Prepublication draft. To appear as:
Divjak, D.S., Sharoff, S. and Erjavec, T. (2017) Slavic corpus and computational linguistics. *Journal of Slavic Linguistics*. ISSN 1068-2090 (In Press)

# Slavic Corpus and Computational Linguistics

*After all, language does look different if you look at a lot of it at once*
(J. M. Sinclair)

## 1 Introduction

The Slavic languages provide a fertile ground for corpus-based and computational investigations. For one, the Slavic languages combined count more than 315 million speakers (Sussex and Cubberley 2006) who have produced and continue to produce massive amounts of data on a daily basis. The Slavic languages also display peculiar typological features, which make them more challenging from a computational perspective in comparison to other commonly studied languages of the Indo-European family. For example, due to the richness of their morphology, Slavic languages exhibit a relatively extensive freedom in word order in comparison to the Germanic and Romance languages. From a computational viewpoint, this property is challenging because it results in a greater data sparsity, i.e., there are considerably more surface realisations for a given underlying linguistic phenomenon, be it the number of forms a word can display, or the number of places the subject or object of a sentence can occupy relative to the verb. Take, for example, the number of morphological forms of a verb: if participial forms are counted, Russian transitive verbs yield about 80 different forms, while English verbs only have at most five different forms. Another example concerns the increased number of surface patterns for such pairs as Verb-Direct Object, because the position of the Direct Object can be quite flexible. Therefore, sparsity needs adequate representation in computational research on Slavic languages. Yet, the same morphological richness and regularity in inflection impacts computational studies of Slavic languages positively: it is possible to predict with reasonable precision the Part-of-Speech (PoS) category and the syntactic function of word forms from their endings, something that is considerably more difficult to achieve in the Germanic and Romance languages. Finally, the high regard in which both linguistics and mathematics are held in Slavic countries has yielded remarkable results. One of the earliest examples of research with Slavic corpora is the seminal paper by Andrei Markov, which concerned predictions of word sequences on the basis of *Eugene Onegin* (Markov 1913; Hayes et al., 2013). This study led to development of Markov models, which are commonly used in modern computational linguistics for predicting a linguistic phenomenon from the adjacent context.

Both corpus and computational linguists use corpora to study languages but there are fundamental differences between and within the two groups. As Renouf (2005) put it: “Corpus linguistics is essentially an arts-based discipline, while computational linguistics has a mathematical heritage, and though the latter is now increasingly engaging in textual study, the approaches remains philosophically distinct." Yet, even within the group of corpus linguists, some will use corpora to address theoretical questions, while others will turn to corpora to collect the data they need to produce an accurate description of a phenomenon, and others still are fascinated by how corpora should be compiled and structured to be representative of a population of language speakers and useful to a group of corpus users. The same difference can be observed within the computational linguistic community, where some researchers will use corpora to design and improve machine learning models that can provide insights in how language works, while others focus on solving practical problems (related to e.g. information retrieval or automatic translation) or on improving the corpus as a resource. In this chapter, we will focus on corpus-linguistic studies that address theoretical questions and on computational linguistic work that makes it possible to annotate corpora, thereby making them useful for linguistic work.

# 2 A corpus-linguistic perspective

In this section, we reflect on the essence of corpus linguistics. We will discuss why and how the approach was discredited by generative linguists in the second half of the 20th century, how it made a comeback through advances in computing and was adopted by usage-based linguistics at the beginning of the 21st century.

## 2.1 Chomsky vs. corpus linguistics

Corpus linguistics was discredited in the early 1950s by Chomsky. From a theoretical point of view, he suggested that the corpus could never be a useful tool for the linguist, because the linguist must model language competence rather than performance. A corpus is by its very nature a collection of externalised utterances - it contains performance data and is therefore a poor guide to modelling linguistic competence. A second major criticism was levelled at two - admittedly erroneous - assumptions that many early corpus linguists are said to have held (McEnery and Wilson 1999), i.e. that 1) the sentences of a natural language are finite and 2) the sentences of a natural language can be collected and enumerated. Chomsky countered that the only way to account for a grammar of a language is by description of its rules - not by enumeration of its sentences. Corpora are necessarily "finite and somewhat accidental (Chomsky 1957: 17). The syntactic rules of a language, on the other hand, are finite but give rise to infinite numbers of sentences.

Up to this day, Chomsky remains at odds with corpus linguistic approaches, even in their more recent guises, as becomes clear from excerpts of an interview with him (Andor 2004) – italics ours.

> "Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this. But maybe they're wrong. Maybe the sciences should just

> collect lots and lots of data and try to develop the results from them. [...] We'll judge it by the results that come out. So if results come from study of massive data, rather like videotaping what's happening outside the window, fine—look at the results. I don't pay much attention to it. I don't see much in the way of results. My judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to *accumulate huge masses of unanalyzed data and to try to draw some generalization from them*. The modern sciences, at least since Galileo, have been strikingly different. What they have sought to do was to *construct refined experiments which ask, which try to answer specific questions that arise within a theoretical context as an approach to understanding the world*." [italics ours]

Although Chomsky was wrong implying that scientists do not draw generalizations from looking at massive amounts of data (the early astronomers made vast numbers of detailed observations and modern astrophysicists very literally do "videotape what's happening outside the window", or on the sun for that matter), he does make two relevant points here that have been debated in linguistic circles more generally: 1) the fact that disagreement exists about what corpus linguistics is - a theory, a discipline, or a method – as well as about the type of things a corpus linguist is interested in and 2) what information a corpus can contribute to (theoretical) linguistic inquiry. We would like to add a third issue that was not raised by Chomsky but that has been debated extensively over the past 10 years, i.e. the lack of awareness of and training in proper data handling techniques that is widespread among linguists using corpora.

## 2.2 What is Corpus Linguistics? A theory, a discipline, or a method?

A rather innocent announcement for a Bootcamp in Quantitative Corpus Linguistics, posted on August 12 2008 to the Corpora List[1], sparked a discussion that continued for nearly three weeks and revealed a deep divide in the corpus linguistic community: disagreement exists about whether corpus linguistics is or should be a theory, a discipline, or a method. We will consider each of the options in turn.

### 2.2.1 Corpus-theoretical approach

A theory is a system of ideas, intended to explain something. Ideally, a theory is based on general principles independent of the thing to be explained. The corpus-theoretical approach rests on the cyclical principle of minimal assumptions. Such a theory is constantly in the making and develops according to the requirements of the data. A corpus-theoretical approach would rely on what Tognini-Bonelli (2001) termed "corpus-driven" work, as opposed to work that is merely "corpus-based".

[1] http://mailman.uib.no/public/corpora/2008-August/007064.html

*Corpus-based* approaches bring models of language which are believed to be fundamentally adequate to the analysis and analyse the corpus data through these categories. Examples of such categories are things that most linguists take for granted, such as words or part-of-speech tags, for example. *Corpus-driven* approaches derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context. Linguists should only accept parts of speech to the extent that the data they are examining supports these distinctions. Here, the question arises of how practical such an extreme data-driven approach is: does querying everything, time and again, really make for better analyses? Mahlberg (2005: 32) proposed, by way of compromise that "[f]or the time being, a corpus linguistic theory seems to be best regarded as the rationale that governs the investigation of data and that determines how the results of the investigation are accommodated within a language description."

There is also a third strand, so-called *corpus-illustrated* work (Tummers et al. 2005), which subjects examples extracted from corpora to a largely traditional introspective analysis and does not require statistical analysis. The vast majority of corpus linguistic work on Slavic languages takes this approach.

### 2.2.2 Corpus linguistics as a discipline

A discipline is an area of study and a branch of learning that has developed a terminology and methodological conventions for the study of something and has accumulated a body of knowledge, stored and accessible in the form of published inventories, analyses and descriptions

Over the past 30 years corpus linguistics has, no doubt, established itself. There are now plenty of conferences, journals, book series and discussion lists at or in which the results of corpus linguistic studies can be presented. Some examples of these would be the biennial Corpus Linguistics conference held since 2001 in the UK, or its American counterpart, the conference of the American Association for Corpus Linguistics, that has been held since 1999. Several publishers offer corpus linguistics journals, e.g. the oldest journal, ICAME, has been around since 1978. The *International Journal of Corpus Linguistics* has been published by John Benjamins since 1996, *Corpus Linguistics and Linguistic Theory* since 2005 by De Gruyter and *Corpora* has been published by Edinburgh University Press since 2006. Many publishing houses also offer book series devoted to corpus linguistics, e.g. *Studies in Corpus Linguistics* by John Benjamins and *Language and Computers: Studies in Practical Linguistics* from Brill lists more than 75 volumes each. Finally, the Corpora List brings together all researchers interested in developing or using corpora. But, does this make corpus linguistics a proper discipline? Renouf (2005) aptly states that, as a discipline, corpus linguistics would still be "an amalgam of great precision and best endeavours; a somewhat undisciplined discipline. Yet as a branch of empirical study, this is ultimately its purpose, for empiricism precludes any a-priori assumptions".

### 2.2.3 Corpus linguistics as a method

Could corpus linguistics be a method then? And what would this method achieve? Chomsky even doubted that corpus studies would qualify as method. In his opinion, corpus linguistics is like observing the tides:

> "If you want to use hints from data that you acquire by looking at large corpuses [sic], fine. That's useful information for you, fine. I mean, Galileo might have gotten some hints from looking at events that were happening in the world. In fact, he did. He observed the tides—that's like corpus linguistics. You're observing the tides." (Andor 2004: 99)

Yet, this picture reveals a stubborn misconception: corpus linguists do not merely observe. Yes, the corpus linguist selects a random sample from a representative and balanced collection of texts that represent one or more varieties of the language s/he is studying. S/he sticks to that sample, and is not allowed to include additional sentences that would nicely illustrate his/her point, nor to remove sentences that disprove his/her account. In traditional studies, this object of study is typically presented in the form of KWIC (keyword in context) concordance lines, and this presentation inclines the researcher to scan the item serially within an ordered, usually alphabetical context. But, at the heart of a corpus-based study lies the (often still manual) annotation of examples. In order to do this in a verifiable way, the corpus linguist needs to operationalize linguistic parameters so that they can be applied consistently to a large number of examples. This introduces rigour and objectivity into the analysis, and makes it possible to feed the annotated sample into a computer for statistical analysis.

Unfortunately, corpus linguistics remains (or used to remain) silent on the mechanics underlying this process: there is no specified convention for matching a hypothesis against textual reality, or vice versa, or even a requirement for an articulated hypothesis. Corpus linguistics does not espouse particular statistical methods, or demand statistical rigour, even though some statistical measures (e. g. relative frequency, chi-square) are commonly applied (Renouf 2005).

Is there a way in which we could move from (being accused of) *"accumulat[ing] huge masses of unanalyzed data and try[ing] to draw some generalization from them"* to using corpus linguistic methods to *"answer specific questions that arise within a theoretical context as an approach to understanding the world"*? This is the question we explore in Section 2.3.

## 2.3 Corpus linguistics and linguistic theory

Over the past ten to 15 years, there has been a successful rapprochement between corpus linguistics as a method and usage-based theoretical approaches such as cognitive linguistics.

Cognitive Linguistics (henceforth CL) is a usage-based model of language structure (Langacker 1987: 46), a "data-friendly" theory, with a focus on the relationship between observed form and meaning. CL rejects that idea that language would be innate and imposed top-down by Universal Grammar, but instead explores the hypothesis that language would be built bottom-up, from exposure to actual usage. Cognitive linguists do not make a distinction between competence and performance: all aspects of grammatical knowledge are derived from the language users' experience with frequent strings of concrete linguistic expressions.

### 2.3.1 Frequency and usage-based linguistic theory

Corpora are very attractive sources of data for linguists working in the usage-based tradition and the probabilistic turn in grammar research was influenced by the rise of corpus linguistics and the development of new statistical and computational tools for the analysis of quantitative data.[2] Corpus linguists are not interested in (the frequency of) sentences - they look for (frequencies of) patterns (Stefanowitsch 2005: 295). Corpora give access to form frequency and distribution in naturalistic settings. Frequency is among the most robust predictors of human performance (Hasher and Zacks 1984) and human beings extract frequency information automatically from their environment. Given this, they can use statistical properties of linguistic input to discover structure, including sound patterns, words and the beginnings of grammar (Ellis 2002). The ability to extract the distributional characteristics of natural language plays a key role in linguistic development: what we learn may well be a probabilistic grammar grounded in our language experience. The extent and strength of influence of frequency of occurrence on processing supports a dynamical model of grammar: the frequency with which linguistic forms are experienced forms the core of our grammatical knowledge. In other words, pattern extraction abilities in humans could circumvent the need for environmental linguistic triggers to set parameters specifying a fixed set of mutually exclusive linguistic properties, as Generative Grammar assumes.

### 2.3.2 Corpora and the form-meaning relationship

Usage-based linguistics thus provides corpus linguists with a theoretical framework that generates hypotheses that can be tested against corpus data, and this is particularly relevant for colleagues working on Slavic languages. As Divjak, Kochańska and Janda (2007) argued, from its early days, cognitive linguistics has attracted the attention of linguists with research interests in Slavic languages, to name but a few: Cienki (1989), Dąbrowska (1997), Janda (1986), Rudzka-Ostyn (1992). This is not surprising, for at least two reasons. Politics have played a crucial role in bringing the Slavic linguistic tradition and the cognitive paradigm close to each other. The Cold War era was the time when Eastern European linguists in general and Russian linguists in particular were largely isolated from theoretical discussions in the West, due in part to the political writings of Chomsky, which led to his entire oeuvre being censored. As a consequence, East-European linguists were never forced to experiment with autonomous theories of language, but rather maintained focus on the form-meaning relationship and how it is embedded in the larger reality of human experience. Some of their theories and models became known in the West. These include the Russian Meaning↔Text framework, first developed by Mel’čuk (1988) in Moscow and the Natural Semantic Metalanguage theory formulated by Wierzbicka (see Wierzbicka (1972) for the first book-length treatment).

Most of the work done in Eastern Europe, however, never made it to the other side of the Iron Curtain, which is all the more regretful since one of the founding assumptions of cognitive and functional linguistic theories that are currently holding sway in the West has been present

[2] Note that proponents of the “corpus as a theory” do not consider this a happy marriage: “For cognitive linguists, meaning is in the individual, monadic minds of speakers and hearers; for corpus linguists, meaning is in the discourse (or the corpus, as a sample thereof)” (Teubert, 14 August 2008, corpora list).

in Slavic linguistics all along: Slavic linguists have always recognized the fundamentally symbolic nature of language and hence the fact that diverse formal aspects of language exist for the purpose of conveying meaning. Another, and very vigorous tradition of formal and then computational research on Slavic languages, is the Prague tradition with its attention to linking the form, which is computationally tractable, to its function, which needs to be inferred (Sgall, 1995). On the basis of this Functional Syntax theory a large syntactically annotated corpus, the Prague Dependency Treebank, was manually annotated (Hajičová 1998; Hajič 1998), and used in various research, as well as a number of other annotated corpora and tools for processing various registers of the Czech language.

### 2.3.3 Linguists and data handling

Unfortunately, linguists, including those taking a corpus-based approach, have been lacking in data handling hygiene.

For one, they have not necessarily been very concerned about *formulating hypotheses to test their theory* (Divjak 2015; Dąbrowska 2016). Yet we need to formulate hypotheses, derive testable predictions from our hypotheses, carry out the tests, and use the results to refine the hypotheses - and the theory - when necessary. This is part and parcel of the scientific cycle (Kuhn 1962) and corpora offer unrivalled opportunities to test theoretical hypotheses on naturally occurring data.

Early corpus linguistics required data processing abilities that were simply not available at that time, and quantitative insights that had not yet spread beyond the sciences. Yet, it wasn't until the turn of the century that the principle of "total accountability" began to be adhered to in corpus linguistic work and researchers started to respect the requirement to look at all available examples, not just cherry-picking the ones that look nicest in their argument. Typically, the corpus was (and often still is) considered a repository of examples, and there is no systematic approach to addressing all the evidence. Arnold et al. (2000) were among the first to analyze post-verbal word order variation in English across the total set of instances from the corpus search, and the results were subjected to quantitative analysis (significance testing, correlation and regression). Over the past 15 years, tremendous progress has been made and more and more studies are being published that respect the principle of total accountability. With this, the insight has come that raw frequencies are not in and of themselves, or directly, of any scientific interest. While as little as 10 years ago the fact that quantitatively sophisticated corpus-based argumentation was required remained something for which the case had to be made (cf. the discussion about the observed-frequency fallacy versus the expected frequency epiphany (Stefanowitsch 2005: 296)), we can now speak of a Quantitative Turn. Janda (2013) presents a unique selection of seminal articles that together have caused the Quantitative Turn in (Cognitive) linguistics. But we can go further still: Milin et al. (2016) recently made the case for relying on modelling techniques that are based on biologically and psychologically plausible learning algorithms if we are to use a quantitative approach to advance our understanding of how knowledge of language emerges from exposure to usage.

The sum up, the criticisms that Chomsky levelled against using corpora for linguistic research have been addressed by the corpus linguistic community, and by cognitively and functionally minded corpus linguists in particular. They have shown how corpus data can be used to "*construct refined experiments which ask, which try to answer specific questions that arise within a theoretical context as an approach to understanding the world"*. Of course, this work would not have been possible without large collections of text that are annotated with relevant information. The automatic annotation of text collections, is one of the core tasks of computational linguistics.

# 3 Computational perspective

Many computational linguistic tasks can be seen as a process of text annotation. Such tasks are especially relevant for corpus linguistics, as they result in corpora that are annotated with linguistic information that can be then profitably used in further linguistic research. This section therefore gives an overview of the most necessary and common annotation layers and the issues that are encountered when performing such automatic annotation, with special emphasis on Slavic languages. We start, however, with an excursion into history to explain how the need for large annotated text collections arose.

## 3.1 Rules vs. Machine Learning

Up until the 1990s the approaches for computational processing of languages, including Slavic ones, were based on developing a system of rules. The rule-based approach was, especially in the U.S., primarily applied to syntax following Chomskyan grammar, while in Europe, including Slavic speaking countries, this approach was more evident in the computational formalisms for morphological analysis. These were typically based on finite-state methods and their implementations, such as Intex (Silberztein 1993), with manually constructed rules, e.g., to derive all acceptable surface forms for a set of lemmas (Vitas et al. 2003) or to generate possible interpretations of a given word form. A Russian word form such as *душа* can have several interpretations, i.e., the nominative case of a feminine noun ('soul'), the genitive case of a masculine noun ('shower') or the gerund form of a verb ('strangle').

With more data becoming available in electronic form, towards the end of the 1990s the dominant rule-based paradigm shifted towards the use of Machine Learning (ML) (Manning and Schütze 1999), i.e., a set of methods to find and exploit regularities in data. In the case of supervised ML, this procedure is based on existing — typically manual —annotations on the desired level of linguistic description. For example, instead of a linguist formulating syntactic rules, a ML method can detect statistical patterns in a large number of manually syntactically annotated sentences, such as the patterning of the PoS tags of words, and produce a model for their automatic annotation. Some ML methods produce weighted or ordered rules that can be inspected (e.g., decision trees), while others, and these are now becoming a majority, are, to a greater or lesser degree, "black box" systems, with the models being very large matrices of statistical values over combinations of the defined features. Such systems are very useful in practice, as they can be used to automatically annotate corpora or texts, i.e. they are used as language technology tools for application purposes, such as PoS tagging, parsing, or machine translation. It is also possible to draw statistical inferences without a manually annotated corpus, i.e., using Unsupervised Machine Learning. An early example of this approach for linguistic purposes is Biber's Multi-Dimensional Analysis (Biber 1988), which detects statistically significant grouping of features, such as the greater amount of noun phrases and nominalisations vs. the use of personal pronouns and stance verbs.

The focus in development of tools for annotating (Slavic) languages has thus moved from manually developing rules (grammars) and lexica for their processing to manually annotating corpora with the phenomenon under investigation, and relying on (combinations of) largely generic ML methods. Still, to engineer the required features for the problem and use the optimal set of parameters for training the model, linguistic insights into the problem are necessary.

Current ML methods can, for most annotation tasks, already take into account much more contextual information than manually constructed rules ever could. For example, rule-based Machine Translation systems were much less successful in dealing with ambiguities in language when compared to the current ML-based approaches, since too many fairly subtle rules are needed for resolving the ambiguities, while Statistical MT easily "plagiarises" a large number of translation examples. A rule-based approach needs a lot of information to translate ambiguous words, e.g., *конёк* in Russian as 'small horse', 'skate', 'seahorse', 'hobby' or 'roof ridge', while a Statistical MT can efficiently memorise the most frequent contexts of use.

Another reason for their success is that ML methods are largely language-agnostic, so a working model can be built from nothing but a large number of examples (Sharoff and Nivre 2011).

## 3.2 Part of Speech tagging

One of the early applications of computational linguistics is the automatic detection of morpho-syntactic properties of words in text (Nikolaeva 1958; van Halteren 1999). This task is commonly known as Part-of-Speech (PoS) tagging, mainly because it was first developed for English, where the main problem is to disambiguate between the PoS of words in context, e.g., to determine whether "walk" in "He likes to walk" or "He took a walk yesterday" is a verb or a noun. While typical English PoS tag sets also take into account some other lexical (e.g., common or proper noun) and inflectional (e.g., singular or plural) properties of words, English, as an inflectionally poor language, has only a few of the latter, so the tag sets for English, as well as for most Western European languages, are rather small, with about 20 – 100 different tags, e.g., NN for singular common nouns, NNS for plural, etc.

The situation is quite different for Slavic languages, where all the morpho-syntactic properties are typically encoded in the tag sets. In the multilingual MULTEXT-East specifications (Erjavec 2012), Slavic languages (apart from Bulgarian and Macedonian) have a tagset of over 1,000 "PoS" tags to cover different inflectional categories. Given the wealth of information encoded in tagsets covering Slavic inflections, it is better to call such tags morpho-syntactic descriptions (MSDs), a practice we adopt in this paper. While English typically used "synthetic" tags, where each tag had a legend explaining what it means, a more structured approach is needed for the large Slavic MSD tagsets. A commonly used approach, first proposed in the scope of the EAGLES (Expert Advisory Group on Language Engineering Standards) project (EAGLES 1996) is to use a position-based encoding, where each attribute is given a position in the MSD string and its value is a character, with the specifications giving the mapping between attribute-value pairs and their encoding. So, for example, the MULTEXT-East specifications define that the MSD `Vmen` corresponds to the feature-structure `Verb, Type=main, Aspect=perfective, VForm=infinitive`. Some approaches then dispense with the MSD tagsets altogether, and only retain the features. The latest and most important development in this respect is the Universal Dependencies (UD) project (Nivre et al. 2016), which has the ambition to cover all languages, not only for morpho-syntax, but also for dependency syntax.

As with other levels of annotation, in order to automatically tag a corpus with PoS tags or MSDs, modern approaches rely on machine learning. To train a tagger for a new language, three components are usually required:

1. a training corpus, in which each word is marked with its morpho-syntactic features in a given context, the following Slovenian example can be represented as:[3]

   *Dogodek v Ankaranu je bila dramatična nesreča*
   'The event in Ankaran was a dramatic accident'

| **Form** | **Lemma** | **UD PoS** | **MSD** | **UD features** |
|---|---|---|---|---|
| Dogodek | dogodek | NOUN | Ncmsn | Case=Nom, Gender=Masc, Number=Sing |
| v | v | ADP | Sl | Case=Loc |
| Ankaranu | Ankaran | PROPN | Npmsl | Case=Loc, Gender=Masc, Number=Sing |
| je | biti | AUX | Va-r3s-n | Mood=Ind, Negative=Pos, Number=Sing, Person=3, Tense=Pres, VerbForm=Fin |
| bila | biti | VERB | Va-p-sf | Gender=Fem, Number=Sing, VerbForm=Part |
| dramatična | dramatičen | ADJ | Agpfsn | Case=Nom, Degree=Pos, Gender=Fem, Number=Sing |
| nesreča | nesreča | NOUN | Ncfsn | Case=Nom, Gender=Fem, Number=Sing |
| . | . | PUNCT | Z | |

2. a separate lexicon with compatible morpho-syntactic annotations to cover cases not present in the training corps, for example:

| | | | | |
|---|---|---|---|---|
| Ankarana | Ankaran | PROPN | Npmsg | Case=Gen, Gender=Masc, Number=Sing |

3. a tool for building disambiguation models from such annotations, for example, for learning that the elements of noun phrases agree in case, number and gender:

| | | | | |
|---|---|---|---|---|
| dramatična | dramatičen | ADJ | Agpfsn | Case=Nom, Degree=Pos, Gender=Fem, Number=Sing |
| nesreča | nesreča | NOU | Ncfsn | Case=Nom, Gender=Fem, Number=Sing |

[3] From the Slovenian Universal Dependencies Treebank Version 1.4, http://hdl.handle.net/11234/1-1827

N

The Machine Learning frameworks for PoS tagging learn the probabilities of sequences of tags in the traditional taggers (Brants 2000; Schmid 1994) or the matches between the morpho-syntactic features in their more modern versions (Müller et al. 2015). For example, in the Russian example of *душа* discussed above, the context provides sufficient information to choose the right interpretation via the sequence of probabilities:

*У него преобладает не интеллект, а душа.* (coordination of two nominative cases) 'For him, reasoning is more important than soul.'

*нельзя было помыться после игры из-за отсутствия душа.* (genitive often follows a noun) 'it was impossible to wash after the game since there were no showers.'

*Враги катались по земле, душа друг друга*. (availability of a direct object) 'The enemies rolled over the ground, trying to strangle each other'

# 4 Research activity

In this final section, we survey the types of research requiring corpora that Slavic linguists are involved in world-wide, and the resources they have at their disposal.

## 4.1. Self-reported data on research activity based on corpora

Overall, linguistic research for Slavic languages that relies on corpora is well-represented. The information we received in response to our survey, posted on the Corpora list July 2016, is visually summarized in the map in Figure 1 below. The dots are proportional to the number of linguists who filled out the survey per country. The US is not pictured due to space constraints. In total, linguists from 28 different countries responded. Those with 5 or more responses are listed in Table (1) below.

| **Country** | **Number of self-reported corpus and computational linguists** |
|---|---|
| Poland | 21 |
| Czech Republic | 19 |
| Germany | 18 |
| USA | 14 |
| Russian Federation | 13 |
| Bulgaria | 6 |

| | |
|---|---|
| Croatia | 5 |
| Norway | 5 |

Table (1): Countries with 5 or more self-reported corpus or computational linguists.

Three linguists each from Slovenia and the UK (excluding the current authors) self-reported, two each from Austria, Canada, Italy, Slovakia, Ukraine and one each from Belarus, Denmark, Finland, France, Georgia, Ireland, Macedonia, Romania, Serbia, Switzerland,
However, it should be noted that this is self-reported data - the fact that a country is not represented does not mean that those countries do not employ corpus or computational linguists and the numbers reported are not official counts.

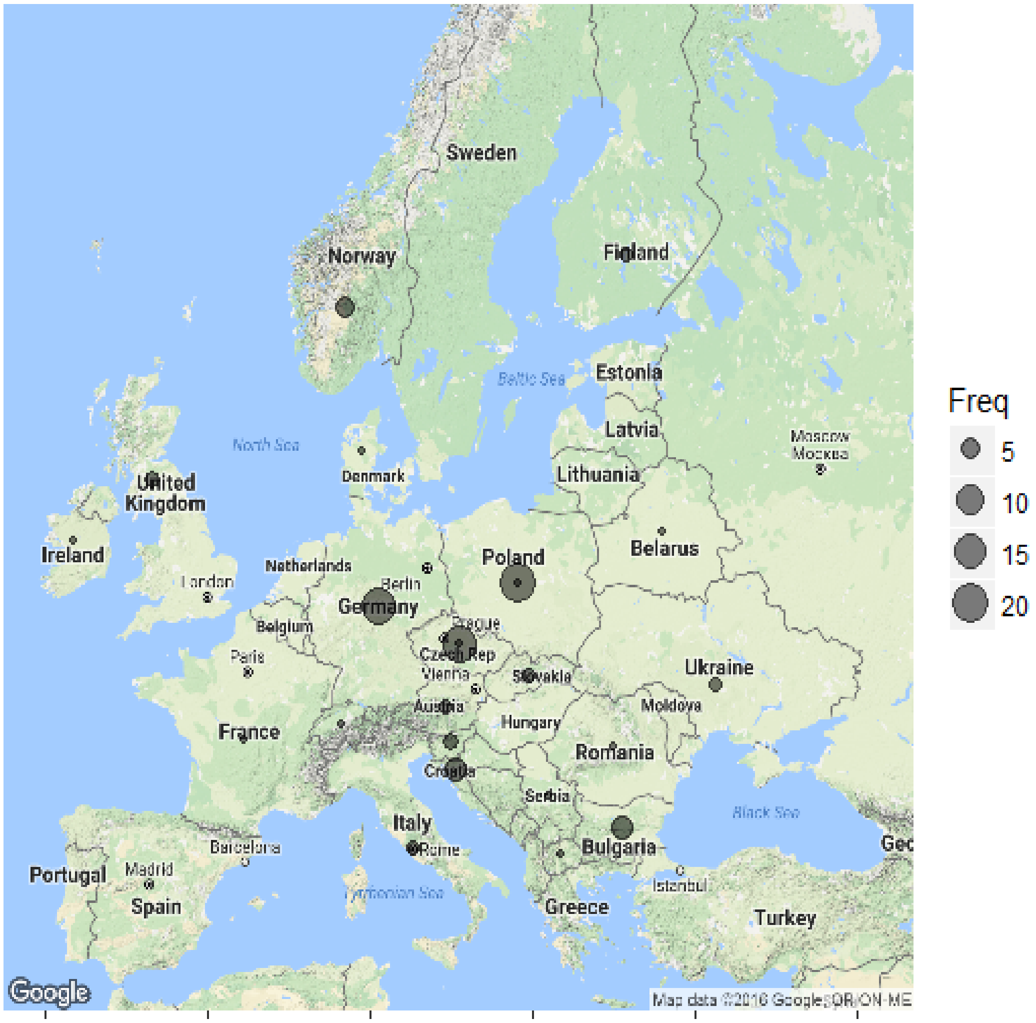

Figure 1: Map depicting number of (self-reported) corpus linguists working on Slavic language across Europe. The dots are proportional to the number of linguists who filled out the survey per country.

Overall, three types of research are being conducted:

1. **Computational linguistic research**, focusing on compiling corpora (historical and contemporary; individual, comparative, parallel; translation; learner) and developing the NLP tools to annotate and mine them morphologically, syntactically and semantically (treebanks, ontologies, wordnet, word sense disambiguation, construction identification); machine translation; information retrieval; sentiment analysis; topic modelling
2. **Linguistic research that uses corpora** as rich sources of authentic examples in a wide variety of domains, from the core areas of linguistics (morphology, syntax, semantics) branching out into discourse-pragmatics (politeness, anaphora resolution) and sociolinguistics (including dialectology, language variation and change); historical linguistics (grammaticalization); cultural linguistics; bilingualism (and language contact); teaching methodology;
3. **Quantitative corpus-based studies within a usage-based framework**; mostly study "rival forms", variation of some kind, phenomena that resist being captured by a rule. Very few respondents work within a formal framework (LFG, HPSG or generative grammar; semantics, pragmatics). Chomsky's spell still binds ...

In the next sections we focus on the last category, the "quantitative corpus-based studies within a usage-based framework" because it is this type of work that refutes Chomsky's criticism.

## 4.2. Corpus linguistics studies

We will highlight a few strands of research that use corpus linguistic methods to *"answer specific questions that arise within a theoretical context as an approach to understanding the world"* (Section 4.1.1.1) and to do so in a methodologically sound way that introduces new techniques to the field (Section 4.1.1.2). Because of this, these studies have found resonance outside the domain of Slavic linguistics in which they originated. Interestingly, the overwhelming majority of these studies relies on insights from Cognitive Linguistic Theory - as argued before, Cognitive Linguistics is a data friendly theory, and as such, ideally suited to be tested against data from corpora.

### 4.2.1. Answering theoretical questions

Work by (Divjak  2003a/b) focused on capturing how verbs are used in context to determine their exact meaning and distinguish between even the most semantically similar words such as near-synonyms. Some of these studies are described in Section 4.2.2. Here we will focus on the fact that these studies put Behavioral Profiles (Divjak 2006; Divjak and Gries 2006) on the corpus linguistic agenda and spawned diverse types of studies on Linguistic Profiling.

The BP studies start from two basic assumptions, rooted in Cognitive Linguistics. On the one hand, all levels of linguistic analysis - morphology, syntax and semantics - are expected to convey meaning and are therefore potentially relevant for determining a word's lexical core. Because of this, until more knowledge has accumulated, we should not focus a priori on one level, e.g. co-occurrence semantics, discarding the other levels. The effects of incorporating one or more levels in the analysis were illustrated in Divjak (2006). On the other hand, all annotation should be *naive*, that is, directly accessible to speakers and should not require linguistic abstraction. For example, BPs do not expect native speakers to be able to identify an inanimate subject or a past tense, but we do expect them to know whether something is alive or whether an event has already happened. In the annotation, the linguistic labels (e.g. "past tense") corresponding to the experience (of an event that has happened) were used, but merely as shorthand; no linguistic knowledge on the part of the speaker is implied. That is, speakers do not need to be able to label something as a past tense to be sensitive to the experience of "pastness".

Members of the CLEAR group at the University of Tromsø in Norway[4] have focused on individual dimensions of the Behavioral Profile. Solovyev and Janda (2009) analyzed 500 corpus sentences for each of six Russian synonyms for 'sadness' and five synonyms for 'happiness'. They annotated the dataset with properties capturing the nouns' *constructional profiles*. More concretely, they tracked the statistical distribution of case marking on the noun and the presence (or absence) of prepositions. Their data revealed that each noun has a unique constructional profile.

Janda and Lyashevskaya (2011) analyzed verbs in terms of their *grammatical profiles*, the "relative frequency distribution of the inflected forms of a word in a corpus" (Janda and Lyashevskaya 2011: 719). In their study, they collected data on the distribution of Tense-Aspect-Mood across 1,575 pairs of verbs, representing 5,951,250 verb forms in the Russian National corpus. The result show that there is a strong attraction between certain lexical items and specific constellations of TAM markings, and that this attraction is motivated semantically. Their study also contributed to an important general discussion in corpus methodology, i.e. the level of granularity at which annotation should be carried out, concluding that the appropriate level is determined by the language and the linguistics phenomenon under scrutiny. Grammatical profiles have been used extensively in historical linguistic work as well (Eckhoff and Janda 2014; Nesset, Janda, and Eckhoff 2014a/b).

Janda and Lyashevskaya (2013) explore the power of *semantic profiles* to test the hypothesis that Russian verbal prefixes express meaning even when they are used to create a purely aspectual pair. This contradicts the traditional assumption that prefixes in this function are semantically empty. Relying on the semantic tags provided by the RNC, they analyze 382 perfective partner verbs with *po-, s-, za-, na-* and *pro-*, five of the most common verbal prefixes in Russian. They found evidence of a significant relation between semantic tags and prefixes and were able to pin down the meaning of each prefix by relying on the semantic tags it attracts or repels. This confirms their hypothesis that verbs choose the prefix that best fits their lexical meaning when forming a perfective counterpart.

Kuznetsova (2015) presents an overview of studies in linguistics profiling, including diachronic profiles. As Kopotev, Lyashevskaya and Mustajoki (forthcoming) conclude, "the list of profiling types can be easily continued if we take into account word order, syntactic and

[4] https://en.uit.no/forskning/forskningsgrupper/gruppe?p_document_id=344365

semantic roles, narrator's viewpoint or any other kind of linguistic features.” Yet, original BP studies did *not* separate out the dimensions in order to establish whether corpus data can be used as a shortcut to a cognitively realistic representation of language knowledge: analyzing the morphology, syntax, semantics, pragmatics separately is something that may suit linguists, but there is no proof that this is the approach taken by speakers of a language, hence caution is needed when a dimensional approach is taken within a cognitive linguistic framework.

### 4.2.2. Introducing statistical techniques to the field

Much of the methodological progress made in the analysis of corpus data involves the use of ever more advanced statistical techniques. The used of binary logistic regression is well-attested in Slavic corpus linguistics, such as the study by Sokolova, Lyashevskaya, and Janda (2012), explained in more detail in the Chapter on Cognitive Linguistics by Stephen Dickey and Laura Janda. The idea of “rival forms” is one that permeates morphological work on Slavic languages as well. The morphological richness these languages display have yielded an (over)abundance of choice in places, and questions of what might motivate native speakers in choosing between two options is something that corpus linguistic approaches are particularly well-suited to answer.

In collaboration with Baayen, Slavic linguists have explored and compared a number of statistical techniques to predict which of a number of forms, typically two, will be chosen given a range of properties (Baayen et al. 2013). Choices between more than two option are possible too, and again Slavic linguists were one of the first to do so: the choice between 6 near-synonyms expressing TRY was modelled using polytomous logistic regression model (Divjak 2010a). This study is described in more detail in the Chapter on Cognitive Linguistics by Stephen Dickey and Laura Janda.

Possibly due to the complexity of the languages they analyze, linguists working on Slavic languages have been instrumental in embedding statistical analysis techniques within corpus linguistics, and in expanding the range of techniques used. Importantly, Slavic linguists turned to more sensitive models early on.

Divjak turned to mixed effects binary logistic regression to study the relationship between aspect and modality in modal chunks of the type “modal word + infinitive” in Russian (Divjak 2009) and Polish (Divjak 2010b) in a custom-made 1-million word parallel corpus, now part of the Parallel Corpus of Slavic and Other Languages [http://www.slavist.de/]. The results of this study conrm that the general linguistic hypothesis linking perfective aspect to deonticity needs reversing for Russian, where the imperfective goes hand in hand with deonticity, while the perfective favours dynamicity. In addition, it was found that the relation between aspect and modality is mediated by a variable that outperforms even the reversed hypothesis in predicting aspectual choice in modal contexts. The meaning of this variable, State of Affairs applicability (generic vs. specic), predicts aspectual choice in modal constructions better than the lexical meaning of modality (dynamic vs. deontic) since the concepts captured by the ‘generic vs. specic’ parameter are an abstraction of the major constraints on aspectual form in Russian.

Mixed effects modeling method was also applied by Janda, Nesset and Baayen (2010) to study an ongoing suffix shift in Russian verbs, a diachronic change in which the suffix -a is being replaced by the productive suffix -aj. Corpus data show that the Russian suffix shift is not taking place uniformly. Using insights from cognitive linguistics, Janda, Nesset and Baayen (2010) approached the paradigm as a prototypical category with centre and periphery and found evidence for the fact that the peripheral forms of a paradigm, such as the gerund, are more affected by language change than the prototypical forms, such as the 3sg, which is insulated from change.

In both studies, mixed-effects modeling served as a tool to take the complex interdependencies in language data into account, such as multiple observations per author, per word or per paradigm, which results in more reliable inferential models.

### 4.2.3 Bridging the gap between corpus and computational studies

Recently a series of studies has appeared that bridges the gap between corpus and computational work, that is, work that uses computational linguistic techniques to extract data from corpora to answer specific linguistic questions. The Needle-in-a-Haystack Method (NHM), elaborated by Fidler and Cvrček, provides a quantitative method for text analysis. Fidler and Cvrček (2015), for example, use corpus-linguistic methods to examine the relationship between language usage patterns and divergence in text interpretation in Czech. They analyse a set of texts (Czechoslovak presidential New Year's addresses from 1975 to 1989) and contrast the texts statistically with corpora from two different periods: one from the totalitarian period and the other from the contemporary (post-totalitarian) period. The comparison was based on the Difference Index, an effect-size estimator, which was used to enhance the interpretation of keyword analysis outcomes. The two analyses yield significantly different results: the data from the analysis using the contemporary corpus were commensurate with contemporary readers' impressions; those from the analysis using the totalitarian corpus fluctuated in tandem with (and sometimes in anticipation of) political and social changes during the 15-year period and suggested an interpretation of the texts by a reader more familiar with totalitarian texts.

# 5. Language resources

The term "language resources" refers to any digital language data — in particular various types of language corpora and lexicons — that can be, inter alia, used for research on language. Such resources are the basis for corpus linguistics, but these resources must be, to be truly useful, made available to other researchers. This has long been the case for large, national reference corpora, typically automatically annotated with MSD tags and lemmas, which are available for a number of Slavic languages, e.g., the Russian National Corpus (Sharoff 2005),[5] the National

5http://www.ruscorpora.ru/

Corpus of Polish (Przepiórkowski et al. 2008),[6] the Czech National Corpus,[7] and many others. Parallel corpora are also available, e.g., ParaSol.[8] However, these corpora are only available for on-line searching with specialised concordancers, but not for downloading, mostly due to issues of copyright. Yet, having access to the complete corpus for downloading and local data crunching is a prerequisite for many more sophisticated linguistic analyses and, esp. in the case of manually annotated corpora, for training machine learning annotation software.

Probably the first publicly downloadable and manually annotated corpora were produced for the Czech language, starting with their morpho-syntactically and syntactically annotated Prague Dependency Treebank (PDT) (Hajič 1998; Hajič et al. 2006). At almost the same time, the MULTEXT-East project (Erjavec 2012) released the first version of its multilingual annotated corpus. Because it only contained the novel "1984" by George Orwell in the original and its many translations, it was of somewhat limited use for linguistic investigations. Yet, the project also yielded medium-sized morphosyntactic lexicons, which were the first to cover many Slavic languages.

In an effort to map this landscape, the META-NET project[9] has edited a series of (bilingual) books that cover the EU languages and give the landscape of their language resources and technologies in the digital age[10]. According to the series, the resources and technologies available for the Slavic languages are still considered to be "fragmentary" or "weak". The only exception here is the Czech language, which has a long tradition in (computational) linguistics and has also substantially benefited from the country's membership in the EU, with many projects over the years investing into building corpora and tools for its processing. Although not as vigorous as for Czech, other Slavic languages also have a long tradition in computational linguistics, where most of the research initially concentrated on producing inflectional lexica for the languages, e.g., for Bulgarian (Paskaleva et al. 1993) or Polish (Vetulani et al. 1998).

The situation, with some corpora available for on-line exploration and very few language resources available for download persisted well into the 21 century, when it slowly began to change. The main reasons for this were that it became increasingly obvious that producing language resources, the use of which is subsequently limited only to their developers, leads to great duplication of effort, where researchers are forced again and again to repeat work already done by others, rather than focusing on interesting analyses, thereby wasting time and money. Furthermore, experiments undertaken on closed resources cannot be duplicated and checked, which goes against the basic tenet of scientific research. This opening up came with the rise of Wikipedia and the Creative Commons licences and is by no means limited to language resources, but affects all sciences. So, for example, in Horizon 2020 projects funded by the European Union it is mandatory for all research results as well as publications to be openly accessible. Of course, it is not only legal issues that prevent the dissemination of language resources: to be truly useful, the resources need to be stored in well-documented and standard

[6] http://nkjp.pl/

[7] https://ucnk.ff.cuni.cz/

[8] http://www.slavist.de/

[9] http://www.meta-net.eu/

[10] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

formats, and interested researchers must be able to find them. One of the first projects to approach this wider view of accessibility was the already mentioned META-NET, which also developed META-SHARE, a system of digital repositories to provide the infrastructure for describing and documenting, storing, preserving, and making language resources publicly available in an open, user-friendly and trusted way.

Of greater interest to linguists than the technologically oriented META-NET is CLARIN[11] (Common Language Resources Infrastructure), a European-wide research infrastructure centered on language, but targeted towards humanities and social sciences scholars. While CLARIN has a central node, it is essentially a distributed infrastructure, with centres now existing in 19 EU countries, including Bulgaria, Czech Republic, Poland and Slovenia. National centres have, for the most part, established trusted and long-term digital repositories, which host various language resources for the languages of the countries, often under permissive Creative Commons licences. Of course, they also enable depositing of new resources, and allow for harvesting of their metadata by other research data repositories, making their holdings widely known. Furthermore, some centres also offer openly accessible Web applications and services, such as annotation toolchains, which enable linguists to annotate their own corpora using remote applications.

The other aspect of modern corpus collection activity is that social media offer unprecedented access to studying the language of everyday interaction. For example, in 2016 Facebook reported generating more than 4 million posts every minute.[12] However, apart from technical issues connected with capturing this amount of information from various social media platforms, there is a problem of variation in spelling, morphology and syntax, because of the wide population from which data is drawn and the lack of gatekeepers (Selegey et al. 2016). While social media solve the problem of having access to sufficient data, the data they provide pose new problems for computational linguists.

# 6 Conclusions

The Slavic languages present a particular challenge to linguists of all feathers and stripes. Rich nominal morphology, free word order and aspectual pairs can introduce challenges for both computational processing and corpus analysis. Fortunately,  resources, both in the form of corpora and tools for processing them, are now becoming available for an increasing number of Slavic languages.

From its early beginnings, computational and corpus linguistics has, for all Slavic languages, significantly advanced and become more diversified, as can also be evidenced by regular international conferences organised in a number of countries with Slavic languages: the TSD (Text Speech and Dialogue) conferences in the Czech republic, the RANLP (Recent Advances in Natural Language Processing) conferences in Bulgaria, the LTC (Language Technology Conference) conferences in Poland, the HLT (Human Language Technologies) conferences in Slovenia, Slovko conferences in Slovakia etc. There is also SIGSLAV, a Special

[11]https://www.clarin.eu/

[12]http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/

Interest Group for the Slavic languages at the ACL (Association for Computational Linguistics).

The range of issues encountered when applying corpus methods to Slavic languages is now matched by the range of approaches for dealing with them. One of the important trends of the last decade concerns the harmonisation of text annotation. Even if each Slavic language has its own set of categories, the principles for representing the categories can be relatively generic with a shared set of conventions for representing each category and for choosing the specific values of categories (Erjavec and Džeroski 2004; Nivre et al. 2016; Zeman et al., 2012), as well as for converting between the conventions (Zeman, 2008). A related important trend is the availability of models shared between Slavic languages, for example, starting from a better resourced one to improve models for a lesser resourced language, for example, by adapting part-of-speech taggers (Babych and Sharoff, 2016) or parsers (Agić et al. 2014).

In short, this is a good time for Slavic corpus linguistics, as more and more resources are becoming available, be it as on-line searchable corpora or as downloadable resources, and these often come with standardised encoding and are stored in long-term repositories. In addition web-based tools are becoming available for the annotation and exploration of texts, not requiring the knowledge and high-end hardware often necessary to install and run such tools locally.

# References

Agić, Ž., Tiedemann, J., Dobrovoljc, K., Krek, S., Merkler, D., and Može, S. (2014). "Cross-lingual Dependency Parsing of Related Languages with Rich Morphosyntactic Tagsets". *Proc EMNLP 2014 Workshop on Language Technology for Closely Related Languages and Language Variants*.

Andor, J. (2004). "The master and his performance: An interview with Noam Chomsky." *Intercultural Pragmatics* 1(1): 93–111.

Arnold, E. J., T. Wasow, A. Losongco and R. Ginstrom. 2000. "Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering". *Language* 76 (1): 28-55.

Baayen, R. H., A. Endresen, A. Makarova and T. Nesset. (2013). "Making choices in Russian: Pros and cons of statistical methods for rival forms". Space and Time in Russian Temporal Expressions, a special issue of *Russian Linguistics* 37(3): 253-291.

Babych, B., and Sharoff, S. (2016). "Ukrainian part-of-speech tagger for hybrid MT: Rapid induction of morphological disambiguation resources from a closely related language." *Proc Fifth Workshop on Hybrid Approaches to Translation (HyTra)*.

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge: Cambridge University Press.

Brants, T. (2000). "TnT - a statistical part-of-speech tagger." *Proc. of 6th Applied Natural Language Processing Conference*, pages 224–231, Seattle.

Chomsky, N. (1957). *Syntactic structures*. Mouton & Co: 's Gravenhage.

Cienki, A. (1989). *Spatial Cognition and the Semantics of Prepositions in English, Polish, and Russian*. Munich: Verlag Otto Sagner.

Dąbrowska, E. (1997). *Cognitive Semantics and the Polish Dative*. Berlin: Mouton de Gruyter.

Dąbrowska, E. (2016). "Cognitive linguistics' seven deadly sins." *Cognitive Linguistics* 27(4): 479–491.

Divjak, D. (2003). "On trying in Russian: a tentative network model for near(er) synonyms." "Belgian Contributions to the 13th International Congress of Slavicists, Ljubljana, 15-21 August 2003". Special Issue of *Slavica Gandensia* 30: 25-58.

Divjak, D. (2003). "Слова о словах. К вопросу о неточных синонимах «умудриться», «ухитриться», «исхитриться» и «изловчиться." T. Soldatjenkova and E. Waegemans, eds. For East is East. Liber Amicorum Wojciech Skalmowski. Leuven-Paris: Peeters, 345-365. [Orientalia Lovaniensia Analecta].

Divjak, D. (2006). "Ways of Intending: Delineating and Structuring Near-Synonyms." St. Gries and A. Stefanowitsch, eds. *Corpora in cognitive linguistics. Corpus-based Approaches to Syntax and Lexis.* Berlin-New York: Mouton de Gruyter, 19-56. [Trends in Linguistics 172].

Divjak, D. (2009). "Mapping between domains. The aspect-modality interaction in Russian." *Russian Linguistics* 33 (3): 249-269.

Divjak, D. (2010a). *Structuring the Lexicon: a Clustered Model for Near-Synonymy.* Berlin: De Gruyter. [Cognitive Linguistics Research 43].

Divjak, D. (2010b). "Corpus-based evidence for an idiosyncratic aspect-modality interaction in Russian." Dylan Glynn and Kerstin Fischer, eds. *Quantitative Methods in Cognitive Semantics: Corpus-driven Approaches.* Berlin: De Gruyter, 305-330. [Cognitive Linguistics Research 46].

Divjak, D. (2015). "Four challenges for usage-based linguistics." J. Daems, E. Zenner, K. Heylen, D. Speelman and H. Cuyckens, eds. *Change of Paradigms: New Paradoxes. Recontextualizing Language and Linguistics*, 297–309. Berlin: Walter de Gruyter.

Divjak, D. and St. Th. Gries (2006). "Ways of Trying in Russian. Clustering Behavioral Profiles." *Journal of Corpus Linguistics and Linguistic Theory* 2 (1): 23-60.

Divjak, D., A. Kochańska and L.A. Janda. (2007). "Why cognitive linguists should care about the Slavic languages and vice versa." D. Divjak, D. and A. Kochanska, eds. *Cognitive Paths into the Slavic Domain*, 1–19. Berlin: Mouton de Gruyter.

EAGLES (1996). "Expert Advisory Group on Language Engineering Standards." http://www.ilc.pi.cnr.it/EAGLES/home.html.

Eckhoff, H. M. and L. A. Janda. (2014). "Grammatical Profiles and Aspect in Old Church Slavonic." *Transactions of the Philological Society* 112(2): 231-258.

Ellis, N. C. (2002). "Frequency effects in language processing." *Studies in second language acquisition* 24(2): 143–188.

Elworthy, D. (1995). "Tagset design and inflected languages." *Proc. From Texts to Tags: Issues in Multilingual Language Analysis SIGDAT Workshop at EACL-95)*, 1–10, Dublin.

Erjavec, T. (2004). "MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora." *Proc LREC*, Lisbon.

Erjavec, T. (2012). "MULTEXT-East: morphosyntactic resources for Central and Eastern European languages." *Language Resources and Evaluation*, 46(1):131–142.

Erjavec, T. and Džeroski, S. (2004). “Machine learning of morphosyntactic structure: Lemmatizing unknown Slovene words.” *Applied Artificial Intelligence*, 18(1):17–41.

Fidler, M. and V. Cvrček. (2015). “A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis.” *Journal of Slavic Linguistics* 23 (2): 197-240.

Hajič, J. (1998). “Building a Syntactically Annotated Corpus: The Prague Dependency Treebank.” E. Hajičová (ed.), *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, 12–19. Prague Karolinum, Charles University Press.

Hajič, J., Panevová, J., Hajičová, E., Pajas, P., Sgall, P., Štěpánek, J., Havelka, J., and Milkulová, M. (2006). Prague Dependency Treebank 2.0. Catalog Number LDC2006T01.

Hajičová, E. (1998). “Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation.” *First Workshop on Text, Speech, Dialogue*, pages 45–50, Brno, Czech Republic.

Hasher, L. and R. T. Zacks. (1984). “Automatic processing of fundamental information: the case of frequency of occurrence.” *American Psychologist* 39(12): 1372–1388.

Hayes, B. et al. (2013). “First links in the Markov chain.” *American Scientist*, 101(2): 92.

Janda, L. (1986). *A Semantic Analysis of the Russian Verbal Prefixes za-, pere-, do-, and ot-*. Munich: Otto Sagner.

Janda, L. (ed.) 2013. *Cognitive Linguistics – The Quantitative Turn: The Essential Reader.* Berlin: De Gruyter Mouton.

Janda, L. and O. Lyashevskaya (2011). “Grammatical profiles and the interaction of the lexicon with aspect, tense and mood in Russian.” *Cognitive Linguistics* 22(4): 719-763.

Janda, L. and O. Lyashevskaya (2013). “Semantic Profiles of Five Russian Prefixes: po-, s-, za-, na-, pro-". *Journal of Slavic Linguistics* 21(2): 211-258.

Janda, L., T. Nesset and R. Harald Baayen. (2010). “Capturing Correlational Structure in Russian Paradigms: a Case Study in Logistic Mixed-Effects Modeling". *Corpus Linguistics and Linguistic Theory* 6: 29-48.

Josselson, H. H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*. Detroit: Wayne University Press.

Kopotev, M., O. Lyashevskaya, A. Mustajoki. (Forthcoming) Russian challenges for quantitative research. M. Kopotev, O. Lyashevskaya and A. Mustajoki, eds. *Quantitative Approaches to the Russian Language*. London: Routledge.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago press.

Kuznetsova, J. (2015). *Linguistic Profiles: Going from Form to Meaning via Statistics*. Berlin: De Gruyter. [Cognitive Linguistics Research 53].

Langacker, R. (1987). *Foundations of Cognitive Grammar: volume I, theoretical prerequisites*. Stanford: Stanford University Press.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manning, C. D. (2011). “Part-of-speech tagging from 97% to 100%: is it time for some linguistics?” *Proc Computational Linguistics and Intelligent Text Processing*, pages 171–189.

Markov, A. (1913). “An example of statistical investigation of the text of *Eugene Onegin* concerning the connection of samples in chains.” *Bulletin of the Imperial Academy of Sciences of St. Petersburg*, 7(3):153–162. (In Russian).

McEnery, T. and Wilson, A. (1999). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mel’čuk, I. (1988). *Dependency Syntax: Theory and Practice*. New York: SUNY Press.

Micklesen, L. (1958). “Russian-English MT.” *American contributions to the Fourth International Congress of Slavicists*, Moscow.

Milin, P., Divjak, D., Dimitrijevic, S., and Baayen, R. (2016). “Towards cognitively plausible data science in language research.” *Cognitive Linguistics* 27(4): 507–526.

Müller, T., Cotterell, R., Fraser, A., and Schütze, H. (2015). “Joint lemmatization and morphological tagging with Lemming.” *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon.

Nesset, T., L. A. Janda and H. M. Eckhoff. (2014a). “Old Church Slavonic *byti* part one: grammatical profiling analysis.” *Slavic and East European Journal* 58(3): 482-497.

Nesset, T., L. A. Janda and H. M. Eckhoff. (2014b). “Old Church Slavonic *byti* part one: constructional profiling analysis.” *Slavic and East European Journal* 58(3): 498-525.

Nikolaeva, T. (1958). “Soviet developments in machine translation: Russian sentence analysis.” *Mechanical Translation*, 5(2): 51–59.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). “Universal Dependencies v1: A multilingual treebank collection.” *Proc LREC*, Portoroz.

Paskaleva, E., Simov, K., Damova, M., and Slavcheva, M. (1993). “The Long Journey from the Core to the Real Size of Large LDBs.” *Proceedings of the ACL Workshop on Acquisition of Lexical Knowledge from Text*, pages 161–169, Columbus, Ohio.

Przepiórkowski, A., R. L. Górski, B. Lewandowska-Tomaszczyk and M. Łaziński. (2008). “Towards the National Corpus of Polish.” *Proc LREC*, Marrakech.

Renouf, A. (2005). “Corpus linguistics: past and present.” N. Wei, W. Li and J. Pu (eds.). *Corpora in Use: In honour of Professor Yang Huizhong.* [s.l.]

Rudzka-Ostyn, B. (1992). “Case relations in cognitive grammar: Some reflexive uses of the Polish dative.” *Leuvense Bijdragen* 81:327–373.

Schmid, H. (1994). “Probabilistic part-of-speech tagging using decision trees.” *Proc International Conference on New Methods in Language Processing*, Manchester.

Selegey, D., T. Shavrina, V. Selegey and S. Sharoff. (2016). “Automatic morphological tagging of Russian social media corpora: training and testing.” *Proc. Dialogue, Russian International Conference on Computational Linguistics*.

Sgall, P. (1995). “Formal and computational linguistics in Prague.” E. Hajičova, M. Cervenka, O. Leska and P. Sgall (eds.). *Prague Linguistic Circle Papers, Volume 1*, 23–35. John Benjamins.

Sharoff, S. (2005). “Methods and tools for development of the Russian Reference Corpus.” D. Archer, A. Wilson and P. Rayson (eds), *Corpus Linguistics Around the World*, 167–180. Amsterdam: Rodopi.

Sharoff, S. and J. Nivre. (2011). “The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge.” *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo.

Shteinfeld, E. (1963). *Chastotnyj slovarj sovremennogo russkogo literaturnogo jazyka (Frequency dictionary of modern Russian literary language)*. Tallin.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Collection informatique linguistique. Masson.

Sokolova, S., O. Lyashevskaya and L. A. Janda. (2012). “The Locative Alternation and the Russian 'empty' prefixes: A case study of the verb gruzit' 'load'". D. Divjak and St. Th Gries, eds. *Frequency effects in language representation*, 51-86. Berlin: Mouton de Gruyter. [Trends in Linguistics. Studies and Monographs. 244.2]

Solovyev, V. and L.A. Janda. 2009. "What Constructional Profiles Reveal About Synonymy: A Case Study of Russian Words for SADNESS and HAPPINESS". *Cognitive Linguistics* 20(2): 367-393.

Stefanowitsch, A. (2005). “New York, Dayton (Ohio), and the raw frequency fallacy.” *Corpus linguistics and linguistic theory* 1(2): 295–301.

Sussex, R. and P. Cubberley. (2006). *The Slavic languages (Cambridge Language Surveys*. Cambridge University Press.

Tognini-Bonelli, T. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tummers, J., K. Heylen and D. Geeraerts. (2005). “Usage-based approaches in Cognitive Linguistics: A technical state of the art.” *Corpus Linguistics and Linguistic Theory* 1 (2): 225–261.

van Halteren, H. (1999). *Syntactic Wordclass Tagging*. Text, Speech and Language Technology. Springer.

Vetulani, Z., Walczak, B., Obrebski, T., and Vetulani, G. (1998). Unambiguous coding of the inflection of Polish nouns and its application in electronic dictionaries — format POLEX. *Poznan: Wydawnictwo Naukowe UAM.*

Vitas, D., Pavlović-Lažetić, G., Krstev, C., Popović, L., and Obradović, I. (2003). “Processing serbian written texts: An overview of resources and basic tools.” *Workshop on Balkan Language Resources and Tools*, 97–104.

Wierzbicka, A. (1972). *Semantic Primitives*. Frankfurt: Athenäum.

Zeman, D. (2008). “Reusable tagset conversion using tagset drivers.” *Proc LREC*, Marrakech.

Zeman, D., Marecek, D., Popel, M., Ramasamy, L., Stepánek, J., Zabokrtsǩy, Z., and Hajic, J. (2012). “Hamledt: To parse or not to parse?” *Proc LREC*, Istanbul.