Is the Distribution of L-Motifs Inherited from the Word Lengths Distribution?

Jiří Milička

1. Abstract

The distribution of L-motifs (measured on a text T) is similar to the L-motifs distribution measured on the pseudotext T' constructed by random transposition of all tokens within the text T. This inspires the suggestion that the distribution of L-motifs is inherited from the word length distribution (or, by other words, that the word length distribution of a text implies the distribution of L-motifs). The paper clearly shows that despite of the similarity, an L-motifs structure, independent of the word length distribution, can be detected.

2. Introduction

An increasing number of papers[1] shows that word length sequences can be successfully analyzed by means of *L-motifs*, which are a very promising attempt to discover the syntagmatic relations of the word lengths in a text.

The L- motif[2] has been defined by Reinhard Köhler (2006a) as:

> (...) the text segment which, beginning with the first word of the given text, consists of word lengths which are greater or equal to the left neighbour. As soon as a word is encountered which is shorter than the previous one the end of the current L-Segment is reached. Thus, the fragment (1) will be segmented as shown by the L-segment sequence (2):
>
> Azon a tájon, ahol most Budapest fekszik, már nagyon régen laknak emberek.
>
> 2,122,13,2,12223

The main advantage of such segmentation is that it can be applied iteratively, i.e. L-motifs of the L-motifs can be obtained (so called LL-motifs). Applying the method several times results in not very intuitive sequences,

which, however, follow lawful patterns[3] and they are even practically useful, e.g. for automatic text classification (Köhler – Naumann 2010).

3. Hypothesis

However it needs to be admitted that the fact that the L-motifs follow lawful patterns does not imply that the L-motifs reflect a syntagmatic relation of the word lengths, since these properties could be merely inherited from the word length distribution in the text, which has not been tested yet. The paper focuses on the most important property of L-motifs – the frequency distribution of their types and tests the following hypothesis:

The distribution of L-motifs measured on the text T differs from the distribution of L-motifs measured on a pseudotext T'. The pseudotext T' is created by the random transposition of all tokens of the text T within the text T.[4]

4. Data

The hypothesis was tested on three Czech and six Arabic texts:

Table 1. The list of texts.

Tag	Author	Title	Cent.	Lang.	# of Tokens
[Zer]	Milan Kundera	Žert	20	Czech	88435
[Kat]	Kohout	Katyně	20	Czech	99808
[Bab]	Božena Němcová	Babička	19	Czech	70140
[Ham]	al-Ḥāzimī al-Hamadānī	Al-ʾIʿtibār fi ʾn-nāsiḫ wa-ʾl-mansūḫ	15	Arabic	71482
[Sal]	ibn aṣ-Ṣallāḥ	Maʿrifatu ʾanwāʿi ʿulūmi ʾl-ḥadīṯ	13	Arabic	54915
[Zam]	ibn abī Zamanīn	Uṣūlu ʾs-sunna	11	Arabic	18607
[Maw]	al-Mawwāq	Tāǧ wa-l-ʿiklīl 2	15	Arabic	274840
[Baj]	al-Bāǧī al-ʿAndalūsī	Al-Muntaqī 2	11	Arabic	301232
[Bah]	Manṣūr al-Bahūtī	Šarḥ muntahīyu ʾl-irādāt 2	17	Arabic	263175

The graphical word segmentation was respected when determining the number of syllables in the Arabic texts. In the Czech texts zero syllabic words (e.g. *s*, *z*, *v*, *k*) were merged with the following words according to

the conclusion in Antić et al. (2006), to maintain the compatibility with other studies in this field (e.g. Köhler 2006b).

5. Motivation

One of those texts [Kat] was randomized for one million[5] times and the rank–frequency relation (RFR) of L-motifs was measured for every randomised pseudotext. Then these RFRs were averaged. This average RFR can be seen on the following chart, accompanied by the RFR of L-structures measured on the real text:

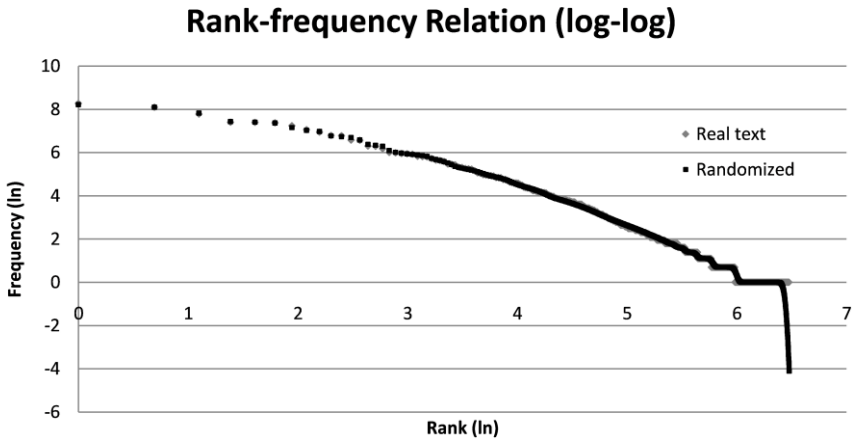

Figure 1. RFR of the L-motifs, [Kat].

Visually, the RFR of the L-motifs distribution for the real text does not differ from the average pseudotext RFR of the L-motifs very much. This impression is supported by the Chi-squared discrepancy coefficient $C = 0.0008$.[6] Also the fact, that both the real text's L-motifs RFR and the randomized text's L-motifs RFR can be successfully fitted by the Right truncated Zipf-Alexeev distribution with similar parameters[7] encourages us to assume that the RFR of L-motifs is given by the word length distribution in the text.

Very similar results can be obtained for LL-motifs, LLL-motifs[8] etc. (the Zipf-Mandelbrot distribution fits the distribution of the higher orders L-motifs better than the right truncated Zipf-Alexeev distribution).

But these results do not answer the question asked. The next section proceeds to the testing of the hypothesis.

6. Methods

Not only L-motifs as a whole, but every single L-motif has the distribution of its frequencies within those one million randomized pseudotexts. For example the number of pseudotexts (randomized [Bab]), where the L-motif (1, 1, 2, 2, 2) occurred 72 times, is 111. From this distribution we can obtain confidence intervals (95%) as depicted on the following chart:

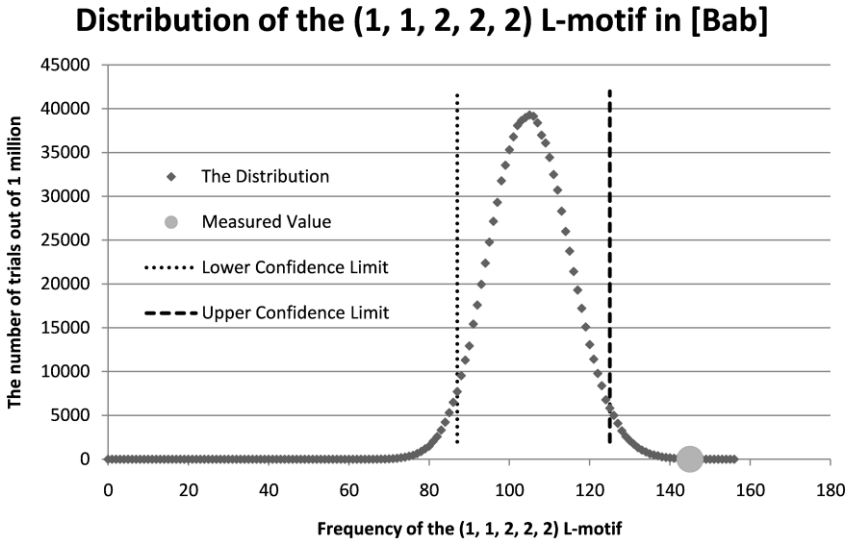

Figure 2. Distribution of one of the L-motif types in one million pseudotexts (randomized [Bab]) vs. the frequency of the L-motif in the real text.

In this case, the frequency of the motif (1, 1, 2, 2, 2) measured on the real text [Bab] is 145, which is above the upper confidence interval limit (in

this case 125). But the frequencies of many other L-motifs are within these intervals, such as the motif (1, 1, 1, 2, 2):

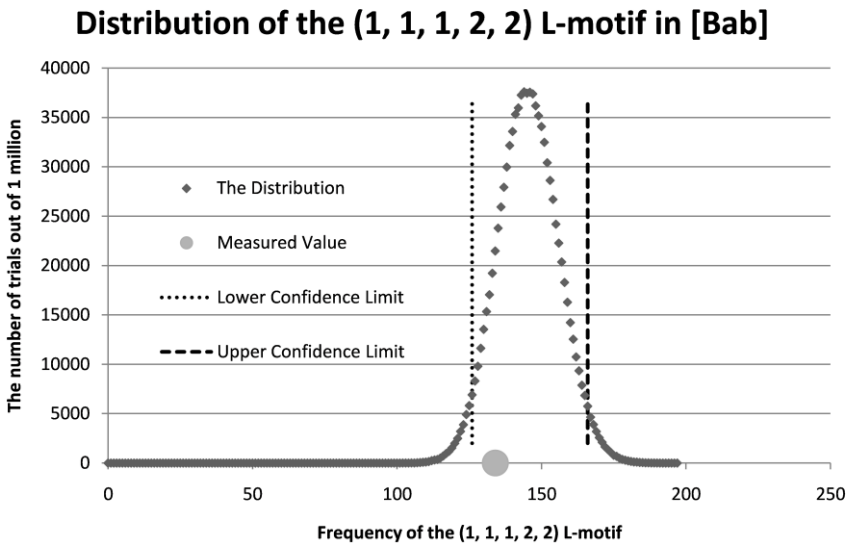

Figure 3. Distribution of one of the L-motif types in one million pseudotexts (randomized [Bab]) vs. the frequency of the L-motif in the real text.

The fact that the frequencies are not independent from each other does not allow us to test them separately as multiple hypotheses, and moves us to merge all values of the distribution into one number. The following method was chosen:

1. The text is many times randomized (in this case 1 million times) and for each pseudotext frequencies of L-motifs are measured. The average frequency of every L-motif is calculated. The average frequency of the motif (indexed by the variable i, N is the maximal i) will be referred as $\bar{m}_i$.
2. The total distance (D) between the frequencies of each motif (m_i) in the text T and their average frequencies in the randomized pseudotexts ($\bar{m}_i$) are calculated:

$$D = \sum_{i=1}^{N} |\bar{m}_i - m_i|$$

3. All total distances (D') between the frequencies of each motif (m'_i) in one million pseudotexts T' (these pseudotexts must be different from those that were measured in the step 1) and their average frequencies in the randomized pseudotexts ($\overline{m}_i$) (must be the same as in the previous step) are calculated:

$$D' = \sum_{i=1}^{N} |\overline{m}_i - m'_i|$$

4. The distribution of the D' distances is obtained.
5. The upper confidence limit is set. The distance *D* significantly lower than the distances D' would mean that the real distribution is even closer to the distribution generated by random transposing tokens than another distributions measured on randomly transposed tokens. This would not reject the null hypothesis. Considering this, the lower confidence limit is senseless and the test can be assumed to be one-tailed.
6. *D* is compared with the upper confidence limit. If *D* is larger than the upper confidence limit, then the null hypothesis is rejected.

An example result of this method follows (applied on L-motifs of [Bab]):

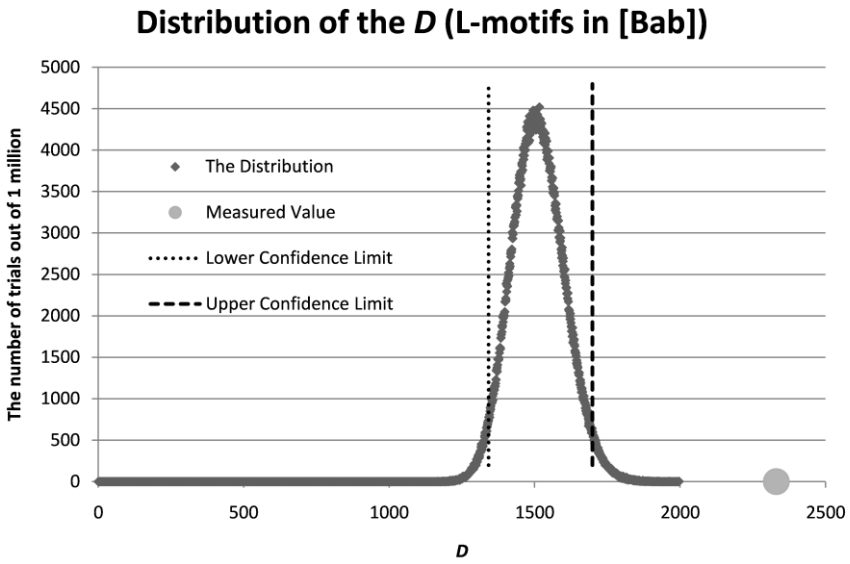

Figure 4. The distribution of the variable D in one million pseudotexts (randomized [Bab]) vs. the value of the D variable in the real text. Here, for example, 4371 out of one million randomized texts have D' equal to 1500.

As the *D* is larger than the upper confidence limit, we shall assume that the distribution of the L-motifs measured on [Bab] is more distant from the average distribution of L-motifs measured on pseudotexts (derived by the random transposition of tokens in [Bab]), than the distribution of L-motifs measured on another pseudotexts (also derived by the random transposition of tokens in [Bab]).

7. Results

In the following charts, one column represents the $\overline{D'}$ values compared to the measured *D* value (like in the Fig. 4, but in a more concise form) for 7 orders of motifs. Confidence limits of the 95% confidence intervals are indicated by the error bars.

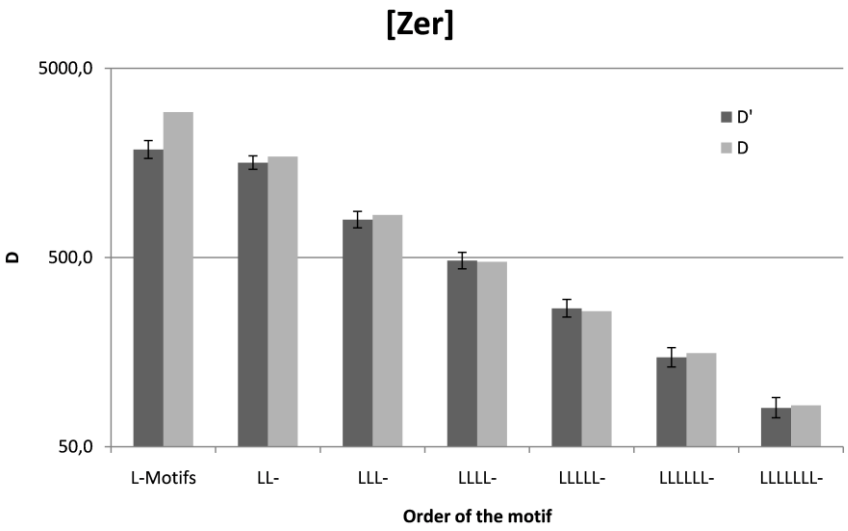

Figure 5. The D-value of the distribution of L-motifs (in the [Zer]) is significantly different from the D'-value measured on randomly transposed tokens of the same text. Notice that the LL-motifs distribution D-value is also close to the upper confidence limit.

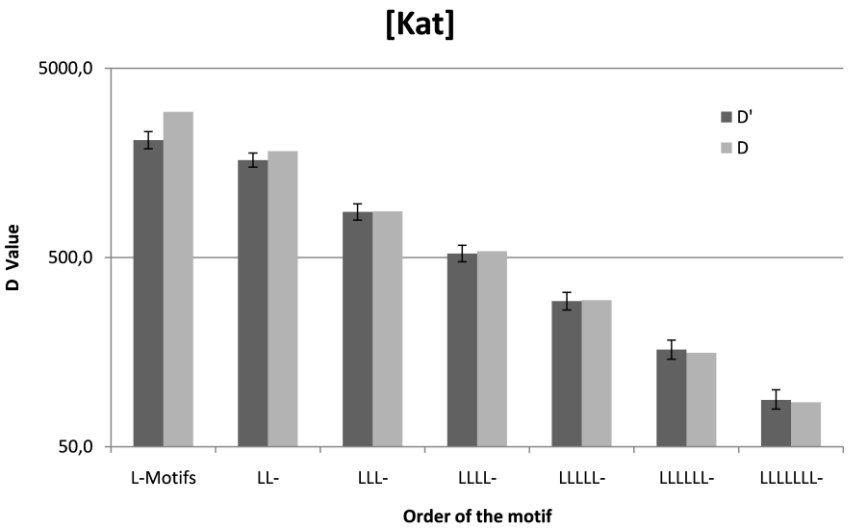

Figure 6. The D-values of the distributions of L-motifs and LL-motifs (in the [Kat]) are significantly different from the D'-values measured on randomly transposed tokens of the same text. Notice that the LL-motifs distribution D-value is very close to the upper confidence limit.

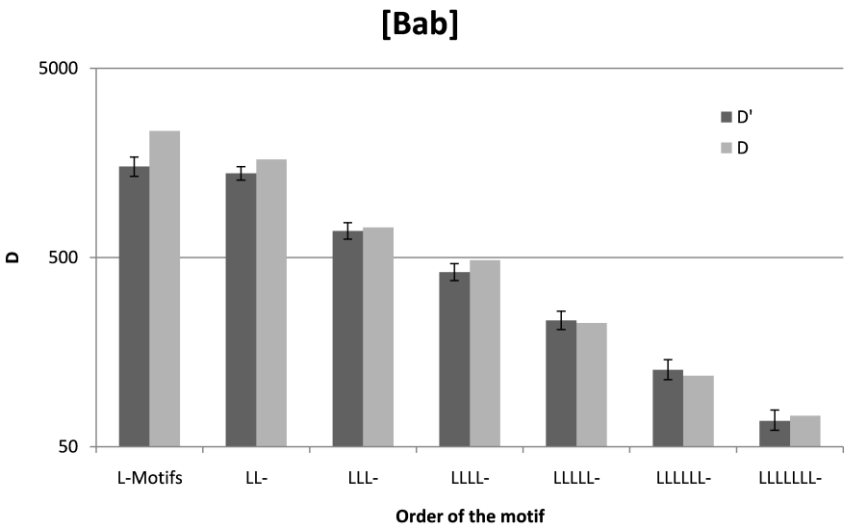

Figure 7. The D-values of the distributions of L-motifs, LL-motifs and LLLL-motifs (in the [Bab]) are significantly different from the D'-values measured on randomly transposed tokens of the same text. Consider that the LLLL-motifs distribution can be different just by chance.

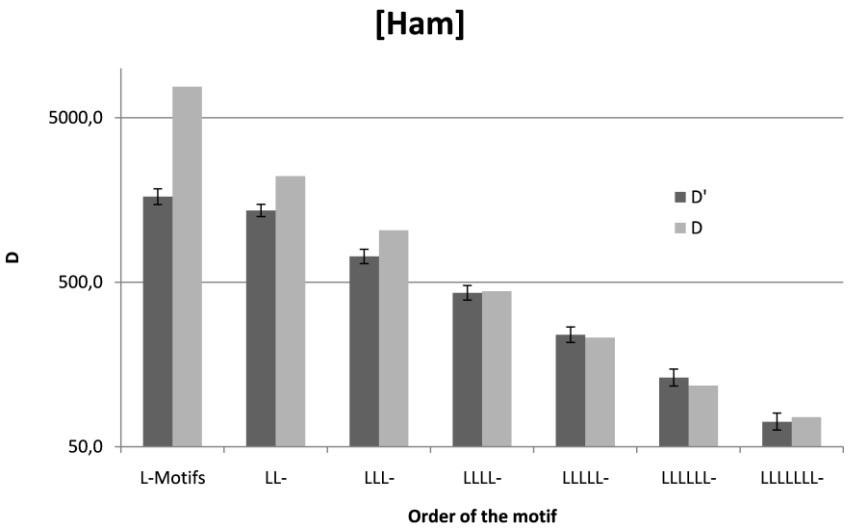

Figure 8. The D-values of the distributions of L-motifs, LL-motifs and LLL-motifs (in the [Ham]) are significantly different from the D'-values. The ratios between the D-values and the upper confidence limits are more noticeable than those measured on the Czech texts (the y-axis is log scaled). As the size of these texts is comparable, it seems that the L-motif structure is more substantial for the Arabic texts than for the Czech ones.

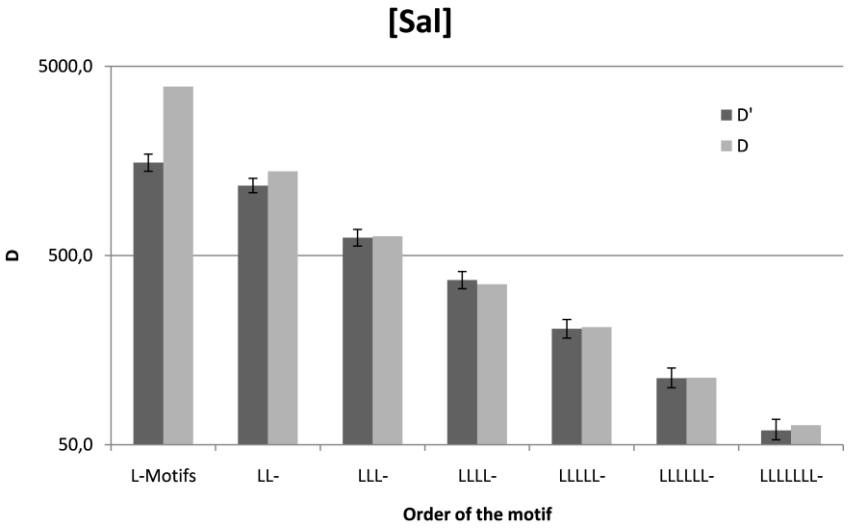

Figure 9. The D-values of the distributions of L-motifs and LL-motifs (in the [Sal]) are significantly different from the D'-values.

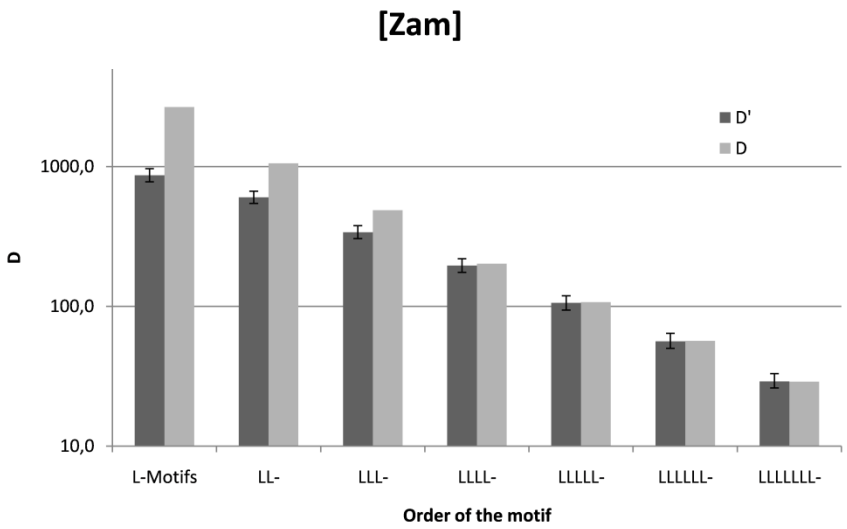

Figure 10. The D-values of the distributions of L-motifs, LL-motifs and LLL-motifs (in the [Zam]) are significantly different from the D'-values despite of the fact, that the text is relatively short.

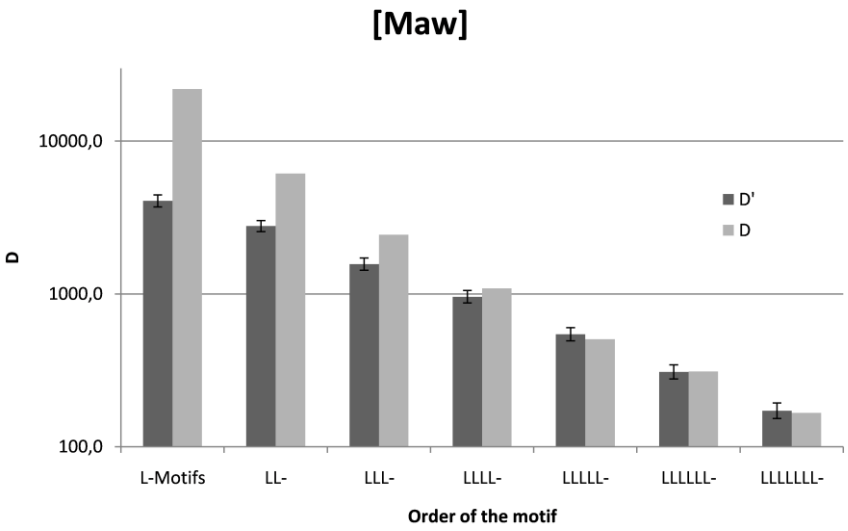

Figure 11. The D-values of the distributions of L-motifs, LL-motifs, LLL-motifs and LLLL-motifs (in the [Maw]) are significantly different from the D'-values despite of the fact, that the text is relatively large and incoherent.

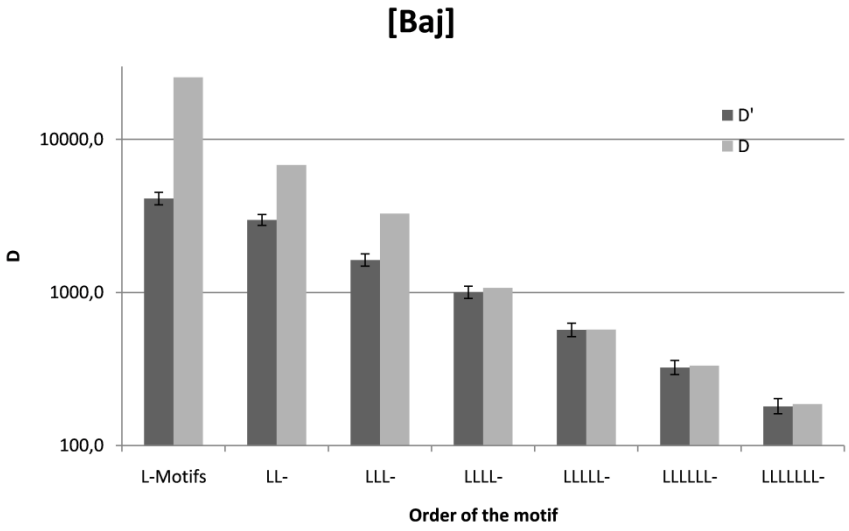

Figure 12. The D-values of the distributions of L-motifs, LL-motifs and LLL-motifs (in the [Baj]) are significantly different from the D'-values.

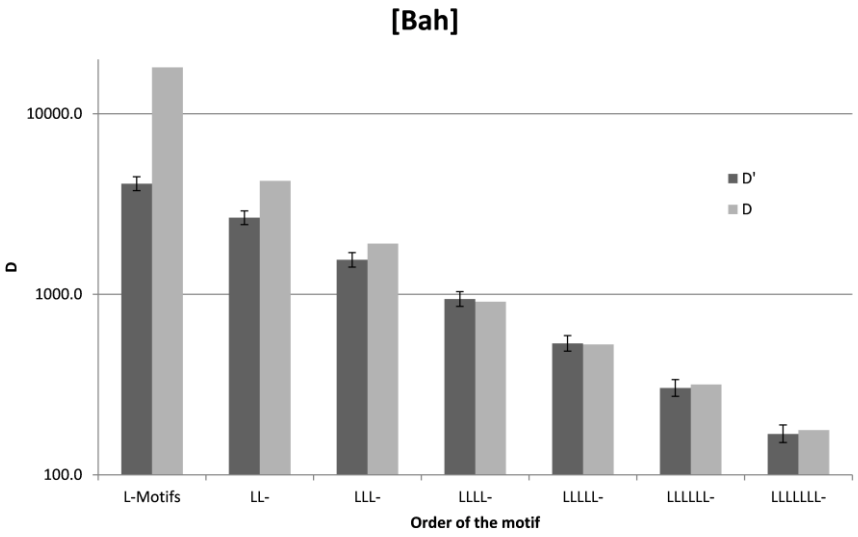

Figure 13. The D-values of the distributions of L-motifs, LL-motifs and LLL-motifs (in the [Bah]) are significantly different from the D'-values.

Table 2. Exact figures presented in the charts. U stands for the upper confidence limit. All D; $D > U$ are marked bold.

		[Zer]	[Kat]	[Bab]	[Ham]	[Sal]	[Zam]	[Maw]	[Baj]	[Bah]
L-motifs	D	**2930**	**2943**	**2331**	**7712**	**3909**	**2673**	**21899**	**25437**	**18070**
	$\overline{D'}$	1860	2084	1512	1659	1547	867.6	4065.1	4108.6	4094.7
	U	2072	2312	1699	1851	1716	967	4448	4506	4472
LL-motifs	D	1705	**1825**	**1649**	**2207**	**1394**	**1057**	**6133.9**	**6798.9**	**4245.5**
	$\overline{D'}$	1584	1635	1391	1364	1171	603.2	2779.0	2980.9	2654.6
	U	1719	1780	1510	1486	1278	666	3023	3234	2895
LLL-motifs	D	839.1	874.5	721.5	**1035**	631.0	**487.2**	**2444.2**	**3279.7**	**1909.4**
	$\overline{D'}$	792.8	869.4	690.6	716.9	620.5	339.2	1568.0	1627.4	1551.8
	U	876	959	764	792	687	378	1719	1785	1701
LLLL-motifs	D	474.3	539.3	**483.5**	441.1	352.1	201.9	**1085.3**	1072.0	911.3
	$\overline{D'}$	481.3	525.0	417.7	430.9	370.5	195.6	957.3	1000.2	942.0
	U	532	580	463	477	411	219	1051	1097	1034
LLLLL-motifs	D	260.3	297.2	225.6	230.5	209.0	107.3	504.3	572.2	527.3
	$\overline{D'}$	269.4	293.6	232.3	239.4	204.8	105.7	544.0	569.4	534.7
	U	300	327	260	267	229	119	601	629	591
LLLLLL-motifs	D	156.1	156.6	118.7	117.7	113.1	56.6	310.2	332.8	316.8
	$\overline{D'}$	148.5	163.0	127.4	131.6	112.2	56.1	308.4	322.8	303.1
	U	167	183	144	148	127	64	343	359	337
LLLLLLL-motifs	D	82.7	85.9	73.0	75.6	63.3	28.9	166.2	186.5	177.1
	$\overline{D'}$	80.1	88.4	68.5	70.8	59.6	29.0	171.7	180.1	168.6
	U	91	100	78	80	68	33	193	202	189

1. Conclusion

The null hypothesis was rejected for the L-motifs (all texts) and for LL-motifs (except [Zer]) and was not rejected for L-motifs of higher orders (LLL-motifs etc.) in Czech, but was rejected also for LLL-motifs in Arabic (except [Sal]). As type-token relation and distribution of lengths are to some extent dependent on the frequency distribution, similar results for these properties can be expected, but proper tests are needed. Our methodology can be also used for testing F-motifs and other types and definitions of motifs.

It needs to be said that non-rejecting the null hypothesis does not mean, that the L-motifs of higher orders are senseless – even if their distribution was inherited from the distribution of word lengths in the text (which is still

not sure), it still could be used as a tool mediating to see the distribution of the word lengths from another point of view. However, it turns out that if we wish to use the L-motifs to examine the syntagmatic relations of the word lengths, the structure inherited from the word length distribution must be taken into account.

Notes

1. See (Köhler – Naumann 2010), (Mačutek 2009), (Sanada 2010).
2. Former term was *L-segments*, see (Köhler 2006a).
3. E.g. the rank-frequency relation of the L-motifs distribution can be successfully described by the Zipf-Mandelbrot distribution, which is a well established law for the word types rank-frequency relation.
4. The null hypothesis is: "The distribution of L-motifs measured on the text *T* is the same as the distribution of L-motifs measured on a pseudotext *T'*. The pseudotext *T'* is created by random transposition of all tokens of the text *T* within the text *T*."
5. 1 million has been arbitrarily chosen as a "sufficiently large number", which makes it the weakest point of our argumentation.
6. $C = \mathrm{X}^2/N$, where *N* is the sample size (Mačutek – Wimmer 2013).
7. For the real data: $a = 0.228$; $b = 0.1779$; $n = 651$; $\alpha = 0.1089$; $C = 0.0066$. For the randomized pseudotexts: $a = 0.244$; $b = 0.1761$; $n = 651$; $\alpha = 0.1032$; $C = 0.0047$. Altmann Fitter was used.
8. For the LL-motifs: $C = 0.0009$; for the LLL-motifs: $C = 0.0011$.

Acknowledgement

The author is grateful to Reinhard Köhler for helpful comments and suggestions. This work was supported by the project *Lingvistická a lexikostatistická analýza ve spolupráci lingvistiky, matematiky, biologie a psychologie*, 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

Antić, Gordona & Emmerich Kelih & Peter Grzybek. 2006. Zero-Syllable Words in Determining Word Length. In Peter Grzybek (ed.), *Contributions to the science of text and language. Word length studies and related issues*, 117–156. Dordrecht: Springer.

Köhler, Reinhard. 2006a. Word length in text. A study in the syntagmatic dimension. In Sibyla Mislovicová (ed.), *Jazyk a jazykoveda v prohybe*, 416–421. Bratislava: VEDA vydavateľstvo SAV.

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

Köhler, Reinhard & Sven Naumann. 2009. 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*, 81–89. Wien: Praesens Verlag.

Köhler, Reinhard & Sven Naumann. 2010. A contribution to quantitative studies on the sentence level. In Reinhard Köhler (ed.), *Issues in Quantitative Linguistics*, 34–57. Lüdenscheid: RAM-Verlag.

Mačutek, Ján. 2009. Motif richness. In Köhler, R. (ed.), *Issues in Quantitative Linguistics*, 51–60. Lüdenscheid: RAM-Verlag.

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

Sanada, Haruko. 2010. Distribution of motifs in Japanese texts. In Peter Grzybek, Emerich Kelih & Ján Mačutek (eds.), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives*, 183–193. Wien: Praesens Verlag.