Glottometrics 20, 2010, 1-11

The type-token relationship in Slavic parallel texts

Emmerich Kelih[1]

Abstract. The aim of the paper is to analyse the statistical regulation of the type token relationship in Slavic parallel texts. Furthermore it is shown that this relationship in parallel texts can be explained due to morphological and typological characteristics.

Keyords: type-token relationship, Slavic languages, corpus, parallel texts

0. Introduction

Parallel texts are a reliable empirical resource for the cross-linguistic analyses of typological and morphological features of languages. One seemingly trivial characteristic of parallel texts is text length. The main purpose of this paper is to show that text length of parallel texts is a property that can be used for the measurement of the typological "closeness" of languages. Furthermore, an attempt to model the relation between text length (number of tokens) and vocabulary size (number of types), including a typological interpretation of parameters derived from the model used, is offered. The empirical data base consists of the translations of the Russian novel "How the steel was tempered" (N.A. Ostrovskij) into eleven Slavic Standard Languages (Slovenian, Croatian, Serbian, Bulgarian, Macedonian, Slovak, Czech, Polish, Upper-Sorbian, Belarusian and Ukrainian).

1. Parallel Text Corpora and the Type Token Ratio

1.1. Text material and linguistic input

For the examination of the relationship between text length and vocabulary size, a multi-parallel corpus was built, which contains ten chapters of the Russian novel "Kak zakaljalas' stal'/How the steel was tempered" in twelve Slavic Standard Languages, i.e., Russian, Slovenian, Croatian, Serbian, Bulgarian, Macedonian, Ukrainian, Belarusian, Polish, Upper-Sorbian, Czech and Slovak. The novel was written by N.A. Ostrovskij in the years 1932-1934. For details about the translations used and general linguistic problems of parallel corpora research cf. Kelih (2009a, 2009b).

The calculation of text length in terms of the number of tokens (*N*) and the number of types (V) was performed by means of specific software (Wordsmith 5.0). Linguistically, the word form is defined using purely orthographical criteria, e.g. every sequence of letters/ graphemes between two blanks[2] (for further details cf. Antić/Kelih/Grzybek 2006) is treated as

[1] Address correspondence to: emmerich.kelih@uni-graz.at .

[2] The orthographical criterion is considered to be the simplest definition of the word and word form respectively. For a detailed and illuminating discussion of the "graphematical" word in German cf. Fuhrhop (2008). In Kelih (2007) it has been shown that the use of different word definitions (based on

one word form. The hyphen is understood as punctuation mark which has a delimitative function. The number of tokens (*N*) is the total number of word forms occurring in the whole text. Not taking into account the frequency of word forms in the text, we obtained the number of word form types (*V*), which are sometimes called tokemes (cf. Andersen, Altmann 2006). The texts are thus analysed without any lemmatisation, i.e. the analysis focuses on the lexical level, including morphological and morphosyntactical information.

1.2. Description: Number of Tokens (N) vs. number of Types (V) in translated texts

The analysis of the text length, i.e. the number of tokens and types respectively, of parallel corpora has so far not been in the focus of quantitative and synergetic linguistics. Leaving aside practical problems concerning the calculation of a translator's remuneration based on text length, it is also of interest from a theoretical point of view. Without a full list of text length related linguistic problems, one has to consider at least two different influence factors:

1. According to contemporary translational studies, one of the main strategies of translation is the simplification of the translated text. Empirically, this can be seen not only at the syntactical level (e.g. shorter sentence length in the target text than in the source text), but on the lexical level too. In this regard, a smaller number of tokens and types of the target texts can be understood as a lexical simplification (cf. Baker 1995, 1996). Furthermore, the well-known translators' strategy of explicitation, e.g. the process of rendering information that is only implicit in the source text explicit in the target text (cf. Frankenberg-Garcia 2009: 48), plays a crucial role in text length. It is reasonable to assume that explicitation produces the effect that target texts are longer than the source text (cf. Teich 2003). Therefore in parallel text research a priori simplification as well as explicitation has to be taken into account.
2. For a comparison of the text length in source and target texts, morphological and morphosyntactical characteristics of the languages involved nevertheless have to be taken into consideration. Theoretically, no absolute 1:1 correspondence of the types and tokens level between two languages, but rather deviations from this correspondence, have to be assumed. This will now be demonstrated on the basis of the available translations of the Russian source text in 11 Slavic languages.

First, a brief analysis of the text length of our parallel texts is offered, measured in terms of the token numbers (*N*). Cf. Table 1 with an overview of the different text lengths in translated Slavic texts.

Table 1
Types (*V*) and Tokens (*N*) of „How the steel was tempered"

Language	N	V
Russian	49672	15053
Ukrainian	49612	14645
Belarusian	49874	14814
Slovak	52093	14025

orthographical, orthographical-phonetical and phonological criteria) leads to a systematic shift of the calculated word length and related empirical parameters.

Czech	52180	14136
Polish	52735	14979
Upper-Sorbian	58480	14465
Slovenian	62646	13940
Serbian	56227	13637
Croatian	56415	13830
Bulgarian	57165	12303
Macedonian	58819	11437

Let us start with a comparison of text length (*N*) between the source text and the languages that are traditionally treated as the members of the South Slavic group. While the Russian source text contains 49672 tokens, the Slovenian translations – the language with the highest number of tokens – has over 62600 tokens, i.e. the Slovenian text has approximately 12000 tokens (= 26%) more than the Russian source text. A relatively similar picture is to be obtained if one compares the Russian source text with another south Slavic language, such as, Macedonian, in which the translations consist of 58819 tokens; similarly large differences can be obtained in a comparison of the Russian texts with the Serbian and Croatian translations, which have approximately a 12% higher text length than the source text; for Bulgarian it is even slightly more (15%).

The same, i.e. translated texts being longer than the source, holds true for the Western Slavic languages: All of these (Slovak, Czech, Polish, Upper-Sorbian) are longer than the original Russian text, e.g. the Slovak, Czech and Polish translations of ”How the steel was tempered“ have approximately 2600 tokens more than the Russian one; only the Upper-Sorbian translation with its 58480 tokens has a behaviour similar to that of South Slavic languages in this regard.

Furthermore a phenomenon was observed that is even more important than these cross-linguistic comparisons, namely that the text length of genetically close languages is almost the same: For instance, the Eastern Slavic languages (Russian, Ukrainian, and Belarusian) share approximately the same text length, both in respect to the number of types and tokens. The difference in relation to Russian of 63 tokens (Ukrainian) and 335 tokens (Belarussian) is relatively marginal, i.e. no striking differences in text length are to be obtained. The analysis of text length of Western Slavic languages yields a relatively similar picture: For Slovak, Polish and Czech a more or less similar text length (≈ 52300 Tokens) was obtained. This especially holds true for Slovak and Czech, with a difference of only 82 tokens. The genetically “very close” languages Croatian (56424 tokens) and Serbian (56230 tokens) again do not show any notable differences whatsoever; similar small differences regarding the number of tokens can be obtained for the Bulgarian and Macedonian translations.

Generally, it can be seen that the text length (*N*) appears to be a rather robust characteristic of parallel texts (in Slavic languages). Interestingly enough, it roughly coincides with the areal and geographical affiliation of the languages under examination. Only the Slovenian and the Upper-Sorbian translations show a somehow different behaviour in this respect: The Slovenian text – the longest text with 62646 tokens – is slightly above the south Slavic average (Croatian, Serbian, Bulgarian, Macedonian) with a mean text length of approximately 58000 tokens. The same holds true for the Upper-Sorbian text (58480 tokens), which is “above” the West Slavic average of approximately 53000 tokens.

The differences obtained in text length are not caused by stylistic preferences of the translators or by an over-explicitation of the Russian original text (cf. Kelih 2009a, 2009b), but by the typological, especially morphological or morphosyntactical, features of the languages under examination. This will be discussed in more detail in 1.3.

Similar differences in text length are obtained at the type-level (*V*) too, albeit to a lesser extent than at the token level. In this respect, the Russian text has the largest number of types (= 15053), whereas the Slovak and Slovenian translations have approximately the same *V* (approx. 14000 types), i.e. 1000 types less than the Russian source text. Only the Macedonian and the Bulgarian texts are, with respect to the number of types in relation to other Slavic languages, clearly outliers, inasmuch as they have "only" 11437 and 12327 types respectively. This can in particular be reasonably explained by the commonly known extremely morphologically limited case flexion system of these two south Slavic languages.

Finally, the relation between types and tokens has to be analysed. As can be seen from figure 1, there is no particular (strong) statistical correlation between the number of tokens (*N*) and types (*V*) in Slavic languages. Evidently, there are no mechanisms of compensation in the way that a small vocabulary (*V*) of one language is accompanied by a long text length in the number of tokens.

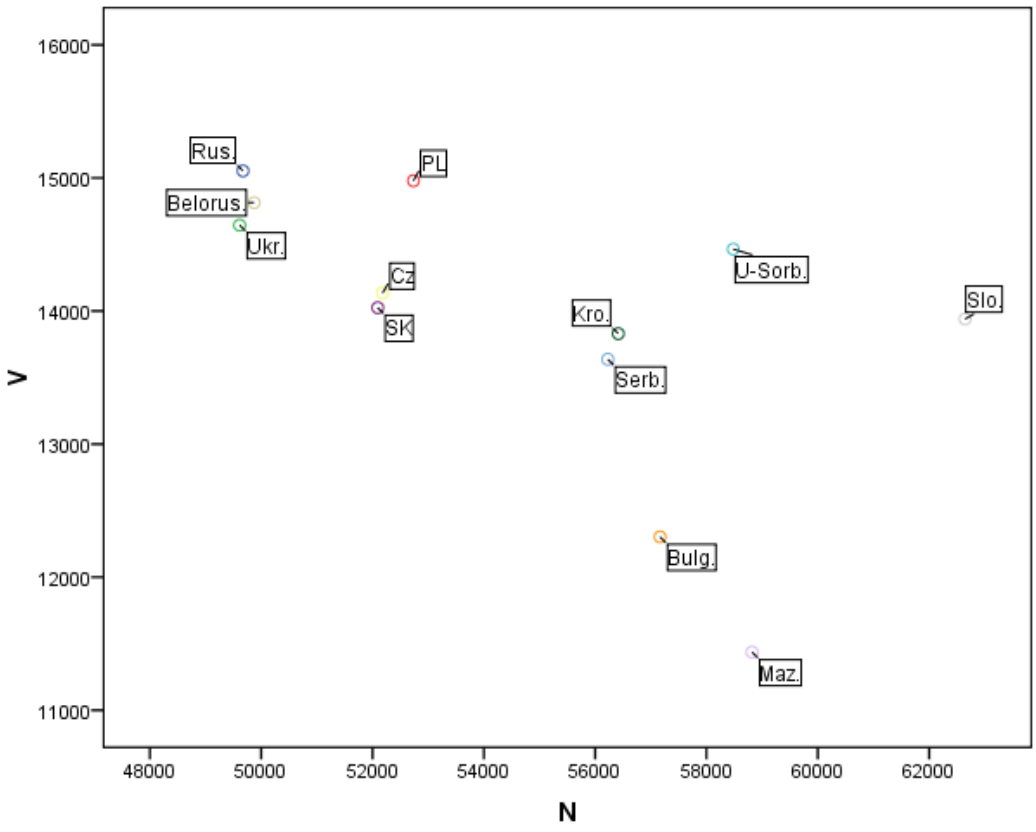

Figure 1: Relation between number of tokens (*N*) and types (*V*) in KZS

Although there are no systematic inter-lingual relations at text length level, one more analysis has to be performed: The relation between the number of types (*V*) and tokens (*N*) within the individual chapters of the translated novel, i.e. within the different languages. It is well known from quantitative linguistics that the relation between the number of tokens (*N*) and number of types (*V*) is systematically organised and controlled by a statistical mechanism, which has to be analysed in two respects: On the one hand in regard to the intra-textual regulation in the various chapters of the novel, and on the other hand again cross-linguistically, i.e. the control of types and tokens, in dependency of the morphological characteristics (productivity of the inflection system, morphological expression of the reflexivity, analytic or synthetic expression of temporal forms etc.) of the languages.

1.3. Modelling Text length (*N*) and Vocabulary Size (*V*): Intra-textual perspective

For the modelling of the relation between text length (*N*) and vocabulary size (*V*) many different continuous and discrete models have been proposed in the past. It is beyond the scope of this paper to present and discuss these different models in detail (cf. Altmann/ Altmann 2008: 107ff. and Fan (2008) for an extensive overview). Generally, not only is the relation between text length (*N*) and vocabulary size (*V*) discussed, but different indices and coefficients are derived from these two text characteristics such as the Type Token Ratio (TTR), calculated as *V*/*N* or *N*/*V*. These indices are usually understood as measurements of the vocabulary richness and style (cf. Herdan 1966: 77) of a text. However, the TTR clearly depends on the type of the analysed texts and the text length (cf. Tuldava 1974, 1993, 1995), and thus an interpretation of the TTR as a stylistic feature must be approached with caution.

A slightly different interpretation of the TTR – which is indirectly related to the absolute number of types and tokens too – is offered by Altmann/Altmann (2008: 107), who understands the TTR as a measurement of the information flow within one particular text. The more often one token is repeated within one text/chapter, the slower the spread of new information. It goes without saying that again this information regulation is text type specific and clearly dependent on the length[3] of the texts under examination, but nevertheless this interpretation is much more reasonable than the stylistic one offered above.

However, both interpretations of TTR, i.e. as the control of the information flow within one text and as the measurement of the vocabulary/stylistic richness of one text, seem to fail in the analysis of parallel texts. Differences in the number of types and tokens between the source text and the target text can, at least for the KZS corpus, be explained mainly[4] by morphological and morphosyntactical features of the translated languages, as shown below.

Thus, from our point of view the number of types and tokens in parallel-text research represents the morphological productivity of the languages under examination: The more productive the inflection system (especially of verbs and nouns), the greater the number of different morphological forms that are generated by the language system. For a language that has a highly productive flexion system, the probability of the appearance of the same word form type seems to be lower than for a language with a less productive flexion system. Thus, there must be a direct impact on the number of types and tokens of translated texts.

Furthermore, differences in text length at the type and token level in parallel-text research can, as already pointed out by Popescu/Altmann (2008: 371) and Popescu/ Mačutek/Altmann (2009: 99), be connected with the discussion of the significance of the number of Hapax Legomena for language typology issues, used as an indicator for the degree of analyticity/ syntheticity of the languages under examination. Thus – stylistic variation by the translators has to be excluded for the sake of simplicity – it can be stated that the longer the translated text in relation to the source text, the higher the analytic expression of the morphological and morphosyntactical categories of the target languages. To at least illustrate this with one example, the expression of reflexivity in Slavic languages, for instance, has to be kept in mind: In Russian it is part of an (orthographically defined) word form (e.g. podnjat'sja – to stand up), whereas in Slovenian the reflexive marker appears as a separate word form (vzdigniti se – to stand up); similar phenomena can also be obtained for the analytic and synthetic morphological expression of the temporal system. Without going into much more

[3] A quite similar problem has been discussed in quantitative linguistics in connection with Zipf's law and Zipf's size *U*, a hypothetical length of a text constructed by a speaker. Cf. Orlov (1982) and Orlov/Boroda/Nadarejšvili (1982).

[4] For the sake of simplicity, the well known phenomena of simplification and explication of translated texts are not taken into consideration here.

detail, it is now quite clear that the analysis of the number of types and tokens comes far behind the interpretation of the information flow and the problem of vocabulary richness of texts.

To give a more elaborate insight into the statistical behaviour of the relation between the number of types and tokens (and not of the TTR in general), the intra-textual behaviour now has to be analysed and modelled. For all the translations, the dynamic relationship between *V* and *N* of the entire novel in twelve languages was computed at a chapter-spaced interval. The results are in Table 2.

Table 2
Type and Token in cumulated chapters: 12 Slavic languages

	Russian		**Ukrainian**		**Belarusian**		**Slovak**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	4107	1907	4119	1895	4145	1916	4275	1895
1-2	8243	3538	8279	3491	8322	3524	8600	3472
1-3	14567	5591	14561	5498	14689	5536	15096	5405
1-4	18300	7003	18325	6865	18480	6890	18981	6701
1-5	22070	7956	22080	7751	22271	7802	22843	7568
1-6	29604	9822	29622	9533	29818	9646	30864	9260
1-7	35624	11388	35621	11061	35881	11217	37201	10677
1-8	40976	12864	40983	12534	41243	12674	42982	12020
1-9	44270	13635	44261	13309	44555	13447	46394	12735
1-10	49672	15053	49612	14645	49874	14814	52093	14025
	Czech		**Polish**		**Upper-Sorbian**		**Slovenian**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	3925	1773	4348	1970	4851	1976	5209	1955
1-2	8306	3375	8716	3625	9663	3547	10408	3490
1-3	14976	5360	15410	5668	17085	5496	18379	5420
1-4	18896	6693	19413	7084	21541	6884	23166	6737
1-5	22748	7559	23410	8010	25813	7768	27886	7565
1-6	30865	9295	31347	9849	34608	9506	37432	9188
1-7	37255	10717	37695	11399	41666	10992	44952	10624
1-8	42993	12086	43448	12839	47982	12406	51744	11990
1-9	46444	12840	46949	13600	51832	13157	55849	12696
1-10	52180	14136	52735	14979	58480	14465	62646	13940
	Serbian		**Croatian**		**Bulgarian**		**Macedonian**	
chapter	**N**	**V**	**N**	**V**	**N**	**V**	**N**	**V**
1	4579	1899	4582	1900	4653	1709	4810	1636
1-2	9235	3435	9271	3458	9387	3127	9708	2944
1-3	16328	5330	16431	5386	16611	4807	17178	4509
1-4	20618	6639	20747	6718	20916	5991	21602	5619
1-5	24859	7494	25002	7586	25193	6710	26027	6295
1-6	33425	9085	33555	9169	33866	8170	34941	7624
1-7	40241	10436	40396	10574	40858	9421	42094	8788
1-8	46270	11746	46471	11913	47100	10605	48508	9873
1-9	50019	12427	50231	12613	50887	11202	52358	10421
1-10	56227	13637	56415	13830	57165	12303	58819	11437

For the statistical modelling of the relation between types and tokens, the power model $V = aN^b$ was used. As a result, this model is appropriate without any exception for all languages analysed here, attaining in all cases a $R^2 > 0.99$. A graphical representation of the modelled interrelation of four selected Slavic languages is given in Figures 1a-1d. Table 3 shows the goodness-of-fit results and the parameters *a* and *b* of the model used.

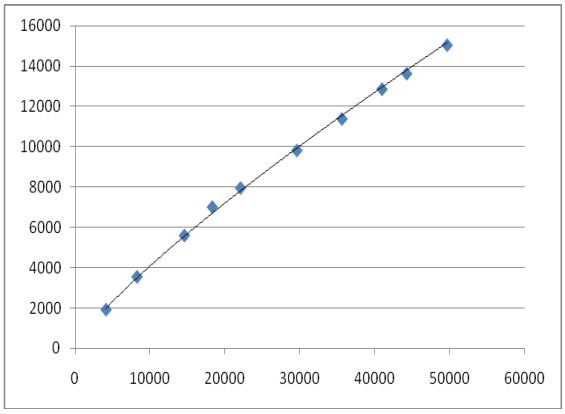

Figure 1a.: Russian: *N* vs. *V*

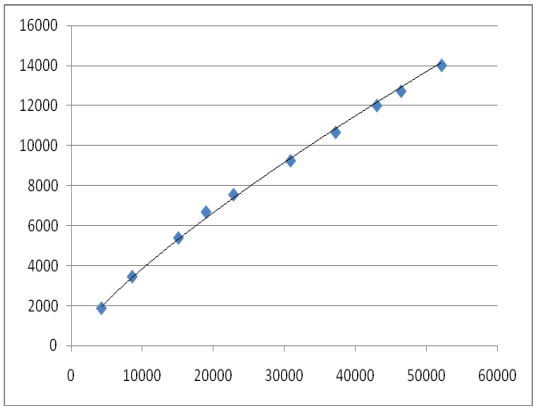

Figure 1b: Slovak: *N* vs. *V*

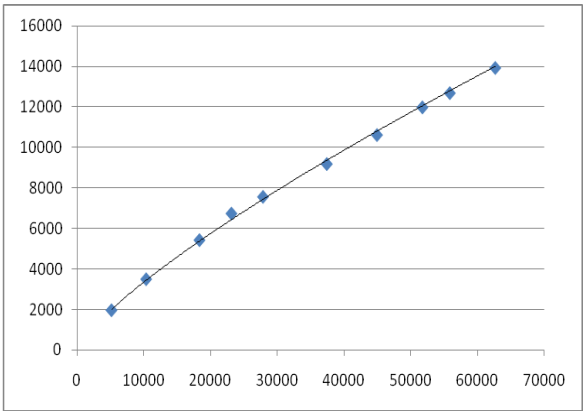

Figure 1c: Slovenian: *N* vs. *V*

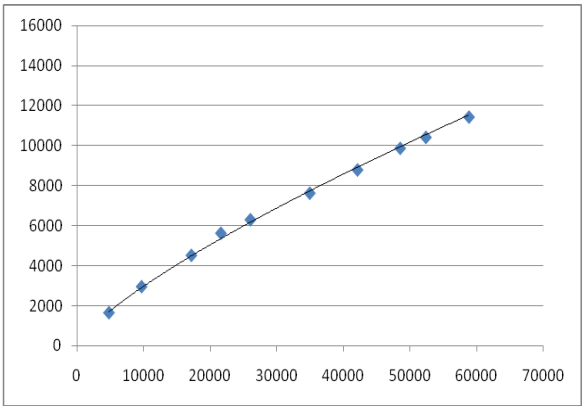

Figure. 1d: Macedonian: *N* vs. *V*

Table 3
Goodness of fit and parameters

Language	Parameter b	Parameter a	R^2
Russian	0.7974	2.6990	0.9993
Slovak	0.7630	3.5115	0.9991
Slovenian	0.7636	3.0114	0.999
Macedonian	0.7454	3.1736	0.999
Czech	0.7678	3.3560	0.9991
Serbian	0.7520	3.6417	0.999
Croatian	0.7569	3.4909	0.9989
Bulgarian	0.7514	3.2640	0.9990
Belarusian	0.7980	2.6274	0.9994
Ukrainian	0.7932	2.7479	0.9992
Polish	0.7810	3.0616	0.9993
Upper-Sorbian	0.7753	2.9102	0.9993

As can be seen in Figure 2, which displays the dependence of the combined theoretical values of *V* on the number of tokens (*N*) in five Slavic languages (Russian, Slovak, Slovene, Macedonian and Upper-Sorbian), control of text length (*N*) and the vocabulary size (*V*) are evidently strongly determined by the morphological status of the language. There is no overlap of the analysed languages, and the supposed typological relevance in the comparison of text length of translated texts/languages can be shown empirically. A striking characteristic of the relation of types and tokens is the different increase of the fitting curves, which again can be interpreted as important typological information: The slower the increase of the curve, the more analytic the language.

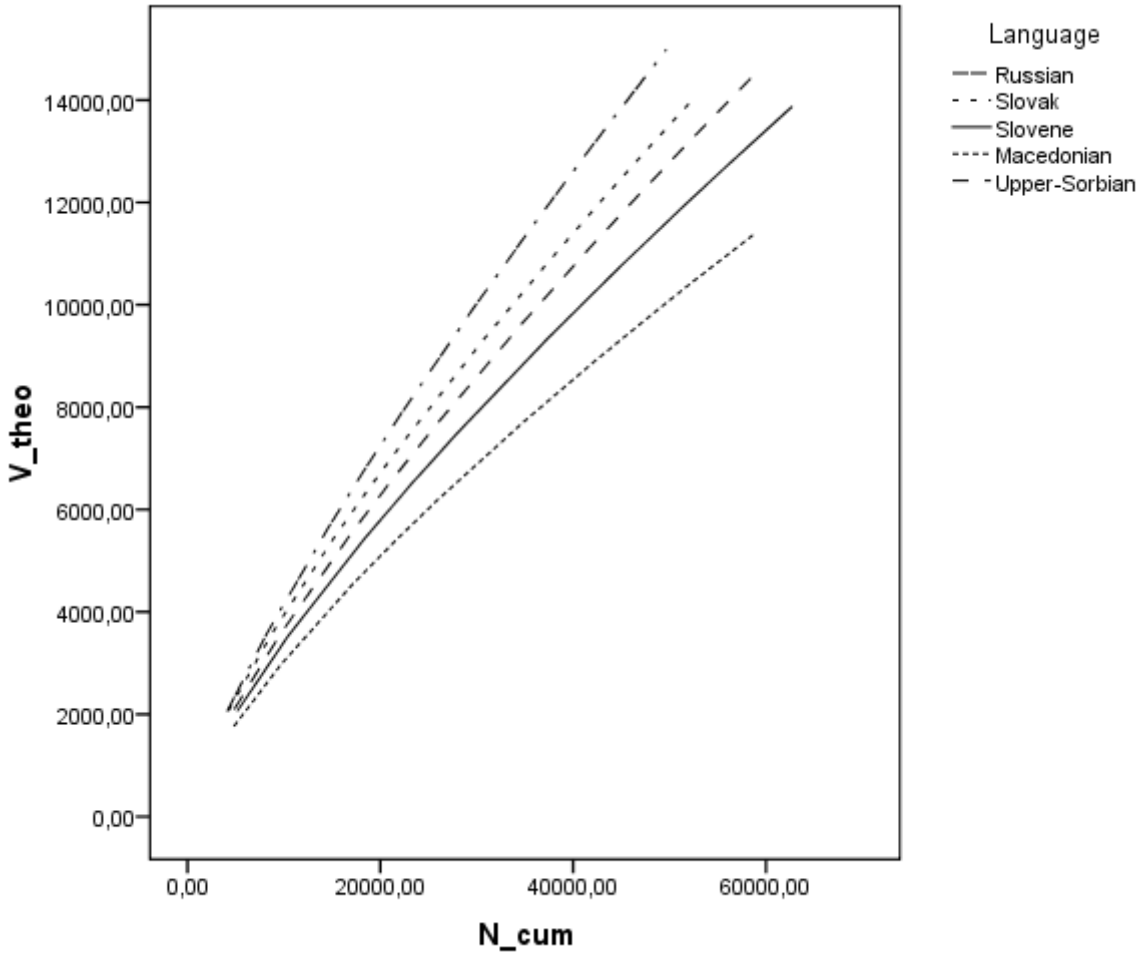

Figure. 2. Interrelation between *N* und *V* (theoretical values) in five Slavic languages

Furthermore, the conclusion can be drawn from Figure 2 that with increasing text length the discrepancy, i.e. the distance between the analysed languages (= texts) becomes wider and wider. In other words, to reach maximal comparability in parallel corpora research, the texts that are analysed should not be too short.

1.3.1. Significant difference in the increase?

Having found a simple and appropriate model for the relation between number of tokens (*N*) and number of types (*V*) for different Slavic parallel texts, it can finally be tested whether there is a significant difference in the increase of the fitting curves, i.e. of the regression coefficients. To do so, the abovementioned model is linearised by logarithmisation and transformed into *log(V) = log(a) + blog(N)*, and it is tested whether there are significant differences in the regression coefficients. Cf. Zöfel (2002: 146) for details of the procedure used.

In Figure 3 the calculated regression coefficients of the analysed texts (= languages) are plotted in ranked order. Russian is at the upper most level and Macedonian at the lower most level. In particular, there is no need for complicated and systematic comparisons between the texts, but rather in a first step it is sufficient to compare the languages, with the maximal and minimal regression coefficients only.

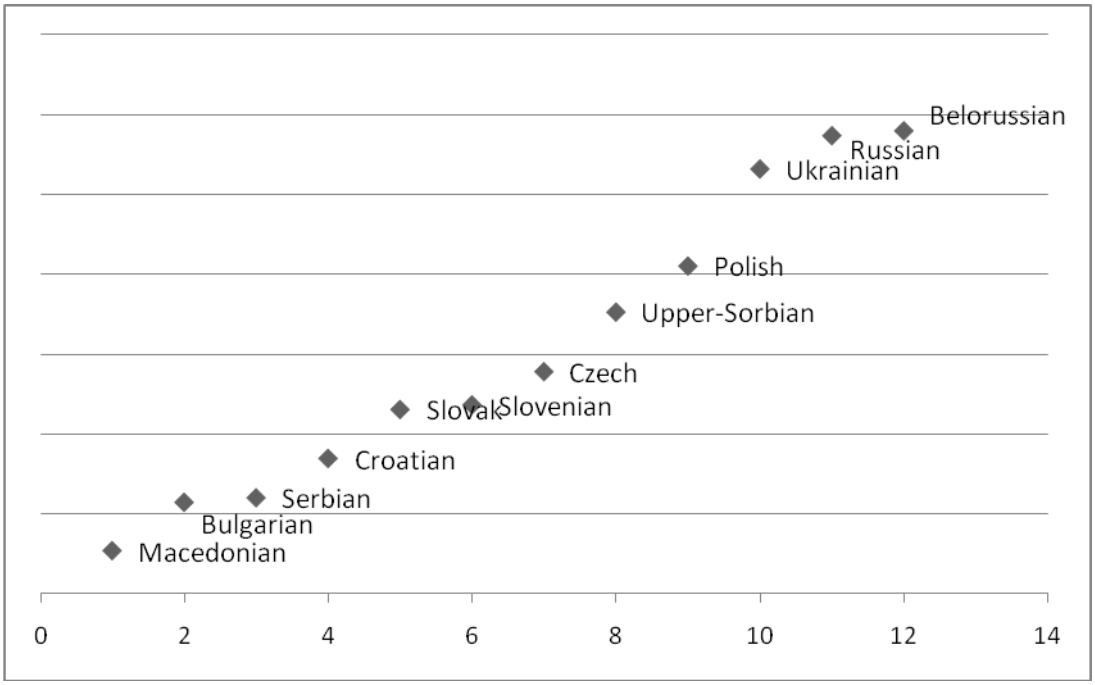

Figure 3. Regression coefficient *a* (ranked values)

Indeed, the comparison of the Russian and Macedonian coefficients shows no significant differences ($p > 0.05$). As long as there are no statistical significant differences between these two turning points, the same results are naturally to be obtained for other languages. Cf. Table 3 with some selected comparisons, for instance Russian-Bulgarian with a $p = 0.9961$, Russian-Slovak with a $p = 0.9973$, etc.

Table 4
Results of *t*-distributed test statistics (selected languages)

pairs of comparison		t-value	DF	p
Russian	Macedonian	0.0008	26486	0.9993
	Bulgarian	0.0048	27352	0.9961
	Slovak	0.0033	29074	0.9973
	Slovene	0.0043	28989	0.9965
	Belarusian	0.0007	29863	0.9994

Interestingly enough, there are, in general, no significant differences in the increase of the relation between text length (*N*) and vocabulary size (*V*) obtained above. Nevertheless, the order of the regression coefficients ranked above is particularly interesting for language typology: Russian, Belarusian, Polish, Upper-Sorbian, Czech, Slovak, Slovene, Croatian, Serbian, Bulgarian and Macedonian. This rank order roughly represents the degree of syntheticity/ analycity of the Slavic texts/languages under examination. It starts with synthetic Russian and ends with the analytic south-eastern Slavic languages, i.e. Macedonian and Bulgarian. For the time being, it is not known whether or not this typological order can also be obtained for other Slavic parallel text corpora. For future research it is necessary to take also other parameters into consideration, such as, the number of Hapax Legomena, the productivity of the nominal flexion, etc.

2. Summary

Let us summarize the results of the present study. The relation between text length (*N*) and vocabulary size (*V*) in parallel text research is considered to be an efficient tool for the analysis of morphological features of the examined texts from 12 Slavic languages. Furthermore, it has been shown that selective parameters of appropriate models capturing the relation between *V* and *N* of the parallel Slavic texts can be interpreted as the degree of analyticity/ syntheticity of the analysed texts/languages. Nevertheless, further parallel texts have to be analysed in this regard.

References

Altmann, V.; Altmann, G. (2008). *Anleitungen zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM-Verlag..

Andersen, S.; Altmann, G. (2006). Information content of words in text. In: Grzybek, P. (ed.), *Word length studies and related issues: 93-117*. Boston: Kluwer.

Baker, Mona (1995). Corpora in translation studies: An overview and some suggestions for future research. In: *Target 7(2), 223–245.*

Baker, Mona (1996). Corpus-based translations studies: The challenge that lie ahead. In: Somers, Harold (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager; 175-186*. Amsterdam: Benjamins.

Fan, Fengxiang (2008). A Corpus-Based Study on Random Textual Vocabulary Coverage. *Corpus Linguistic and Linguistic Theory 4(1), 1–17.*

Fuhrhop, Nanna (2008). Das graphematische Wort (im Deutschen): Eine erste Annäherung. *Zeitschrift für Sprachwissenschaft 27, 198-228.*

Frankenberg-Garcia, Ana (2009). Are Translations Longer Than Source Texts? A Corpus-Based Study of Explicitation. In: Beeby, Allison; Rodríguez Inés, Patricia; Sánchez-Gijón, Pilar (eds.), *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate: 47-58*. Amsterdam: Benjamins.

Kelih, E. (2007). Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen. In: Kaliuščenko, Volodymir; Köhler, Reinhard, Levickij, Viktor (eds.), *Problems of Typological and Quantitative Lexicology: 91-105*. Chernivtsi: Ruta..

Kelih, E. (2009). Slawisches Parellel-Textkorpus: Projektvorstellung von „Kak zakaljalas' stal' (KZS)". In: Kelih, E.; Levickij, V.V.; Altmann, G. (eds.), *Methods in Text Analysis:* 106-124.. Chernivtsi: Ruta.

Orlov, Ju.K. (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Guiter, H.; Arapov, M.V. (eds.*), Studies on Zipf's law*: *154-233.* Bochum: Brockmeyer.

Orlov, Ju.K.; Boroda, M.G.; Nadarejšvili, I.Š. (1982), *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Popescu, Ioan-Iovitz; Altmann, Gabriel (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics 15(4), 370-378.*

Popescu, I.-I.; Mačutek, J.; Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Teich, Elke (2003). Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts. Berlin, New York: Mouton de Gruyter.

Tuldava, Ju.A. (1974). O statističeskoj strukture teksta. In: *Sovetskaja pedagogika i škola* 9, *5-33*. [English translation in: Tuldava, Ju. (1995): *Methods in Quantitative Linguistics*. Trier: WVT.]

Tuldava, J. (1993). The statistical structure of a text and its readability. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 215-227.* Trier: WVT.

Tuldava, Ju.A. (1995). The ratio of words forms and lexemes in texts. In: Tuldava, Ju. (1995), *Methods in Quantitative Linguistics*. Trier: WVT.