Slavic Corpus Linguistics: The Historical Dimension

Tromsø, Norway, April 21–22, 2015

UiT The Arctic University of Norway

ABSTRACTS OF THE TALKS

INVITED TALKS

David Birnbaum

A synthetic theory of digital edition

One of the most exciting promises of digital text processing is *multipurposing*, the idea that the same digital object can be used to ask and answer different questions in different ways. The selection and organization of information in a paper publication is immutable, which means that it is designed and implemented to support the purposes intended by the editor, and those purposes are constrained both intellectually (by the editor's interests and imagination) and physically (by the need to maintain equilibrium between richness of information and legibility of presentation). Digital editions (the examples in this presentation involve medieval Slavic manuscripts, but the theory and method are generally applicable), in which the presentation on the screen can be understood (and implemented) as one of many possible views of the same underlying composite digital object, have the potential to mitigate both types of limitations. The shape of the presentation becomes the result of a negotiation between the editor and the reader, and is no longer solely the purview of the editor. And the dynamic properties of digital publication make it possible to incorporate information into an edition without the traditional risk that the edition would become cluttered or confusing, as would happen on paper. Building on earlier collaborative work with Lara Sels about reconciling diplomatic editions of manuscripts with critical editions of texts for digital publication, the present report explores how to enrich a digital edition still further through the incorporation of linguistic information about the texts and manuscripts.

Leonid Iomdin

SynTagRus: A Brief History and Current State of the Russian Deeply Annotated Corpus

SynTagRus is the first Russian corpus annotated with full morphological and syntactic structures and a number of other linguistically important features. The corpus is constructed in two stages, the automatic parsing and manual correction by experts. The talk will be focused on the linguistic ideas and background underlying the creation of the corpus, including the syntactic dependency component of the Meaning ⇔ Text theory, the theory of lexical functions, the theory of ellipsis and other types of syntactic gaps, and lexicographic approaches to polysemy. The principles and techniques of corpus construction will be discussed. Special emphasis will be given to the theoretical and computational impact of the corpus, which is used in the construction of statistical parsers, as a source of statistical data for parser modification, and as a resource for parsing regression tests.

Roland Meyer

A variationist analysis of relativizers and subordinators in Middle East Slavic

Historical links between relative and complement clauses are typologically well-established (cf. van Gelderen 2004, 81ff; 89ff; Harris und Campbell 1995; Hopper und Traugott 2003; Roberts und Roussou 2003; Axel 2009 for Germanic and Romance). According to the majority view in the literature, demonstratives developed into uninflected relative markers on the one hand, and into subordinate complementizers on the other; Axel (2009), however, has advanced an alternative view, arguing that subordinate clauses and their complementizers were actually derived from relative clauses with uninflected relativizers. As argued in this paper, Slavic generally provides solid evidence for the latter development, i.e., from inflected relative pronouns via uninflected relativizers to complement clause subordinators. Nevertheless, we still find remarkable variation in the syntax of relativization and subordination in 16th/17th c. Middle East Slavic, on the one hand, whereas e.g. in historical Polish, the system has remained more or less stable since the 15th c. The East Slavic variation is clearly related to the overlay of Church Slavonic and different vernacular systems, and most probably, also to influence from Polish on the chancellary language. A useful analysis should therefore disentangle, or at least do justice to, this mixture of variants to the greatest possible extent. The present paper establishes and discusses ways of solving this task in a variationist analysis, systematically relating the distribution of relative and subordinate clause markers to other features of register. Since the overlay of several systems is, of course, the rule rather than the exception in the history of East Slavic languages, this has implications also for other areas of diachrony.

References

Axel, Katrin (2009): Die Entstehung des dass-Satzes – ein neues Szenario. In: V. Ehrich, C. Fortmann, I. Reich & M. Reis (eds.): *Koordination und Subordination im Deutschen.* Buske, Hamburg.

van Gelderen, Elly (2004): *Grammaticalization as Economy*. John Benjamins, Amsterdam.

Harris, Alice C. und Lyle Campbell (1995): *Syntactic Change in Cross-Linguistic Perspective.* Cambridge University Press.

Hopper, P. und E. Traugott (2003): *Grammaticalization*. Cambridge University Press.

Roberts, Ian und Anna Roussou (2003): *Syntactic Change. A Minimalist Approach to Grammaticalization*. Cambridge University Press.

Pavel Petrukhin

The periphrastic form '*byti* + active present participle' in the history of Russian language

The paper is dedicated to the participial construction made up of the auxiliary verb *byti* in form of the imperfect or aorist and the active present participle (cf. *Kain" že bě dělaja zemlju* 'Cain was a tiller of the ground'). The auxiliary may take various tense forms, but the most frequent ones are the aorist and the imperfect.

The construction is testified both in Old Church Slavonic and in Old Russian. There is no scholarly consensus as to the origins of this construction. According to some scholars it is a calque of the corresponding Greek form, according to others it pertains to the original Slavic verbal system. In the Early East Slavic literature the form in question is mostly used in two types of texts: 1) biblical translations; 2) original narrative texts, such as chronicles and *žitija*.

The paper has two main goals:

1. Analysis of the semantics of the construction in the earliest East Slavic narrative texts. Traditionally the participial form is said to have the "progressive" meaning, such as that of the English *to be doing* form. However, it can also express the habitual and stative semantics; moreover, in original East Slavic texts such as the Russian Primary Chronicle the habitual and stative forms prevail over the progressive ones. The investigation will take advantage of the recent progress in analysis of the corresponding Greek participial form.

2. Analysis of the diachronic development of the form in the history of Russian written language from the earliest texts until the 17th century. In course of the time the construction '*byti* + active present participle' underwent various morphologic, syntactic and semantic changes. Tracing this development helps understand mechanisms of adoption and reanalysis of a "bookish" morphosyntactic construction in a written language tradition. Linguistic corpora may be very helpful in this work.

CONTRIBUTED TALKS

Using historical data to define regularity

Aleksandrs Berdičevskis[1], Alexander Piperski[2,3]
[1] UiT The Arctic University of Norway
[2] Russian Academy of National Economy and Public Administration
[3] Russian State University for the Humanities

One of important linguistic notions that are intuitively understandable, but notoriously difficult to define and quantify is morphological regularity. Some theories, however, do provide a criterion for establishing regularity. Under the dual-mechanism model of morphology (Pinker 1999), regular forms are the ones that are constructed using rules and not normally stored in memory. This view allows to draw a line between regularity and irregularity using psycholinguistic experiments (Markus et al. 1995), but, obviously, only for synchronic descriptions of modern languages. Another wide-spread approach is to equate irregularity with low type frequency, cf. the list of examples in Stolz et al. (2012: 15–19), but this solution depends on determining the boundaries of each type and is thus highly subjective and inconclusive. The approach we attempt to pursue in this paper is based on the assumption that purely morphological change is virtually unidirectional and leads to regularization, even though separate examples of the opposite exist (Maiden 1992). Irregularity is introduced by changes of other kinds, primarily the phonological ones. Hence, in the absence of phonological change we may assume that an expanding morphological pattern is likely to be regular, while a declining morphological pattern is likely to be irregular.

We test whether this theory fits with intuition on the domain of Russian verbal morphology. The relationship between verbal stems in Russian has remained almost untouched by phonological change since Old Russian (basically, since the phonemicization of palatalization). For this reason, any changes occurring in this domain (the most salient one being the so-called suffix shift, see Nesset & Kuznetsova 2011) can be treated as regularization and thus can be expected to show what was regular and what was not. Using data obtained from Old Russian and Modern Russian subcorpora of the Russian National Corpus (Berdicevskis & Piperski 2014), we build a statistical model to estimate the degree of regularity of each verb throughout the history of Russian.

References:

Berdicevskis, Aleksandrs & Alexander Piperski. 2014. What do we regularize and what is regular: Russian verbs through the centuries. Paper presented at SCLC-2014.

Maiden, Martin. 1992. Irregularity as a determinant of morphological change. *Journal of Linguistics* 28.2: 285-312

Markus, Gary F., Ursula Brinkmann, Harald Clahsen, Richard Wiese & Steven Pinker. 1995. "German inflection: the exception that proves the rule". *Cognitive Psychology* 29.3: 189–256.

Nesset, Tore & Julia Kuznetsova. 2011. Stability and complexity: Russian suffix shift over time. *Scando-Slavica* 57(2): 268–289.

Pinker, Steven. 1999. *Words and rules: the ingredients of language*. New York: Basic Books.

Stolz, Thomas, Hitomi Otsuka, Aina Urdze & Johan van der Auwera (eds.). 2012. *Irregularity in morphology (and beyond)*. Berlin: Akademie Verlag.

Conjugate, decline and spell like years ago. A corpus-based morphological analyzer of 19th century Polish

Joanna Bilińska, Witold Kieraś, Magdalena Derwojedowa
University of Warsaw

In the talk we will present a project on automatic inflectional analysis of some historical texts. The project consists of two intertwined parts: a tool for inflectional analysis of Polish texts in their historical spelling, declension and conjugationand 1M tokens corpus of the 2nd half of the 19th century Polish texts. Although technically not particularly demanding and linguistically rather limited, this program is indispensable step towards a bundle of tools for historical corpora of Polish. It is also intended to be a first stage of building an analyzer capable to encompass all stages of inflection development. The corpus itself is also designed to be a valuable stand-alone resource for further research.

The analyzer is based on the tool for contemporary Polish (cf. Woliński 2014), which consists of ≈250,000 entries, some of them as old as last quarter of 18th century (cf. Saloni et al. 2012). In the project we expand analyzer's lexicon with 19th century vocabulary and modify the inflectional patterns to make them fit for obsolete and archaic forms attested in the corpus. These forms will be time-stamped. The entire process consists of extracting data from the corpus, classification of faults, making modifications to the analyzer, evaluation of the modified tool. A final output should keep the current efficiency rate around 97% for both contemporary and historical texts.

The corpus is the main resource of linguistic facts and it is used to test effectiveness and efficiency of the program. To assure stylistic and temporal variety, it is a set of 1,000 words long samples, grouped into five genre sub-corpora of 200,000 tokens each. All samples are excerpted from digital resources, in most cases from on-line libraries. In particular we would like to discuss difficulties and obstacles we encountered when using these collections. we will give more details on corpus' structure, sample's meta-data and some numerical data (variety in length, number of analyzed tokens, most popular authors, titles or sources etc.) for the current stage of development of the corpus (≈75% of the total size).

In conclusion we would like to talk a little about future applications of the analyzer and/or the corpus, e.g. part-of-speech tagging, an overall description of Polish diachronic inflection, tracing dynamics of spelling and inflectional changes, following adaptation stages of loanwords, as well as marking neologisms or disued words with date.

References

Saloni Z., Woliński M., Wołosz R., Gruszczyński W. i Skowrońska D., *Słownik gramatyczny języka polskiego* [*Grammatical dictionary of Polish*], ed. II, Warszawa 2012, CD.

Woliński M., *Morfeusz Reloaded*, [w:] *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, red. N. Calzolari, K. K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk i S. Piperidis, s. 1106--1111, ELRA, Reykjavík, Iceland 2014, ISBN 978-2-9517408-8-4, URL http://www.lrec-conf.org/proceedings/lrec2014/index.html.

The use of stylometry in historical linguistics: a case study in recent diachronic change in Polish

Maciej Eder and Rafał L. Górski
The Institute of the Polish Language, Polish Academy of Sciences

In the last decades, quantitative linguistics (following exact and social sciences) has developed a great number of statistic methods providing an insight into measurable phenomena of natural language. Although to a lesser extent, it also applies to the analysis of diachronic changes.
A significant drawback of many of the methods applied so far is a tacit assumption that the researcher knows in advance which elements of the language are subject to change. In other words: the method of, say, plotting and inspecting the trend for a given phenomenon may be applied only to verify hypotheses stipulated earlier by traditional diachronic linguistics. A real challenge, however, is to develop such a method that would allow to trace chronological change in the language without a prior knowledge which linguistic features are responsible for the change. Promising results may be expected using some of the time-proven stylometric techniques based on the statistic analysis of style, especially the so-called multidimensional methods. These methods include the Principal Components Analysis (PCA), Multidimensional Scaling (MDS), Cluster Analysis (CA), Burrows's Delta, and many others.
Certainly, the diachronic process can be traced via lexical changes. However, an interesting question arises what if we disregard words and examine grammatical features instead? Obviously, the usage of archaic vs. modern inflected forms will differentiate *per se* texts written in two distinct (still close) periods. But what is less obvious, is whether processing only POS-tags can show the dynamics of language change. Note that the sequences of POS-tags are a good approximation of syntax, even if they cannot replace parsing.
To scrutinise the above research question, we performed a number of stylometric tests using different combinations of lexical and grammatical features' n-grams. If, say, MDS is applied to bigrams of POS-tags drawn from Polish novels from the 19th and the 20th century, we would like to observe a clear chronological pattern. In fact the plot seems to scramble the texts. However the plane is divided into two parts - one of which is occupied by texts written after 1918, the other one by older texts, with some few outliers.

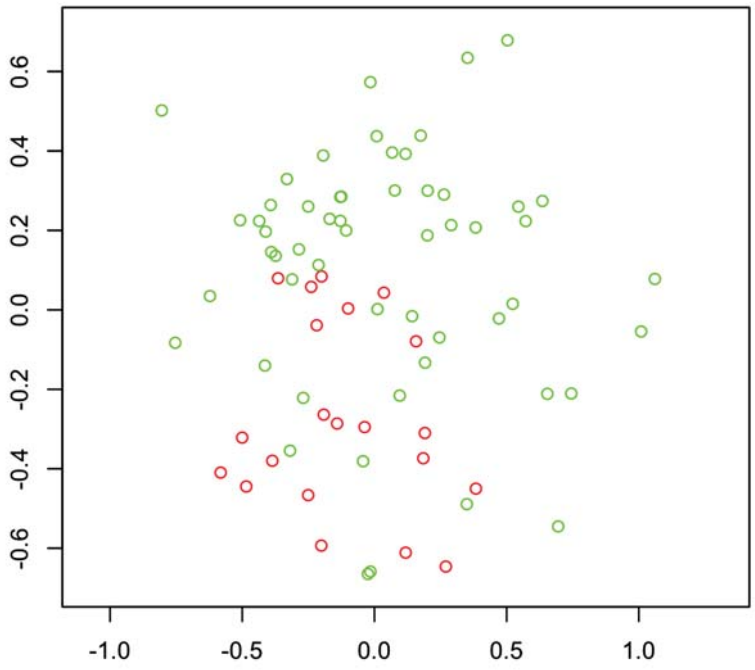
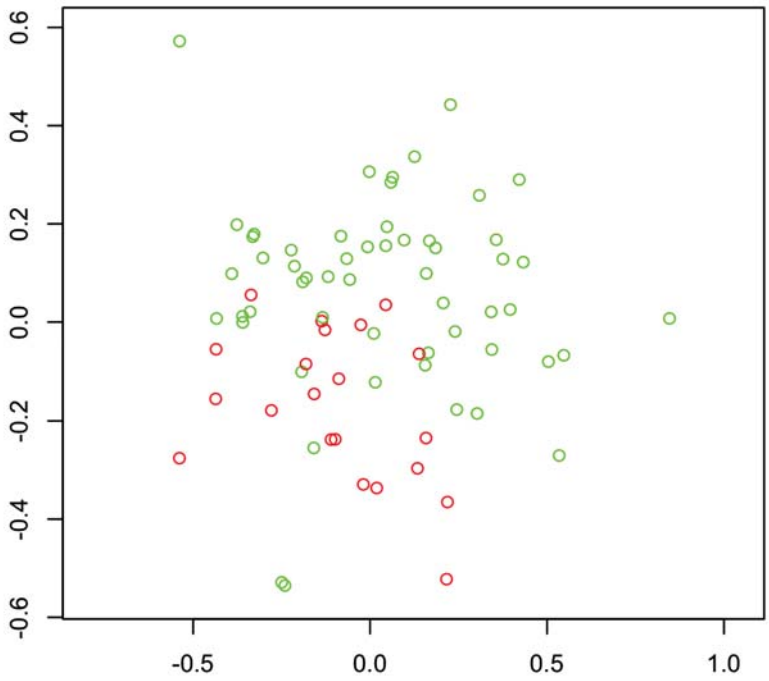


Fig. 1. Stylometry and chronology, i.e. Multidimensional Scaling of 76 Polish novels from the 19th and the 20th century (examined features: relative frequencies of 100 most common bi-grams of POS-tags): the figure shows a strong, yet not perfect, division into the novels created before the First World War (red) and afterwards (green).



1000 frequent tag bi-grams (1st and 2nd part of the tags)

This supports a well established claim of Polish historical linguistics that the year 1918 divides two subperiods of the Modern Polish (Klemensiewicz 1972, Walczak 1995). At the same time it does not support another claim, namely that 1939 (or 1945) is another borderline. Now, one can ask what discriminates these texts. There is no simple and obvious answer, it is rather a bunch of features. Each of them can easily be overlooked in close reading, however in a mass they make a text sound somewhat strange to a modern reader.

Klemensiewicz, Z (1972): Historia języka polskiego, Warszawa: PWN.
Walczak, B. (1995): Zarys dziejów języka polskiego, Poznań: Kantor Wydawniczy SAWW.

The electronic corpus of the 17th and 18th century Polish texts (up to 1772) – aims, methods, current state, problems and prospects for development

Włodzimierz Gruszczyński*

* Instytut Języka Polskiego PAN
Al. Mickiewicza 31, 31-120 Kraków

Maciej Ogrodniczuk†

† Instytut Podstaw Informatyki PAN
Ul. Jana Kazimierza 5, 01-248 Warszawa

Abstract

Since 2013, two institutes of the Polish Academy of Sciences (Institute of the Polish Language and Institute of Computer Science) are co-operating on *The electronic corpus of the 17th and 18th century Polish texts* project (code name: KORBA = KORpus BArokowy 'Baroque Corpus'), planned to finish in 2018. The project aims at creation of a fairly balanced corpus of Polish texts dating between 1601 and 1772, with size planned to around 12 million tokens. KORBA will make a historical subcorpus of the National Corpus of Polish (Pol. Narodowy Korpus Języka Polskiego = NKJP, http://nkjp.pl/).

The corpus design features multi-layer description of:

- structural annotation – with rich bibliographic, stylistic, genologic and structural metadata, intended to enable refined search and provide page-aware location of each segment in original text
- linguistic annotation – of all foreign elements, with respect to language identification (which is important due to specificity of the period, when Latin, and later French interjections and quotations from various languages, such as Turkish, interlarded Polish passages)
- morphosyntactic annotation – of a 0.5 million token-size subcorpus, planned to be manually annotated and then used to train automated tagger to be applied to the remaining part of the corpus (the outline of the process will be presented during our talk).

The project benefits from the experience gained in the process of building NKJP, but historical material calls for additional solutions. Most of the texts in our corpus are transliterated (based on old prints or manuscripts), yet Polish orthographies of 17th and 18th centuries were so inconsistent that their automated lemmatization is practically impossible. To solve this problem, texts are automatically transcribed using a dedicated application and each text features two parallel forms: transliterated and transcribed. The majority of further processing will be carried out on the latter form, while search results will be retrieved in transliteration. Corpus management will use modified version of NKJP tools: Morfeusz morphological analyser (http://sgjp.pl/morfeusz/index.html.en), Polita tagger (http://zil.ipipan.waw.pl/PoliTa), new version of Poliqarp search engine etc.

The work is further complicated by heterogeneity of sources. While most of the texts are transcribed, some need to be acquired from later (19th, 20th or contemporary) editions, mainly because originals do not longer exist or are extremely difficult to obtain. Most of them use modernized spelling, with transcription orthography effective in each given publication period. Such texts will be represented in the corpus with only one form, since reconstruction of their transliterated version is impossible.

At the moment the project has collected over 200 texts of the time amounting to 6 million tokens. Most of them have been manually transcribed from originals within the current task or taken over from the recently finished IMPACT project. The process of transcription (and, at the same time, structural and linguistic annotation) is text editor-based and follows detailed guidelines. Results are automatically converted to TEI XML and uploaded to an integrated corpus workflow management environment which also features simple search, making the resource immediately available for ongoing work on the *Electronic dictionary of 17th and 18th century Polish* (http://sxvii.pl/).

The data of the resulting corpus is already used in the parallel task of compiling the *Electronic dictionary of 17th and 18th century Polish* and will be later used to create the diachronic model of Polish inflection. The corpus will be freely available and is planned to be further developed after the end of the project (possibly within DARIAH research infrastructure).

Łukasz Jędrzejowski

On the origin and the development of infinitival *wh*-complements in the history of Polish

In this talk, I will examine dependent infinitive complements introduced by a *wh*-phrase in the history of Polish and investigate their emergence circumstances, individual development steps and role in the Polish complementation system in general. The main focus will be on two patterns: (i) embedded infinitival questions (= EIQs) and modal existential *wh*-constructions (= MECs). Diachronically, I will show that although both patterns emerged under similar circumstances, they developed into two different directions.

Modern Polish EIQs and MECs share two main properties: (i) they are introduced by a *wh*-phrase and (ii) they consist of an infinitive. [1] and [2] illustrate both patterns:

[1] *Człowiek* *nie* *wiedział* [*gdzie* *uciekać*] [EIQ]
human.being NEG know.3SG.M.*l*-PTCP where run.away.INF
'One didn't know where to run away to.'
(NKJP, *Express Ilustrowany*, 28/7/2001)

[2] *Nie* *mam* [*gdzie* *zaparkować*] [MEC]
NEG have.1SG where park.INF
'There is no place where I could park (my car).'
(NKJP, *Dziennik Zachodni*, 26/6/2001)

At first sight, [1] and [2] seem not to differ on the surface. Their matrix predicates (*wiedzieć* 'know' in [1] and *mieć* 'have' in [2]) are under the scope of the negation operator *nie* and the *wh*-clauses are headed by the *wh*-operator *gdzie* 'where' followed by infinitives (*uciekać* 'run away' in [1] and *zaparkować* 'park' in [2]). However, there are also many differences between EIQs and MECs. Firstly, Polish MECs can only be introduced by two existential predicates: *mieć* ('have') and *być* ('be'). EIQs, in turn, are embeddable under various matrix predicate classes: verbs of retaining knowledge (e.g. *wiedzieć* 'know'), decision verbs (e.g. *decydować* 'decide'), verbs of one-way communication (e.g. *wyjaśniać* 'explain'), see Bhatt (2006) for an overview. Secondly, EIQs - but very seldom MECs - can have both existential and universal force. Thirdly, according to Šimík (2011) Polish MECs cannot be headed by *kiedy* ('when'), *jak* ('how') and *dlaczego* ('why'), admitting only the *wh*-operators belonging to the first group on the *wh*-hierarchy given in [3]:

[3] {*what*, *who*, *where*} > {*when*, *how*} > *why*

Based on the data extracted from the National Corpus of Polish, I will illustrate that this assumption does not hold for Polish, though. I will argue that Polish MECs can be headed by all *wh*-phrases except for *dlaczego*. This makes them different from EIQs occurring with all *wh*-phrases given in [3] (cf. Jędrzejowski 2014). The major claim is that all these differences between EIQs and MECs come from two different development paths, which they went through in the history of Polish.

I will illustrate that both EIQs and MECs emerged in the Middle Polish period and that one structural prerequisite has to be met for both patterns to arise. The matrix verb has to be under the scope of a negation operator (cf. [1] and [2]). But on the other hand, I do not claim that it was not possible in Old Polish to express similar kinds of attitude towards what is embedded. In other words, I argue that EIQs as well as MECs occurred already in Old Polish. The only difference is that Old Polish EIQs/MECs did not embed infini-

tival forms. Instead, they used to select for subjunctive complements. [4] illustrates such a use for MECs:

[4] *Toć ubogi Krolewiec był, iże nie imiał*
but poor King be.M.SG.*l*-PTCP that NEG have.M.SG.*l*-PTCP.AOR
[*gdzie by swoję głowę podkłonił*]
where COND.CL his head.ACC put.M.SG.*l*-PTCP

'However, the King was so poor that he did not get any place where he could have passed the night.' (Ksw IV, 6: 26-7)

As for the *wh*-hierarchy depicted in [3], I shall demonstrate that EIQs were more progressive than MECs: The former - but not the latter - started to allow in the 19th century the *wh*-phrase *dlaczego*.

As it turns out, the diachrony of EIQs and MECs provide more empirical evidence underpinning the view that both patterns cannot be brought down to a common denominator, i.e. to the *wh*-movement and the presence of infinitives.

On the syntax of possession in Old Church Slavonic (on the basis of historical corpora)

Iliana Krapova (University of Venice Ca' Foscari),
Tsvetana Dimitrova (Institute for Bulgarian Language)

The task of our work is to outline, observing the available data from three corpora, the development in Bulgarian of constructions involving dative pronominal clitics which may express possession when surfacing either in the nominal or in the clausal domain and have ambiguous interpretation between possessive and affected dative. The construction, also known in the literature as *doubly bound Dative* (Minčeva (1964: 29-30) following Melьničuk (1958: 283), can be traced back to Old Church Slavonic (OCS) and had its parallels in NTGreek where a similar construction involving Extraposed Genitive instead of Dative, was used to show a possessive alongside an affected interpretation (Gianollo 2010 for a review). Since the Greek construction has been analyzed as crucial for the Genitive-Dative syncretism in Greek, we will compare it with the doubly bound Dative in OCS and in later texts of Middle Bulgarian up to damaskins to reveal its role for the Genitive-Dative syncretism in Bulgarian where the Dative prevailed over the Genitive.

Our study is corpus-based covering texts from three corpora containing early OCS and later (Middle) Bulgarian texts. The *PROIEL* (Pragmatic Resources in Old Indo-European Languages) corpus (Haug, Eckhoff 2011) is used for quantitative observations as the data is annotated, accessible and available for replica. It contains the gospel text acc. to Codex Marianus (following the edition of Jagić, 1883), verses of the gospel text acc. to Codex Zographensis (Jagić, 1879) that are missing in Codex Marianus and texts from Codex Suprasliensis. Texts are annotated (lemma, POS, morphological and syntactic annotation). Variant readings that we need are consulted on the *TITUS* database which gives access to parallel and sometimes aligned data (but not annotated). Gospel text acc. to Cod. Mar., Cod. Zogr., Cod. Assemani, and Codex Sabbae is organized with aligned corresponding passages across OCS texts and the Greek NT. We discuss different variants according to the texts in different manuscripts (e.g., genitive or dative pronouns, pronominal clitics post- and pre-NP, etc.). For later texts, we turn to another available databas – the Historical Corpus of Bulgarian Language (HCBL) for further sketching the process. The HCBL is a collection of texts which are not annotated but concordances can be build using an external service. We will also discuss the benefits and limitations of each of these three electronic resources on our research task.

References:

Gianollo, C. 2010. *External Possession in New Testament Greek*, in: *Papers on Grammar IX*, pp. 102-129.

Haug, D., H. Eckhoff. 2011. *The PROIEL Corpus as a Source to Old Church Slavic: a Practical Introduction,* in: *Pisьmennoe Nasledie i Sovremennye Informacionnye Tehnologii,* Izhevsk 2011, pp. 37–55.

Jagić, V. 1879. *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus.* Berlin.

Jagić, V. 1883. *Quattuor Evangeliorum versionis palaeoslovenicae Codex Marianus Glagoliticus.* Saint Petersburg.

Melьničuk, O.S. 1958. *Istoriya vživaniya davalьnogo, bezpriymennikovogo vidminka v ukrainsьkiy movi,* in: *Doslidženiya z sintaksisu ukrainsьkoi movi*, Kiiv.

Minčeva, A. 1964. *Razvoj na datelnija pritezatelen padež v blgarskija ezik,* Sofia.

Julia Kuznetsova and Alexander Rubin

Russia in the Twentieth Century: the Corpus Linguistics Perspective

Authors portray situations they consider mostly important, so texts of the past reflect significant historical events. Historical corpora make it possible to extract words denoting such events. Using keyword analysis (Scott & Tribble 2006, Baker & Ellece 2011), we investigate Russian Google Books corpus and for each year in the twentieth century determine a word that is most strongly associated with that year. As a result we have received the list of the most important events of the twentieth century highlighted by the corpus.

The quantity of the books in the Google Books archive differs for different years. In order to compensate for that we use a standardized year-word frequency – word frequency in that year divided by the amount of the books attested for that year. For each year we identify a list of words that have their peak of standardized year-word frequency in that year. These are words most characteristic for the year. In order to exclude the influence of very infrequent words we only use words that are attested in at least fifty books that year. In order to compensate for word's overall frequency we use an attraction ratio (similarly to Gries and Stefanowitsch 2003) – we divide the standardized year-word frequency by the overall frequency of the word. The word that has the highest attraction ratio is considered to be the most characteristic for that year.

We have received the list of 100 words that tell us the story of the twentieth century Russia. These words emphasize the events that were mostly described during the year. The word of the year 1912 is *Napoleon"* 'Napoleon' featured in the articles referring to the anniversary of the French invasion. The word of the year 1942 is *gitlerovskoj* 'Hitler's' reflecting the invasion of Russia during the WWII. The word of the year 1980 is *Killanin* due to the references to the Baron Killanin – the President of the International Olympic Committee during the Olympics that were held in Moscow in 1980.

Our approach connects research in history and corpus linguistics, showing that corpus linguistic methods can help us to automatically cull the words pointing out the most significant historical events. This approach is within the newly developing area of digital humanities that is concerned with the intersection of computing and traditional humanity disciplines.

References

Baker, Paul and Sibonile Ellece. 2011. *Key terms in discourse analysis*. New York/London: Continuum International.

Scott, Mike and Christopher Tribble. 2006. *Textual patterns: Keyword and corpus analysis in language education*. Amsterdam/Philadelphia: John Benjamins.

Stefanowitsch, Anatol and Gries Stefan .2003. Collostructions: investigating the interaction between words and constructions. *International journal of corpus linguistics*. 2003. 8(2).

Parallel Historical Corpora – a New Method in Standardization Research?

Marija Lazar
Sächsische Akademie der Wissenschaften zu Leipzig

The current contribution deals with the motivation for the parallel historical corpus of legal texts and its design. The advantages of its application in linguistic research on two types of standardization processes will be discussed; in the language as a system of centre and periphery and in the legal language as a professional domain.

Žilinská pravná kniha (*Žilina Law Book*, 1378-1561) serves as a basis for the parallel historical corpus and includes, among others, the copy of *Sachsenspiegel (Mirror of the Saxons)* in Middle High German as well as its Slovak translation (book editions Piirainen 1972; Kuchar 2009). Thus the corpus is put into the larger context of *Ius Maideburgense* which is a legal source with profound effect in the Middle and Eastern European legal systems (Lück 1996: 37, 42-46). Nevertheless, the translations of *Ius Maideburgense* has been little involved in the systematic linguistic research yet and this gap has to be closed in the ongoing project. The methodology developed for the architecture of historical and parallel corpora is unified in the corpus design, its detailed description will be provided.

The corpus application will be demonstrated on the study of automatically extracted bi- and tri-grams from a test corpus sample. The structural and lexical evidence from Middle High German and Slovak legal texts will be compared and contrasted with Bily's (2015) outcome from the terminology and formula analysis of Czech *Práwa saszká* (*Saxon Laws*, 1473). The findings will be integrated into the discussion about the Slovak standard language development and will shed light on the standardization of the legal language in the Lands of the Bohemian Crown and Slovakia. Finally, the methodology for comparable settings (central versus peripheral language usage in the community) will be discussed.

Bibliography

Bily, Inge; Carls, Wieland; Gönczi, Katalin 2015 (forthcoming). Sächsisch-magdeburgisches Recht in Tschechien und in der Slowakei. Untersuchungen zur Geschichte des Rechts und seiner Sprache. (=IVS SAXONICO-MAIDEBVRGENSE IN ORIENTE. Bd. 5).

Kuchar, Rudolf 2009. Žilinská právna kniha. Preklad Magdeburského práva. Zápisy právnych úkonov Žilinských mešťanov. Bratislava: VEDA – vydavatel'stvo Slovenskej akademie vied.

Lück, Heiner 1996. Die Verbreitung des Sachsenspiegels und des Magdeburger Rechts in Osteuropa. In: Der sassen speyghel. Sachsenspiegel - Recht - Alltag. Bd. 2: Aus dem Leben gegriffen - ein Rechtsbuch spiegelt seine Zeit. Beiträge u. Katalog zur Ausstellung 'Aus dem Leben gegriffen - ein Rechtsbuch spiegelt seine Zeit'. Hrsg. von Mamoun Fansa. 2., verb. Aufl. (=Archäologische Mitteilungen aus Nordwestdeutschland. Beih. 10). Oldenburg, S. 37-49.

Piirainen, Ilpo Tapani (Hrsg.) 1972. Das Stadtrechtsbuch von Sillein. Einleitung, Edition und Glossar. Berlin, New York: de Gruyter. (=Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker. N. F. 46/170).

Práwa saszká: http://psp.cz/kps/knih/prawa/prawa.htm (accessed on 22.01.2015).

The loss of referential null subjects in Russian: what subordinate clauses can tell us

Silvia Luraghi and Erica Pinelli, University of Pavia

As well known, Old Russian was a null subject language. The loss of referential null subjects can be observed in texts during the history of Russian. Careful analysis taking into account stylistic factors and differences between the frequency of 1st/2nd person pronouns as opposed to 3rd person pronouns shows that the frequency of the latter present a slow but steady increase, not only in the past tense, but also in the present (Meyer 2011). Moreover, a recent corpus study has shown that the rate at which referential null subjects decrease is much faster in subordinate than in main clauses (Claudi 2014). In particular, in texts from the 12th to the 17th century the percentage of 3rd person null subjects in subordinate clauses remains stable and close to 100% until the beginning of the 16th century, and drops suddenly to 63% at the end of the century and 25 % in the course of the 17th century, while in the same corpus 3rd person null subjects in main clauses remain above 70% until the end of the period considered. Notably, a higher frequency of overt subjects in subordinate clauses has also been detected in Early Germanic languages in the process of losing null referential subjects (see Håkansson 2013 on Old Swedish, Walkden 2013 on Old English). This finding challenges various assumptions, among which the Constant Rate Effect, which holds that that "the rate of change in different surface contexts reflecting a single underlying parameter change is the same" (Kroch 2001), as well as the idea that subordinate clauses are more conservative than main clauses and tend to preserve older patters (Bybee 2001 among others). In our paper, we will try to understand the reasons for the observed development, also in the light of ongoing changes in the system of subordination that were taking place in Russian at the same time (Borkovskij 1979).

References

Borkovskij, V. I. 1979. *Istoričeskaja grammatika russkogo jazyka: sintaksis - složnoe predloženie*. Moskva: Nauka.

Bybee, Joan. 2001. Main clauses are innovative, subordinate clauses are conservative: consequences for the nature of constructions. In J. Bybee and M. Noonan (eds.) *Complex sentences in grammar and discourse*. Amsterdam: John Benjamins,1-17.

Claudi, Tommaso. 2014. *The status of subject pronouns in Old Russian. A diachronic analysis*. MA thesis, University of Pavia.

Håkansson, David. 2013. Null referential subjects in the history of Swedish. *Journal of Historical Linguistics* 3/2: 155-192.

Kroch, Anthony. 2001. Syntactic change. In Baltin and Collins, eds., *Handbook of Contemporary Syntactic Theory*. London: Blackwell.

Meyer, Roland. 2011. *The history of null subjects in East Slavonic. A corpus based diachronic investigation*. Habilitation thesis, University of Regensburg.

Walkden, George. 2013. Null subjects in Old English. *Language Variation and Change* 25: 155–178.

Peripheral Czech modal verb + infinitive constructions

František Martínek, Charles University in Prague, Faculty of Arts

New electronic sources filled with historical Czech texts (see below) make it possible to describe small shifts in the language development of Czech and in this way to understand language changes.

Various Czech constructions of finite modal verb with infinitive, expressing volitive modality,[1] are described already.[2] In this paper, I propose to outline the development of one less frequent full verb to modal (*hodlat*) and one contrary development (*uspět*).

The verb *hodlat* (together with its earlier form *hodlovati*; original 'to adjust sth., [to tailor sth.] to make sth. appropriate', related to adjective *vhodný* 'appropriate')[3] underwent the following semantic development. In the first phase, the verb broadens its meaning from designating a specific "constructional activity"[4] to more general 'to prepare'. This semantic bleaching responds to a wider collocability – instead of concreta, the verb may newly be combined with abstract substantives. In the second, grammaticalizational phase, taking place in the 19th century, the bleached light verb becomes modal, stabilizes in this function and its collocability radically changes: now it collocates with verbal infinitive only.

The opposing process of degrammaticalization is illustrated on the verb *uspět* 'to succeed in sth.'; formerly also 'to achieve, to manage sth.' and '[to manage] to flee, to escape'. This verb belonged to modals[5] in the Czech of the 19th and the first half of the 20th century. However, during the 20th century, it has lost its collocability with infinitive and also its wider meaning scale has been reduced to an exclusive meaning 'to succeed' ('to fail' in negation, respectively). This verb is firstly attested only in Jungmanns dictionary (1834-39), althougth more derivates of this stem with similar meaning, like *prospěti*, already occur in Old Czech.[6] Exactly thanks to its non-firm position in the system as well as through the influence of the equally old substantive *úspěch* 'success' its semantical and functional shift can probably be explained.

On both these examples from the last 200 years, one can see that new sources can help to describe the recent grammar development of Czech.

Sources:

Český národní korpus – Diakorp [on-line]. Available under https://kontext.korpus.cz.
Příruční slovník jazyka českého [on-line]. Available under http://psjc.ujc.cas.cz.
Vokabulář webový [on-line]. Available under http://vokabular.ujc.cas.cz.

References:

Grepl, Miroslav (1973): Vyjadřování modalitní kategorie záměru. *Slovo a slovesnost* 34/1, 74-77.
Karlík, Petr – Štícha, František (2011): Infinitivní fráze v některých typech syntaktických struktur. In Štícha, F. (ed.): *Kapitoly z české gramatiky*. Praha: Academia, 928-944.
Kolářová, Ivana (1999): Některé významy a funkce sloves podívat se, koukat/kouknout se. *Naše řeč* 82/2, 65-71.
Němec, Igor (1987): Předložka *po* se substantivem jako základ sloves a jejich etymologické východisko. *Naše řeč* 70/4, 177-184.
PMČ (1995): *Příruční mluvnice češtiny*. Praha: NLN.
Rejzek, Jiří (2002): *Český etymologický slovník*. Praha: Leda.

---

[1] Which includes intention, necessity, possibility and ability (see for instance Grepl 1973 and PMČ).
[2] For example, Karlík and Štícha describe the infinitive in a syntactic structure with verbs *být* and *mít* expressing possibility (*Mám/Je kde spát.* ‚I have / There is a place for sleeping.'; Karlík – Štícha 2011: 941-944), and Kolářová outlines the usage of several verbal forms of *(po)dívat se*, *hledět* and *koukat (se)*, all 'to see', as modal verbs and modal particles (Kolářová 1999).
[3] The synonyme Old Czech verbs *(při)hotovati/(při)hotoviti* 'to adjust, to prepate sth.' were able to combine with infinitive, on the contrary.
[4] This term was used by I. Němec for many Old Czech verbs, mostly denominatives, first specialized for designating of an concrete activity, later generalized to an action verb (see for example Němec 1987).
[5] To the "modal verbs in wider sense", respectivelly, in the terminology of PMČ.
[6] In Rejzek (2002), *uspět(i)* is classified as a loan word from Eastern or Southern Slavonic languages.

Ekaterina Mishina
Vinogradov Institute of Russian Language RAS, Moscow
kmishina@mail.ru

On the study of verbal aspect system in Old Russian and Old Church Slavonic

The problem of aspect in Old Russian and Old Church Slavonic is one of the most controversial matters. There have been many different opinions stated that vary between two opposite views: some scholars were inclined to identify the ancient state of the category of aspect in Old Russian with its current state in modern Russian (A. Vajan 1948, Ju. Maslov 1951, G. Khaburgaev 1997), whereas others deny it and suggest a quite late provenance for aspect in Russian (Bermel 1997, Norgard-Sorensen 1997).

In our opinion, although aspect was not totally grammaticalized in Old Church Slavonic and Old Russian (first of all, due to the fact that there still were quite a lot of biaspectual verbs), we have reasons to argue that some verbs were alredy perfective or imperfective at the early stage, at least their aspectual behavior in language is very similar to the behavior of perfective and imperfective verbs in modern Russian. There were, probably, few aspectual pairs in early period, however their derivation was already in progress (e. g., *dati – dajati*, *pustiti – pushati* etc.)

The modern approach to determination of the aspectual meaning of a verb in early period should combine both morphological and semantic criteria. Morphological criterion, which was already declared by A. Vajan 1948 and Koshmider 1934, is quite significant. The perfective verbs should not have forms of the present participles and imperfects, whereas the imperfective forms should not have the past participles. The forms of participles are more demonstrative here in comparison with the opposition *aorist:imperfect* that is more complicated, as imperfective verbs can be used in aorist (this form is not forbidden although less common for them). Nevertheless, the use of morphological criterion solely is not sufficient. Despite the main tendency, some forms of imperfects and the present participles of the perfective verbs as well as some forms of the past participles of the imperfective verbs attested both in Old Church Slavonic and Old Russian monuments. These forms (some uses of perfective present can be added as well) are not accidental, although most of them have been brought as the proof of abcence of aspect in Old Russian in many works. It is important to take these peculiar uses into account while building the verbal aspect system in Old Russian an Old Church Slavonic.

I believe that this combinatorial method allows the researcher to discover which aspect a verb had in Old Russian or Old Church Slavonic (or to state its biaspectuality) in most cases. According to their aspectual behavior, all verbs in this period could be classified into five groups (perfective, mostly perfective, biaspectual, mostly imperfective, imperfective).

I am going to demonstrate the use and the advantages of the declared approach on the data from two electronic resources: Russian National Corpus (http://www.ruscorpora.ru) – for Old Russian, and Old Russian Treebank (https://nestor.uit.no) – for Old Church Slavonic. It is interesting that the results are not always the same for both langueges.

Today's arguments are yesterday's circumstantials: a corpus-study of Russian valency patterns

Maria Ovsjannikova (Institute for linguistic studies, Russian academy of sciences, St. Petersburg)

The semantic arguments of some predicates in Russian are syntactically expressed by the prepositional phrase in which the choice of the preposition is assigned by the predicate (e.g. *spasti ot* 'save from', *smotret' na* 'look at'). Synchronically, the meaning of the preposition in such combinations can be regarded as corresponding to the participant role determined by the definition of the subcategorizing lexeme, cf. [Apresjan 1974], or induced on the basis of the common meaning shared by a semantically coherent group of lexemes, cf. [Zolotova 2006].

This study is concerned with the development of valency patterns with prepositional encoding of arguments in Russian of the XVIII–XX cc. The data are taken from the Russian National Corpus.

I argue that it is misleading to base the judgments on the argumental / circumstantial status of a participant in the texts of the earlier epochs on the semantic intuition of modern speakers even if the participant encoding strategy is stable over time. Instead, I propose to use the lexical distribution of the encoding strategy and the degree of syntactic bondedness of the preposition with the head as a proxy to determine whether the valency pattern should be considered more or less "fixed" (lexically subcategorized for) for a particular period.

The comparison of lexical distributions for different periods reflects whether a particular means of encoding is more or less productive, i.e. can be combined with an open class of lexical heads. To assess the lexical distribution for each preposition under study a random sample of examples from each period (usually about 1000) were manually searched for a particular meaning of a preposition. E.g. the sample for *ot* 'from' was searched for the examples where *ot* is used to encode the "undesirable" participant, as in *izbavit'sja ot* 'get rid of'. The resulting sample of about 100 instances for each encoding means for each period was then analyzed in terms of its type and token frequency structure to assess the degree of productivity of the encoding means, using the measures proposed in [Baayen 2009; Goto, Say 2009]. Some encoding strategies become less productive and over time become centered around specific lexemes, whereas some other are consistently unproductive through the three periods (e.g. *na* with verbs of watching like *smotret'*).

The degree of syntactic bondedness was mainly assessed comparing the median distance from the head to the dependent encoded by a preposition. For the prepositional strategies studied so far this measure correlates with the changes in the productivity characteristic of the encoding strategies.

Viewed from the perspective of the preposition (and other flagging devices), the development of valency patterns can be modeled as follows. In the course of grammaticalization the preposition (or case) acquires a new meaning whereby the range of its contexts of use is widening. In some of these contexts it yields to newly grammaticalized means with similar semantics, whereas in other it becomes lexicalized as a valency pattern and such prepositions "lose their independence from the verb and are somehow subsumed under its meaning" [Lehmann 1982/1995: 89].

References

Apresjan 1974 — Ю. Д. Апресян. Лексическая семантика. Синонимические средства языка. М.: Наука.

Baayen 2009 — H. R. Baayen. Corpus linguistics in morphology: morphological productivity // A. Luedeling, M. Kyto (eds.). Corpus Linguistics. An international handbook. Mouton De Gruyter, Berlin. P. 900–919.

Goto, Say 2009 — К. В. Гото, С. С. Сай. Частотные характеристики русских рефлексивных глаголов // К. Л. Киселева, В. А. Плунгян, Е. В. Рахилина, С. Г. Татевосов (ред.). Корпусные исследования по русской грамматике. М.: Пробел-2000, 2009. С. 184–223.

Lehmann 1982/1995 — Ch. Lehmann. Thoughts on Grammaticalization. München: Lincom Europa.

Zolotova 2006 — Г. А. Золотова. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М.: УРСС.

Vladimir Plungian, Anna Urmanchieva

THE PERFECT IN OLD CHURCH SLAVONIC: A CORPUS-BASED STUDY IN GRAMMATICAL SEMANTICS

Periphrastic perfect is a notoriously difficult form of OCS verb, as far as it remains constantly reluctant to any coherent semantic description. While the majority of OCS texts are (very literal) translations, readily calquing both lexical and grammatical features of Hellenistic Greek, the OCS perfect is almost unique in deviating drastically from this common trend.

The present paper attempts to tackle the semantic puzzle of OCS perfect making a greater emphasis on various corpus data. Our main source was an OCS segment within PROIEL corpus (http://www.tekstlab.uio.no:3000) consisting of *Codex Marianus* and *Codex Suprasliensis*. To this, data from other documents were added (for comparative purposes), namely, *Psalterium Sinaiticum* and *Euchologium Sinaiticum*.

A primary look at the examples indicates that one can hardly speak of OCS perfect as a unified grammatical value with one and the same range of uses in all available texts. It would be more profitable to establish the patterns of perfect use for individual documents, then proceeding in turn to comparison of these patterns in order to reveal possible common features.

Moreover, it is often the case that even different fragments of one document show different patterns of perfect uses. Thus, Codex Suprasliensis distinguishes the rules for perfect choice in *Vita* and in *Homily* (this fact was pointed out as early as in Večerka 1993). Similarly, *Codex Marianus* has slightly different rules in Matthew and Mark on the one hand, and in Luke and John, on the other hand.

When looking for a common denominator of various OCS perfect uses, long-established notion of "current relevance" (see Dahl & Hedin 2000, inter alia) may seem appropriate. It should be noted, however, that such notion remains inoperative without further specification, since it lacks predictive power in explaining aorist/perfect choice. We argue that for different groups of OCS texts this general notion has different interpretations. It should be emphasized, that we propose a different approach to the notion of "current relevance": not an aspectual, but a pragmatic one (which explains a wide range of its divergent uses driven by the speaker's pragmatical intention). To be pragmatically relevant, it is insufficient for a past situation to maintain its resulting phase and even to predetermine some present or future conditions. The current relevance in a pragmatic semce implies that the speaker, using an utterance with a perfect form, expects some *reaction* from his interlocutor. Thus, one can compare Mt. 20:12 with a perfect form requiring a definite reaction, with Mt 14:31 with an aorist in a rhetorical question. Different factors, predetermining the choice between competing perfect and aorist forms in different OCS texts will be discussed in the paper and illustrated by various examples.

REFERENCES

Dahl, Östen & Eva Hedin. 2000. Current relevance and event reference. In: Östen Dahl (ed.). *Tense and aspect in the languages of Europe*. Berlin: Mouton de Gruyter, 386-401.

Večerka, Radoslav. 1993. *Altkirchenslavische (altbulgarische) Syntax*. II. *Die innere Satzstruktur*. Freiburg i. Br.: Weiher.

Anna Ptentsova
Moscow State University

On part-of-speech attribution and grammatical tagging of Old Russian *krivo*

In this talk I will discuss the difficulties in part-of-speech attribution of Old Russian **кривο** (krivo). My research is based primarily on the Old Russian subcorpus of the Russian National Corpus (RNC). In addition, I consider the following sources: historical dictionaries, tagged texts from the collection of Old Russian manuscripts (http://www.lrc-lib.ru/), and also a number of offline sources.

I analyze the following types of contexts:

(1) **да аще кто ѿ руси или ѿ грекъ створи кривο да оправлѧетъ** – da ašče kto ot rusi ili ot grekъ stvori krivo da opravlyaetъ (Povest' Vremennykh Let, Hypatian Codex, year 945); "If any of the Russians or Greeks broke the rules (done awry), let them correct [it]".

In this kind of cases the dictionaries, as well as the RNC and lrc-lib.ru, consider **кривο** an adverb. It is quite likely, however, that it is a noun (compare to **добро** (dobro) 'good' and **зълο** (zъlo) 'evil'). Such an example is found in Ptchela, 287: **вѣрны͡х мни не тѣхъ иже по твоемоу словоу молвѧть но иже противѧть сѧ гл҃емымъ тобою по кривоу** – vernyx mni ne texъ iže po tvoemu slovu molvyaty no iže protivyatьsya glagolemymъ toboyu po krivu "Consider faithful not those who repeat your words, but those who disagree with your untrue sayings (saying by wrongfulness)".

Consider also the following example:

(2) **поутѧта пьсалъ . даче кривο да исправите а не кльните** - putyata pьsalъ dače krivo da ispravite a ne klьnite (Menaia XI c., 135 r. - a margin note by scribe Putyata); "Putyata wrote (this). If (it is) crooked, correct (it), but do not curse".

It appears that in cases like (1) and (2) **кривο** can either be classified as an adverb meaning 'making a mistake in action' or as a noun meaning 'a wrongful deed'. (Note that the first sense is retained in modern colloquial Russian: *Они как-то криво договорились и не смогли встретиться - Oni kak-to krivo dogovorilis' i ne smogli vstretit'sa* "They made an awry arrangement and failed to meet".)

It's very important, however, that in contexts like (3) we may not classify **кривο** as an adverb:

(3) **дажь въ нѫ поѫ обрѫще кривο а вьсе** <...> - daže v nya poya obryašče krivo a vьse <...> (Novgorod Menaion 1095-1097, September, 176 r.– a margin note by scribe Domka) "Should (anyone) conducting a sermon by them (these books) find a mistake" <...> (fragment ends).

Обрѫще кривο does not signify 'makes a mistake in the process of finding something';

here **кривѡ** is an object, and therefore a noun, which votes for the noun interpretation in (2) and, evidently, in (1).

In a number of contexts the homonymy is impossible to resolve, and this should be reflected in the morphological markup in Corpus and other electronic resources. The solution may be to introduce double tags for these cases.

The same attribution problem is observed in similar Old Russian words ending in **–ѡ**, e.g. in **правѡ** (pravo). According to the Corpus, however, ambiguity is much less common for these lexemes.

Corpus as a tool in real-time sociolinguistics: the spread of an innovation in the texts of Russian 19th-century writers

Sergey Say (Institute for linguistic studies, Russian Academy of Sciences)

Background. Our understanding of the unfolding of language change in real time is undermined by the shortage of sufficiently deep longitudinal studies. In this study I investigate the potential of the Russian National Corpus (RNC, www.ruscorpora.ru) as a tool in real-time sociolinguistics. The variationist quasi-longitudinal approach has already been applied to literary written texts (see [Arnaud 1998] for an early example), but such studies are few and, as far as I know, have never been carried out on Russian material.

Data and method. I focus on the spread of the syncopated Instr.Sg. inflection *-oj* instead of older bisyllabic *–oju* in Russian *a*-class nouns (cf. *ruk-oj*/*ruk-oju*). This grammatical phenomenon clearly undergoes a classical S-shape development, with its rapid stage taking place in the 19th century: the ratio of *-oju*, p(*-oju*), dropped from 0.71 in 1801-1820 to 0.14 in 1881-1900. I studied the use of the alternative forms in texts of 50 Russian writers who are well represented in the RNC. The independent variables in the study are the date of creation, author, and author's year of birth; the dependent variable is p(*-oju*) in the writings of a particular writer from a particular period. I also propose a measure of the conservativeness of a writer in a particular period: a standard score (z-score) of their p(*-oju*) in a period as compared to the distribution of p(*-oju*) of other writers in the same period.

Results

1. Both the date of creation and the date of birth are very strong predictors of the dependent variable, but the former correlation is stronger. In linguistic terms it means that the writers tend to reflect the current use at the time of their writing rather than to simply stick to the pattern typical of their generational cohort.

2. There is a huge dispersion among individual writers, even if they belong to the same generation and write at the same time (cf. e.g. Leo Tolstoy and Buslaev in Fig. 1 below).

3. The ratio of *–oju* typically is not stable during writer's lifespan, i.e. it changes far beyond the so-called "critical age" (there are notable exceptions, though; see Leo Tolstoy's almost flat curve in Fig. 1 below).

4. On average, individual writers' p(*–oju*) values tend to decline during their lifetimes, but at the pace which approximately two times lower than in the community in general (Saltykov-Schedrin, see Fig. 1, shows a fairly average pattern of change). These data discredit the "apparent-time hypothesis" in its uttermost form (generational differences faithfully reflect stages of language change). The observed scenario is somewhere between generational change and communal change (see [Sankoff 2002] for these types of patterns).

5. As a consequence, there is an almost exceptionless pattern: irregardless of whether individual writer's absolute p(*–oju*) is high or low, and even if it declines over time, the relative conservativeness of an individual writer (z-score) increases with age.

6. Conformity of individual writers to the general dynamic trend correlates negatively with age: on average, older writers move in the general direction slower that younger writers, and more frequently turn to more archaic patterns in absolute terms (cf. Goncharov's curve in Fig. 1).

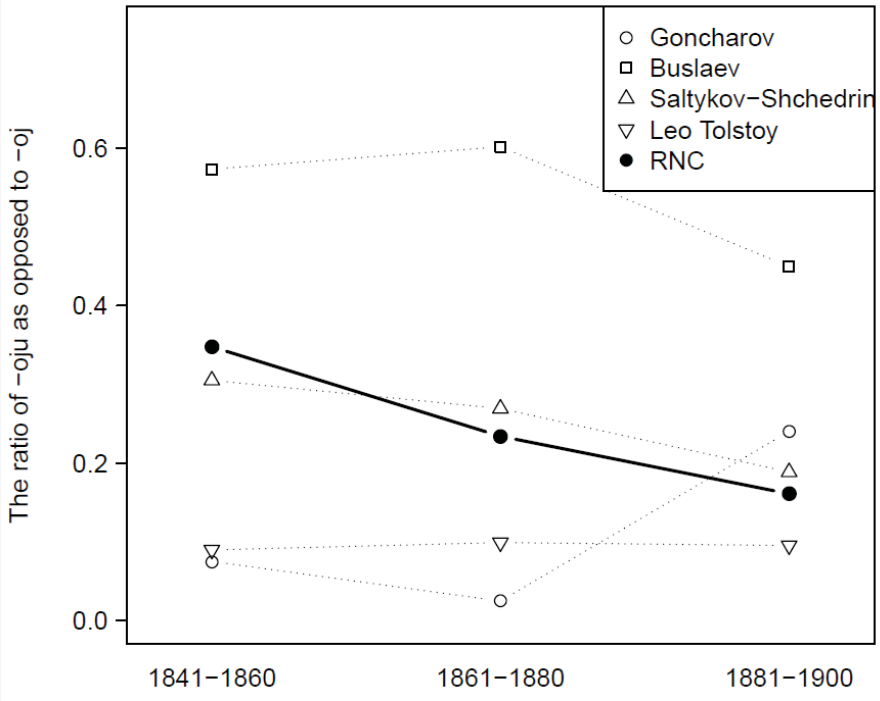


Fig. 1. The ratio of *–oju*: four selected writers and RNC in general (1841-1900)

References

Arnaud, R. 1998. The development of the progressive in 19th century English: a quantitative survey. *Language variation and change*, 10: 123-152.

Sankoff, G. 2005. Cross-sectional and longitudinal studies in sociolinguistics. *U. Ammon et al. (eds.). Sociolinguistics: an international handbook of the science of language and society*. Vol. 2. NY, Berlin: de Gruyter. 1003-1013.

Animacy-driven Differential Object Marking in Russian. Diachrony.

Ilja A. Seržant

(Johannes-Gutenberg-University Mainz / University of Vilnius)

1. Introduction

Modern Russian synchronically has four distinctive differential argument marking systems. All four are based on the alternation between NOM/ACC (the unmarked alternate) versus GEN (semantically the marked alternate): (i) the animacy-driven Differential Object Marking (DOM), (ii) the referentiality-driven DOM with verbs under negation and (iii) quantification/(in)definiteness-riven DOM. The present paper is devoted to the first, major type (i):

(1) *Ja* *uvidel* *otc-**a***
I see.pst.m.sg father-**acc=gen**
'I saw a/the father.'

(2) *Ja* *uvidel* *vyxod-#*
I see.pst.m.sg exit-**acc=nom**
'I saw a/the exit.'

The historical development of the A-DOM is somewhat perplex and has two main motivations, as was first recognized by Klenin (1983):

- the discriminatory function (after the conflation of the nominative and accusative due to phonetic changes), and
- it is the result of a major process of the penetration of the accusative marking into the domain of the genitive (originally partitive genitive) direct objects and vice versa.

The discriminatory function, i.e. the function of distinguishing subjects from objects, plays an important role here. This is most evident if the chronology of the expansion of the animate genitive direct objects is taken into account: it gradually affects the NPs from the left to the right of the Animacy Hierarchy:

(3) First & Second person > 3rd person pron. > Proper names > Common nouns, human > Common nouns, Anim. > Common nouns, Inanim.

Notably, the progress has a number of exceptions.

2. Discussion

It is commonly assumed that the reason of the genitive marking in the direct object position is the result of a major functional conflation of accusative and genitive direct objects in the course of development (first suggested in Klenin 1983). This conflation enabled the penetration of the accusative into the genitive domain (typically partitive objects, objects under negation, objects of intensional verbs, etc.) as well as the penetration of the (former partitive) genitive into the accusative domain. The only question that remains to be answered is **how the genitive marking, typically associated with decreased referentiality could have been transferred to highly prominent NPs**, namely, humans denoting NPs in Old Russian, while not affecting the low prominent NP types (indefinite pronouns, NPs with inanimate reference).

I will claim in this paper that this penetration has been enabled by ambiguity contexts in which the semantics of *decreased referentiality* typical for genitive (direct) objects has been neutralized. These ambiguity contexts have been created by the (former partitive) genitive

becoming just a syntactic rule intruding into the case frame of the respective predicate already during the Old Russian period:

i. the genitive under negation which became a syntactic rule applying to indistinguishably all types of direct objects (including those with inherent prominence),

ii. intensional verbs (such as *iskati* 'to seek', *žьdati* 'to wait') have generalized the genitive case into the only option of encoding their objects.

I will present statistic data from Codex Laurentius and Codex Assemanianus to corroborate my claims.

References

Klenin, E. 1983: *Animacy in Russian: A new interpretation*. Columbus, Ohio: Slavica Publishers, Inc.

Frequency-oriented diachronic approach to the study of prefixes variation in the aspectual system of the Russian language

Valery Solovyev

Kazan Federal University, Russia

The aspectual system of the Russian language description with the emphasis on the Natural and Specialized perfectives rises up a principle issue of these two perfectives types separation. The difference between the two refers to the match or mismatch of their lexical semantics with the semantics of the basic imperfective. However, the semantics is a thing of a very delicate and informal nature, the property that makes the distinction very complex and leads to the "diffuse zone" between the Natural and Specialized perfectives (according to Janda, et al, 2013, "Why ..."). This is well illustrated by the "Questionary" (Gorbova, 2011, Voprosy Jazykoznanija) through the professional linguists' opinion dispersion. We find great variation in the dictionaries as well. For example, the Ozhegov's Dictionary construes as Natural perfectives from the word "бить" (beat) only "побить", "разбить", "пробить". In Shvedova's Semantic Dictionary these are added with "забить" and "сбить". And in Ushakov's Dictionary also - "убить" and "прибить". The purpose of this paper is to introduce a certain formal approach, which could serve as a working tool for the detection of Natural perfectives.

We use diachronic corpus Google Books Ngram (https://books.google.com/ngrams) with the service of graph plotting for frequency of words and phrases use (hereinafter referred to as time series). It contains more than 67 billion Russian words and covers over two centuries. Our basic assumption is the following.

The main hypothesis. Semantics of imperfective and corresponding perfective coincide (up to their aspectual meaning) if and only if their time series forms also coincide. Here, the key point is, that if a perfective is of essentially different meaning as compared with an imperfective, the frequency of their use cannot change synchronously during a long time period. At some point the additional meaning of a perfective would be either more or less in demand as compared with the meaning of an imperfective, and forms of their time series will vary significantly.

This is well illustrated by the following examples: "анализировать" – "проанализировать" (analyze), "фотографировать" – "сфотографировать" (photograph), "нагреть" – "нагревать" (heat up) и "гореть" – "загореть" (burn – get a tan). The data is presented on a random sampling of perfectives from the "Exploring Emptiness" database (http://emptyprefixes.uit.no/book.htm). We discuss the possibility of complete formalization of the concept "similarity of time series form". The limitations of the method and difficult cases of its application related to the words polysemy are discussed. Separation of the meaning in question is possible through the context fixation (usually bigram) and comparing of bigrams' time series. Detailed analysis is given to the perfectives from the verb "бить".

This methodology can also be used in educational process. It allows specifying the difference in semantics of alternative prefixes and helps to make a correct choice of the appropriate prefix. Developed approach supports the aspectual cluster conception (Janda, 2007, Studies in language) and non-emptiness of aspectual prefixes, as well as specifies these concepts for the diffuse zone structure between Natural and Specialized perfectives. This is an interesting example of how a diachronic corpus can be used for purely synchronous problems solution. In future it is intended to be used for study of the aspectual system evolution in the Russian.

Orthography as a window to diachrony

Barbara Sonnenhauser

The present paper relates to a project which aims at establishing a linguistic resource that may serve both as a digital edition and a searchable diachronic corpus. This poses challenges not only for the technical make-up, but also for the preparation of the data. While this is self-evident for morphosyntactic and pragmatic aspects, it is less obvious for orthography.

From a contemporary perspective, orthography may easily be considered a secondary phenomenon with only marginal linguistic relevance and hence of little importance for digital resources. As regards pre-standardised texts, however, matters are different, as will be shown here for 17th-19th century Balkan Slavic. With traditional norms vanishing and new norms only gradually evolving, orthography is more or less subject to an individual's performance and thus to a large degree rhetorically conditioned, making overt what Chafe (1988) calls 'written language prosody'. Thereby, orthography promises – at least partial – access to actual language usage.

Discussing the pragmatic import of punctuation in Old Russian texts, Gvozdanović (1995: 177) concludes that "the use of modern punctuation in philological editions fails to do justice to the language of the manuscripts". This paper argues that orthography has an even wider relevance, in that it permits insight into the diachronic development of morphosyntactic regularities. Imposing contemporary orthography in modern editions may cause this important resource to be overlooked, as is indicated by the following observations:

— Inserting full stops in order to divide texts into sentence-units may mistake functionally relevant 'thetical' elements (Kaltenböck et al. 2011) for anacolutha.

— Contemporary punctuation may impose distinctions that might not have been relevant to the same degree in older stages, e.g. hypotaxis vs. parataxis, direct vs. indirect speech.

— Inserting spaces may obscure differences that are potentially relevant as concerns word order preferences or clitic placement. The integration of *se* into the phonological word, (1a), or not, (1b), in Punčo's manuscript may thus be significant, but is lost in the edition :

(1) a. inevarnus*e* (Punčo) > i ne vrьnu *se* (Angelov 1958)
b. i onь *se* čudeše (Punčo) > i onъ *se* čudeše (Angelov 1958)

— Judging from the edited text, the nominative *edna žena* as object to *vide* in (2a) is wrong. But *edna žena* could also have been intended to function as a subject of a main clause, (2b). Possibly, evidence is provided by the punctuation used in the manuscript.

(2) a. i vide *edna žena* kato edna carica (Demina 1971)
b. i vide : *edna žena* kato edna carica

— Orthography in the manuscripts is largely phonetically conditioned and thus an important source of evidence for dialectological research, especially as regards variation.

These observations raise the question of how to manage, philologically as well as technically, the trade-off between faithfulness to the data (making the corpus truly diachronic) and generalization and abstraction by means of normalization (making the corpus accessible).

Angelov, B. 1958. *Iz starata bălgarska, ruska i srăbska literatura*. Sofija
Chafe, W. 1988. Punctuation and the prosody of written language. *Written Communication* 5, 395-426
Demina, E.I. 1971. *Tixonravovskij damaskin. Bolgarskij pamjatnik XVIIv. Issledovanie i tekst*. Sofija
Gvozdanović, J. 1995. Parameters underlying the organization of mediaeval Russian texts. Andersen, H. (ed.). *Historical linguistics 1993*. Amsterdam, 177-190
Kaltenböck, G. et al. 2011. On thetical grammar. *Studies in Language* 35/4, 852-897
Punčo: *Pop Punčov Sbornik*. 1796

Reconstructing functional change based on a parallel corpus: The rise of DO+Genitive in North Slavic

Ruprecht von Waldenfels

We use ParaSol (www.parasolcorpus.org), a tagged, lemmatized, sentence and word aligned corpus of translations in all major Slavic languages as a data basis. Building on a methodology outlined in Waldenfels (2014), we classify all prepositional phrases in the parallel text into cognate prepositional classes such as DO, V+ACC, V+LOC, K, NA+ACC, NA+LOC, ZA+INS, ZA+ACC, U+GEN, and others. Since the corpus is word aligned, we can directly compare the use, and thus, the function of these preposition classes in a large number of instances across translations into many Slavic languages.

The corpus based comparison gives us a handle to judge functional similarity bottom-up and affords diachronic insights, adding to traditional descriptions such as Kopečný (1973). We **focus on** DO+GEN, which takes over the functions of V+ACC in an areal reaching from Sorbian, Czech and Slovak in the west, where this change probably originated, towards Polish, Ukrainian and Belarusian, where this change is least pronounced. At the same time, we see a similar tendency of DO to take over the functions of K in a subset of these languages, i.e., the change is not coextensive. Other changes seem more isolated, as the functions of K in Slovenian, which has taken a development very different from K in other Slavic languages. These findings are arrived at by analyzing visualizations of the aggregate data supplemented by specific corpus searches to validate and refine hypotheses.

Problematic *iže*

Miriam Zumstein

The current policy concerning the annotation of 'relatives with *-že*' in the TOROT corpus is to count them as one token and not as adverb/pronoun + clitic. At first sight, the decision to follow the editions in these cases is sensible because the main aim in the annotation process is to be consistent, and it is fairly easy for the annotators to observe this rule. Closer inspection of the output shows, however, that this 'rule of thumb' may lead to confusion about sentence boundaries and the relationship between the sentences involved. This is not astonishing when considering the following problems:

1. The homonymy between the anaphoric pronoun *i* emphasized by the particle *že* and the relative pronoun *iže* (cf. 1)
2. The different…
   a) types of relative clauses (e.g. restrictive, nonrestrictive, pseudo-relatives = connectio relativa, correlatives) and the oscillation of some subtypes between parataxis and hypotaxis (cf. Gołąb and Friedman 1972; Mitrenina 2012)
   b) degrees of clausal linkage (cf. Lehmann 1988)

In the first part of the talk I will focus on syntactic environments in which the interpretation of the *i-že* sequence as a relative pronoun is either impossible, e.g.:

A) *i-že* in participial clauses modifying the predicate of the main sentence (XADV)
   – iff the argument expressed by *iže* is not shared between the participle and the main verb as in 1.

(1) и списаша книгы многы. и сниска $\emptyset_{\text{OBJ}}$ [имиже пооучаѣсѧ] вѣрнии людье наслажаютсѧ. оученья бжтсвнаго. (Laur., but also Rad, Aka, Ipa, Xle)

or debatable, e.g. in 2 where the conjunction *no* 'but' contrasts the predicat in the *iže*-clause with the following sentence (β) which doesn't exhibit any connection to sentence α.

(2) служать бо wпрѣснoки рекше wплатки.$_\alpha$ || ихъже бъ не преда. || но пове хлѣбомъ служити.$_\beta$ (Laur.)

In cases like these, *i-že* should not be tagged as a relative pronoun but as a sequence of an anaphoric pronoun and an emphatic particle – *although* it is rendered as one form in the editions used.

In the second part of the talk I will provide some statistics for the frequency of the different subtypes of *i-že* clauses and attempt to design a checklist for annotators based on my findings.

References

Zbigniew Gołąb and Victor A. Friedman. The relative clause in slavic. In P. Peranteau, J. Levi, and G. Phares, editors, *The Chicago Which Hunt. Papers from the Relative Clause Festival*, page 30–46, Chicago, 1972.

Christian Lehmann. Towards a typology of clause linkage. In J. Haiman and S. A. Thompson, editors, *Clause combining in discourse and grammar. Proceedings of a conference at Rensselaerville Institute, Albany*, pages 181–225, Amsterdam, Philadelphia, 1988. John Benjamins.

Olga V. Mitrenina. The syntax of pseudo-correlative constructions with the pronoun kotoryj ('which') in middle russian. *Slověne*, 1:61–73, 2012.