllables) to decrease with the increasing motif length (measured in the number of words) is obvious in all five texts investigated, cf. Table 2.

We modelled the relation by the function

$$y(x) = \alpha x^b \tag{1}$$

where $y(x)$ is the mean length of words which occur in motifs consisting of x words; α and b are parameters. Given that

$$y(1) = \alpha,$$

we replaced α with the mean length of words from motifs of length 1, i.e., motifs consisting of one word only (cf. Kelih 2010, Mačutek and Rovenchak 2011). In order to avoid too strong fluctuations, only motif lengths which appeared in particular texts at least 10 times were taken into account (cf. Kelih 2010). The appropriateness of the fit was assessed in terms of the determination coefficient R^2 (values higher than 0.9 are usually considered satisfying, cf., e.g., Mačutek and Wimmer 2013). The numerical results (values of R^2 and parameter values for which R^2 reach its maximum) are presented in Table 2.

Table 2: Fitting function (1) to the data. *ML*- motif length, MWL_0- observed mean word length, MWL_t theoretical mean word length resulting from (1).

	Matesis		**Michailidis**		**Milliex 1**		**Milliex 2**		**Xanthoulis**	
ML	MWL_0	MWL_t	MWL_0	MWL_t	MWL_0	MWL_t	MWL_0	MWL_t	MWL_0	MWL_t
1	2.30	2.30	2.32	2.32	2.34	2.34	2.39	2.39	2.42	2.42
2	2.13	2.15	2.08	2.13	2.13	2.18	2.16	2.22	2.13	2.19
3	2.08	2.07	2.00	2.03	2.08	2.09	2.11	2.12	2.02	2.06
4	2.04	2.01	1.97	1.96	2.04	2.02	2.09	2.05	2.01	1.97
5	1.96	1.97	1.90	1.90	1.99	1.98	2.01	2.01	1.96	1.91
6	1.94	1.94	1.88	1.86	1.94	1.94	1.94	1.97	1.92	1.86
7	1.98	1.91	1.87	1.83	1.94	1.91	1.96	1.88	1.74	1.82
8	1.81	1.88	1.69	1.80	1.84	1.88				
9			1.84	1.77						
	b=–0.096		b=–0.123		b=–0.105		b=–0.109		b=–0.147	
	R^2=0.9195		R^2=0.9126		R^2=0.9675		R^2=0.9582		R^2=0.9312	

3.2 MA law for word length motifs in random texts

The results presented in the previous section show that longer motifs contain shorter words and vice versa. The relation between the lengths (i.e., the MA law) can be modelled by a simple power function. One cannot, however, apriori exclude the possibility that the observed regularities are necessary in the sense that they could be only a consequence of some other laws. In this particular case, it seems reasonable to ask whether MA law remains valid if the distribution of word

length is kept, but the sequential structure of word length is deliberately forgotten.

Randomization (i.e., random generating of texts – or, only some properties of texts – by means of computer programs) is a useful tool for finding answers to questions of this type. It is slowly finding its way to linguistic research (cf., e.g., Benešová and Čech, this volume, pp. 57-69, and Milička, this volume, pp. 133-145 for other analyses of the MA law; Liu and Hu 2008 applied randomization to refute claims that small-world and scale-free complex language networks automatically give rise to syntax).

In order to get rid of the sequential structure of word lengths, while at the same time preserving the word length distribution, we generated random numbers from the distribution of word length in each of the five texts under our investigation. The number of generated random word lengths is always equal to the text length of the respective real text (e.g., we generated 47,852 random word lengths for the text by Matesis, as it contains 47,852 words, cf. Table 1). Then, we fitted function (1) to the randomly generated data. The outcomes of the fitting can be found in Table 3. The generated data were truncated at the same points as their counterparts from real texts, cf. Section 3.1.

Table 3: Fitting function (1) to the data. *ML* - motif length, MWL_r - mean word length from randomly generated data, MWL_f - fitted values resulting from (1).

	Matesis		**Michailidis**		**Milliex 1**		**Milliex 2**		**Xanthoulis**	
ML	MWL_r	MWL_f	MWL_r	MWL_f	MWL_r	MWL_f	MWL_r	MWL_f	MWL_r	MWL_f
1	2.41	2.41	2.36	2.36	2.44	2.44	2.42	2.42	2.42	2.42
2	2.27	2.19	2.20	2.13	2.28	2.20	2.29	2.20	2.26	2.18
3	2.11	2.07	2.06	2.01	2.15	2.07	2.15	2.08	2.12	2.05
4	2.02	1.99	1.96	1.93	2.02	1.99	2.07	2.00	1.98	1.97
5	1.94	1.93	1.87	1.87	1.94	1.92	1.97	1.94	1.89	1.90
6	1.86	1.88	1.84	1.82	1.83	1.87	1.86	1.89	1.82	1.85
7	1.81	1.84	1.77	1.78	1.82	1.83	1.72	1.85	1.76	1.81
8	1.78	1.81	1.69	1.74	1.73	1.79				
9			1.66	1.71						
	b=–0.138		b=–0.146		b=–0.148		b=–0.137		b=–0.150	
	R^2=0.9685		R^2=0.9682		R^2=0.9553		R^2=0.8951		R^2=0.9595	

It can be seen that the MA law holds also in this case; however, parameters b in random texts are different from the ones from real texts. The parameters in the random texts have always larger absolute values, which means that the respective curves are steeper, i.e., they decrease more quickly.

As an example, in Fig. 1, we present data from the first text (by Matesis, cf. Section 2) together with the mathematical model (1) fitted to the data.

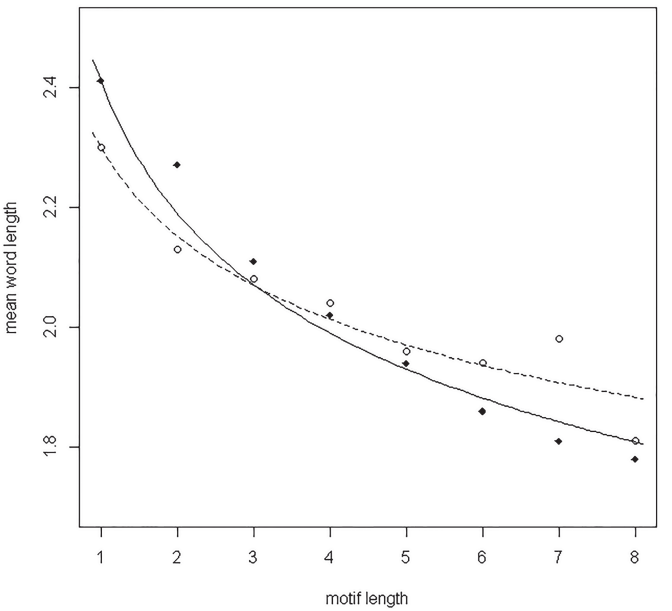

Fig. 1: Data from the text by Matesis (circles) and from the respective random text (diamonds), together with fitted models (dashed line – the model for the real text, solid line – the model for the random text)

The other texts behave similarly.

The validity of the law in random texts (created under condition that their word length distribution in the same as in the real ones) can be deductively explained directly from the definition of word length motifs (cf. Section 1). A word length motif is ended at the place where the sequence of word lengths decreases. Hence, if a long word appears in a motif, it is likely to be the last element of the motif (the probability that the next word would be of the same length – or even longer – is small, because long words occur relatively seldom). Consequently, if a long word appears soon, the motif tends to be short in terms of words, but the mean word length in such a motif will be high (because the length of one long word will have a higher weight in a short motif than in a long one).

Differences in values of parameters b in real and randomized texts indicate that, regardless of the obvious impact of the word length distribution, also the

sequential structure of word lengths plays an important role in the MA law for word length motifs.

4 Conclusions

The paper brings another confirmation that word length motifs behave in the same way as other, more "traditional" linguistic units. In addition to the motif frequency distribution and the distribution of motif length (which were shown to follow the same patterns as words), also the MA law is valid for word length motifs, specifically, the more words a motif contains, the shorter mean syllabic length of words in the motif.

The MA law can be observed also in random texts, if word lengths distributions in a real text and in its random counterpart are the same. The validity of the law in random texts can be explained deductively from word length distribution. However, parameters in the exponents of the power function which is a mathematical model of the MA law are different for real and random texts. The power functions corresponding to random texts are steeper. The difference in parameter values proves that not only word length distribution, but also the sequential structure of word lengths has an impact on word length motifs.

It remains an open question whether parameters of the MA law can be used as characteristics of languages, genres or authors. In case of the positive answer, they could possibly be applied to language classification, authorship attribution and similar fields.

Acknowledgments

J. Mačutek was supported by VEGA grant 2/0038/12.

References

Benešová, Martina & Radek Čech. 2015. Menzerath-Altmann law versus random models. This volume, pp. 57-69.

Cramer, Irene M. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An international handbook*, 659–688. Berlin & New York: de Gruyter.

Liu, Haitao & Fengguo Hu. 2008. What role does syntax play in a language network? *EPL* 83. 18002.

Kelih, Emmerich. 2010. Parameter interpretation of the Menzerath law: Evidence from Serbian. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and language. Structures, functions, interrelations, quantitative perspectives*, 71–79. Wien: Praesens.

Köhler, Reinhard. 2006. The frequency distribution of the lengths of length sequences. In Jozef Genzor & Martina Bucková (eds.), *Favete linguis. Studies in honour of Viktor Krupa*, 145–152. Bratislava: Slovak Academic Press.

Köhler, Reinhard. 2008a. Word length in text. A study in the syntagmatic dimension. In Sybila Mislovičová (ed.), *Jazyk a jazykoveda v pohybe*, 416–421. Bratislava: Veda.

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

Köhler, Reinhard. 2015. Linguistic motifs. This volume, pp. 89-108.

Köhler, Reinhard & Sven Naumann. 2008. Quantitative text analysis using L-, F- and T-segments. In Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme & Reinhold Decker (eds.), *Data analysis, machine learning and applications*, 635–646. Berlin & Heidelberg: Springer.

Köhler, Reinhard & Sven Naumann. 2009. A contribution to quantitative studies on the sentence level. In Reinhard Köhler (ed.), *Issues in quantitative linguistics*, 34–57. Lüdenscheid: RAM-Verlag.

Köhler, Reinhard & Sven Naumann. 2010. A syntagmatic approach to automatic text classification. Statistical properties of F- and L-motifs as text characteristics. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and language. Structures, functions, interrelations, quantitative perspectives*, 81–89. Wien: Praesens.

Mačutek, Ján. 2009. Motif richness. In Reinhard Köhler (ed.), *Issues in quantitative linguistics*, 51–60. Lüdenscheid: RAM-Verlag.

Mačutek, Ján & Andrij Rovenchak. 2011. Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In Emmerich Kelih, Victor Levickij & Yuliya Matskulyak (eds.), *Issues in quantitative linguistics 2*, 136–147. Lüdenscheid: RAM-Verlag.

Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitatitive linguistics. *Journal of Quantitative Linguistics* 20(3). 227–240.

Milička, Jiří. 2015. Is the distribution of L-motifs inherited from the word lengths distribution? This volume, pp. 133-145.

Sanada, Haruko. 2010. Distribution of motifs in Japanese texts. In Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives*, 183–194. Wien: Praesens.

Wimmer, Gejza & Gabriel Altmann. 1999. *Thesaurus of univariate discrete probability distributions*. Essen: Stamm