Preliminary Analysis of a Slavic Parallel Corpus

Emmerich Kelih

Institut für Slawistik, University of Graz, Austria

Abstract. The focus of this paper is on a detailed description of a newly-developed parallel corpus of Slavic languages. It consists of 11 Slavic translations of the well-known Russian socialist realist novel "Kak zakaljalas' stal'/How the steel was tempered" (KZS), written by N.A. Ostrovskij in the years 1932-34. The KZS contains the Slovene, Croatian, Serbian (ekavian), Macedonian, Bulgarian, Ukrainian, Belorussian, Slovak, Czech, Polish and Upper Sorbian translations. Thus, for the first time a parallel text of almost all Slavic standard languages is available. In addition to the discussion of some text-specific issues of KZS, an explorative statistical analysis and a linguistic interpretation of text length and the Type-Token Ratio is offered.

1 Introduction

Parallel texts and parallel text corpora play a crucial role in corpus linguistics, linguistic typology and text linguistics. A parallel text is a text or part of a text placed alongside its translation in one or many[1] languages ([19, 47ff.], [17, 121ff] and [25, 73]). Parallel corpora are explored in general linguistics and language processing. With respect to Slavic languages in particular, the well-known "Multext East" project (cf. [6] and ([7]), which contains many translations of George Orwell's "1984" into Slavic languages (Bulgarian, Croatian, Czech, Resian, Russian, Serbian and Slovene) and the ambitious "Regensburg Parallel Corpus of Slavic Languages" [28], which includes different translated texts from and into Russian, Belorussian, Croatian, Serbian, Slovak, Czech and Ukrainian, have to be mentioned.

According to our knowledge, however, despite the availability of these "larger" parallel corpora focussed on Slavic languages, no parallel text corpus[2] with one original text in a large number of Slavic standard languages exists. To overcome this deficiency, a new parallel text corpus, containing in sum 11 Slavic translations of the well-known Russian socialist realist novel "Kak zakaljalas' stal'/How the steel was tempered" (hereafter KZS) has been compiled by the author for a systematic cross-linguistic quantitative analysis of Slavic languages from a synergetic and quantitative point of view ([15] [16]). In our paper only a few selected problems can be outlined. Firstly, a short overview of

[1] [5, 95] recently introduced the term "massively parallel texts", defined as a huge text corpus with translations preferably in many (genetically) diverse languages, such as, protocols of the European Parliament in over 30 languages, translations of the "Universal Declaration of Human Rights" and of the Bible into more than 100 languages, etc.

[2] For more parallel-corpora projects dealing with Slavic languages cf. [28, 123] and [8], with a project on parallel word lists of Western Slavic languages. [23] mention different available – not solely Slavic – translations of "Le Petit Prince".

linguistic applications of parallel texts is given; secondly, a description of the project in progress is presented in detail; and finally the initial results of a preliminary statistical analysis of the sample size and the Type-Token Ratio of the translations is given.

2 Parallel texts and their linguistic applications

From a linguistic point of view,[3] parallel texts are an important empirical data base. They can be used for many linguistic purposes, such as:

1. For typological and cross-linguistic analyses on the phonological, morphological, syntactical and lexical level (cf. [11], [27], [5] and [29]); in particular, parallel texts are a valuable basis for the study of quantitative features of language cf. ([2, 63-64]).
2. For the analysis of the quality and the linguistic structure of translations (cf. [1], [11], [12], [9], [20] and [24]).
3. For the examination of hypotheses from quantitative and synergetic linguistics. Parallel texts provide a deeper understanding of self-regulation mechanisms of translated texts; in particular, it is not known whether these special kinds of languages abide by language laws, such as the Zipf and the Menzerath law.
4. For comparative text linguistic analyses, including the investigation of stylistic features.
5. For the study of inter-lingual readability – a research field that thus far has hardly made use of parallel texts.

Despite this broad and appealing spectrum of various applications of parallel texts, some remarkable arguments have been raised in the past concerning the limited linguistic and methodological usefulness of parallel texts have been raised. Sometimes it is argued that translated texts lack "authenticity" and "quality" (cf. [22, 102], [29, 128]). Thus some kind of "unnaturalness" of translated texts is postulated and intensively discussed in corpus linguistics as the problem of "translationese" (cf. [18] [19, 49]). This problem is hardly to be avoided in general, but except for this legitimate objection, the outstanding attractiveness of parallel texts lies in the comparision of semantically-identical or nearly-identical texts from different languages cf. [29, 130]. Furthermore parallel texts are at least written in a grammatically and morphologically "correct" way, and hence they are, depending on the examined linguistic hypotheses, furthermore an adequate and powerful empirical data base for cross-linguistic studies.

3 "How the steel was tempered (KZS)" in twelve Slavic languages

The KZS is the empirical database of an ongoing research project on quantitative phonology, namely the investigation of interrelations between the size of phoneme inventories,

[3] The paper focuses on a description of some basic quantitative features of the KZS, and no further language processing issues are discussed. For further information about sentence alignment and POS-Tagging of parallel texts cf. [27].

phoneme frequency, distribution and syllable structure in Slavic languages. In view of the lack of available and accessible parallel texts in many Slavic languages, the Belorussian, Ukrainian, Czech, Polish, Slovak, Upper-Sorbian[4], Bulgarian, Croatian (ijekavian), Macedonian, Serbian (ekavian) and Slovenian translations of the novel "Kak zakaljalas' stal'/How the steel was tempered" (KZS) were recently collected, scanned and submitted to OCR. All text files were then manually proofread. They are now available as plain text.

The particular choice of this "highly" ideological text, however, is connected with some problematic issues, which should be discussed briefly here. Firstly, the authorship of the novel is disputed; secondly, the influence of reversions by the editors has to be discussed; and thirdly, the base of the Russian original for the translations must be specified.

However, today it is accepted that the novel was doubtlessly written by N.A. Ostrovskij. The final version of the novel was however "checked" by many editors, who "polished" the text from a stylistic, and especially ideological, viewpoint ([10, 121] and [4, 8]). In this respect – as highlighted in the detailed study by [10] – the canonical monographic issue of KZS from 1934 differs slightly from the chapters previously (1932-34) published separately in the Soviet literary journal *Molodaja Gvardija*. But the changes in the monographic issue of 1934 in comparison with the versions published earlier are primarily related to some negligible ideological details, whereas the macrostructure of the novel, e.g. the division into two parts, with nine chapters in each part, remained unchanged. As a rule, the canonical text, i.e. the monograph from 1934 and it translations into the different Slavic languages, were used for the KZS. Cf. [14, 124] with a more detailled list of the scanned translations and further bibliographical information.

Leaving aside the somewhat problematic production history, edition and translation of the KZS, the novel – the literary and aesthetic quality can be considered low – is a quite interesting mixture of different styles. According to [10, 140], for KZS a simple, linear sentence structure, a high frequency of colloquial elements and in part a "declamatory" register ('obščestvennaja reč") is characteristic. Additionally, a high proportion of oral speech, mixed with a few narrative sequences and some (typical) Soviet abbreviations and acronyms (especially in respect of the political "lexicon") are obtainable. Furthermore, the novel is characterised by an "internal heterogeneity": Throughout the novel many poems, diary entries, letters and public announcements can be found. However, the novel is not literarily "deformed" in the strictest sense, but rather a representative example of a literary prose text, written in a style typical for the socialist realist writing of the 30s, in which several sub-registers of written language in the original and translated texts can be analysed linguistically.

As already mentioned, at the macro level the KZS is divided into 18 chapters. Because of the time-consuming procedure of scanning and proofreading, only 10 chapters of 18 (nine chapters of the first part and one chapter of the second part, with a total of approximately 240 printed pages) of the KZS were processed for further analysis. A

[4] Special attention was paid to the Sorbian languages, at least in one of the standard languages. [26] note in their analysis of some Slavic translations of the novel *Harry Potter* that unfortunately no translations of this text are available for Sorbian.

more detailed explanation for this decision of a size limitation goes beyond the scope of the present paper. For our purposes, i.e. a systematic analysis of the quantitative structure (esp. phonological and syllable structure), the material seems to be sufficient and should moreover be understood as case study material.

4 Quantitative characteristics of KZS

This chapter provides some promising results[5] of the analysis of the text length (number of types and tokens) and the Type-Token Ratio (hereafter TTR) of the KZS. It will be shown that a quantitative characteristic as seemingly trivial as the sample size can already provide more detailed information about the morphological structure of the languages under examination. Furthermore it will be demonstrated that the TTR is an appropriate parameter for a typological ordering of Slavic languages.

4.1 Sample size: Number of types and tokens

An important characteristic of parallel texts is the text length, measured by the number of "words"; to be more precise in this context, by the number of tokens and types. Both linguistic entities are here defined as orthographical units of a written text, where the space has the function of a delimitation marker, i.e. all alphabetical signs between two spaces are defined as a token/type. The results of the counts of tokens and types in the KZS are summarised in Table 1 (see p. 179), where the Slavic languages are already arranged according to their genetic/areal affiliation. Fig. 1[6] represents the number of types and tokens.

The noticeable variety of the sample size[7] in the parallel texts, especially in respect of the number of tokens, can probably be explained by different morphological and syntactical characteristics of the Slavic languages. To start with the Eastern Slavic languages (Russian, Ukrainian, Belorussian), it appears that they share approximately the same text size with respect to the number of types and tokens. The difference in relation to Russian of 63 tokens (Ukrainian) and 335 tokens (Belorussian) is relatively

[5] For a comparative analysis of KZS grapheme frequencies cf. [13].

[6] Slo = Slovene, Cro = Croatian, Serb = Serbian, Bulg = Bulgarian, Mz = Macedonian, Sorb = Upper Sorbian, Rus = Russian, Ukr = Ukrainian, Belorus = Belorussian, Cz = Czech, Sk = Slovak and Pol = Polish

[7] The parallel texts have not yet been annotated and tagged. For Russian, Serbian and Slovene a sentence alignment (with the help of Duško Vitas, Belgrade) has already been performed. The author is grateful for any cooperation with specialists of tagging, annotation and lemmatisation of Slavic texts. Even if the KZS has thus far not been processed adequately, some quantitative studies can at least be performed. It can also be claimed that all texts bear approximately equal semantic information, and that all texts do not differ in terms of the original version used for the translations. On the one hand all scanning and proofreading has been done manually and on the other hand the text length of all ten chapters of each language was statistically correlated with the chapter text length of the Russian original. The interrelation of the text length in translated and original texts can be described by simple linear models, with a sufficient $R^2 > 0.98$ in all examined cases. Thus, it is likley to conclude that for all translated texts the same source was used and there are no significant stylistic modifications in the translations.

№	Language group	Language	Tokens	Types	TTR
1	South Slavic	Slovene	62655	13946	4.4927
2		Serbian	56230	13642	4.1218
3		Croatian	56424	13737	4.1074
4		Bulgarian	57174	12308	4.6453
5		Macedonian	58837	11465	5.1319
6	Eastern Slavic	Russian	49675	15053	3.3000
7		Ukrainian	49612	14645	3.3876
8		Belorussian	50010	14858	3.3659
9	Western Slavic	Czech	52180	14136	3.6913
10		Slovak	52099	14027	3.7142
11		Polish	52737	14978	3.521
12		Sorbian	58484	14574	4.0129

Table 1. Text length (types, tokens) and TTR of KZS

marginal, so it can be concluded that the Eastern Slavic languages do not show notable differences between the text lengths, either at the tokens or at the types level.

A relativley similar picture is obtained for the Western Slavic languages: The Slovak, Polish and Czech translations have a more or less similar text length (approx. 52300 tokens). This especially holds true for Slovak and Czech, with a difference of just 82 tokens. The text length of Polish is somewhat higher, for instance in relation to Czech, which has 550 tokens more. A clear outlier within the Western Slavic languages is the Sorbian translation, which has 6304 tokens more than the Czech text. This difference can perhaps be explained by morphological differences between Czech and Sorbian, such as the analytic form of the past tense, the formal expression of the reflexivity, etc. This is a hypothesis which of course must be studied in more detail in the future. However it is worth mentioning that on the types level the difference is not so for the Sorbian text, which has only 438 types more than the Czech text.

Further evidence for our claim that morphological characteristics are responsible for the differences in text length can be found in South Slavic languages. The genetically "very close" languages of Croatian (56424 tokens) and Serbian (56230 tokens) do not show any notable differences at all; similarly small differences regarding the number of types and tokens can be found for the Bulgarian and Macedonian texts. Again, just one language, namely Slovene, demonstrates a behaviour that is in some way specific regarding the number of tokens. Slovene, in comparision with other Slavic languages, shows a relativley "normal behaviour" on the types level, on the tokens level it is one of the longest texts (62655). This striking feature – at least on the tokens level in relation to the Russian text (49675 tokens) – can again be explained linguistically: In Slovene, the analytical form of the past tense by means of the auxiliary verb "to be/je", the intensive usage of the analytical past perfect ("je bil vedel/be was known") and the frequent use of relative clauses are relatively typical and commonly used. Russian neither makes use of auxiliary verbs nor has a complex analytical past tense form and thus for the Slovene text a frequent use of synsemantic words and a high number of tokens can be observed.

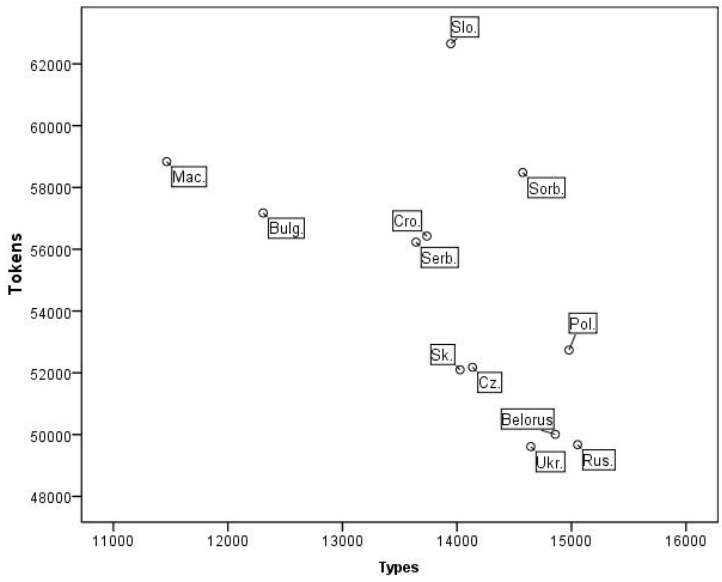

Fig. 1. Types and tokens in Slavic parallel texts

All in all, a fairly simple analysis of the text length already provides some in-depth information about the morphological and syntactical structure of the parallel text corpus under examination.

4.2 Type-Token Ratio

In addition to a simple description and interpretation of the number of types and tokens of the parallel texts, a more detailed examination of the so-called Type-Token Ratio (TTR = Tokens/Types) is required. Whereas in the past this index was primarily understood as an indicator of lexical richness (cf. the overview on possible interpretations of the TTR in [3, 108]), the TTR – especially in comparative studies of parallel texts – can be introduced as a measurement of the morphological richness of word forms and the productivity of the flexion system. The more word form types a text (=language) has, the greater is the variety of morphological forms that it contains and thus a low TTR in parallel-text research can be interpreted as an indicator of the synthetism of one language. For more in-depth research of word frequencies in language typology cf. [21]. The TTR for all examined Slavic languages is graphically represented in Fig. 2 and already ranked from its maximum to its minimum.

It can clearly be seen from Fig. 2. that, based on the TTR, a typological order of the Slavic languages can be obtained: It starts with the strongly analytic Southeast Slavic languages, Macedonian and Bulgarian, and continues with Slovene, Serbian and Croatian. They are followed by the Western Slavic languages (order of languages: Sor-

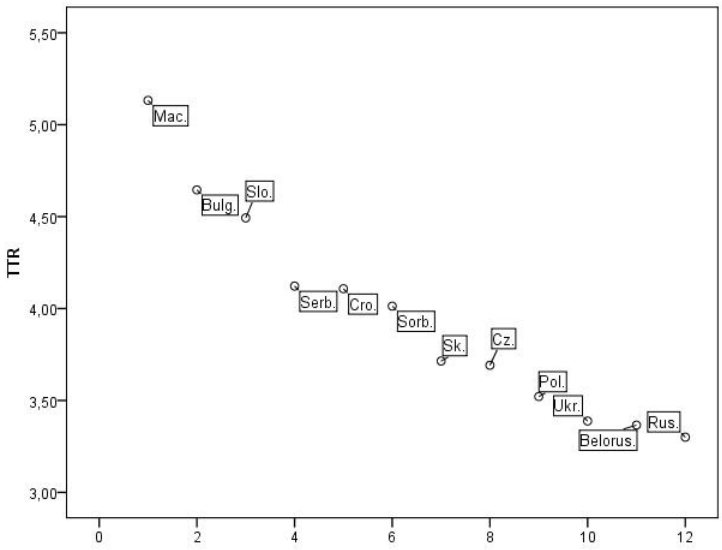

Fig. 2. TTR in Slavic parallel texts

bian, Slovak, Czech and Polish), which appear as one "group"[8]. Finally, the East Slavic languages (Ukrainian, Belorussian and Russian) have the lowest TTR and also appear as one typological "row". In sum, the TTR of parallel texts can be interpreted alongside the morphological richness as an indicator of the degree of analytism/synthetism of the languages.

5 Conclusion

The paper has the main function of presenting the KZS parallel-text corpus, which should, of course, be refined in the future, especially from the viewpoint of language processing (tagging, alignment, and lemmatisation). As shown in our exploratory discussion, the text size (number of types and tokens) and the Type-Token Ratio of parallel texts appear to be a simple yet efficient tool for obtaining language-specific morphological behaviour of the parallel texts analysed. Moreover, strong evidence for the general usefulness of parallel texts for language typology is given.

References

[1] Altenberg, B. and Aijmer, K. (2000). The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies. In Ch. Mair and Chr.

[8] In future more rigorous methods such as cluster methods must be applied.

Hundt, editor, *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, pages 15–33, Amsterdam. Rodopi.

[2] Altmann, G. and Lehfeldt, W. (1973). *Allgemeine Sprachtypologie*. Fink, München.

[3] Altmann, V. and Altmann, G. (2008). *Anleitungen zu quantitativen Textanalysen. Methoden und Anwendungen*. RAM-Verlag, Lüdenscheid.

[4] Anninskij, L. (1989). Obručennyj s ideej. In Ostrovskij, N. A., editor, *Sobranie sočinenij v trech tomach. Tom 1. Kak zakaljalas' stal'*, pages 7–28, Moskva. Molodaja Gvardija.

[5] Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. *Sprachypologie und Universalienforschung*, 60(2):95–99.

[6] Dimitrova, L., Ide, N., Petkevič, V., Erjavec, T., and Tufiș, D. (1998). Multext-East: Parallel and Comparable Corpora and Lexicons and Lexicons for six Central and Eastern European Languages. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics. Volume 1*, pages 315–319, Montreal and Quebec.

[7] Erjavec, T., Ide, N., Petkevič, V., and Véronis, E. (1995). Multext-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. In *Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995)*, pages 88–97.

[8] Garabík, R. and et al. (2007). A Cross-linguistic Database of Children's Printed Words in Three Slavic Languages. In Levická, J. and Garabík, R., editors, *Slovko 2007. Fourth International Seminar. Bratislava, Slovakia, 25–27 October 2007*, pages 51–64, Bratislava. Tribun.

[9] Gellerstamm, M. (1996). Translations as a source for cross-linguistic studies. In Aijmer, K., Altenberg, B., and Johansson, S., editors, *Language in Contrast: Papers from a symposium on Text based Cross-linguistic studies. Lund, March 1995*, pages 53–62, Lund. Lund University Press.

[10] Guski, A. (1981). N. Ostrovskij: Kak zakaljalas' stal': biographisches Dokument oder sozial-realistisches Romanepos? *Zeitschrift für slavische Philologie*, 42:116–145.

[11] Johansson, S. (1998). On the role of corpora in cross-linguistic research. In Johansson, S. and Oksefell, S., editors, *Corpora and Cross-Linguistics Research*, pages 3–24, Amsterdam. Rodopi.

[12] Johansson, S. (2003). Reflections on Corpora and their Uses in Cross-linguistic research. In Zanettin, F., Bernardini, S., and Stewart, D., editors, *Corpora in translator education*, pages 135–144, Manchester. St. Jerome Publisher.

[13] Kelih, E. (2009a). Graphemhäufigkeiten in slawischen Sprachen: Stetige Modelle. *Glottometrics*, 18:53–96.

[14] Kelih, E. (2009b). Slawisches Parellel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In E. Kelih, E., Levickij, V. V., and Altmann, G., editors, *Methods in text analysis*, pages 106–124, Černivci. Ruta.

[15] Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

[16] Köhler, R. (2005). Synergetic linguistics. In Köhler, R., Altmann, G., and Piotrowski, R. G., editors, *Quantitative Linguistik/Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*, pages 760–774, Berlin. de Gruyter.

[17] Lemnitzer, L. and Zinsmeister, H. (2006). *Korpuslinguistik: eine Einführung*. Narr, Tübingen.

[18] Mauranen, A. (2002). Will 'translationese' ruin a contrastive study? *Languages in Contrast*, 2:161–186.

[19] McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book*. Routledge, London.

[20] Mohanty, P. (2008). The Semantic Differential Technique and Measurement of Translational Meaning. In Altmann, G., Zadorozhna, I., and Matskulyak, J., editors, *Problems of General, Germanic and Slavic Linguistics. Papers for the 70th anniversary of Professor V.V. Levickij.*, pages 215–225, Černivtsi. Knichi XXI.

[21] Popescu, I.-I. and Altmann, G. (2008). Hapax Legomena and Language Typology. *Journal of Quantitative Linguistics*, 15(4):370–378.

[22] Stolz, T. (2007). Harry Potter meets Le petit prince: On the usefulness of parallel corpora in crosslinguistic investigations. *Sprachtypologie und Universalienforschung*, 60(2):100–117.

[23] Stolz, T., Stroh, C., and Urdze, A. (2007). Nicht ganz ohne [...]. In Grzybek, P. and Köhler, R., editors, *Exact Methods in the Study of Language and Text. Dedicated to Professor Gabriel Altmann on the Occasion of His 75th Birthday.*, pages 633–646, Berlin and New York. de Gruyter.

[24] Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In Altenberg, B. and Granger, S., editors, *Lexis in Contrast*, pages 189–214, Amsterdam. Benjamins.

[25] Teubert, W. and Čermáková, A. (2007). *Corpus linguistics: A short introduction*. Continuum, London.

[26] van der Auwera, J., Schallea, E., and Nuyts, J. (2005). Epistemic possibility in a Slavonic parallel corpus – a pilot study. In Hansen, B. and Karlík, P., editors, *Modality in Slavonic languages. New perspectives*, pages 201–217, München. Sagner.

[27] Véronis, J. (2000). From the Rosetta Stone to the Information Society: A Survey of Parallel Text Processing. In J. Véronis, editor, *Parallel Text Processing. Alignment and Use of Translation Corpora*, pages 1–25, Dordrecht. Kluwer.

[28] von Waldenfels, R. (2006). Compiling a Parallel Corpus of Slavic Languages. Text strategies, Tools and the Question of Lemmatization in Alignment. In B. Brehmer and V. Zhdanova and R. Zimny, editor, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pages 123–138, München. Sagner.

[29] Wälchli, B. (2007). Advantages and disadvantages of using parallel texts in typological investigations. *Sprachypologie und Universalienforschung*, 60(2):118–134.