Branko Tošović, Arno Wonisch (ur.)

# Srpski pogledi na odnose između srpskog, hrvatskog i bošnjačkog jezika

## Die serbische Sichtweise des Verhältnisses zwischen dem Serbischen, Kroatischen und Bosniakischen

## I/1

Institut für Slawistik der Karl-Franzens-Universität Graz
Beogradska knjiga
2010

Branko Tošović (Graz)

# The distance between standard Slavic languages

Diese Arbeit setzt sich aus vier Teilen zusammen, wobei der erste eine Betrachtung der Distanz aus theoretischem Blickwinkel vornimmt. Der zweite Teil beschreibt die Methodologie der Messung der Distanz zwischen den slawischen Standardsprachen. Nach Meinung des Verfassers würde sich für die Durchführung einer solchen Untersuchung eine Heranziehung elektronischer Parallelkorpora bestens eignen. Im dritten Teil wird ein Überblick über gegenwärtig vorhandene Korpora zu den einzelnen Sprachen gegeben. Abschließend erfolgt im letzten Teil eine Auseinandersetzung mit psycho- und soziolinguistischen Aspekten.

**0.** The concept of distance is understood as the relationship between object A and object B, referring to the degree of separation as well as the space between these two objects. Accordingly**,** language A may possess characteristics that create distance as well as a space between itself and the languages B, C, X, etc. Every language occupies a specific position on a scale from close“ to distant“ The unit of measurement to be used to measure this amount of separation will be referred to as distance – which, among other things, will comprise a certain number of structural features such as quality, quantity, intensity, level, degree, cause, and direction **(A → B, B → A, A ↔ B**).

There are different kinds of distance – such as structural, discipline-specific distance (be it intellectual, cultural, mathematical, mathematical-linguistic, political, psychological, socio-linguistic, ethnic, etc.); quantitative distance (be it minimal, imperceptible, insignificant, minute, large, huge, etc.); intentional distance (be it purposeful, progressive, or regressive); or, inter alia, evaluative distance (be it analytical, relevant, noticeable, optimal, predicted, divergent, convergent, real, perceptive, convenient vs. inconvenient, expected vs. unexpected, anticipated vs. unanticipated, measurable vs. immeasurable, etc.). Accompanied by two forces – namely, centrifugal and centripetal force – distance is associated with various processes. Among those processes, acceleration, deceleration, attraction and repulsion, as well as convergence and divergence are especially noteworthy. Distance influences the level of difficulty in acquiring a new language as well as the threshold of intelligibility between speakers of different languages.

William Mackey differentiates between several different types of distance between languages: (1) the distance between language systems and the distance in discourse; (2) static distance (which paradigmatically seems to distinguish the differences between elements and constructions of different languages) vs. dynamic distance (which syntagmatically represents the results of concrete speech acts); (3) distance as the discrepancy between two language systems (or subsystems) vs. distance as the conversion from one langue into another; (4) taxonomic vs. integral distance; (5) distance in form and content (the same form

can have different meanings); and (6) distance in intensity (diversity, intensity, and efficiency of linguistic differentiation), as some categories are more strongly differentiated in some languages than in others (Mackey 1971: 105–106). The measurement of distance and the previously mentioned distinctions allow (1) judgements on the contrastive relationships between languages in contact as well as (2) indirect predictions about the difficulty and distance in communication between speakers of the respective languages to be made (Mackey 1971: 106–107). Additionally, Mackey cites examples illustrating possible ways of determining the amount of distance with the aid of calculations based on a quantitative indicator.

In his definition of polycentric languages, Ulrich Ammon distinguishes between three types of linguistic distance: little distance (which is typical in standard variations of polycentric languages, such as the distance between Austrian Standard German and its German counterpart); medium distance (which denotes minimal linguistic distance between standard variations of different languages, so-called „Ausbausprachen“ or „languages by development“ such as the distance between Luxembourgish and Germany's Standard German); and great distance (which is observable between any variations that constitute different languages, also known as Abstandsprachen“ – Ammon 2005: 1538). Analogous to the scheme outlined above, Ammon also differentiates between the different degrees of influence that distance has on mutual comprehension. For example, a large distance allows for no comprehension, a medium amount allows for understanding with considerable difficulties, and a small distance allows for problem-free communication.

S. E. Jachontov is of the opinion that linguistics, especially sociolinguistics, is in need of its own scale for measuring similarities between languages based on practical implications for speakers and researchers (Jachontov 1980). Such a model would consist of five levels: (a) speakers of different idioms communicate freely with one another; (b) speakers of different idioms communicate with each other without great effort, even if some individual items are not understood; (c) speakers of different idioms cannot converse freely; (d) communication is not possible; and (e) only experts can discover the relationship between the two idioms.

**1.** Relying on corpus-based analysis, the distance between the following languages should be examined: Bosnian/Bosniak (Bs), Bulgarian (Bg), Croatian (Hr), Macedonian (Mk), Montenegrin (Mo), Polish (Pl), Russian (Ru), Serbian (Sr), Slovene (Sl), Slovak (Sk), Sorbian (Ls), Czech (Cs), Ukrainian (Uk), and Belarusian (Be). The relationship of these languages to German (De) would also be examined. Additionally, should it prove to be feasible, so-called Slavic microlanguages – such as Burgenland Croatian (Hg), Kashubian (Ks), and Rusyn (Rs) – would also be integrated into the study.

**2.** The distance between SSL covers a broad spectrum of values – ranging from cases where there is very little distance between languages (such as between

Bs, Hr, Mo and Sr as well as between Mk and Bg) to cases where two languages are occupying polarised positions in relation to one another. Light can also be shed on differences in opinion concerning the sociolinguistic classification of SSLs: are the languages (a) entirely separate languages, (b) variations of a polycentric language, (c) dialects of a national language, (d) one language with various (politically determined) names, etc.? The lack of concrete research that would reveal the relationships between SSLs and provide relevant information to clarify all fundamental questions encourages subjective and biased interpretations of those relationships as well as politically tinged explanations of interlinguistic correlations – a phenomenon in which not only linguists, but also non-experts in the field, especially politicians, get involved. In the 1990s, with the military conflicts in south-eastern south-eastern Europe and the fall of three federal Slavic states (the USSR, Czechoslovakia, and Yugoslavia), the linguistic circumstances grew more intense. The situation was exacerbated by differing interpretations regarding the legitimacy or illegitimacy of the official designation and codification of several SSLs in the 20$^{th}$ century (take for instance Mk in the 1940s and Bs, Hr, and Sr in the 1990s). The change in status of several SSLs made interslavic linguistic relationships more complicated. Thus, languages of former autonomous republics within larger states – such as (a) Cs and Sk, (b) Be and Uk, and (c) Sl – became the national languages of newly founded nations. On the other hand, Ru lost its status as a language of interregional communication in Belarus and the Ukraine, leaving the Russian-speaking population of these states in a conflict situation, in which tendencies towards the complete repression of Ru in nearly all spheres of communication were intensified. By the same token, Serbo-Croatian also lost its status as the lingua franca of the former Yugoslavia, giving rise to three separate codified languages (Bs, Hr, and Sr) and a fourth (Mo), which was extolled in the framework of the 2007 Montenegrin constitution and is currently in the process of being standardised. By contrast, several of the so-called microlanguages in the newly established nations demand a change in their status (such as Ru in the Ukraine).

Moreover, with the fall of several nations, some SSLs became languages of national minorities (such as Hr in Serbia or Sr in Croatia). A certain, and in some cases substantial, portion of Russian-speakers found themselves outside the borders of the Russian Federation and, with the expansion of the European Union (in the nations of Estonia, Latvia, and Lithuania), entered into the EU.

Currently, there are five SSLs represented in EU member states: Bg, Cs, Pl, Sk, and Sl. Hr will soon gain that status; and Bs, Mk, Mo, Sr as well as others are also moving in that direction, which means that there could soon be ten SSLs represented in the EU. Hence, it is worth noting that a considerable number of languages in countries trying to enter into the EU can be classified into groups of closely related languages whose speakers can communicate freely without the help of an interpreter (Bs, Hr, Mo, and Sr representing a typical example of such a relationship). For this

very reason, discussions are currently taking place as to whether the provision of translation services between those languages would be sensible.

Additionally, there is an array of linguistic studies examining the distance between specific languages. One such study was carried out by O. Revzina, who set out to measure the distance between related systems of Slavic languages (Revzina 1970). According to the amount of distance between them, Rovzina divided the languages into five main groups of related systems: a Polish group (including Upper Sorbian and Slovak); a Russian group (including Ukrainian); a Serbian group; a Slovene group; and a Czech group. According to Revzina, the sharpest contrast made itself evident between the Polish type and the Russian type.

It is interesting that the Czech type goes only slightly in the direction of the Polish type, while the two South Slavic types – namely Slovene and Serbian – occupy a relatively symmetrical position between the Polish and Russian types, the Slovene type diverges further than the other two types […] With regard to gender (masculine, feminine, or neuter), a relatively symmetrical position with considerable deviations is characteristic of the Serbian and Czech languages, which developed no new genders. For the Polish and Russian types – languages in which the decomposition of the old genders and the corresponding expansion of their inventories have progressed the most – not only is asymmetrical positioning of the genders in the system typical; so is a reduction in the distance between the genders (Revzina 1970: 30).

In a comparison of Serbo-Croatian and Russian, Pavle Ivić carried out an analysis of the genetic distance between Slavic languages on a phonological level (Ivić 1998: 66–67). To accomplish this, he set out to analyse the following monosyllabic words of Ur-Slavic origin that form part of the languages' common lexical heritage: *sin – сын*, *list – лист*, *lek – лек*, *red – ряд*, *led – лед*, *naš – наш*, *luk – лук*, *bok – бок*, *san – сон*, and *lan – лен*. In his research, Ivić came to the conclusion that a sound system of 10 vowels could nearly cover the entire stock of late Ur-Slavic phonemes.

In a study published in 2007, Ginsburgh, Ortuño-Ortin and Weber analysed the distance between languages in relation to the usefulness of learning them, whereby the usefulness of learning languages increased with the distance between them. In addition, the difficulty of acquiring languages was depended solely on the distance between them: the less distance between languages, the less difficulty speakers would have learning one another's languages. The authors also ascertained that distance between languages also influences success in learning.

US linguist Morris Swadesh elaborated a 100-word comprehensive list of core lexical elements, which originally consisted of 215 words, and was of the opinion that the upper limit would be around 300 elements (Swadesh 1999, Swadesh_lists1-www). The following elements were included as part of this lexical

core: pronouns, numbers, names of body parts, geographic features and certain rudimentary natural phenomena as well as activities particular to humans that are of universal meaning and are expressed in every society and language. Not to be overlooked, terms for *I, you* (*all*), *we, this, that, who, what, no*/*not,* and *everything* were also included. Swadesh hoped to create a directory usable for all languages, which would prove to be impossible.

The basics of Swadesh's theoretical approach can be summarised as follows: (1) in a dictionary of any given language, there is a certain part which includes the basic, everyday terms and can be viewed as rudimentary and stable; (2) in any given language, there are ideas that are categorically expressed with words from this inventory. By comparing the percentage of words from this central stock, Swadesh attempted to estimate the amount of time that had passed since the two languages had parted ways. In doing so, he deduced that this fundamental, core stock of vocabulary has changed continually at a steady pace. This contradictory method was especially popular in the 1960s and 70s and has even recently been taken up again, as researchers attempt to develop new concepts based on this very method. For example Kromer (2004, 2005) examined the regularities named by Swadesh in consultation with his own specifying methods.

Kromer's concept consists of four points. First of all, in addition to the factor of divergence among dictionaries and in accordance with Swadesh's postulates, it is assumed that words from the stock of fundamental vocabulary are simultaneously and erratically replaced. This modified method thereby makes it possible to examine Pidgin as well as Creole languages. In that same way, it became possible to reconstruct a language or multiple languages for every language group – which were likewise defined according to this new method – leading directly to protolanguage(s) for each group. The results garnered thus far show that, amongst Celtic languages, Breton is closest to the protolanguage; amongst the Germanic languages, Danish and German occupy this position; and amongst the Slavic languages, Slovene is the closest to the protolanguage. With regard to linguistic tree diagrams, it can be maintained that not only divergence, but also convergence can make a second measurement necessary. Kromer's second point is that it is necessary to clarify de facto distortions in linguistic tree diagrams because, in practice, mixed languages are also drawn upon to a greater or lesser extent.

The distance between languages was also examined with regard to interference. In line with this approach, some specialists are of the opinion that a smaller amount of typological distance between related languages, that is to say a high degree of similarity with minimal differences, is more likely to lead to interference. In this way, a greater amount of distance – such as between genetically unrelated languages – would reduce not only the frequency of errors but also the frequency of the automatic acquisition of new vocabulary experienced by learners (Dmitrijeva-www).

In order to evaluate the distance between languages on a semantic level as well as to measure assessment of the message of lexical units, Osgood's differential (Osgood et al. 1957) can be used as confirmed by Šipka's analysis (2008), in which this method was used to determine differences between Hr and Sr with regard to speakers' interpretation of words as being either foreign or part of their own native tongue.

Primarily dedicated to explaining differences between closely related languages as well as between variations of standard languages, sociolinguistics also deals with linguistic distance. The work of Ulrich Ammon is especially noteworthy, particularly his work dealing with genetic, typological, and linguistic distance (1987b). According to Ammon, genetic distance is important for genetic classification, while linguistic distance (that is to say distance based on grammar) is crucial for typological classification.

Expressed heterogeneously in the distinct SSLs, purism also has a great impact on the amount of distance between SSLs. Some SSLs such as (Bg, Ru and Sr) show no pronounced purist tendencies, while marked purism is traditionally inherent in other SSLs (such as Hr and Sl), which promotes an increase in the amount of distance between languages. Unfortunately, in the current literature on the subject, purism is generally only examined intralinguistically, that is to say within the context of one single language, without considering its influence on other languages. Traditional norms as well as rules of linguistic etiquette are generally also thereby assigned essential roles.

Even in exact sciences, there is an array of methods for measuring the distance between languages. One such method is based on so-called editing distance, indicating the amount of work required to translate a string of characters from one language into another. Levenshtein distance (Levenshtein 1965), which is also based on the conversion of one string of characters into another, is considered the simplest and most prevalent form of calculating distance. With this method, three operations – namely, erasing, replacing and adding – are used. Distance is determined based on the number of operations that have to be carried out in order to transform the string of characters into another. For example, in order to transform the Russian noun *диалог* (dialogue) into *одеяло* (cover), the following steps are necessary: add **о**, replace **и** with **е**, replace **а** with **я**, and erase **г**, yielding a Levenshtein distance of four. It is advisable to write the words phonetically in order to evaluate the actual linguistic distance independently, not on the basis of differences in traditional orthography. Levenshtein's method is primarily used to measure the distance between dialects in the field of dialectometry, in which Wilbert Heeringa's dissertation (2004) and his joint article together with Charlotte Gooskens (2004) are particularly noteworthy.

The Wagner-Fischer distance model can be used to achieve a more in-depth evaluation of the distance between languages. Using this method, the process of

transformation is further defined by the distance between the original phonemes and the phonemes replacing them. For example, the replacement of a vowel by another vowel represents a less drastic transformation than the replacement of a vowel by a consonant (Bērziņš 2006). Using the Wagner-Fischer distance method, an eight-part phonetic model for all phonemes of Latgalian and Latvian was developed (Bērziņš/Grigorjevs 2007). All Slavic phonemes could also be classified within this system.

In the categorisation of texts, William Cavnar and John Trenkle suggested citing sequences of characters within the framework of an n-gram model (Cavnar/Trenkle 1994). In accordance with Zipf's law, characters, consonant clusters (blends), and words, among other things, can be ordered according to the frequency of their occurrence. The aforementioned researchers recommend establishing frequency lists and n-grams for different texts so that the category in which the text belongs can be determined based on those lists and n-grams. In using such a method, the language, coding, and topic of a text can be determined with high recall. Bērziņš also suggests using frequency lists and n-grams to measure the distance between languages (2004a, 2004b). In such cases, source data is provided by random, unmarked text corpora from the languages under examination. Though still unpublished, Bērziņš has obtained positive results in assessing linguistic distance by means of phonograms. Untranscribed, multilingual audio recordings of numerous speakers are drawn upon and analysed using the Hidden-Markov model. Furthermore, a number of researchers rely on n-grams (for example, Cavnar/Trenkle 1994, Cavnar/Vayda 1992, Cavnar/Vayda 1993, Kondrak 2005).

In 1992, to analyse the distance between 95 European languages, a distance matrix of 200 basic terms with common roots was set up (Dyen/Kruskal/Black 1992). The distance between De und Ru was valued at 0.76. The distances between De and other SSLs are listed as follows: Sl at 0.73; Cs and Sk at 0.74; Pl at 0.75 and Uk at 0.76. The distance between Ru and other SSLs looked considerably different: Sl at 0.39; Cs and Sk at 0.26; Pl at 0.27 and Uk at 0.22. With regard to European languages, the distance between De and Dutch (162), Danish (293), Swedish (305), and English (422) was minimal, while the distance between De and Finnish (1000) and Greek (812) was at a maximum. The distance between De and French (756), Spanish (747), Portuguese (753) and Italian (735) is moderate.

Stecjuk suggests a specific method by which all the shared characteristics between the two languages in a language pair are assessed (Stecjuk-www). Using a formula from the field of logic, the linguistic distance between two languages is compared. From the analysis of language pairs separated by great linguistic distance (such as De-Ru) to the analysis of language pairs separated by much less linguistic distance (such as Ru-Uk), this method is particularly interesting in analysing Slavic and non-Slavic relations, revealing the number of shared features within Germanic-Slavic and Slavic-Slavic language pairs.

The methods used in Arapov and Cherc's attempt at determining the age of individual languages in the 1970s proved to be similar to the methods used in measuring the distance between languages (Arapov/Cherc 1974). Both researchers took on the task of developing a model of the changes in dictionary inventories on the basis of which the dependence of the point in time at which a word emerged and its position in a frequency dictionary could be garnered (Arapov/Cherc 1974:3).

Yet another method, which is based on a lexical database and includes a semantic matrix, aims to calculate an algorithm determining the distance between two Russian words connected with their English equivalents. The number of English equivalents that can be assigned to both words at the same time is thereby determined (Potemkin-www).

In literature in the field of contact linguistics, interpretations regarding loan words and linguistic interference prove useful. In these examinations in particular it is pointed out that, with regard to distance, the presence or lack of direct or indirect contact is of crucial significance. Closer and more intense contact naturally leads to a lowering of the threshold of intelligibility.

Publications dedicated to psycholinguistic aspects of linguistic distance also prove to be important, especially those exploring the perception and comprehension of languages that are first and foremost closely related. Weinreich ascertained that linguistic contact could only be understood in a broader psychological and culturological context (Weinreich 1953). He is of the opinion that meaningful results are to be expected when efforts are made on the part of representatives from diverse disciplines, integrating Linguistics, Psycholinguistics, Sociolinguistics, etc. in an interdisciplinary approach.

**3.** Currently, there is only small number of parallel corpora available for SSLs. In view of its general structure and scope, the Gralis-Corpus (Gralis-Korpus-www) – developed in the course of the FWF Project P19158-G03 (2006–2009) and used in the analysis of Bs, Hr, and Sr – represents one such corpus (Tošović 2008a, 2008b). This corpus comprises not only written texts in the form of a text corpus, but also spoken language in the form of a speech corpus. In the text corpus (Wonisch 2008b), texts of all functional styles – including literary-artistic, publicity, academic, and administrative texts (see Tošović 2002) – were incorporated. As of February 2009, this corpus contained around three million tokens. The texts were furnished with basic metalingustic and grammatical annotations.

The speech corpus is subdivided into three subcorpora – namely, a word corpus, a fix corpus, and a free corpus (Forić 2008, Wonisch/Just 2008). The word corpus represents a selection of individually pronounced words. The fix corpus consists of audio recordings of shorter texts that present no lexical or grammatical differences between Bs, Hr, und Sr – such as the text Jutro", which contains 18

sentences. Meanwhile, the free corpus contains about 300 recordings of free and spontaneous spoken language. On a related note, the comparable Russian-Slovak parallel corpus is also worth mentioning.

With respect to Russian-non-Slavic parallel corpora, the English-Russian parallel corpus in the collection of the „National Corpus of the Russian Language" (Ruscorpora), which is currently in the early stages of development, is worth mentioning. On a smaller scale, original literary works and their translations are incorporated into this corpus. In the framework of investigation by the name of „Opus", a collection of freely accessible parallel tests (including technical documentation, a corpus of subtitles, etc.) was set up. The multilingual corpus with translations from the Old Russian text „The Tale of Igor's Campaign" and „Lilabar" (Lilibar-www), an English-Russian corpus of parallel sentences with 8,500 sayings and 130,000 phrases, are also worth mentioning here. At the School of Modern Languages and Translation Studies at Tampere University (Finland), another notable corpus was developed by the name of „ParRus", which consists of artistic, Russian-Finnish parallel texts. With around 1.5 million tokens for each language, the Russian-English corpus of 19th and 20th century Russian literary works and their English translations presents an even larger volume of resources. Also worth mentioning, the Institute of Slavic Studies at the University of Regensburg (Germany) is home to the „Regensburg Parallel Corpus (RPC)" (RPC-www) – which includes not only English and German texts, but also texts composed in several Slavic languages, including Bg, Be, Bs, Cs, Hr, Pl, Ru, Sk, Sr and Uk. Within the framework of the „National Corpus of the Russian Language", a Russian-German and a German-Russian corpus is in the same stage of development along with a Russian-German corpus of parallel texts within the Austrian Academy Corpus from the Austrian Academy of Sciences, which contains only one single novel (F.M. Dostoevsky's „The Idiot" from 1868–1869) together with its German translation. Another German-Russian corpus by the name of „Traumdeutung" („Dream Interpretation"), which consists of Sigmund Freud's text by the same name and its translation, is also currently being prepared (Traumdeutung-www).

Among the non-Slavic parallel corpora, the „Europarl Parallel Corpus" (EPC-www) is noteworthy. It includes subcorpora for the following language pairs: Danish-English, German-English, Greek-English, Spanish-English, Finnish-English, French-English, Italian-English, Dutch-English, Portuguese-English, and Swedish-English. At the Centre for Translation Studies at the University of Leeds (Great Britain), the „Leeds Corpus" was developed. It consists of the following languages: Chinese, German, English, French, Italian, Japanese, Russian, and Spanish. Developed at the University of Augsburg (Germany), the „MAASTR" parallel corpus contains the English as well as the Dutch version of the Maastricht Treaty. The Franco-German Project „Collocations in Context"to the development

of a corpus containing not only German texts with their French translations but also French texts with their German translations. Last but not least, yet another corpus to be added to the list is the freely accessible „Parallel Corpus of Portuguese and English“, which is available online (Compara-www).

The first system of correlations, the intercorrelational system, only includes relationships within one of the examined languages. The current analysisinvestigates not only (a) dynamic processes, but also (b) static processes. Over the last 40 years – from 1970 to 2010, a time period which, for the purposes of this study, would be further subdivided into two shorter time periods of 20 years each (1970 to 1990 and 1991 to 2010 respectively) – linguistic changes have occurred in all the languages to be examined.

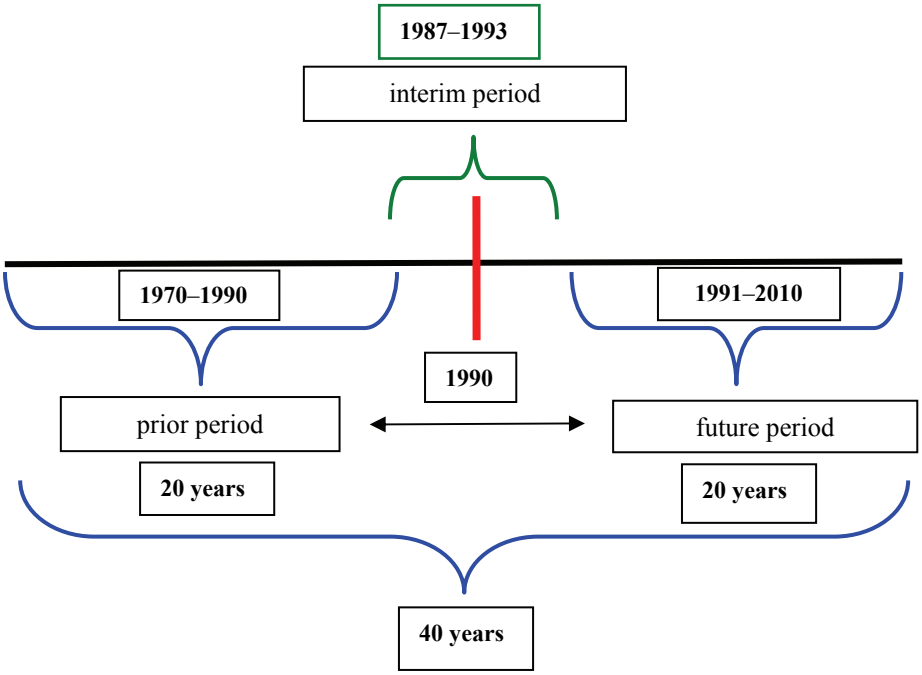



Fig. 1: The intercorrelational system

Dynamic processes are analysed by assessing the degree to which these changes have influenced interlingual distance and what other concrete effects they may have. Moreover, these changes can be expressed as either an increase or a decrease in distance. They can lead to a lack of understanding of neologisms or a change in the perception of the respective languages (whether positive or negative). They can also raise or lower the threshold of intelligibility between the speakers of different languages. As previously mentioned, static processes are also to be examined in this study. Furthermore, special attention would be given to the

assessment of the degree to which structural-typological characteristics that arose before the period of time in question affect interlingual distance.

The corpus material could be compiled in two phases. First of all, (a) monolingual texts from the two examined time periods (1970–1990 and 1991–2010) would be selected, unveiling changes that have influenced (or could have influenced) the amount of distance between the examined languages. Secondly, (b) the following data would be gathered: the character of the discovered changes; the reason and nature of their occurrence (considering contributing factors such as the efficiency of expression, spontaneity or purposefulness, political agendas, linguistic convergence, etc.); and the nature of the processes that produced these changes (whether relevant, coincidental, spontaneous, planned, purposeful, etc.).

In the intracorrelational system, for each language examined, a year in which certain events of particular linguistic significance took place should serve as a temporal break. For Be, Ru, and Uk, this break was marked by the fall of the Soviet Union in 1989; for Cs and Sk, it was the collapse of Czechoslovakia in 1993; for Bs, Hr, Mo, and Sr, it was the fall of the former Yugoslavia in 1991. For De, it was the reunification of Germany in 1990. In cases where such pivotal events did not occur during the given time span, other points in time would be used to separate the periods of time to be analysed. For example, 1993 would be used for the West Slavic languages Pl and Ls, while 1992 would be used for the South Slavic language Bg. Thus, the temporal dividing line, which could be used to help assess the influence that linguistic changes in one language had on interlingual distance, is to be drawn between 1989 and 1993. In order to determine intracorrelational distance occurring due to changes in Slavic languages from 1970 to 2010, translations from each of the examined languages existing in two versions – one produced between 1970 and 1990 and one produced between 1991 and 2010 – could be examined.

The second correlation system, the intercorrelational system, describes the relationship between very closely related SSLs such as (a) Bs, Hr, Mo, and Sr; (b) Bg and Mk; as well as (c) Cs and Sk. Research previously carried out on group (a) gives testimony to the fact that the process of diversion between Bs, Hr, Mo, and Sr was greatly intensified after the fall of the former Yugoslavia (Tošović 2008a). This divergence was intensified in particular by radical purist tendencies in the individual nations; by the growing tension created by some social processes (such as growing nationalism and chauvinism as well as the breakout of armed conflict, etc.); and by emotional factors (such as hatred towards other people and languages). On this level of analysis, the objective would not be to investigate the distance between SSLs within each individual intercorrelation system (in other words, within a, b, and c), but to determine the distance between the three previously mentioned groups – that is to say, to determine which distance is greater: the distance between Hr and Sr, the distance between Bg and Mk, or the distance between Cs and Sk.

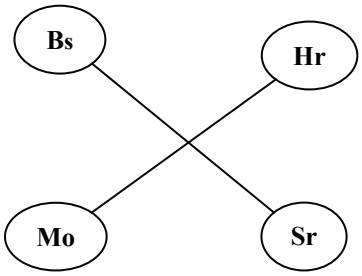



Fig. 2: Intercorrelational

The supracorrelation system is defined as the relationship between geographically close SSLs. It is divided into three parts: (a) East (Be, Ru, and Uk); (b) West (Cs, Ls, Pl, and Sk); and (c) South (Bg, Bs, Hr, Mk, Mo, and Sr). The goal of analysing the supracorrelations would be to determine the distance between languages within the same relationship system (a, b, or c), such as the distance between Ls, Pl, Sk, and Cs, for example.

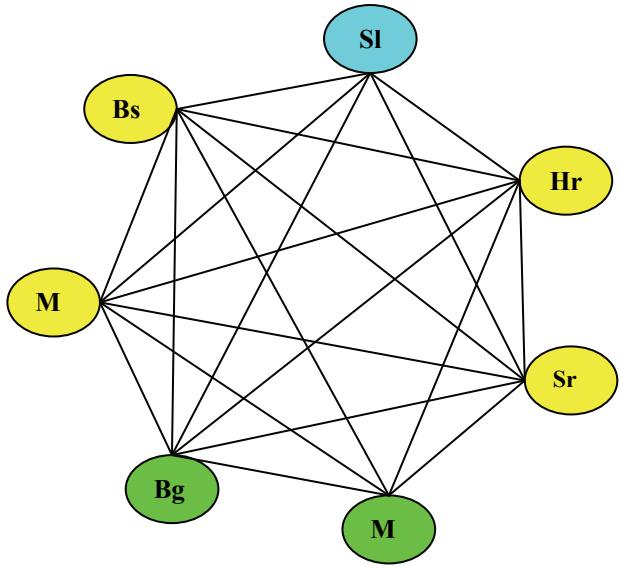



Fig. 3: Supracorrelation

On this level, the question is raised as to whether languages from a certain supracorrelation are closer to each other or closer to languages from another supracorrelation, as seen in the following comparisons: Bg, Mk ↔ Cs, Sk; Cs, Pl ↔ Ru, Uk; Be, Ru ↔ Mk, Sl.

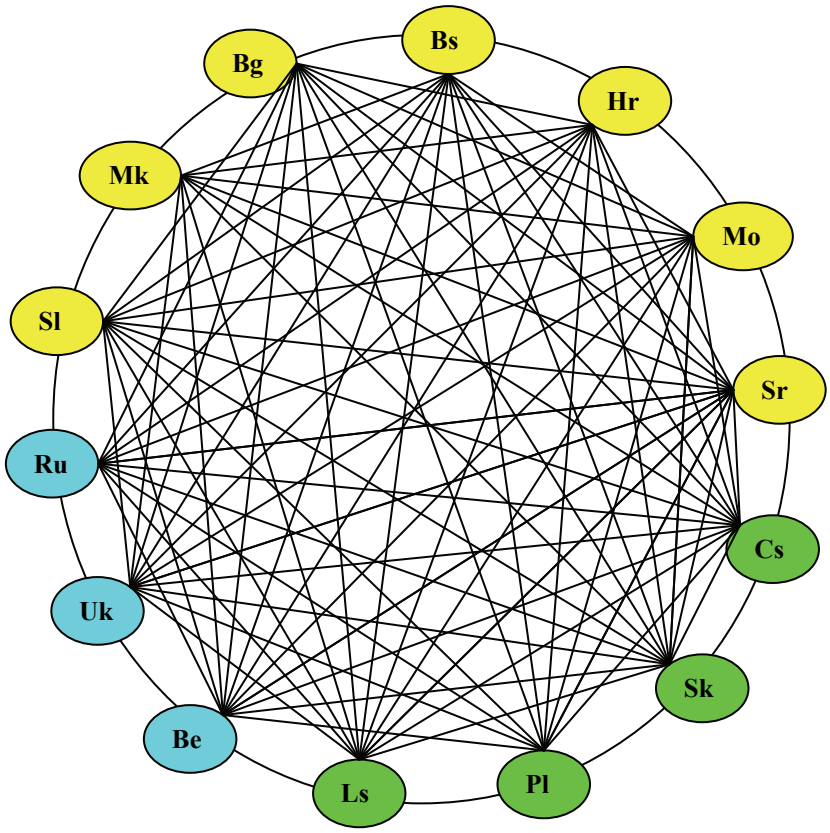



Fig. 4: Supracorrelation

Within the supercorrelation system, SSLs belonging to different language groups and territories would be compared. Firstly, East and West Slavic languages are to be compared in two groups: (a) Be, Ru, and Uk and (b) Ls, Pl, Sk, and Cs. The same is true for East and South Slavic languages: (a) Be, Ru, and Uk and (b) Bg, Bs, Hr, Mk, Mo, Sl, and Sr. Last but not least, the same also applies to West and South Slavic Languages: (a) Cs, Ls, Pl, and Sk and (b) Bg, Bs, Hr, Mk, Mo, Sl, and Sr. A unique feature of the analysis carried out in this system lies in the following question: to what extent do intercorrelational changes influence supercorrelational distance? In this context, it seems sensible to test the accuracy of the hypothesis that an increase in intercorrelational distance influences the character of supercorrelational distance. Analysis carried out as described above can provide information revealing whether some processes in Hr – namely, processes leading to conscious movement away from Sr in the form of an increase

in intercorrelational distance – lead to convergence between Hr and Ru in the form of a reduction in supercorrelational distance.

In order to determine inter-, supra-, and supercorrelational distance, texts are compiled that, if possible, are translated into all the examined languages or at least into the most possible. However, not every Slavic language is as well represented as it should be.

In the extracorrelation, De and the SSLs enter into a special system of relationships. Special attention is to be given to the distance between De and the SSLs in the form of intercorrelations (Bs, Hr, Mo, Sr, etc.); supracorrelations (Cs, Ls, Pl, Sk, etc.); and supercorrelations (Sl and Uk, Pl and Be, Mk and Cs, etc.). This part of the research investigates which SSL is closest to De and which is separated by the greatest distance. In this way, the assumption that direct geographic contact in the form of common borders has an effect on the reduction of distance between the languages can also be verified. In order to determine the extracorrelational distance between the examined languages, two sets of texts are to be drawn upon: (a) Slavic texts together with their German translations (that would have already been incorporated into the corpus in order to ascertain inter-, supra-, and supercorrelational distance) and (b) German texts together with their translations into SSLs.

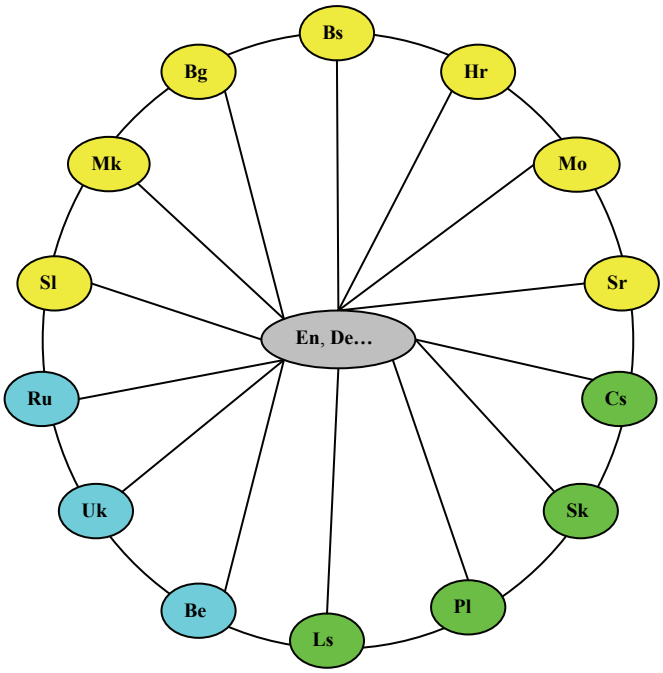



Fig. 5: Extracorrelational

Drawing upon translations produced between 1970 and 2010, linguistic distance within the framework of the entire system of correlations would be investigated. At least one of the two or more translations of any given text must originate from the examined time frame. Furthermore, in the process of selecting texts, it would be taken into consideration which texts are translated into the examined languages most often. Relying upon translations of works of prose, the linguistic distance between different literary-artistic styles would be assessed. Texts with shared content from online sources (such as „The Southeast European Times“, „Deutsche Welle“, and „The Voice of America“ among others) would be used in the analysis of distance within the genre of publicity texts. Given that these texts are only available in some of the examined languages, translations covering those languages lacking representation would have to be produced by research team members. In order to examine the distance in academic style, translations from strictly academic publications would be used, whereby the time of publication (between 1970 and 2010) and the number of languages in which translations are available (the more, the better) must also be considered. In order to analyse the distance in administrative style, different versions of basic documents from international organisations – such as the United Nations; the European Union; and the United Nations Educational, Scientific and Cultural Organization (UNESCO) – are to be used.

In order to be able to analyse colloquial style, a „SlavSpeech-Corpus“ with voice recordings of spoken speech is planned. The corpus is to consist of three parts: a word corpus, a fix corpus and a free corpus. Both the word corpus and the fix corpus are intended to be used in phonetic and prosodic analysis. The word corpus is to consist of recordings of individual words shared by all SSLs. Audio files containing coherent strings of words and sentences that are the same in all SSLs or have a similar structure could be incorporated into the fix corpus. These entries should generally be short (not longer than 20 sentences). The third subcorpus, the free corpus, is planned to be used in analysing distance on a textual and stylistic level and is to include spontaneous statements about certain topics, which should make it possible to measure distance independently of external factors (such as censorship). Test persons would be presented with drawings or sequences of pictures and would then be asked to describe what they have seen in their own words. A program by the name of „Gralis-Akzentarium“, which is based on a relational database, is to serve as the basic component of the speech corpus and is to be used in the analysis of distance in relation to prosody (for more information on the structure for Bs, Hr, and Sr, see Tošović 2008a: 770–776).

Two methods are to be used in the analysis of oral language materials: (1) an auditory method and (2) an acoustic method. With the aid of programs such as „Praat“, basic parameters such as the length of spoken sounds, changes in intonation, and the distance between the examined languages can be measured.

Each of the previously explained correlational systems should have its own subcorpora: intra-cor for one language, inter-cor for very closely related languages, supra-cor for languages that belong to the Slavic-speaking world, super-cor for languages from different language families, and extra-cor for SSLs and De. Only the first subcorpus (intra-cor) is to be monolingual. This corpus is to consist of parallel texts, which should illustrate the development of the languages within the given time frame. All the remaining subcorpora are to have standardised content and are to be functionally aligned, in order to make possible the analysis of distance in all the systems of relationships previously cited (inter-, supra-, super-, and extracorrelational) within the framework of all elementary elements.

The line of research concerning the development of the corpus comprises the gathering, processing, and annotating of materials. In doing so, the main task is to create a parallel corpus for all the languages in question. The material is to be processed in three phases.

In the first phase, texts in every SSL are to be prepared by adding metalinguistic, lexical-semantic, and grammatical annotations. Metalinguistic annotations would consist of information on the source such as author(s), chapter, pages, place and year of publication, publisher, and information about translations. Lexical-semantic annotations are used to record essential lexical and semantic characteristics of the words at hand. Grammatical annotations are solely used to indicate what exist with other words on a sentence and sytagmatic level. Given that the research required in this investigation (to find information in versions of the texts composed in all given language) demands entirely standardised annotations, these annotations have to be adapted as specifically as possible to the peculiarities of the languages in question. Previous experience as well as the steps in the procedure that have been realised thus far in the process of preparing the parallel corpus for Bs, Hr, and Sr (-corpus) provide testimony for the case that the „Multext-East“ corpus (Multilingual Texts and Corpora for Eastern and Central European Languages – multilingual dataset for language engineering research and development: Erjavec 2004, 2006), which was developed as a standard by a group of researchers led by Tomaž Erjavec in 2004, is best suited for this investigation. For the Multext-East corpus, a system of morphosyntactic annotations was developed for (only) some Slavic languages (including Bg, Cs, Hr, Ru, Sl, and Sr) with the idea that, in the course of the research, the corpus and its annotation system could be expanded to include all SSLs.

After the process of annotation has been completed, the second phase would begin. In this phase, texts could be subdivided into sentences, resulting in a system of sentences clustered together with their corresponding translations in languages a, b, c, etc. Should a paragraph in one language consist of five sentences in one language while consisting of only three sentences in another paragraph, it is necessary to compensate for this imbalance. In order to automate this process, either (a) previously existing strategies (perhaps from other studies) would be

borrowed and/or adapted or (b) new programs can be developed. In the third phase, lists of linguistic units are extracted from the corpus texts, which are then converted into relational databases using the program „MySql“, so that different types of dictionaries can be produced as a result.

In generating the corpus, „IMS Corpus Workbench (CWB)“ would be used as software and the web-based workflow manager „Asset Management Systems (AMS)“ would be relied upon for modelling the information. IMS Corpus Workbench is a multifunctional tool used in the administration, preparation, and realisation of searches in large text corpora with linguistic annotations. The main component of this workbench consists of the user-based, comprehensive browser CWB (Corpus Query Processor), which included the following elements: tools for coding, indexing, contracting, decoding and presenting frequencies; a complete register, in which all the information about the corpus (name, attributes, location) is saved; and the browser CQP with its syntax from regular expressions.

In addition to preparing texts and annotating metadata in XML-format, data materials in audio formats (mp3, wav, etc.), which would be used to measure linguistic distance on a phonetic as well as on prosodic level, also have to be gathered in the course of creating the corpus. The Asset Management System provides the processed data a central IT-structure and thus makes it possible to archive those materials in way that is sustainable and can be adjusted and modulated. The Open Source Project „Fedora“ (Flexible Extensible Digital Object Repository Architecture), could be used as a central tool in transferring the materials and administrating as well as protecting web resources.

In addition to various linguistic questions and aspects, the influence of language policy, standardisation, and codification as well as their impact on increasing or decreasing the distance between the examined languages would be discussed in the analysis of the corpus as well. In doing so, special attention could be dedicated to purism and loan words. „Gralis-Akzentarium“ – an online program for creating, carrying out, and processing surveys, which was developed in the course of the FWF Project P19158-G03 (Thomann 2008) – would be used as an additional tool in the collection and analysis of linguistic data.

Based on the written and spoken language material as well as the materials gathered in the corpus, the following question can be addressed: can the degree of verbal communication possible between speakers of these different languages and the distance between them be quantified? The criteria of comprehension as well as the influence of distance on code switching are central to this analysis. Furthermore, not only mere comprehension, but also the level (or degree) of comprehension as indicated on a set scale would be treated with special attention.

**4.** Various methods must be used in order to assess the distance between the SSLs examined on different levels. Phonetic-phonological distance is to be

measured using lists of phonemes and databases extracted out of the spoken language corpus. In the course of the study, the results of the analysis of phonetic distance between different languages and dialects would be considered (Wildgen 1977, Nerbonne/Hinrichs 2006, Nerbonne/Heeringa 1997, Tambovtsev 2002). On the lexical level, semantic distance (that is to say, the assessment of meaning and semantic similarity), which is typically only examined with one single language, would take center stage (see Osgood et al. 1957, Wildgen 1977). Osgood's semantic differential can be used to determine the semantic distance between two or several languages, assessing the subjective evaluation of the content of a lexical unit (Osgood et al. 1957). Danko Šipka put this method to the test by using it to determine the differences between Hr and Sr based on the classification of words as familiar (forming part of one's native language) or foreign. Grammatical annotations are to be used in order to carry out the grammatical analysis portion of the study. On the basis of the grammatical annotations, lists of all tokens would be generated automatically, facilitating the determination of interlingual distance. Additionally, at the outset, a database of all the tokens of each and every language would be set up, which serves as the basis for the paradigmatic and sytagmatic analyses. In the course of the paradigmatic analysis, the frequency of individual tokens would be assessed. Meanwhile, in the course of the contextual analysis, the objective would be to assess distance on a sentence level between two or more languages. In doing so, empirical and expert knowledge from previous studies would be taken into account (especially Nerbonne/Hinrichs 2006, Nerbonne/Wiersma 2006).

In the course of the orthographic analysis, the current spelling rules for the examined languages would recorded in a database by the name of „Gralis-Präskriptarium". Using an interface in this program – which is subdivided according to language, rules, and keywords (such as „comma") – searches for specific rules can be carried out in all the sets of spelling rules. If for any one given language there are multiple authorities on spelling rules, the one which seems most relevant would be relied upon. Sociolinguistic as well as psycholinguistic aspects would be analysed on the basis of annotations in the corpus and online surveys.

## Literature


Ammon 1987: Ammon, Ulrich. Language – Variety/Standard Variety – Dialect. In: Ammon, Ulrich et al. (ed.). *Sociolinguistics*. *An International Handbook of the Science of Language and Society*. Vol. 1. Berlin – New York. S. 316–335.

Ammon 2005: Ammon, Ulrich. Pluricentric and Divided Languages. In: Ammon, Ulrich et al. (ed). *Sociolinguistics*, Vol. 2, Berlin/New York. S. 1536–1543.

Arapov/Cherc1974: Арапов, М. В.; Херц М. М. *Математические методы в исторической лингвистике*. Moskau.

Bērziņš 2004a: Berzinch, A. A. La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes. In: *Identification des langues et des variétés dialectales par les humains et par les machines*. Paris. S. 103–104.

Bērziņš 2004b: Берзиньш, А. А. Сравнение балтийских языков методом n-грамм. In: *Труды международной кофереции „Корпусная лингвистика“ – 2004*. Sankt Peterburg. S. 65–71.

Bērziņš/Grigorjevs 2007: Берзиньш, А. А.; Grigorjevs J. Latviešu izloksnēs sastopamo fonēmu telpa. In: *Iesniegts publicēšanai Linguistica Lettica 2007. gadā*. Rīga. In: http://ansis.lv/raksti/endz2007.pdf

Bosák 1998: Bosák, Jan. *Slovenský jazyk*. Opole.

Cavnar/Trenkle 1994: Cavnar, William B.; Trenkle, John M. Ngram-based text categorization. In: *Proceedings of SDAIR-94*. *3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas. S. 161–175.

Cavnar/Vayda 1992: Cavnar, William B.; Vayda, Alan J. Using superimposed coding of N-gram lists for Efficient Inexact Matching. In: *Proceedings of the Fifth USPS Advanced Technology Conference*. Washington D.C.

Cavnar/Vayda 1993: Cavnar, William B.; Vayda, Alan J. Ngram-based matching for multi-field database access in postal applications. In: *Proceedings of the 1993 Symposium on Document Analysis and Information Retrieval*. Las Vegas: University of Nevada.

Dmitrijeva-www: Дмитриева, Ю. В. *Проблемы двуязычия и интерференции*. In: http://nakhodka.wl.dvgu.ru/forum/section7/7-08.htm

Dimitrova 1997: Dimitrova, Stefana (ed.). *Български език*. Opole.

Duličenko 1981: Дуличенко, А.Д. *Славянские литературные микроязыки: Вопросы формирования и развития*. Tallinn.

Dyen/Kruskal/Black 1992: Dyen, Isidore; Kruskal, Joseph B.; Black, Paul. An Indo-European Classification: a Lexicostatistical Experiment. In: *Transactions of the American Philosophical Society*. Nr. 82 (5).

Erjavec 2004: Erjavec, Tomaž. Multext-East Version 3: multilingual morphosyntactic specification, lexicons and corpora. In: Lino, Maria Teresa; Xavier, Maria Francisca (ed.). *Fourth international conference on language resources and evaluation*. Lisbon, 26th, 27th & 28th May 2004. Proceedings: held in memory of Antonio Zampolli. Paris. S. 1535–1538.

Erjavec 2006: Erjavec, Tomaž. Multext-East Morphosyntactic specifications and XML. In: Slavcheva, Milena; Simov, Kiril, Angelova, Galia. *Readings in multilinguality: selected papers for young researchers*. Sofia. S. 41–48.

Faska 1998: Faska, Helmut. *Serbšćina*. Opole.

Forić 2008: Forić, Sandra. Das Gralis Speech-Korpus. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen*/*Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 755–764.

Gajda 1998: Gajda, Stanisław. *Język polski*. Opole.

Gajda 2000: Gajda, Stanisław (ed.). *Komparacja systemów i funkcjonowania współczesnych języków słowiańskich*. Opole.

Gladrow 1989: Gladrow, Wolfgang. *Russisch im Spiegel des Deutschen*. Leipzig.

Ginsburgh/Ortuño-Ortin/Weber 2007: Ginsburgh, Victor; Ortuño-Ortin, Ignacio; Weber, Shlomo. Why do People Learn Foreign Languages? In: *Journal of Economic Behavior and Organizations*. Nr. 64 (3). S. 337–347.

Gladková/Likomanova 2002: Гладкова, Хана; Ликоманова, Искра. *Языковая ситуация*: *истоки и перспективы* (*болгарско-чешские параллели*). Prag.

Gooskens/Heeringa 2004: Gooskens, Charlotte; Heeringa, Wilbert. Perceptual evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. In: *Language Variation and Change*. 16(3). S. 189–207.

Gutschmidt 2002: Gutschmidt, Karl. *Möglichkeiten und Grenzen der Standardisierung slavischer Schriftsprachen der Gegenwart*. Dresden.

Heeringa 2004: Heeringa, Wilbert J. *Measuring Dialect Pronunciation Differences Using Levenshtein Distance*. Rijksuniversiteit Groningen. [Univ.-Dissertation]

Hinrichs 1999: Hinrichs, Uwe (ed.). *Handbuch der Südosteuropa-Linguistik*. Wiesbaden: Harrassowitz.

Hinrichs/Gerdemann/Nerbonne-www: Hinrichs, Erhard; Gerdemann, Dale; Nerbonne, John. Measuring linguistic unity and diversity in Europe. In: http://www.sfs.uni-tuebingen.de/dialectometry/docs/VW-dialect-proposal.pdf. 31. 1. 2008.

Ivić 1998: Ivić, Pavle. *Rasprave*, *studije članci*: 1. *O fonologiji*. Sremski Karlovci – Novi Sad.

Jachontov-www: Яхонтов, С. Е. Оценка степени близости родственных языков. In: Теоретические основы классификации языков мира. Moskau 1980. S. 148–157. In: www.philology.ru/linguistics1/yakhontov-80.htm.

Jermolenko 1999: Jermolenko, Svitlana. *Українскька мова*. Opole.

Koch/Oesterreicher 1985: Koch, Peter; Oesterreicher, Wulf. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: *Romanistisches Jahrbuch* 36. S. 15–43.

Kofler/Wonisch 2008: Kofler, Stefan; Wonisch, Arno Das Gralis-Rezensarium. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen/Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 803–806.

Kondrak 2005: Kondrak, Grzegorz. N-gram similarity and distance. In: *Proceedings of the Twelfth International Conference on String Processing and Information Retrieval* (*SPIRE 2005*). Buenos Aires. S. 115–126.

Kořenský 1998: Kořenský, Jan (ed.). *Český jazyk*. Opole.

Kromer 2004: Кромер, В. В. Глоттохронологическая ретрогностика языковой системы. In: *Проблемы лингвистической прогностики*. Voronež. S. 136–144.

Kromer 2005: Кромер, В. В. Об одном методе оценивания степени смешанности языков. In: *Актуальные проблемы компьютерной лингвистики*. Минск. S. 104–110.

Kunzmann-Müller 2000: Kunzmann-Müller, Barbara et al. (ed.). *Die Sprachen Südosteuropas heute*. *Umbrüche und Aufbruch*. Frankfurt am Main.

Laškova 1996: Laškova, Lili. On the Phenomenon of Slavic Languages in the Balkans. – In: *Linguistique Balkanique*. Sofija. 38/3. Pp. 231–237.

Lehner 2008: Lehner, Olga. Die technische Entwicklung des Gralis Speech-Korpus. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen/Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 777–779.

Levenshtein 1965: Levenshtein, Vladimir I. Binary codes capable of correcting spurious insertions and deletions of ones. In: *Problems of Information Transmission* 1(1). S. 8–17.

Ljubešić/Nives/Boras-www: Ljubešić, Nikola; Mikelić, Nives; Boras, Damir. Language identification: how to distinguish similar languages? In: http://infoz.ffzg.hr/ljubesic/nlnmdb_iti07.pdf

Lončarić 1998: Lončarić, Mijo. *Hrvatski jezik*. Opole.

Lukašanec et al. 1998: Lukašanec, Aljaksandr. *Беларусская мова*. Opole.

Mackey 1971: Mackey, William F. *La distance interlinguistique*. Quebec: Les Presses de l'Université Laval.

Magocsi 2004: Magocsi, Paul Robert. *Русиньскый язык*. Opole.

Marti 2000: Marti, Roland. Slavische Standardsprachen im Kontakt. Das Neben-, Mit- und Gegeneinander slavischer Standardsprachen. In: Zybatow, L. N. (ed.). *Sprachwandel in der Slavia*. Teil 2. S. 527–541.

Minova-Đurkova 1998: Minova-Đurkova, Liljana. *Македонски јазик*. Opole.

Mokienko/Walter 2008: Mokienko, Valerij; Walter, Harry (ed.). *Komparacja systemów i funkcjonowania współczesnych językow słowiańskich*. Bd. 3: Frazeologia. Opole.

Multext-East-www: Multext East. In: http://nl.ijs.si/ME/V3/.

Nerbonne/Heeringa 1997: Nerbonne, John; Heeringa, Wilbert. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97*: *3rd Meeting of the ACL Special Interest Group in Computational Phonology*. Madrid. S. 11–18.

Nerbonne/Hinrichs 2006: Nerbonne, John; Hinrichs, Erhard. Linguistic Distances. In: Nerbonne, John; Hinrichs, Erhard (ed.). *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics* Sydney. S. 1–6.

Nerbonne/Wiersma 2006: Nerbonne, John; Wiersma, Wybo. Measure of Aggregate Syntactic Distance. In: Nerbonne, John; Wiersma, Wybo (ed.). *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*. Sydney. S. 82–90.

Neščimenko 2003: Нещименко, Г. П. *Языковая ситуация в славянских странах*. *Опыт описания*. *Анализ концепций*. Moskau.

Nikitevič 2003: Никитевич, А. В. К сопоставлению деривационных подсистем глагола в славянских языках. In: *Мовазнаўства*. *Літаратура*. *Культуралогія*. *Фальклярыстыка*. *XIII Міжнар*. *з'езд славістаў* (Любляна, 2003). Minsk. S. 144–158.

Ohnheiser 2003: Ohnheiser, Ingeborg (ed.). *Komparacja systemów i funkcjonowania współczesnych językow słowiańskich*. *Bd*. *1*: *Słowotwórstwo*/*Nominacija*. Opole.

Osgood et al 1957: Osgood, Charles E. et al. *The measurement of meaning*. Urbana: University of Illinois Press.

Potemkin-www: Потемкин, С. Б. Лексическая база данных с наложенной семантической метрикой. In: http://www.philol.msu.ru/~rlc2004/files/sec/19.doc

Radovanović 1996: Radovanović, Milorad. *Српски језик*. Opole.

Revzina 1970: Ревзина, О. Г. *Типологический анализ грамматической категории рода*. (*На материале славянских языков*). Moskau. [Univ.-Dissertation; Zusammenfassung, AKD]

Sawicka 2007: Sawicka, Irena (ed.). *Komparacja systemów i funkcjonowania współczesnych językow słowiańskich*. *Bd*. *2*: *Fonetyka*/*Fonologia*. Opole.

Šipka 2008: Šipka, Danko. Varijantske razlike u semantičkom diferencijalu. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen*/*Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 130–142.

Širjaev 1997: Širjaev, Evgenij. *Русский язык*. Opole.

Stecjuk-www: Стецюк, Валентин. *Лексика как материал для реконструкции исстории языка*. In: http://www.inauka.ru/blogs/

article 71407.html

Stigler 2008: Stigler, Hubert. XML-Frameworks im Korpusmanagement. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen/Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 617–629.

Swadesh 1952: Swadesh, Morris. Lexico-statistic dating of prehistoric ethnic contacts. In: *Proceedings of the American philosophical society*. Nr. 36. S. 452–463.

Tambovtsev 2002: Tambovtsev, Yuri. Comparative typological study of language distances based on the consonants in sound chains of various languages. In: Elliot, John (ed.). *The 5th National Colloquium for Computational Linguistics in the UK*. *Proceedings of the Conference*. 8–9 January. Leeds: University of Leeds. S. 77–80.

Tambovcev 2002: Тамбовцев, Ю. А. Фонологическая схожесть и фонологические расстояния. In: *Гуманитарные проблемы миграции*: *социально-правовые аспекты адаптации соотечественников в Тюменской области*. Tjumen'. S. 274–277.

Thomann 2008: Thomman, Robert. Das Gralis Anketarium. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen/Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 796–802.

Tošović 2001: Tošović, Branko. *Korelaciona sintaksa*. *Projektional*. Graz.

Tošović 2002: Tošović, Branko. *Funkcionalni stilovi*. *Funktionale Stile*. Graz.

Tošović 2008a: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen /Bosniakischen, Kroatischen und Serbischen*. 1/3. Münster et al. Reihe Slawische Sprachkorrelationen. Bd. 1.

Tošović 2008b: Тошович, Бранко. Сопоставительное изучение славянских языков при помощи многоязычного „Гралис-Корпуса“. In: Stanković, Bogoljub (ed.). *Izučavanje slovenskih jezika*, *književnosti i kultura kao inoslovenskih i stranih*. Beograd. S. 336–340.

Vidovič-Muha 1998: Vidovič-Muha, Ada. *Slovenski jezik*. Opole.

Vojvodić 1997: Vojvodić, Dojčil. О елиптичним конструкцијама у словенским језицима. In: *Славистика*. Књ. I. S. 7–14.

Wagner/Fischer 1974: Wagner, Robert A.; Fischer, Michael J. The string-to-string correction problem. In: *Journal of the Association for Computing Machinery*. Nr. 21(1). S. 168–173.

Weinreich 1953: Weinreich, Uriel. *Languages in Contact*. The Hague.

Wildgen 1977: Wildgen, Wolfgang. *Differentielle Linguistik. Entwurf eines Modells zur Bescheibung und Messung semantischer und pragmatischer Variation*. Tübingen.

Wingender 1998: Wingender, Monika. Standardsprachlichkeit in der Slavia. In. *Zeitschrift für Slawistik*. Bd. 43. S. 127–139.

Wonisch 2008a: Wonisch, Arno. Das Gralis Personalium. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen*/*Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 813–821.

Wonisch 2008b: Wonisch, Arno. Das Gralis Text-Korpus. In: Tošović, Branko (ed.). *Die Unterschiede zwischen dem Bosnischen*/*Bosniakischen*, *Kroatischen und Serbischen*. Münster et al. S. 724–749.

Zieniukowa 1992: Zieniukowa, J. (ed.) *Procesy rozwojowe w językach sowiaskich*. Warschau.

Zybatow 1998: Zybatow, Lew N. Zu neuen Horizonten der slavistischen Sprachkontakt und Sprachinselforschung. In: *Die Welt der Slawen*. Jg. XLIII/2. München. S 323–338.

## Parallel (multilingual) corpora

Compara-www: http://www.linguateca.pt/COMPARA/index.php

EPC-www: http://www.statmt.org/europarl

Gralis-Korpus-www: http://www-gewi.kfunigraz.ac.at/gralis/

0.Projektarium/Gralis-Korpus/korpus.html

Kollokation-www: http://www.kokken.go.jp/public/world/mirror/www.ids-mannheim.de/gra/kollokation.html

Leeds-www: http://corpus.leeds.ac.uk

Lilabar-www: http://lilabar.com/index.php

Maastr-www: http://www.philhist.uni-augsburg.de/lehrstuehle/anglistik/sprachwissenschaft/mitarbeiter/stoll/elekhilf

RPC-www: http://www.uni-regensburg.de/Fakultaeten/phil_ Fak_IV/Slavistik /RPC

Ruscorpora-www: http://ruscorpora.ru; http://ruscorpora.ru/search-para.html

SPI-www: http://nevmenandr.net/slovo

Traumdeutung-www: http://www.aac.ac.at/lab_parallel_freud.html

## Sources

Deutsche Welle: http://www2.dw-world.de

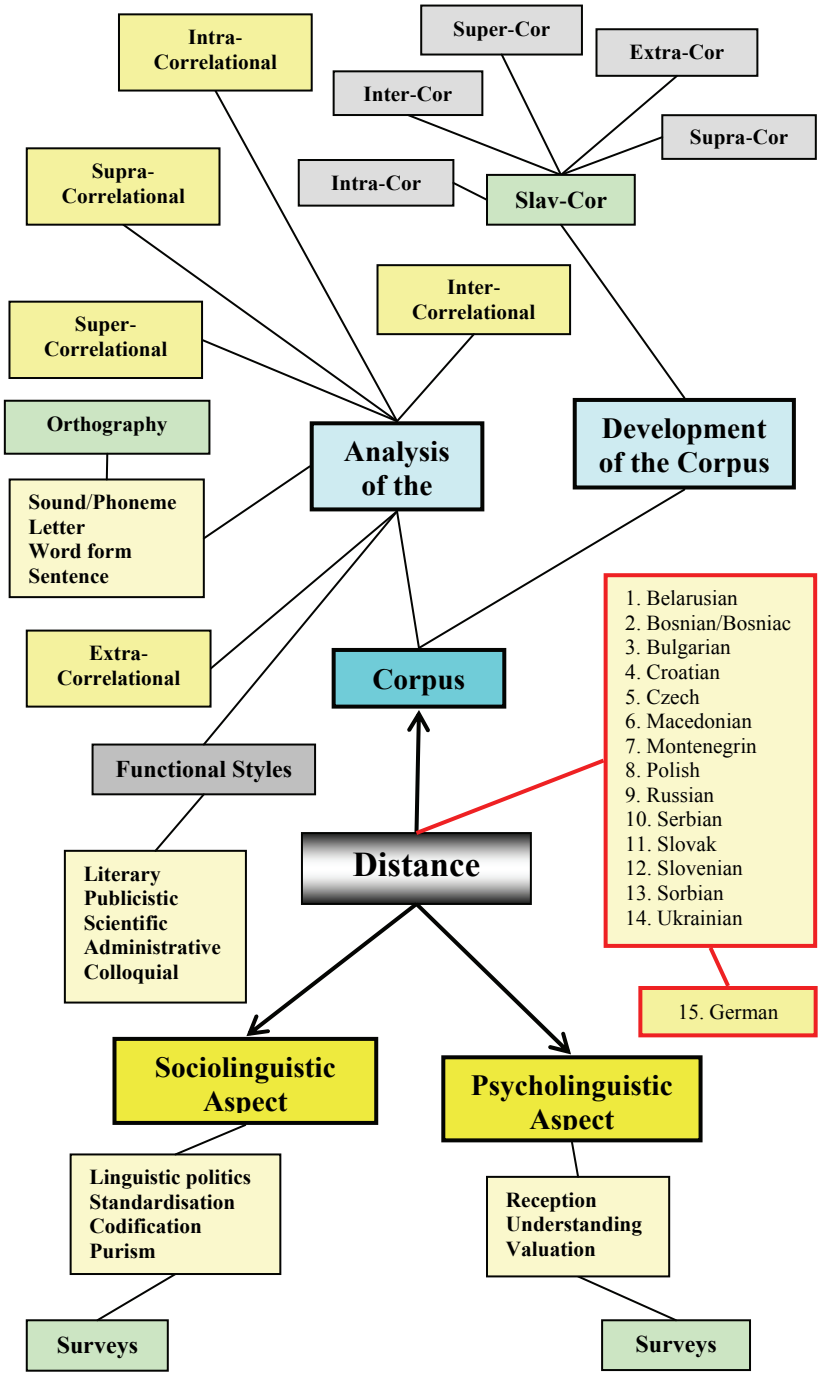
Glas Amerike: http://www.voanews.com/serbian/

Gralis-www: http://www-gewi.uni-graz.at/gralis/

Southeast European Times http://www.setimes.com

## Abbreviations

Be – Belarusian, Bg – Bulgarian, Bs – Bosnian/Bosniac, Cs – Czech, De – German, Hg – Burgenland Croatian, Hr – Croatian, Ks – Kashubian, Mk – Macedonian, Mo – Montenegrin, Pl – Polish, Rs – Rusyn, Ru – Russian, Sk – Slovak, Sl – Slovenian, Sr – Serbian, So – Sorbian, Uk – Ukrainian; SSL – standard Slavic languages


Intra-Correlational
Super-Cor
Extra-Cor
Inter-Cor
Supra-Cor
Supra-Correlational
Intra-Cor
Slav-Cor
Inter-Correlational
Super-Correlational
Orthography
Analysis of the
Development of the Corpus
Sound/Phoneme
Letter
Word form
Sentence
Extra-Correlational
Corpus
1. Belarusian
2. Bosnian/Bosniac
3. Bulgarian
4. Croatian
5. Czech
6. Macedonian
7. Montenegrin
8. Polish
9. Russian
10. Serbian
11. Slovak
12. Slovenian
13. Sorbian
14. Ukrainian
Functional Styles
Distance
Literary
Publicistic
Scientific
Administrative
Colloquial
15. German
Sociolinguistic Aspect
Psycholinguistic Aspect
Linguistic politics
Standardisation
Codification
Purism
Reception
Understanding
Valuation
Surveys
Surveys

Branko Tošović (Graz)

**Distanz zwischen den slawischen Standardsprachen**

Diese Arbeit setzt sich aus vier Teilen zusammen, wobei der erste eine Betrachtung der Distanz aus theoretischem Blickwinkel vornimmt. Unter der Distanz versteht der Autor dabei das Verhältnis zwischen einem Objekt **A** und einem Objekt **B**, indem auf den Grad von denen Nähe/Entfernung wie auch auf den zwischen den Objekten liegenden Raum hingewiesen wird. Der zweite Teil beschreibt die Methodologie der Messung der Distanz zwischen den slawischen Standardsprachen. Nach Meinung des Verfassers würde sich für eine solche Untersuchung einer Heranziehung elektronischer Parallelkorpora bestens eignen. Im dritten Teil wird ein Überblick über gegenwärtig vorhandene Korpora zu den einzelnen Sprachen gegeben. Abschließend erfolgt im letzten Teil eine Auseinandersetzung mit psycho- und soziolinguistischen Aspekten.

Branko Tošović
Institut für Slawistik
der Karl-Franzens-Universität Graz
Merangasse 70
8010 Graz
+43/316/3802522
branko.tosovic@uni-graz.at
http://www-gewi.kfunigraz.ac.at/gralis