https://doi.org/10.51305/ICL.CZ.9788076580336.01
Tackling the Toolkit, Plotting Poetry through Computational Literary Studies. ICL CAS: Prague 2021.

The Menzerath-Altmann Law in Czech Poems by K. J. Erben

Radek Čech
University of Ostrava, Czechia
cechradek@gmail.com
0000-0002-4412-4588

Ján Mačutek
Mathematical Institute,
Slovak Academy of Sciences / Constantine
the Philosopher University in Nitra, Slovakia
jmacutek@yahoo.com
0000-0003-1712-4395

Abstract

The aim of the paper is to test the validity of the Menzerath-Altmann law for Czech poems from K. J. Erben's ballad collection *Kytice z pověstí národních* (A Bouquet of Folk Legends). We focus particularly on the relationship between word length and syllable length. The Menzerath-Altmann law predicts that the mean syllable length will be longer in shorter words. The parameters of the mathematical model of this law for poems are compared with those for prose texts.

1 Introduction

According to Menzerath (1954), longer words in the German vocabulary tend to contain shorter syllables. A similar tendency, now known as the Menzerath-Altmann law (MAL), has been proven valid for multiple languages and several language units. In general, as Altmann observes, "[t]he longer a language construct the shorter its components (constituents)" (1980, p. 1). More specifically, when words are the constructs and morphemes their constituents, the mean length of morphemes decreases as the word length increases (Gerlach 1982). Similarly, when clauses are the constructs and syntactic phrases are the constituents, longer clauses consist of shorter syntactic phrases (Mačutek, Čech, et al. 2017). Overviews of work on this topic can be found in Cramer (2005) and Mačutek, Chromý, et al. (2019).

The most widely used mathematical model of the MAL is the function

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

where x is the size of the construct, $y(x)$ is the mean size of the constituents in constructs of size x, and a and b are parameters. Buk et al. (2007, pp. 43–44) and Mačutek, Chromý, et al. (2019, p. 67) each provide a generalization of this model.

To the best of our knowledge, research on the MAL's application to poetic texts has so far been limited to two cases. A Slovak poem exemplifies the relationship between word length and syllable length in a study by Wimmer et al. (2003, pp. 105–106).[1] In addition, Čech et al. (2011, pp. 54–55) investigate the relationship between the word count of verses and the mean word length. The poetic texts in this case do not rhyme or follow any meter. The verses are also short, often consisting of only one word. In both of the above studies, the relations between unit lengths can be modeled by the MAL.

In what follows, we focus on the relationship between word length (measured in the number of syllables) and syllable length (measured in speech sounds) in a sample of Czech poetry. Our aim is twofold. First, we seek to test the validity of the MAL using a larger volume of "classical" poems that use rhyme and meter. In addition, these poems constitute a relatively homogeneous corpus as works written by one author. Second, we seek to compare the parameters of MAL for these poems with those for Czech prose texts.

2 Language Material and Methodology

Two groups of texts were chosen for our analysis. The first is comprised of 13 poems from the ballad collection *Kytice z pověstí národních* (A Bouquet of Folk Legends) written by Czech poet Karel Jaromír Erben.[2] This collection was first published in 1853. The poems it contains are highly influenced by folk poetry. For comparison, we again used Czech texts, namely eight short stories from the collection *Povídky malostranské* (Tales of the Lesser Quarter, first published in 1877) written by Jan Neruda.[3] We refer to the texts by Erben as E1–E13 and to the ones by Neruda as N1–N8. Table 1 provides an overview of the analyzed texts and their basic characteristics (i.e. the number of tokens and types they contain and the type-token ratio (TTR), which is here the ratio between the number of types and the number of tokens).[4]

In relation to the MAL, words were the constructs while the size of this construct was determined by the number of its syllables. Syllables were the constituents of the words, and their sizes were determined by the number of their speech sounds. We tested the MAL for word types (as opposed to tokens). As such, each word form was taken into account only once.[5]

Each text was analyzed separately. All non-syllabic prepositions were joined to the following word according to the preferred approach in quantitative linguistics, as described by Antić et al. (2006). First, for all of the words in the text, we determined the word length (WL), as calculated by the number of syllables.

[1] It should be noted that this book is written in Slovak and therefore not easily accessible to a broader audience.

[2] For our analysis, we used the texts available at http://www.cist.cz/Poezie/kytice.htm.

[3] We used the digitized edition *Sebrané spisy Jana Nerudy – díl desátý – Povídky malostranské*. Pořádá Ignát Herrmann. Vydal F. Topič v Praze 1893.

[4] For several other approaches to the evaluation of the type-token ratio, see Wimmer (2005) and Mitchell (2015).

[5] We, thus, followed the original approach of Menzerath (1954), who analyzed German vocabulary, i.e. types.

text	title	# tokens	# types	TTR
E1	Dceřina kletba	217	121	0.56
E2	Holoubek	308	236	0.77
E3	Lilie	479	336	0.70
E4	Poklad	2239	1050	0.47
E5	Polednice	197	152	0.77
E6	Štědrý den	651	441	0.68
E7	Svatební košile	1422	713	0.50
E8	Věštkyně	1109	696	0.63
E9	Vodník	923	529	0.57
E10	Vrba	481	312	0.65
E11	Záhořovo lože	2587	1380	0.53
E12	Zlatý kolovrat	1367	648	0.47
E13	Kytice	112	94	0.84
N1	Doktor Kazisvět	1947	1007	0.52
N2	Hastrman	1718	943	0.55
N3	Jak si pan Vorel nakouřil pěnovku	1498	774	0.52
N4	O měkkém srdci paní Rusky	1630	868	0.53
N5	Pan Ryšánek a pan Schlegl	3062	1425	0.47
N6	Přivedla žebráka na mizinu	2207	1128	0.51
N7	Svatováclavská mše	3678	1728	0.47
N8	U Tří lilií	646	397	0.61

Table 1: Texts and their basic characteristics

For all cases except for extra-syllabic consonants (Crystal 2008, pp. 182–183), the number of syllables in the word equaled the number of sonority peaks in the word's sonority profile. A sonority scale consisting of three classes (vowels, sonorants, and obstruents) was applied to construct the sonority profiles. As a next step, word length was also determined based on the number of speech sounds.

Given that we work with mean syllable lengths (measured in speech sounds) in words of certain syllable lengths, the results become unstable if some word lengths do not occur frequently enough (e.g. there is only one five-syllable word in text E2). We therefore pooled some word length groups so that there were at least five words in each category.[6] We can illustrate this approach using text E2 from Table 2. This work has twelve four-syllable words and one five-syllable word. These two word lengths were combined into one category, which was then represented by the mean WL of all the words which it contains

[6] The limit we set (i.e. at least five words) was only our rule of thumb. In contrast, Mačutek and Rovenchak (2011) require 10 words as the minimum frequency per category and their analysis excludes construct lengths that do not satisfy this condition. We note that the results of fitting presented in this paper barely change if the latter approach is taken (the difference between the values of parameter b obtained by the two approaches is always within the limit of 0.02).

(i.e. $(12 \times 4 + 1 \times 5)/13 = 4.08$). The value 2.21 is the mean syllable length of all syllables occurring in words from this category.

3 Results

The MAL in the form of function (1) was fitted to the texts presented in Table 1. The goodness of fit of the model was evaluated in terms of the determination coefficient R^2. A fit is usually considered satisfactory if $R^2 \geq 0.9$; see Mačutek and Wimmer (2013).

It is obvious that if we substitute 1 for x in (1), we obtain $y(1) = a$. Parameter a can, thus, be (in this context) interpreted as the mean length (measured in speech sounds) of monosyllabic words; see Kelih (2010). Parameter b is the value which maximizes the determination coefficient. Fitting was performed using NLREG software (www.nlreg.com).

The results are presented in Table 2, Table 3, and Table 4. In these tables, MSL denotes the mean syllable length, f_{WL} is the frequency of word types of length WL, and a and b are parameters of the MAL; see (1).

As can be seen in Table 2, the fit is mostly very good. Among the texts presented in Table 2, text E7 is the only exception where the value of R^2 falls below 0.9. Since, however, this threshold is only a rule of thumb and the determination coefficient falls only slightly short of it, we may conclude that these texts abide by the MAL.

On the other hand, Erben's poem *Kytice* (text E13) clearly does not tend to reflect the MAL in the same way. This is evident in Table 3.

There are at least two possible explanations for the failure of this poem to follow the MAL. Admittedly, both these accounts are speculative as we do not have similar texts at our disposal which might verify or refute these hypotheses.

First, the poem *Kytice* is the shortest text in our sample, with just 112 word tokens and 94 types (see Table 1). It may be that this text is simply too short, and the "mechanism" behind the MAL is too "weak" to have any impact. We note that Čech (2015) suggests that when studying the structure of word frequencies, the "ideal" text length is somewhere between 200 and 6500 word tokens, and the same recommendation is made by Čech (2016, p. 57) for investigations of the thematic concentration of texts. This poem is the only text that is much shorter than 200 word tokens in our sample (the second shortest one, E5, contains 197 word tokens). In addition, its TTR achieves a very high value (see Table 1), which is typical of short texts.

Second, the poem's word length structure seems quite distinctive in the context of most of the other poetic texts analyzed in this paper. In particular, the ratio of the frequencies of three-syllable to two-syllable word types is roughly 0.79 (with 27 and 34 word types, respectively; see Table 3). The only other poem with this property is *Svatební košile* (E7) where the ratio is 0.68; otherwise, the ratio for works in the sample falls below 0.5. In both these exceptional cases, the iambic meter is often violated by a three-syllable word in an odd position of verse line. As such, these two poems have a less regular iambic character than other poems in the collection. It remains an open question what effect this

E1			E2			E3			E4		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.90	39	1	2.69	52	1	3.04	114	1	2.77	201
2	2.52	48	2	2.35	117	2	2.46	149	2	2.45	603
3	2.30	22	3	2.24	54	3	2.34	64	3	2.32	152
4	2.21	13	4.08	2.21	13	4	2.19	9	4.03	2.22	94
a	2.90		a	2.69		a	3.04		a	2.77	
b	-0.20		b	-0.16		b	-0.25		b	-0.16	
R^2	0.997		R^2	0.950		R^2	0.972		R^2	0.996	
E5			E6			E7			E8		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.87	53	1	2.86	87	1	3.14	182	1	3.05	126
2	2.51	63	2	2.42	236	2	2.45	311	2	2.47	360
3	2.32	31	3	2.32	100	3	2.30	213	3	2.28	163
4	2.15	5	4	2.16	24	4.14	2.32	7	4.02	2.26	47
a	2.87		a	2.86		a	3.14		a	3.05	
b	-0.20		b	-0.20		b	-0.26		b	-0.25	
R^2	0.997		R^2	0.981		R^2	0.883		R^2	0.948	
E9			E10			E11			E12		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.88	112	1	2.75	71	1	3.01	178	1	3.11	172
2	2.38	260	2	2.35	180	2	2.47	727	2	2.41	315
3	2.35	117	3	2.22	40	3	2.32	335	3	2.34	123
4	2.22	40	4.05	2.18	21	4	2.18	125	4	2.20	38
						5	2.16	5			
a	2.88		a	2.75		a	3.01		a	3.11	
b	-0.20		b	-0.19		b	-0.23		b	-0.27	
R^2	0.925		R^2	0.958		R^2	0.967		R^2	0.933	

Table 2: MAL in poems by Erben (texts E1–E12)

peculiarity of the word length structure has on syllable length. The impact may be somewhat neutralized in a relatively long poem like text E7, but text E13 stands out as an exception. Within the current sample, it has the lowest mean syllable length for one- and three-syllabic words and the highest mean syllable length for four-syllabic words.

In Neruda's short stories, the relationship between word length and syllable length also follows the MAL, with R^2 values above the threshold of 0.9 for all eight texts (see Table 4).

The graph in Figure 1 depicts the MAL for texts E6 and N7.

Kytice (text E13)		
WL	MSL	f_{WL}
1	2.50	26
2	2.44	34
3	2.21	27
4	2.34	7
a	2.50	
b	-0.07	
R^2	0.571	

Table 3: MAL in the poem *Kytice* by Erben (text E13)

N1			N2			N3			N4		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	2.99	138	1	2.95	146	1	2.98	124	1	2.95	139
2	2.56	420	2	2.59	392	2	2.56	329	2	2.54	365
3	2.38	289	3	2.39	276	3	2.35	222	3	2.36	250
4	2.33	118	4	2.27	99	4	2.25	87	4	2.35	86
5	2.24	36	5	2.27	24	5.31	2.21	12	5	2.21	21
6	1.81	6	6.17	1.96	6				6	1.84	7
a	2.99		a	2.95		a	2.98		a	2.95	
b	-0.21		b	-0.19		b	-0.20		b	-0.21	
R^2	0.921		R^2	0.955		R^2	0.978		R^2	0.901	
N5			N6			N7			N8		
WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}	WL	MSL	f_{WL}
1	3.08	189	1	3.04	168	1	3.17	212	1	2.94	86
2	2.58	606	2	2.54	496	2	2.57	701	2	2.50	179
3	2.40	407	3	2.37	333	3	2.40	544	3	2.36	93
4	2.28	170	4	2.30	101	4	2.32	211	4	2.16	32
5	2.28	39	5	2.17	24	5	2.18	47	5	2.20	7
6	1.90	8	6.17	1.86	6	6.24	1.90	13			
7	2.12	6									
a	3.08		a	3.04		a	3.17		a	2.94	
b	-0.22		b	-0.23		b	-0.25		b	-0.20	
R^2	0.926		R^2	0.955		R^2	0.968		R^2	0.968	

Table 4: MAL in short stories by Neruda (texts N1–N8)

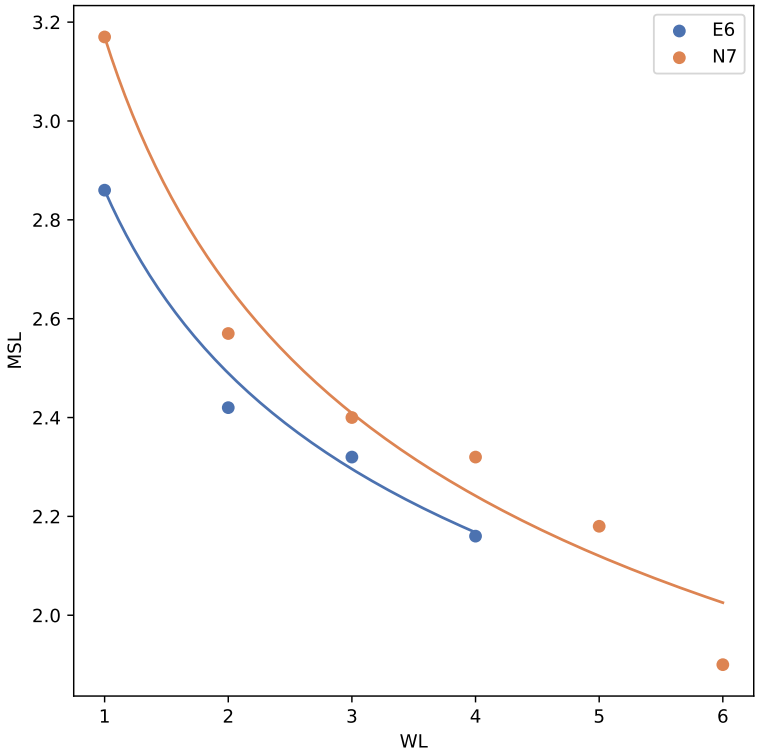

Figure 1: MAL fitted by function (1) for texts E6 and N7

4 Discussion and Conclusion

The results in Section 3 indicate that the MAL is a valid model for the relation between word and syllable lengths not only in prose texts, but also (with one exception, which can be at least partially explained) in poetic texts. Nevertheless, they also reveal several differences between the two groups of texts.[7]

There is a statistically significant difference (a p-value below 0.1) between the values for parameter a, i.e. the mean length of monosyllabic words, in the poetry and prose works. The same is also true of the differences between the mean syllable lengths of two-, three-, and four-syllabic words in the texts by Erben on the one hand and those by Neruda on the other.[8] This suggests a tendency to use shorter words in poetry. At the same time, the difference between parameters b in poems by Erben and short stories by Neruda was not significant. This means that mean syllable length decreases with increasing word length at the same

[7] All data were first tested for normality using the Shapiro-Wilk test. Depending on whether or not the normality hypothesis was rejected, we then applied either the t-test or the Wilcoxon–Mann–Whitney test. The tests were performed in R (www.r-project.org).

[8] Differences in the mean syllable lengths of longer words were not tested because of the insufficient data in these poetic texts.

	Erben				Neruda			
	# types	TTR	a	b	# types	TTR	a	b
# tokens	0.99	-0.74	0.31	-0.19	> 0.99	-0.92	0.94	-0.82
# types		-0.66	0.32	-0.21		-0.92	0.92	-0.81
TTR			-0.44	0.33			-0.82	0.71
a				-0.97				-0.91

Table 5: Pearson correlation coefficients between basic text characteristics and MAL parameters

rate for both groups of texts. However, the curves representing model (1) for prose tend to be higher than the ones for poetry (as can be seen in Figure 1 for two texts).

Another difference relates to the Pearson correlation coefficients between some basic text characteristics (text length in tokens and in types; the TTR) and the parameters of the MAL. In Table 5, we can see that the values of both parameters correlate closely with text length (expressed in terms of both tokens and types) and the TTR in Neruda's short stories. However these correlations are much weaker for the poems by Erben. The correlation between parameters a and b is very strong in all of the texts. These correlations hint towards an interpretation of the MAL parameters for prose texts. The dependence of parameter a (i.e. the mean length of monosyllabic words) on text length conforms with the findings of Kelih (2012), who showed that the mean length of word types increases with increasing text length.

The far weaker correlations between MAL parameters and text length in poems may be explained by the tendency of the words in poems to have shorter syllables than those in prose texts. Shorter syllables have a higher proportion of vowels, which may relate to attempts to achieve euphonic effects.[9] If an author (poet) deliberately opts for shorter syllables,[10] then since we assume a synergetic model of language such as the one suggested by Köhler (2005) where all linguistic properties are interrelated, this choice will be reflected in other changes. One such change may be that text length has a far weaker influence, and other "language forces" (Altmann and Köhler 1996) come into play. Identifying these forces is one of our future challenges. In addition, the MAL in poetic texts may be analyzed at other levels, and phenomena that do not appear in prose (e.g. verses and stanzas) emerge as new candidates for reasonable units. Precisely which properties distinguish poems from other genres remains to be seen.

[9] The idea that the proportions of vowels and consonants differ between prose and poetry can be traced back to at least the second half of the 15th century. A 1830 commentary explicitly mentions that euphony depends partly on this proportionality, and it specifies the values of the proportions for several languages. See Grzybek (2013) for more details.

[10] The words in their poems will, thus, also be shorter than those in prose in terms of speech sounds.

Acknowledgments

J. Mačutek's work on this paper was supported by VEGA grant no. 2/0096/21.

References

Altmann, Gabriel (1980). "Prolegomena to Menzerath's law". In: *Glottometrika 2*. Ed. by Rüdiger Grotjahn. Bochum: Brockmeyer, pp. 1–10.

Altmann, Gabriel and Reinhard Köhler (1996). "'Language forces' and synergetic modelling of language phenomena". In: *Glottometrika 15*. Ed. by Peter Schmidt. Trier: WVT, pp. 62–76.

Antić, Gordana, Emmerich Kelih, and Peter Grzybek (2006). "Zero-syllable words in determining word length". In: *Contributions to the Science of Text and Language*. Ed. by Peter Grzybek. Dordrecht: Springer, pp. 117–156.

Buk, Solomija and Andrij Rovenchak (2007). "Statistical parameters of Ivan Franko's novel *Perekhresni stežky (The Cross-Paths)*". In: *Exact Methods in the Study of Language and Text*. Ed. by Peter Grzybek and Reinhard Köhler. Berlin, New York: de Gruyter, pp. 39–48.

Čech, Radek (2015). "Text length and the lambda frequency structure of the text". In: *Sequences in Language and Text*. Ed. by George K. Mikros and Ján Mačutek. Berlin, Boston: de Gruyter, pp. 71–88.

Čech, Radek (2016). *Tematická koncentrace textu v češtině*. Praha: UFAL.

Čech, Radek, Ioan-Iovitz Popescu, and Gabriel Altmann (2011). "Word length in Slovak poetry". In: *Glottometrics* 22, pp. 44–56.

Cramer, Irene M. (2005). "Das Menzerathsche Gesetz". In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 659–688.

Crystal, David (2008). *A Dictionary of Linguistics and Phonetics*. Malden (MA): Blackwell.

Gerlach, Rainer (1982). "Zur Überprüfung des Menzerath'schen Gesetzes im Bereich der Morphologie". In: *Glottometrika 4*. Ed. by Werner Lehfeldt and Udo Strauss. Bochum: Brockmeyer, pp. 95–102.

Grzybek, Peter (2013). "Historical remarks on the consonant-vowel proportion — from cryptoanalysis to linguistic typology. The concept of phonological stoichiometry (Francis Lieber, 1800-1872)". In: *Glottometrics* 26, pp. 96–103.

Kelih, Emmerich (2010). "Parameter interpretation of Menzerath law: Evidence from Serbian". In: *Text and language. Structures, Functions, Interrelations, Quantitative Perspectives*. Ed. by Peter Grzybek, Emmerich Kelih, and Ján Mačutek. Wien: Praesens, pp. 71–79.

Kelih, Emmerich (2012). "On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts". In: *Synergetic Linguistics. Text and Language as Dynamic Systems*. Ed. by Sven Naumann, Peter Grzybek, Relja Vulanović, and Gabriel Altmann. Wien: Praesens, pp. 67–80.

Köhler, Reinhard (2005). “Synergetic linguistics”. In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 760–775.
Mačutek, Ján, Radek Čech, and Jiří Milička (2017). “Menzerath-Altmann law in syntactic dependency structure”. In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Ed. by Simonetta Montemagni and Joakim Nivre. Linköping: Linköping University Electronic Press, pp. 100–107.
Mačutek, Ján, Jan Chromý, and Michaela Koščová (2019). “Menzerath-Altmann Law and prothetic /v/ in spoken Czech”. In: *Journal of Quantitative Linguistics* 26.1, pp. 66–80. DOI: 10.1080/09296174.2018.1424493.
Mačutek, Ján and Andrij Rovenchak (2011). “Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length”. In: *Issues in Quantitative Linguistics 2*. Ed. by Emmerich Kelih, Victor Levickij, and Yuliya Matskulyak. Lüdenscheid: RAM-Verlag, pp. 136–147.
Mačutek, Ján and Gejza Wimmer (2013). “Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics”. In: *Journal of Quantitative Linguistics* 20.3, pp. 227–240. DOI: 10.1080/09296174.2013.799912.
Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.
Mitchell, David (2015). “Type-token models: A comparative study”. In: *Journal of Quantitative Linguistics* 22.1, pp. 1–21. DOI: 10.1080/09296174.2014.974456.
Wimmer, Gejza (2005). “The type-token relation”. In: *Quantitative Linguistics. An International Handbook*. Ed. by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin, New York: de Gruyter, pp. 361–368.
Wimmer, Gejza, Gabriel Altmann, Luděk Hřebíček, Slavomír Ondrejovič, and Soňa Wimmerová (2003). *Úvod do analýzy textov*. Bratislava: Veda.