Russ Linguist (2015) 39:275–282
DOI 10.1007/s11185-015-9154-5

Corpora and quantitative data in Slavic languages

Корпусы и квантитативные данные в славянских языках

Neil Bermel[1]

Published online: 15 September 2015

The articles in this issue arose from a thematic block of panels at the 2014 conference of the British Association for Slavonic and East European Studies (BASEES Annual Conference, Cambridge, UK, 5–6 April 2014). It was our hope as organizers that these panels would contribute to raising the profile of quantitative approaches to Slavic language data and showcase some of these approaches. What unites them are a focus on the West and South Slavic language groups, their integration of corpus-based data, and their interest in (and exploration of) the use of contrastive methods of analysis to grapple with the ever-increasing amount of data we can collect and access.

1 Background

The work described here is that of three different research teams:

The Czech National Corpus Institute (http://ucnk.ff.cuni.cz) is a department of the Faculty of Arts, Charles University, whose mission is to create, maintain and develop the use of corpora for the Czech language. Starting in 2000, they have released a series of corpora, starting with traditional databases of written texts, annotated, lemmatized and tagged, but now extending as well to oral corpora, historical corpora, specialist text corpora and translation corpora. The paper 'Simplification in translated Czech: a new approach to type-token ratio in this volume is by Institute director Václav Cvrček and staff member Lucie Chlumská; it investigates an issue relating to translation in a corpus of translated and native texts that is examined in Chlumská's doctoral thesis.

Research on this topic was supported by a grant from the Leverhulme Trust (RPG-407).

✉ N. Bermel
n.bermel@sheffield.ac.uk

[1] School of Languages and Cultures, 1 Upper Hanover Street, Sheffield S3 7RA, UK

The MiCReLa project[1] is exploring mutual intelligibility of related languages within the European Union; researchers are working to map it within the Romance, Germanic and Slavic groups using a common methodology and set of tools developed specifically for the project. The article in this issue, 'Mutual intelligibility between West and South Slavic languages' by Jelena Golubović and Charlotte Gooskens, describes the results of the first phase of work on Slavic languages, in which speakers completed a series of online tasks designed to measure their ability to understand a language they had not previously studied.

Sheffield's Slavonic language and linguistics research cluster[2] uses Slavic-language data to investigate usage-based theories of language, typically applying empirical methods involving large textual databases and experiments that provide the data they need to formulate and test hypotheses. The contribution by Dagmar Divjak, Nina Szymor and Anna Socha-Michalik 'Less is more: possibility and necessity as centres of gravity in a usage-based classification of core modals in Polish' derives from earlier work by Divjak and from Szymor's doctoral thesis, tying in with Divjak's ongoing investigation into the cognitive reality of categories. The articles by Bermel, Knittl and Russell 'Morphological variation and sensitivity to frequency of forms among native speakers of Czech' and by Lečić 'Morphological doublets in Croatian: the case of the instrumental singular' form part of the Leverhulme-funded project *Acceptability and forced-choice judgements in the study of linguistic variation*, which explores ways of relating corpus data on variation in language to the performance of native speakers on tasks related to this variation.

2 Topics and methods

Methodologically, the contributions reflect several trends that have gathered pace in the field over the last few years. Four out of five are closely linked to the use of corpora, and four out of five contain experimental data linked to the use of native speakers. In each case, the structure of the papers brings forward contrasts between two or more methods of analysis or modes of exploration. The goal is typically to interrogate a commonly-understood term or belief, operationalize it in one or more ways, and come to a conclusion about the most satisfactory method and explanation.

2.1 Mutual intelligibility

Many subdisciplines within linguistics assume the mutual intelligibility of certain pairs of languages, and mapping this has been a research aim for many years. The Levenshtein distance between two languages or dialects is often used as a yardstick against which empirical research can be measured, using tools to check comprehension (see e.g. Tang and van Heuven 2009; Gooskens 2007). Charlotte Gooskens' collaborative project, part of which is described in the article co-authored with Jelena Golubović in this volume, makes use of three tested methods for verifying mutual intelligibility, implementing them in both written and spoken domains, and contrasting the results to see which are most congruent with each other. Studies of mutual intelligibility across the Slavic domain are remarkably few in number. With the exception of some studies of Czech-Slovak mutual intelligibility (e.g. Nábělková 2008; Sloboda and Nábělková 2013), most of them have relied on purely structural descriptions (e.g. Tafel et al. 2009) or are primarily qualitative in nature, looking foremost at textual or

[1] http://www.let.rug.nl/gooskens/project/?p=home.

[2] www.sheffield.ac.uk/russian/research/slal.

situational documentation where passive bilingualism between two languages or varieties occurs.

2.2 Cognitively real categories

Divjak, Szymor and Socha-Michalik's project takes as its background a well-known area that has been subject to numerous attempts at description and classification: They examine taxonomies of modality and the descriptions that fall out from them, with the eventual goal of determining to what extent such categories are 'cognitively real'. The standard assumption would be that because a linguist is able to perceive and describe a distinction between modalities, it is therefore cognitively real. However, one might expect to see this reality reflected in other ways that can be subject to testing of various kinds, and this is where Divjak, Szymor and Socha-Michalik's innovative exploration begins.

One such reflection might be in corpus data, which is where their first exploration leads them: an exhaustively tagged and annotated dataset of modal words is subject to analysis that shows a patterning very different from that painted in most works on modality. They then subject smaller subsets of data to experimental work with native speakers, trying to see whether the fine-grained definitions of modal categories are upheld by native-speaker behaviour. The conclusion is in effect a plea for the empirical grounding of linguistic analysis: A revised hypothesis, for example, might say that cognitive reality does not follow from the linguist's ability to discern distinctions, and define the former based on data derived from large-scale, multi-author text databases and experimental data. Those interested in other work on this project can consult e.g. Divjak (2009, 2010) on aspect and modality, in which the author raises the question of how clear-cut the various proposed semantic and functional taxonomies of modality actually are when tested against corpus data.

2.3 Characteristics of translated texts

If the work on modality has tested traditionally hypothesized categories and found them wanting, the article by Cvrček and Chlumská on translated texts takes a stride towards confirming a long-posited hypothesis. It has often been said that translated texts are not as complex as originals: that they have undergone a process of simplification due to the fact that they have passed through the hands of a further author—the translator (Baker 1993). This has been anecdotally supported through qualitative work on particular texts: the close examination of discourse in a translated text vs. its original source text and through a variety of work on corpora of translated English by e.g. Laviosa (2002, pp. 59–60, 2003, p. 158). More recently, the hypothesis has been tested using type-token ratios in corpora of translated and untranslated texts. However, a challenge identified over a decade ago has not been remedied across the board, such that large-scale corpora of comparable translated and untranslated texts are seldom available, and even quantitative studies carried out since then have been on a smaller scale (e.g. Yuan and Gao 2008, based on a corpus of just over 2m tokens).

Cvrček and Chlumská's study first problematizes one operationalization of this hypothesis in which a standard type-token ratio can be used. They do so by comparing it to two types of adjusted TTRs, one of which was developed for this project. The data rest on the specially constructed comparable monolingual corpus *Jerome*,[3] which contains Czech texts that are

[3] Jerome—a monolingual comparable corpus of translated and non-translated Czech; available at http://www.korpus.cz.

original and translated, and on a new referential-value TTR measurement that requires information about a much larger superset of texts in addition to those in the relevant subcorpus. The three measures of type-token ratio—TTR, standardized TTR (sTTR) and referential-value TTR (zTTR)—are compared and contrasted, with the conclusion being that the far greater data demands of zTTR are reflected in the greater accuracy of the measure. However, echoing the later path taken by Divjak, Szymor and Socha-Michalik in problematizing distinctions promulgated by linguists, they also raise the question of how relevant 'simplification' is as a dimension of text perceptible to native speakers, as the zTTR scores show only a small to medium effect size for Czech. Experimental work could determine whether simplification can be measured in readers' responses to texts.

2.4 Overabundance

Overabundance is a term from morphology (Thornton 2012) describing a particular sort of variation in which what should constitute a single morphological 'slot' in a paradigm is occupied by two or more morphs. It, along with defectivity, constitutes a challenge to the implicit principle that for one function, there should be one form, not multiple ones. These forms can be called 'variant' or 'competing' forms, or 'doublets'.

One way around the theoretical problem of doubletism is to suggest that the variation is in fact functional, but in a way we have failed to realize. Much of the work on overabundance to date has thus focused on functional variation. Early on Halliday (1991a, 1991b) suggested that a lopsided variation where one variant outnumbered the other by more than 9:1 indicated a simple markedness function (unmarked form–marked form), while variation in more equal proportions of forms indicated that a difference of function had developed between the two forms—in other words, if we partitioned the variant forms using the correct contextual or semantic distinctions, we would eliminate the variation. A convergent finding can be seen in work on construction grammar, where the use of variants is linked to the structures around them (a summary is given in Goldberg 2009, p. 102). Some studies of course do show that variants have a tendency to partition the functional space in consistent ways, for example creating sub-cases (Brown 2007). However, there are nonetheless unexplained instances in which the variation seems unmotivated, and yet is stable (Dąbrowska 2005, pp. 192–194).

The two contributions in this volume (Lečić; Bermel, Knittl and Russell) explore different aspects of such variation, where the functional motivation appears to be missing. Lečić looks at some clear examples of non-functional variation and assesses how the frequency of different variants in a corpus seems to provide the most reliable way of assessing native speakers' judgements of the well-formedness (acceptability) of these variants. While confirming the results of previous studies that showed frequency having an effect, and particularly proportional frequency rather than absolute frequency, Lečič's study in addition uncovers some covariance between the proportional bands—in other words, as the proportion of one ending drops, its acceptability drops while the acceptability of the other variant rises.

Bermel, Knittl and Russell's contribution focuses on inter-speaker variation and sets itself the task of figuring out whether individual characteristics of speakers might contribute to the variability. While the analysis takes in what we might call biographical information (age, gender, education, region of origin) and the way respondents approach questionnaire scales, it eventually focuses in on the relationship between the sorts of acceptability judgements they make as the most relevant feature influencing their choices, using a measure of the strength of respondents' preferences as its yardstick.

The result makes explicit a type of inter-speaker variation that is not easily ascertainable by a simple biographical questionnaire or reading test and is thus difficult to extract: to get it,

one needs not only a full test of respondents' choices, but also a full complement of acceptability judgements on similar data that they can be compared to. In this sense, a parallel can be drawn to Cvrček and Chlumská's findings (see Sect. 3), in that the most accurate measure is only available when a comprehensive background database can be accessed. The greater demands on data and analysis are thus reflected in a more accurate and more nuanced set of findings.

3 Standard languages and standard research practice

The study by Cvrček and Chlumská deals with written language as a separate entity, making use of data from the written sections of the Czech National Corpus. For three of the remaining four, experimental data are used in combination with corpus data, and for the fourth, intelligibility must be assessed as both a written and spoken phenomenon. To what extent, however, do written and spoken data represent the same linguistic system? The issues hinted at here are not unique to Slavic languages, but they are nicely underlined and highlighted by the situations present in these language communities.

It is a truism that while technological advances have to some extent blurred these distinctions, written language remains a different creature from spoken language thanks to the specifics of the medium. If we take a prototypical written text on the screen or in print, it is subject to revision and emendation more readily; it is characterised by a time lag between production and reception; and its representation means all sections of it are available over time, so that it can be reviewed. A prototypical spoken text is created 'on the fly', with an interlocutor present, and it is not available after production for revision or review. In addition to these differences, we could consider a sociocultural aspect: written codes tend to be formally taught in addition to any more 'natural' exposure that the speaker has to them, and the mutual intelligibility of the written utterance is ensured not by feedback from a present interlocutor, but by adherence to formally or informally agreed conventions on various linguistic levels.[4]

To this we could add that the languages of the West and South Slavic group have a long tradition of normative interventionism, resulting in the cultivation of standard varieties that are overseen by national bodies or institutes. In some instances, such as Bulgarian, Slovak and Polish, the codified variety was based on a particular dialect—sometimes, as with Croatian, with additional political-cultural influences shaping the choices and regulatory activity. In others, such as Czech, the standard is more distant from spoken varieties, having been developed from an older written tradition whose distinction from the spoken language has been carefully maintained.

When looking for data about naturally acquired language, we should thus first head to corpora of spoken language. However, corpora of spoken languages are technically demanding to create and present a host of thorny principled problems: in addition to the significant difficulties presented by gathering representative data and the time needed to transcribe recordings, there are well-documented issues with how that data should be graphically represented and coded. Lemmatizing and tagging such corpora present difficulties, so there are further problems once we come to retrieving the data. The result is that even the largest and most thoroughly prepared oral corpora are often less than a hundredth the size of comparable written corpora, and do not allow the same functionality for searching, sorting and analyzing, nor

[4] As an introduction to these distinctions, Vachek (1989) is still an excellent source.

does every language have a publicly available corpus of spoken texts.[5] This means that if we use oral corpora to gather our data, we risk having non-representative samples, gathered on a necessarily limited range of topics, that do not fully reflect the language environment in which the speaker exists.

Three of the four corpus-related studies in this issue grapple with these matters to some extent. In the case of work on Croatian and Czech morphology (Lečić; Bermel, Knittl, and Russell) spoken corpora do not offer enough data to generalize about forms of individual words, as even in the very large written corpora many of these forms are relatively scarce. In the case of work on modality (Divjak, Szymor, and Socha-Michalik), corpus data were drawn from the IPI-PAN corpus, which is primarily composed of written Polish. The result is that while these studies appeal to notions of intuition and instinct in language and try to systematize an explanation of the linguistic structures in our minds, they do so on the basis of the written variants of these languages.

Mutual intelligibility research would seem to come from an entirely different angle, but it too has links to corpus data, and so the question of the modality of these data is relevant. The word lists for the experimental texts were created using frequency lists from the British National Corpus, a 100m-word corpus that is composed of 90 % written and 10 % oral texts. Here a different approach was taken and a variety of tasks were designed that tested intelligibility in both written and aural form. Because the work focused on EU languages, however, there was one significant outlier, which was Bulgarian. The particular difficulties that this posed for Bulgarian, the sole language in the survey written in Cyrillic script, are discussed in the article.

A matter arising from this is whether our corpora of written texts are adequate to the sorts of questions we are now trying to answer. Are our hypotheses correctly framed in terms of the sort of data we can get from a written corpus? If not, what can we do to refine them and incorporate the insights that can be gleaned from oral corpora?

4 Two directions to explore

The range of articles in this issue defies easy recapitulation, but in closing I will point out a couple of themes running through them that highlight two further trends in our field.

First, as alluded to in the previous section, there is a psychological turn in the discipline, with more mainstream linguists trying to relate the structure of language to questions that had up until recently been addressed mainly by psychologists. In part this may be due to the bridge provided by cognitive linguistics, which insists that language is not a separate module but is linked to our other capabilities and perceptions.[6] However, it is also an effect of the spread and accessibility of techniques used by psycholinguists, and our increasing understanding of what sort of conclusions can properly be drawn from corpus data. The

[5]The largest tagged, oral corpus of a Slavic language is the Oral Corpus (Ustnyj korpus) of the Russian National Corpus (Nacional'nyj korpus russkogo jazyka, www.ruscorpora.ru/search-spoken.html). According to Grišina and Savčuk (2009, p. 132), this corpus comprises over 7.5m tokens and is lemmatized and tagged. However, of this, only 10 % (2009, p. 133) is informal speech; the remainder consists of 'public' speech, i.e. radio and television broadcasts, and film dialogue, so only around 700,000 tokens represent natural spoken language. In addition, search and retrieval capabilities are severely limited by the tools of the NKRJa. Numerically speaking, the Slovak oral corpus is, at 5m tokens, the largest publicly available, tagged, lemmatized oral corpus of natural speech for a Slavic language (see http://korpus.juls.savba.sk/shk.html), although the sum total of data available for the three oral corpora of Czech is, at 4.79m tokens, almost the same size.

[6]For a recent summary of the argument against a language module, see Evans (2014, pp. 133–160).

techniques described in our contributions are all on the 'low-tech' side of the linguistics arsenal—there is no brain imaging, not even eye-tracking—but it does demonstrate the extent to which linguists are now trying to answer questions using tools and methods pioneered in other disciplines, and it highlights the rapid changes in one tool at least partly 'home-grown' in our own discipline of linguistics—the corpus—which allow us to use it in exploring some questions of a psychological nature.

Second, all of the papers make use of graphic representations at some level of analysis. This is partially a result of the growth of data referred to earlier: as the amount of material we have grows exponentially, we have to explore different ways of looking at it in order to make sense of it and ensure that it is comprehensible to the reader. Numerous linguists have explored facets of this problem before, and it seems that these explorations will be a growing part of the field as its interpretive and descriptive apparatus becomes ever more varied (see e.g. Eddington 2010). Graphed outputs of multidimensional scales (Divjak, Szymor, and Socha-Michalik; Golubović), classification trees (Bermel, Knittl, and Russell), box-and-whiskers plots and scatter plots (Cvrček and Chlumská) and line charts (Lečić) all appear as key aids to understanding dense tables of figures. In some analyses the graphics in fact are the main output of the analysis: the figures accompany them but are unilluminating without them. A major question for linguists is to ensure that at this level of transmogrification of our data into coloured lines, boxes and points, can we still hold on to the original connection back to what it was we collected? More importantly, these techniques are new enough to our field that we cannot yet assume a general understanding of what underlies them, and thus to a certain extent we assert their usefulness and hope that others are challenged to explore their strengths and weaknesses themselves. These papers cope successfully with this challenge, but as the field moves forward and the number and sophistication of these techniques increases, the need for a greater understanding of them and consensus about them is likely to become more, rather than less acute.

References

Baker, M. (1993). Corpus linguistics and translation studies. Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology. In honour of John Sinclair* (pp. 233–250). Amsterdam, Philadelphia.

Brown, D. (2007). Peripheral functions and overdifferentiation: the Russian second locative. *Russian Linguistics*, *31*(1), 61–76. doi:10.1007/s11185-006-0715-5.

Dąbrowska, E. (2005). Productivity and beyond: mastering the Polish genitive inflection. *Journal of Child Language*, *32*, 191–205. doi:10.1017/S0305000904006609.

Divjak, D. (2009). Mapping between domains. The aspect-modality interaction in Russian. *Russian Linguistics*, *33*(3), 249–269. doi:10.1007/s11185-009-9045-8.

Divjak, D. (2010). Corpus-based evidence for an idiosyncratic aspect-modality relation in Russian. In D. Glynn & K. Fisher (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches* (pp. 305–330). Berlin.

Eddington, D. (2010). A comparison of two tools for analyzing linguistic data: logistic regression and decision trees. *Italian Journal of Linguistics*, *22*(2), 265–286.

Evans, V. (2014). *The language myth. Why language is not an instinct*. Cambridge.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93–127. doi:10.1515/COGL.2009.005.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development*, *28*(6), 445–467. doi:10.2167/jmmd511.0.

Grišina, E. A., & Savčuk, S. O. (2009). Korpus ustnyx tekstov v Nacional'nom korpuse russkogo jazyka: sostav i struktura. In V. A. Plungjan (Ed.), *Nacional'nyj korpus russkogo jazyka: 2006–2008. Novye rezul'taty i perspektivy* (pp. 129–148). St. Peterburg.

Halliday, M. A. K. (1991a). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics* (pp. 30–43). London, New York.

Halliday, M. A. K. (1991b). Towards probabilistic interpretations. In E. Ventola (Ed.), *Functional and systemic linguistics. Approaches and uses* (Trends in Linguistics. Studies and Monographs, *55*, pp. 39–61). Berlin, New York.

Laviosa, S. (2002). *Corpus-based translation studies. Theory, findings, applications* (Approaches to Translation Studies, *17*). Amsterdam, New York.

Laviosa, S. (2003). Corpus and simplification in translation. In S. Petrilli (Ed.), *Translation, Translation* (Approaches to Translation Studies, *21*, pp. 153–162). Amsterdam, New York.

Nábělková, M. (2008). *Slovenčina a čeština v kontakte. Pokračovanie príbehu*. Bratislava.

Sloboda, M., & Nábělková, M. (2013). Receptive multilingualism in 'monolingual' media: managing the presence of Slovak on Czech websites. *International Journal of Multilingualism*, *10*(2), 196–213.

Tafel, K., et al. (2009). *Slavische Interkomprehension. Eine Einführung*. Tübingen.

Tang, C., & van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, *119*, 709–732. doi:10.1016/j.lingua.2008.10.001.

Thornton, A. M. (2012). Reduction and maintenance of overabundance. A case study on Italian verb paradigms. *Word Structure*, *5*(2), 183–207. doi:10.3366/word.2012.0026.

Vachek, J. (1989). *Written language revisited* (selected, edited and introduced by P. A. Luelsdorff). Amsterdam.

Yuan, Y., & Gao, F. (2008). Universals of translation: a corpus-based investigation of Chinese translated fiction. In R. Xiao, L. He, & M. Yue (Eds.), *Proceedings of the international symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2008), 25–27 September 2008. Zhejiang University, Hangzhou.* Retrieved from http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Yuan.pdf (1 June 2015).