SlaviCorp 2018

24–26 September 2018
Charles University, Prague

Book of Abstracts

Table of Contents

Plenaries

Björn Hansen
Universität Regensburg
Bjoern.Hansen@sprachlit.uni-regensburg.de

Edyta Jurkiewicz-Rohrbacher
Universität Regensburg
Edyta.Jurkiewicz-Rohrbacher@ur.de

Zrinka Kolaković
Universität Regensburg
Zrinka.Kolakovic@sprachlit.uni-regensburg.de

Detecting constraints on clitic climbing – with the help of corpora and psycholinguistic tests

The talk aims to show how corpora can be used to study fairly complex phenomena. We will base the discussion on the example of constraints on clitic climbing in Bosnian, Croatian and Serbian (BCS). Descriptively speaking, clitic climbing (CC) "refers to constructions in which the clitic is associated with a verb complex in a subordinate clause but is actually pronounced in constructions with a higher predicate" (Spencer & Luís 2012: 162). An example of CC out of an infinitival complement is given in (1) where the clitical pronoun *ga* 'him' is realised in the second position of the matrix clause (Wackernagel position); in other cases, however, CC does not take place as in (2) where the clitic *ih* stays in the complement clause.

(1)	*Milan*	***ga***$_2$	*mora*$_1$	*vidjeti*$_2$.
	Milan	him.ACC	must.3PRS	see.INF
	'Milan must see him.'			Stjepanović (2004: 179f)

(2)	*Bojim*$_1$	***se***$_1$	*testirati*$_2$	***ih***$_2$.
	afraid.1PRS	REFL	test.INF	them.ACC
	'I am afraid to test them.'			hrWaC v2.2

Although clitics in Bosnian, Croatian and Serbian (BCS) have attracted considerable attention in the syntactic literature (cf. Franks & King 2000, Browne 2014, or Bošković 2004), the syntactic conditions and constraints for CC are seriously understudied in comparison to e.g. Czech (e.g. Junghanns 2002). There are only very few studies on CC in BSC: Stjepanović (2004), Aljović (2004, 2005) mainly deal with theoretical considerations based on a small selection of construed examples.

Jurkiewicz-Rohrbacher et al. (2017a, 2017b), Hansen et al. (2018) are the first descriptions of CC in BCS based on empirical investigations. Basing on the data obtained from massive web corpora {bs, hr, sr} WaC (Ljubešić & Klubička 2014), the raising-control dichotomy of matrix predicates is shown to be a relevant factor of CC. Apart from that, it is found out that reflexivity plays a major role. Kolaković et al. (accepted), on the other hand, tackle the question of register as a relevant factor by comparing results from Forum subcorpus of hrWaC v2.2, Croatian Language Repository (Ćavar & BrozovićRončević 2012) Croatian National Corpus (Tadić 2009)while examining the same types of matrix predicates.

First, the talk presents the results of the corpus based and corpus driven studies mentioned above, discusses in detail the particular steps of a corpus approach, ranging from the formulation of queries, coping with tagging errors, to the statistical analysis of the data. Second, it will show how these results feed into a major psycholinguistic experiment recently carried out in Croatia (7 experiments x 40 participants = 280 participants). The logistic regression mixed models based on data from thespeeded yes-no grammaticality judgment tasks with OpenSesame free software provide the additional evidence for constraints on CC.

Bibliography

Aljović, N. (2004) "Cliticization Domains: Clitic Climbing in Romance and in Serbo-Croatian." In: Crouzet, O. et alii (eds.) *Proceedings of JEL'2004 Domain(e)s,* Université de Nantes, 169-175.

Aljović, N. (2005) "On clitic climbing in Bosnian/Croatian/Serbian". In: Leko, N. (ed.) *Lingvistički vidici* 34:(05). Sarajevo: Forum, 58-84.

Bošković, Ž. (2001) *On the nature of the syntax-phonology interface: cliticization and related phenomena.* Amsterdam: Elsevier.

Browne, W. (2014) "Groups of Clitics in West and South Slavic Languages". In: Kaczmarska, E.; Nomachi, M. (eds.) *Slavic and German in Contact:*

Studies from Areal and Contrastive Linguistics. Slavic Eurasian Studies 26, 81-96
Ćavar, D., Brozović-Rončević, D. (2012) "Riznica: The Croatian Language Corpus". In: Prace filologiczne 63, 51-65.
Franks, S. & King, T. H. (2000) *A Handbook of Slavic clitics*. Oxford: OUP.
Hansen, B.; Kolaković, Z.; Jurkiewicz-Rohrbacher, E.; (2018) "Clitic climbing and infinitive clusters in Bosnian, Croatian and Serbian – a corpus-driven study." In: Fuß, Eric et al., *Grammar and Corpora 2016*. Heidelberg: Heidelberg University Publishing (heiUP).
Junghanns, U. (2002) „Clitic climbing im Tschechischen". In: *Linguistische Arbeitsberichte* 80, 57-90.
Jurkiewicz-Rohrbacher, E.; Kolaković, Z.; Hansen, B. (2017) "Web Corpora – the best possible solution for tracking rare phenomena in underresourced languages: clitics in Bosnian, Croatian and Serbian". In: Bański, P. et al. (eds.): *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section*. Mannheim: Institut für Deutsche Sprache, 49-55.
Jurkiewicz-Rohrbacher, E.; Hansen, B.; Kolaković, Z. (2017) "Clitic climbing, finiteness and the raising-control distinction. A corpus-based study." In: *Jazykovedný časopis* 68:(2), 179-190.
Kolaković, Z.; Hansen, B.; Jurkiewicz-Rohrbacher, E.; (accepted) "Uspon zanaglasnice, dihotomija dizanje – kontrola i stilska varijacija". *6. Hrvatski sintaktički dani - Sintaksa zavisno složene rečenice*, 17.-19.05.2018 Osijek.
Ljubešić, N., Klubička, F. (2014) "{bs,hr,sr}WaC — Web corpora of Bosnian, Croatian and Serbian". In: Bildhauer, F., Schäfer, R. (eds.) *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*. Gothenburg, 29-35.
Spencer, Andrew; Luís, Ana R. (2012) *Clitics. An Introduction.* Cambridge: Cambridge University Press.
Stjepanović, S. (2004) "Clitic Climbing and Restructuring with "Finite Clause" and Infinitive Complements". In: *Journal of Slavic Linguistics* 12: 1, 173-212.
Tadić, M. (2009) "New version of the Croatian National Corpus". In: Hlaváčková, D. et al. (eds.) *After Half a Century of Slavonic Natural Language Processing*. Masaryk University, Brno, 199-205.

••••

Alexandr Rosen
Charles University, Czech Republic
alexander.rosen@ff.cuni.cz

The merits of a parallel corpus and how to get the most out of it

InterCorp, a multilingual parallel component of the Czech National Corpus, has been on-line since 2008, growing steadily to its present size of 1.7 billion words in 40 languages. A substantial share of fiction is complemented by legal and journalistic texts, parliament proceedings, film subtitles and the Bible. The texts are sentence-aligned, tagged and lemmatized. After a brief presentation of the corpus design, content and access options, we will see how useful it can be in linguistic and literary studies, and for practical tasks in fields such as lexicography, teaching or translating. Finally, we will look at the issue of language-specific morphosyntactic annotation, turning a multilingual corpus into a tagset Babylon, and present some solutions.

••••

Ruprecht von Waldenfels
Univesity of Jena
ruprecht.waldenfels@uni-jena.de

Variation on many levels: why and how comparing corpora and (Slavic) languages makes sense

In my talk, I present a number of projects concerned with comparative Slavic corpus linguistics from different perspectives. First, I show how a parallel corpus can be used to identify interesting synchronic contrasts between the Slavic languages that otherwise easily escape attention. Then, I trace some of these contrasts in a diachronic, comparable corpus of West Slavic languages in order to gain historical insight into how these contrasts developed and how they can be explained. I then use a regionally tagged corpus of Ukrainian to investigate the geographical distribution of contrasting forms and put them into perspective in respect to general patterns of variation in standard Ukrainian today. Finally, I outline how a nascent network of dialect speech corpora could be used to help explain this variation. In my talk, I thus attempt a tour de force of several projects that employ different approaches to analyze data from diverse sources and in distinct languages, and aim to show the benefit of such an eclectic approach that becomes increasingly feasible and, I would argue, necessary as more and more corpora of diverse types become available and relatively easy to use for non-computational linguists.

....

Full papers

Magdalena Adamczyk
University of Zielona Góra
m.adamczyk@wh.uz.zgora.pl

A contrastive look at discursive uses of English ‚now' and Polish ‚teraz'

English *now* and its immediate equivalent (but not etymological cognate) in Polish, i.e. *teraz*, both have a double-edged nature in that they can function as temporal adverbs, indicating that events are taking place at the moment of speaking or that they are in progress for a limited period of time, and as discourse markers (DMs), signalling relationships between neighbouring segments of discourse. The present study deals solely with the latter use, which is illustrated below:

Example 1

I'm afraid I can't go today. ***Now****, if you'd asked me yesterday I would have said yes. (Cambridge International Dictionary of English 1995)*

Example 2

Przeglądarka obsługuje grafikę, animacje i multimedia. ***Teraz****, jeśli chcemy przygotować prezentację przenośną, wybieramy z menu Plik polecenie Spakuj na dysk CD…* (*Wielki słownik języka polskiego* [A great dictionary of Polish]; http://www.wsjp.pl)

The browser handles graphics, animations and multimedia. ***Now****, if we want to prepare a portable presentation, we select from the File menu the command Send to CD drive…*

Unlike the DM *now*, which has already been examined in a number of studies (e.g. Schiffrin 1987, Aijmer 2002, Schourup 2011), non-temporal uses of teraz are still an unexplored area and all available information on them comes from lexicographical sources (e.g. *Wielki słownik języka polskiego* [A great dictionary of Polish] online, *Inny słownik języka polskiego* [A different dictionary of Polish] 2000). Although the two markers are intriguing not only from a monolingual perspective but also cross-linguistically, so far they have not been the focus of any contrastive study.

The aim of this paper is to investigate the discourse functions of *now* and *teraz* in order to find out to what extent they overlap and differ. The data for the study have been extracted from the spoken part of two electronic corpora, namely the British National Corpus (BNC) and the National Corpus of Polish (Narodowy Korpus Języka Polskiego (NKJP)). The choice of spoken material was based on a widely acknowledged fact that DMs are to be expected in speech rather than writing. Since there was no way of automatically sifting out the temporal uses of the lexemes from the discursive ones, this task had to be carried out manually. The examples which represented borderline cases (i.e. where the lexemes were neither genuine DMs nor temporal adverbs) were excluded from the final data set.

The results of a preliminary study make it possible to hypothesize that, while there are many contexts in which the items are used identically (e.g. as topic changers or markers of a return to a previous idea), *now* has a broader functional spectrum than teraz, and so the lexemes are not always the exact equivalents of each other. Interestingly, while some functions of *now* can only be reproduced in Polish by means of expressions other than *teraz* (and sometimes zero correspondence will be the optimal solution), for all instances of teraz examined so far *now* is a perfect match in English. The following examples from the BNC illustrate some of the contextual environments of the English marker in which its immediate Polish equivalent would be impossible or sound unnatural:

Example 3

(…) some people are still fighting to get their (…) flights right (…) ***now*** *(≠* **teraz***) aren't they?* (1 J9X)

Example 4

And we understand that he is paid up to 500 a day to carry out those services (…) ***Now*** *(≠* ***teraz****) he has very emotional reasons for carrying out this work.* (132 KRM)

In the paper an attempt is also made to consider plausible reasons for the mismatches between the functional scopes of the markers. One of them could be the presence of the discourse particle *no* in Polish which, sympto-

matically, might be an etymological cognate of *now* (see Auer and Maschler 2016) and in some environments proves to successfully take over its role (although in general it seems to be characteristic of more casual speech than the English marker). Another reason might be related to the fact that although teraz appears in a significantly fewer number of contexts, one could imagine it used as an equivalent of *now* in some settings not found to be shared by the two markers.

References

Aijmer, K. (2002). *English Discourse Particles: Evidence from a Corpus* (Studies in Corpus Linguistics 10). Amsterdam, Philadelphia, PA: John Benjamins Publishing Company.

Auer, P. and Y. Maschler (eds.). (2016). Nu/Nå: A Family of Discourse Markers across the Languages of Europe and beyond (Linguae & Litterae 58). Berlin, Boston, MA: Walter de Gruyter.

Cambridge International Dictionary of English. (1995). (ed. P. Procter). Cambridge: Cambridge University Press.

Inny słownik języka polskiego [A different dictionary of Polish]. (2000). (ed. M. Bańko). Warszawa: Wydawnictwo Naukowe PWN.

Schiffrin, D. (1987). *Discourse Markers* (Studies in Interactional Sociolinguistics 5). Cambridge: Cambridge University Press.

Schourup, L. C. (2011). The discourse marker *now*: A relevance-theoretic approach. Journal of Pragmatics, 43 (8), 2110-2129.

Wielki słownik języka polskiego [A great dictionary of Polish]. Wydawnictwo Naukowe PAN, <http://www.wsjp.pl> (accessed 25 November 2017).

••••

Dorota Adamiec
Institute of Polish Language, Polish Academy of Sciences
dorota.adamiec@interia.pl

Renata Bronikowska
Institute of Polish Language, Polish Academy of Sciences
r.bronikowska@wp.pl

Włodzimierz Gruszczyński
Institute of Polish Language, Polish Academy of Sciences
wlodekiewa@poczta.onet.pl

Emanuel Modrzejewski
Institute of Polish Language, Polish Academy of Sciences
modrzejewski.emanuel@gmail.com

Aleksandra Wieczorek
Institute of Polish Language, Polish Academy of Sciences
aleksandra.e.w@gmail.com

The Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772). The final result

In our speech we will present the final results of the five-year work on *The Electronic Corpus of the 17th and 18th c. Polish Texts (up to 1772)*. Shortly after the ending of the project we can present an open, considerably large corpus of Polish texts from the 17th and the 18th centuries. In the first part of the presentation we will present the main information about the corpus: its content and size (with wide statistic information), the history of its development, the main tools used during the building of the corpus. In the second part we would like to present the corpus as a rich resource for studies on the Polish language of 17th and 18th centuries. Finally, we will tell about our plans of further development of the corpus and its integration with another electronic resources of the Polish language.

The corpus contains Polish texts that had been published from the beginning of the 17th century up to the year 1772, which makes about 13,5M tokens from over 700 sources. Some of the texts were re-written from the original editions, but also for some of them we had to use later (19th and

20th-century) editions. The texts are transliterated (very close to the original spelling, but including some graphical unifications) as well as transcribed. The corpus is morphosyntactically annotated and lemmatized. The annotation and lemmatization were done by the tagger, which had been trained on the manually annotated subcorpus of 0,5M. tokens. The corpus is also annotated with rich metadata, e.g. author, title, release date, region, style, genre. During the works we used some already existing tools and methods for the Polish language processing, which had to be adjusted to the historical material. Other tools had been created for the purpose of this project. The corpus is already used by the authors of *The Electronic Dictionary of the 17th-18th c. Polish* and in the project Chronoflex which will enable representation of changes in the Polish inflection over the span of the history of Polish.

The corpus is searchable with the MTAS search engine. For users that are not very familiar with the query language there was created a tool which allows to create a query by choosing grammatical categories from the list. The user can search for the given string of characters, lemma, grammatical form or word form – on the level of transliteration or transcription, or both of them at the same time. We can limit the search to the parts of texts written in Polish, but we can also search for the given string of characters in the parts written in Latin or other foreign languages. We can also limit the search using the metadata, e.g. choose all texts of the given author or from the given period or area. Using metadata we can choose original editions from the 17th and 18th centuries or later, modernized editions as well. The search engine presents the results as a concordance list. For each result we can see a wider context with all information about the text (metadata) and even with the number of the page in the original edition, which is important to the researchers of the history of the Polish language. The user can choose between the view of results in transliteration or in transcription.

Planned works on the corpus on the one hand will concentrate on its further development. We plan to add new texts from the same period as well as to expand its time range up to the end of the 18th century. We also want to check the possibilities of applying on the corpus other tools created for the Polish language, like syntactic parsers, and adjusting them to the historical material. On the other hand we plan to integrate the corpus with other resources of Polish. The main one is *The Electronic Dictionary of the 17th-18th c. Polish.* The corpus is supposed to streamline and speed up the work on the Dictionary, especially by automation of supplementing the list of entries, filling the tables of grammatical forms of each entry and searching for quotations. The aim of the Corpus is also to supplement the *National Corpus of Polish* (http://nkjp.pl) with a diachronic aspect.

References

Adamiec, D. (2015). Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)". *Prace Filologiczne*, LXVII, 11-20.

Bronikowska, R. (2015). Możliwości przeszukiwania korpusu barokowego – cele i założenia. *Prace Filologiczne*, LXVII, 45-56.

Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M., Woliński, M. (2016). The use of electronic historical dictionary data in corpus design. *Studies in Polish Linguistics*, vol. 11, issue 2, 47-56.

Gruszczyński, W., Adamiec, D., Ogrodniczuk, M. (2013). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.). *Polonica* XXXIII, 311–318.

Kieraś, W., Komosińska, D., Modrzejewski, E., Woliński, M. (2017). Morphosyntactic Annotation of Historical Texts. The Making of the Baroque Corpus of Polish. In Ekštein, K., Matoušek, V. (Eds.), *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31, 2017, Proceedings.* Springer International Publishing, 308-316.

Przepiórkowski, A., Bańko, M., Górsk, R. L., Lewandowska-Tomaszczyk, B. (Eds.) (2012). *Narodowy Korpus Języka Polskiego.* Warszawa: Wydawnictwo Naukowe PWN.

National Corpus of Polish, http://nkjp.pl.

The Electronic Dictionary of the 17th-18th c. Polish, ed. W. Gruszczyński, 2004-, http://sxvii.pl.

Anastasiia Baranchikova
National Research University Higher School of Economics
six3.danika@gmail.com

Anna Dmitrieva
National Research University Higher School of Economics
black-letter@yandex.ru

Mariia Fedorova
National Research University Higher School of Economics
maria.fjodorowa@gmail.com

Aleksandr Klimov
National Research University Higher School of Economics
aleksklimow@gmail.com

Olesya Kisselev
The Pennsylvania State University
olesyakisselev@gmail.com

Mikhail Kopotev
University of Helsinki
mihail.kopotev@helsinki.fi

Svetlana Toldova
National Research University Higher School of Economics
toldova@yandex.ru

Natalia Zevakhina
National Research University Higher School of Economics
natalia.zevakhina@gmail.com

CAT&kittens: a corpus-based text-analytic tool for Russian academic writing

Corpus linguistics has contributed significantly to the study of academic discourse in the past two decades, with studies ranging from descriptions of specific grammatical features (Swales, 1990; Hyland, 1994) to general investigations of linguistic patterns, syntactic or lexical (Biber et al. 2004; Durrant & Mathews-Aydinli, 2011; Gray & Biber, 2013), to the development of specific academic vocabulary lists and academic phrase lists (Simpson-Vlach & Ellis, 2010; Ackerman & Chen, 2013). Similar studies for the Russian academic genre, however, have been lacking. The project described in this proposal intends to fulfil this gap.

The paper describes the development of a representative Russian Corpus of Academic Texts (CAT) outfitted with a built-in data processing tool, which allows for evaluation of texts written by novice writers of Academic Russian, both native and non-native, along a set of criteria in relation to the CAT corpus. Consequently, the goal of this paper is twofold: a) to describe the Corpus, and b) to discuss the criteria, upon which a novice text can be evaluated against the Corpus.

The project is currently being developed by a team of researchers from the Higher School of Economics (HSE) in Moscow, the University of Helsinki, and the Pennsylvania State University. The development of the CAT corpus follows established corpus development procedures (e.g., BAWE). It was collected by extracting recently published texts sourced from textbooks, academic journals, and collecting high-quality master's theses from available sources. All texts entered in CAT are divided into six disciplinary fields: social studies and history, political science and international relations, law, general and applied linguistics, economics, psychology and education science. Every discipline sub-corpus consists of about 300 to 400 thousand tokens, amounting to appr. 2 million tokens in the corpus in general. CAT is supplied with metalinguistic information, as well as morphological and syntactic annotation, carried out with the help of the annotation software RU Syntax (Mediankin et al. 2016). Further corpus improvement is also planned.

Since the main goal of the project is to create a tool that compares novice texts to standard academic texts along the lists of pre-set criteria, the tool will run a series of "error analysis" test. The patterns of deviations are identified along lexical, collocational, morphological, and syntactic planes. Their full list is still under discussion, therefore we present a preliminary set,

1. The general observation of an analyzed novice text includes text readability test, average sentence length, and TTR — all as compared to the CAT.
2. Lexical analysis includes identifying recurring tokens/lemmas in the student texts and comparing their frequencies to the frequency lists based on the CAT corpus. This analysis, based on low-frequency items

and hapax legomena, identifies overuse/underuse of specific vocabulary, highlights terminology that are unattested in the discipline, and suggests alternatives.

3. Collocational analysis. Based on n-gram frequencies, a specific type of errors, namely, non-standard word choice selection, will also be identified, and more standard collocational alternatives will be provided. This part consists of two steps: first we extract domain-specific collocations using standard measures (LL, (p)MI, t-score, etc.). Second, we determine non-standard collocations in a student text and suggest an alternative, based on more regular collocations and on distributionally close alternatives calculated with reference to the word2vec model trained on the semantically similar data.
4. Grammar check. Having morphological and syntactic annotations both in the CAT and in a student text under examination, checking morphological and syntactic errors is a two-step task. Unlike available spell-checkers, our tool is focused on detecting deviations that feature in academic writing— specifically those written by non-native speakers, e.g. genitive chains and ProDrop.

The results of these multidimensional analyses are provided in two ways: the general information about the whole text and highlighted fragments supplied with recommendations for correction. Although the robustness of the proposed analysis and the implementation of the tool require extensive testing, our project and lessons learnt from its development have implications for methodology of corpus linguistics already at this stage. Being a well-developed, deeply annotated representative corpus of Russian academic texts for the fields of Humanities and Social Studies, the CAT provides language researchers studying academic genres with an indispensible data set. Furthermore, the tool will, upon completion, be a useful to Russian teachers and students, who are seeking to improve their writing skills in this specific register.

References

Ackerman, K., & Chen, Y-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes,* 12(4), 235-247.

BAWE (The British Academic Written English), available at: http://www2.warwick.ac.uk/fac/soc/al/research/collections/bawe/how_to_cite_bawe.Last retrieved Feb, 15, 2018.

Biber, D., Conrad., & Cortes, V. (2004). If you look at …: Lexical bundles in University teaching and textbooks. *Applied Linguistics*, 25(3), 371-405.

Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1), 58-72.

Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13, 239-56

Gray, B., & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, 18(1), 109-136.

Mediankin N., & Droganova K. (2016). Building NLP Pipeline for Russian with a Handful of Linguistic Knowledge. In: Proceedings of the Workshop on Computational Linguistics and Language Science, Copyright © CEUR-WS, Aachen, Germany, ISSN 1613-0073, pp. 48-56.

Simpson-Vlach, R., and Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4), 487-512.

Swales, J.M. (1990). Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press

••••

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
vladimir.benko@juls.savba.sk

Radovan Garabík
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
radovan.garabik@juls.savba.sk

Ensemble Tagging Slovak Web Data

In Computational Linguistics, the "ensemble" term is used to describe approaches where several tools are utilized to (independently) perform the same operation, assuming that aggregation of their outputs could improve the overall success rate of the whole process. In the framework of morphosyntactic annotation, we can speak about "ensemble tagging" if more than one tagger is available for a particular language – which typically applies to many "large" languages.

Slovak also belongs to languages with several tools for morphosyntactic annotation available, with two of them being currently available. They have been developed (partially in parallel) within the framework of two Slovak projects – the Slovak National Corpus (Šimková and Garabík, 2014), and Aranea (Benko, 2014), respectively. Despite the fact that both have been based on the same source data – the *Morphological database of the Slovak language*[1] and *Manually morphologically annotated corpus*[2], and also use the same tagset[3] (Garabík and Šimková, 2012), they do not produce the same output. As they are only two, the straightforward method of "voting" is not applicable for aggregation.Due to the nature of web data, tagging presents an additional challenge and any possibility of improving the process is welcome.

The *Slovak National Corpus* (*SNC*) system is based on *MorphoDiTa*[4] (Straková et al. 2014; Spoustová et al. 2009) – an open-source tool for morphological analysis of natural language texts. While mostly language-agnostic, it has been developed with Czech language tagging in mind, and the application for Slovak has been straightforward.

1 http://korpus.sk/morphology_database.html
2 http://korpus.sk/ver_r(2d)mak.html
3 http://korpus.sk/morpho_en.html
4 http://ufal.mff.cuni.cz/morphodita

Slovak version has been trained on a manually annotated corpus (1.2 M tokens) and the tagging process uses separate tokenization and an additional guesser for Out of Vocabulary (OOV) words.

The guesser is based on a simple suffix based word similarity, where the list of tentative lemmas is derived by considering the suffix alteration of known lemmas during inflection process reversing the process for an unknown word form.

Within the framework of the *Aranea* Project (Benko, 2014), *TreeTagger*[5] (Schmid 1994) with a custom model is used for annotation. *TreeTagger* is a "monolithic" tool performing the analysis, disambiguation and guessing the tags for the OOV items by the same program. No functionality is provided for guessing lemmas for OOV word forms, and those lemmas are just copies of the respective word forms.

Two Slovak language models are available for *TreeTagger* – besides "standard" model, an "ASCII-only" model expecting input without diacritics. Both models assign full-diacs lemmas.

Our experiment involved tagging a web corpus of approx. 3 Gigawords by the systems mentioned, with the resulting annotations merged into a single vertical file. From this file, a frequency list was produced, containing lemma, tag and OOV info, that is being subject of further investigation. In the first stage of our work, we decided to concentrate on words either recognized, or being OOV in both tools.

We expect to provide the summary data in the final version of our presentation. At this point, however, we can say the *SNC* tool assigned lemma correctly in most cases, while *TreeTagger* is more reliable in assigning the tag (or at least the word class). By combining both results with the frequency information, a system for (unsupervised) lexicon update can be designed that could enlarge its size be many (hundreds of) thousands of entries.

Acknowledgment

This work has been supported by the Slovak VEGA and KEGA Grant Agencies, Projects *2/0017/17*, and *K-16-022-00*, respectively, as well as by the MŠVVŠ SR, MK SR, and SAV in the *Construction and development of the Slovak National Corpus* Project.

5 http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

References

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): T*ext, Speech and Dialogue. 17th International Conference, TSD 2014. LNCS 8655.* Springer International Publishing Switzerland, 2014.

Garabík, R. – Šimková,M. (2012). Slovak Morphosyntactic Tagset. In: Journal of Language Modeling. Institute of Computer Science PAS, 2012, Vol. 0, No. 1.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester. 1994.

Šimková, M. – Garabík, R. (2014). Slovenský národný korpus (2002 – 2012): východiská, ciele a výsledky pre výskum a prax. In: Jazykovedné štúdie XXXI. Rozvoj jazykových technológií a zdrojov na Slovensku a vo svete (10 rokov Slovenského národného korpusu). Eds. Katarína Gajdošová – Adriána Žáková. Bratislava: VEDA 2014.

Spoustová, D. – Hajič, J. – Raab, J. – Spousta, M. (2009). Semi-Supervised Training for the Averaged Perceptron POS Tagger. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), Athens, Greece, March. Association for Computational Linguistics.

Straková, J. – Straka, M. – Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, June 2014.

••••

Neil Bermel
University of Sheffield
n.bermel@sheffield.ac.uk

Luděk Knittl
University of Sheffield
l.knittl@sheffield.ac.uk

The fate of variant forms in historical corpora: Tracing locative exponents in DIAKON

Slavonic languages have at their disposal a wide variety of morphological material that has persisted through far-reaching reorganisations of their inflectional systems over time. In some cases, a significant degree of variation may remain in the forms speakers use, in the sense of Thornton's (2012) overabundance. This pattern is well-attested in Slavic languages, with their overabundance of morphological material (e.g. Janda 1996 for an overview; Lečić 2014 on Croatian; Brown 2007 on Russian). Our previous research on contemporary Czech shows that current usage is still varied and 'residual' forms that are rarely used still enjoy a reasonable degree of acceptability for native speakers.

We have been investigating a morphosyntactic change in progress in Czech, whose progress we should be able to track easily against the overall statistics regarding its use: the replacement of locative {ě} in the masculine hard inanimate paradigm with {u} (Bermel & Knittl 2012a, 2012b, 2015, 2017). This is one of a number of such processes that are moving through the Czech declensional system, but despite nearly a millennium of coexisting variant forms, do not seem to have proceeded through to completion.

Using the 50 most frequent nouns in Czech that display loc. sg. variation in this paradigm according to SYN2005, we interrogated the DIAKON corpus of historical and synchronic Czech (Kučera, Řehořková & Stluka 2015). The goal was to look at the development of these nouns diachronically, to see to what extent the patterns of change described in the literature on Czech and the general literature on diachronic change could be ascertained. The corpus presents an unparalleled opportunity to explore developments in the history of Czech, especially as regards data from the nineteenth and early twentieth centuries.

Data was extracted using form-by-form searches, manually cleansed to remove errors and non-locative forms, and divided into diachronic "cells". To approximate the sort of easy visual apprehension offered by interfaces like SyD, we measured percentages of each ending in each "cell" against each other, and produced graphs showing the direction of evolution of both endings for each lexeme in our sample. Our results show that:

1. Only high-frequency lexemes are represented often enough in historical corpora to allow us to draw conclusions about their development;
2. For these lexemes, a clear direction of travel away from residual forms can be established, although it does not seem to match the sort of S-curve development frequently proposed for changes in English ('slow, slow, quick, quick, slow' – see inter alia Denison 2003, Croft 2000, Blythe & Croft 2012);
3. Change in these lexemes has proved highly variable, with a significant number of words showing no change, reversible change, or change in the "opposite" direction of travel from that known to hold more generally;
4. Other potential explanations – such as the possibility that our data are non-representative because of phonological, word-formational, etymological, semantic or syntactic/constructional features, were explored thoroughly and do not appear to account for the anomalous developments.

Our most prominent finding has been that the words most commonly used in these slots frequently develop "against the grain", showing a pattern in which the recessive ending strengthens over time. Therefore, even when the number of lexemes employing these competing forms is contracting, the high and sometimes increasing usage levels with higher-frequency lexemes seems to contribute to maintaining the recessive endings as viable options.

References

Bermel, N., & L. Knittl. (2012a). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory*, 8, 241–275.

Bermel, N., & L. Knittl. (2012b). Morphosyntactic variation and syntactic environments in Czech nominal declension: Corpus frequency and native-speaker judgments. *Russian Linguistics*, 36, 91–119.

Bermel, N., L. Knittl & J. Russell. (2015). Morphological variation and sensitivity to frequency of forms among native speakers of Czech. *Russian Linguistics*, 39, 283–308.

Bermel, N., L. Knittl & J. Russell. (2017). Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguistics and Linguistic Theory*. doi: 10.1515/cllt-2016-0032.

Blythe, R.A. & W. Croft. (2012). S curves and the mechanisms of propagation in language change. *Language*, 88, 269-304.

Brown, D. (2007). Peripheral functions and overdifferentiation: The Russian second locative. *Russian Linguistics,* 31, 61–76.

Croft, W. (2000). *Explaining language change: An evolutionary approach.* Harlow: Longman

Kučera, K., Řehořková, A., Stluka, M. (2015). *DIAKORP: Diachronní korpus, verze 6 z 18. 12. 2015.* Ústav Českého národního korpusu FF UK, Praha 2015. Web access: http://www.korpus.cz

Denison, D. (2003). Log(ist)ic and simplistic S-curves. In: Hickey, R. (Ed.) *Motives for language change.* Cambridge: Cambridge University Press, 54-70.

Janda, L. (1996). *Back from the brink: A study of how relic forms in language serve as source material for analogical extension.* Munich and Newcastle: Lincom Europa.

Lečić, D. (2015). Inflectional doublets in Croatian: The case of the Instrumental singular. *Russian Linguistics*, 39, 375–393.

Thornton, A. (2012). Reduction and maintenance of overabundance: A case study on Italian verb paradigms. *Word Structure*, 5, 183–207.

••••

Martina Berrocal
Friedrich Schiller University, Jena
martina.berrocal@gmail.com

A corpus-assisted study of the Presidential interviews of Milos Zeman

The Czech Republic has got some unwelcome attention in international politics and media due to the opinionated and often controversial statements by its current President Miloš Zeman. Lately, the Czech society has been emotionally stirred by the Presidential election which is said to have shown a deep division among the Czechs. The personification of this divide is President Miloš Zeman who has been elected for his second mandate early this year and is known for his critical stance towards the elites and his political adversaries. In both media and social media, he is often described as a person who is rude, especially to groups of people, such as the journalists and intellectuals, as a person who ignores the Constitution and surrounds himself with people who are involved in suspicious networks and deals. Despite the above, he is mostly portrayed as an excellent and well-read speaker.

This study is a part of a larger project which aims to account for the prominent linguistic and discursive phenomena (e.g. topic orientation, discourse structure of populist text and talk, creating of antagonistic dichotomies, language aggression, representing danger and crisis, communication attitudes…) in the Czech populist discourse.

This paper specifically aims to examine Milos Zeman's discourse in the context of 19 weekly interviews "Týden s Prezidenten" (A week with the President) broadcasted on the private commercial TV Barandov and Zeman's opinion platform (total number of 112 971 tokens). In this program, Prezident Zeman enjoy privileged visibility in an uncontested environment where he is able to voice unopposed his views and opinions. The data are uploaded in Sketch-engine and morphologically annotated. This is done with the aim to broaden the scope and depth of the quantitative analysis and to make the data easily manageable.

Methodologically, the study draws on the knowledge and procedures of CADS (Baker 2006, Partington, Duguid, and Taylor 2013, Partington 2013). To start, the analysis of prominent keywords (KWs) is carried out. The list of KWs is generated against the reference corpus SYN 2015 in the application KWords of the Czech National Corpus. The advantage of this application is that it applies common statistical measures (chi2, LL) and for the units for which a statistical significance is determined, a difference index (DIN) is computed which when achieving values 75-100 is indicative of prominence in a particular text (Cvrček/Richterová 2016). In our study, a KW analysis is conducted for each interview and the DIN results are compared. This way, it is possible to get a more comprehensive picture of the keyword distribution than when analysing the data in a single data bulk. These initial results are further examined in a detailed collocation analysis. A special attention is dedicated to a personal pronoun *já*, verbs *vědět* (*know*), *myslet* (*think*), *opakovat* (*repeat*), *říkat* (*say*) a *znamenat* (*mean*) as means of positioning, authority exertion and aggression. In line with that, the presidential Self and the possible antagonistic *Other(s)* are determined.

Keywords: CADS, keyword analysis, Czech political discourse, TV interview

References

Baker, Paul. (2006*). Using Corpora in Discourse Analysis*. London: Continuum.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocations networtks. *International Journal of Corpus Linguistics*, 20(2), 139-173.

Cvrček, Václav and Olga Richterová (eds). (2016) "manualy: kwords[Internet]. Příručka ČNK, September 9, 2016 <http://wiki.korpus.cz/doku.php?id=manualy:kwords&rev=1473408294> [Retrieved February 15, 2018]

Partington, Alan. (2013). Corpus Analysis of Political Language. In Carol Chapelle. (ed.) *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell Publishing.

Partington, Alan, Alison Duguid, and Charlotte Taylor. (2013). Patterns and Meanings in Discourse. Theory and Practice in corpus-assisted discourse studies (CADS). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Katja Brankačkec
Slovanský ústav AV ČR v.v.i.
brankatschk@slu.cas.cz

Productivity and Meaning of the Prefix nad- in the Word-Formation of Upper Sorbian, Lower Sorbian and Czech in a Diachronic Perspective: Evidence from Corpora

The prefix *nad-* seems to be rather periphery in Slavic word-formation system (Šlosar 1981, 70), as in most Slavic languages, there are relatively numerous prefixes concurring with it in its different meanings (cf. Oertle 2016, 177ff.). Among others, in many Slavic languages it competes with the prefix *na-*, which is natural, as *nad-* is probably a derivate of *na-* in Slavic (ibid. 177). Moreover, its usage in Western Slavic languages often seems to be motivated by language contact with Latin *super-* (Martínek et al., 2017) and German *über-* (e.g. Oertle 2016, 177; 179). On the other hand, obviously only a part of the derivates with this Latin and German prefix is generally transferred with the help of *nad-* into Western Slavic (Martínek et al., 2017): Derivates with the meaning 'surfeit' are transferred with the prefix *pře-/pše-*.

In our investigation we want to compare its role in three Western Slavic languages (US: Upper Sorbian, LS: Lower Sorbian and Cz: Czech) from a diachronic point of view. On the basis of the diachronic corpora Hotko, Dotko and Diakorp we want to address possible meanings of the prefix in the three languages, their frequency in the lexicon of the three languages as well as the usage of particular derivates with *nad-* in the course of time. In doing so, we will report on the possibilities and problems with an analysis on the basis of these corpora and compare our results with dictionaries of these languages.

While we take into account possible influence from Latin, German and other contact languages on the derivation with *nad-*, we are trying to describe the establishment of complex words with this prefix in the lexicon of these languages. While derivates with the concrete local meaning of *nad-* show a quite stable meaning in the corpora, the derivates with a more abstract meaning are more variable. Besides numerous loan translations with *pře-/pše-* with the meaning ‚surfeit'[1] there is a number of doublets with the prefix *na-* in US and LS. Especially confusing is the usage of the verbs with the meanings 'to have an idea' and 'to attack' in both US and LS: Both meanings occur with derivates with both *nad-* and *na-* parallelly. While the conversion *nadpad/napad* is used in both meanings as well in both languages, the situation with further derivates differs. In contemporary Czech, on the other hand, both meanings are expressed with the lexem *napadnout/-at*, resp. the deverbal noun does not differ in the prefix used: *nápad* 'idea', *napadení* 'attack'. That means, in contrast to contemporary German, where the meanings are represented in separate lexems, i. e. *einfallen* 'to have an idea'[2]; *auffallen* 'to stand out, to attract attention'[3] on the one hand and *überfallen*, *angreifen* 'to attack' for *nadpadnyć* on the other hand, there is only one polysemous verb in Czech[4]. Parallely, in both Old Czech and Sorbian, there is a partly synonymous derivate cz. *nadběhnúti*/us. *nadběhować* 'to attack', which is evident already in Old Czech texts. In our paper, we are going to concentrate on this semantic field of 'attack' and its realizations in the material investigated.

References

Martínek et al. (2017). F. Martínek, K. Brankačkec: Lehnübersetzungen lateinischer Wörter mit dem Prafix super- ins Alttschechische. In I. Janyšková, H. Karlíková & V. Boček (Eds.), *Studia etymologica Brunensia* 22, Praha: Lidové noviny, 59–73.

Oertle, S. (2016). *Die slavischen Verbalpräfixe und Präpositionen: Polysemie und Grammatikalisierung*. Herne: Gabriele Schäfer Verlag.

Pohončowa et al. (2009). A. Pohončowa, J. Šołćina & S. Wölkowa: *Z labyrinta serbšćiny. Bjesady wo rěči*. Bautzen: Domowina-Verlag.

Šlosar, D. (1981). *Slovotvorný vývoj českého slovesa*. Brno: Univerzita J. E. Purkyně v Brně.

1 The meaning of 'measuring' for the prefix *pře-* is documented already for Old Church Slavonic (Šlosar 1981, 81). Therefore, it is not surprising that the prefix *nad-* did not develop such a meaning in Sorbian and Czech, as it is the case for Latin *super-* and German *über-* (similarly to English *over-*).

2 Note that, in older states of German, this verb was also used in the meaning 'to attack the territory of another state' more often than in contemporary German.

3 The prefix *na-* is often used to transfer German *auf-* in calques.

4 Compare also Pohončowa et al. 2009, 117f., who is discussing these two verbs in a popular linguistic guidebook.

Corpora and Web-sources:

Vokabulář webový: Vokabulář webový: webové hnízdo pramenů k poznání historické češtiny [online]. Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka. © 2006–2018. Verze dat 1.1.4 [cit. 1. 3. 2018]. Dostupné z: http://vokabular.ujc.cas.cz

Diakorp: Kučera, K. – Řehořková, A. – Stluka, M.: *DIAKORP: Diachronní korpus, verze 6 (18. 12. 2015).* (Diachronic corpus, version 6) [on-line]: http://www.korpus.cz Praha 2015.

Dotko: Sorbisches Institut, Abteilung Niedersorbisch Cottbus (Sorbian institute, Department of Lower Sorbian, Cottbus): *DOTKO: dolnołužyski textowy korpus, werzija 1 (20. 12. 2010) (*Lower Sorbian text corpus, version 1) [on-line]*:* http://www.korpus.cz, Praha 2010.

Hotko: Sorbisches Institut Bautzen (Sorbian institute Bautzen): *HOTKO: hornjołužiski textowy korpus, werzija 1 (6. 3. 2013).* (Upper Sorbian text corpus, version 1) [on-line]: http://www.korpus.cz, http://www.serbski-institut.de/cms/os/48/hornjoserbski, Praha 2013

••••

Kat Dziwirek
University of Washington, United States
dziwirek@uw.edu

To taste is to live and love: Verbs of taste in Polish and English

This paper is part of a larger project where I examine the lesser verbs of sensory perception in English and Polish. The gist of my argument is that despite claims to the contrary metaphorical extensions of smell, touch and taste cannot be considered universal, but must be filtered through lexical semantics and cultural values. Thus, while *smell* is inherently negative in English, that is to say *It smells* is to say that something smells bad, the exact opposite is true in Polish. The verbs of touch provide another such contrast, based in lexical semantics. Metaphorical extensions of taste, on the other hand, are quite similar in the two languages and the main differences lie in frequencies of certain constructions due do different cultural discoursive scripts. The paper uses data from the Corpus of Contemporary American English (COCA) and Narodowy Korpus Języka Polskiego (NKJP).[1]

The main metaphorical extension of the nouns *taste and smak* is liking, preference, partiality. They instantiate the metaphor Liking is tasting. The Polish and English expressions are given below.

Polish expression	**meaning**	**English expression**	**meaning**
do smaku	to someone's liking	to someone's taste	to someone's liking
w smak	to someone's liking		
nabrać smaku do	to begin to like	to develop a taste for	to begin to like
		to have a taste for	to like
obejść się smakiem	to have to do without		
dobry/zły smak	sense of decorum	good/bad taste	sense of decorum

1 The balanced version of NKJP consists of 300 million words, while the entire Polish corpus has 1,500 million words, compared to COCA's 560 million. Comparisons are made in words per million. COCA is a balanced corpus equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. NKJP contains texts from classic literature, daily newspapers, specialist periodical and journals, internet texts and transcripts of conversations. The conversations represent both male and female speakers of various ages and from various regions.

The agentive verbs *taste*, *posmakować* and *zasmakować* metaphorically mean to experience. The main differences are the complements: English uses adjectives to describe taste, while Polish speakers most often talk about something having a certain taste. Also, as with other perception verbs, English uses *like* clauses to describe the quality of the perception much more frequently than Polish. For example, *smell like* is used over 10 times more frequently than *pachnieć jak* (a rate of 0.54/0.36 per million words in NKJP versus 5.8 times per million words in COCA). I argue that this follows form the cultural discoursive script proposed here using Wierzbicka's (1999, 2008) Natural Semantic Metalanguage.

American English:
People think:
I have to say something about X
It is good if I say that X is like something else

Finally, looking at corpus data allows us to see language change in progress. The agentive verbs *skosztować, spróbować, posmakować* and *zasmakować*. occur with the object in the genitive case, as their meaning is partitive (to eat or drink a little of something), but are occasionally found with the accusative in NKJP (e.g. 40 out of 590 tokens of *skosztować*), though only in the literal sense of tasting food or drink. This looks like the beginning of a shift to the accusative, which also affects other genitive governing Polish verbs. Based on the corpus data we can propose that case shifting language change starts with literal meanings, while the metaphorical extensions hold on to the original case.

References

Wierzbicka, A. (1999). *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge University Press.

Wierzbicka, A. (2006). *English: Meaning and Culture*. Oxford University Press.

••••

Hanne Martine Eckhoff
University of Oxford
hanne.eckhoff@mod-langs.ox.ac.uk

Aleksandrs Berdičevskis
Uppsala University
aleksandrs.berdicevskis@lingfil.uu.se

Marius Jøhndal
University of Oslo
marius.johndal@ifikk.uio.no

From diachronic treebank to dictionary resource: the Varangian Rus project

In this paper we present the Varangian Rus' dictionary resource, which is based on the Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no, torottreebank.github.io), a diachronic treebank of Russian containing a balanced selection of 11th–17th century Old East Slavic and Middle Russian texts. The treebank is lemmatised and has detailed morphological and syntactic annotation. With simple glossing of the word meanings found in the treebank, we are able to generate a dictionary resource with rich grammatical information. The dictionary resource is released as a part of the Syntacticus treebank browsing interface (http://syntacticus.org/). The Syntacticus interface generates dictionaries for every language represented in the treebank collection, but glosses have not been systematically added to the other dictionaries, for instance the Old Church Slavonic one.

TOROT[1] is the only existing treebank of Old East Slavic and Middle Russian texts. There are other tagged resources, such as the Old Russian subcorpus of the Russian National Corpus and the Manuskript corpus, but none of them currently provide syntactic annotation. The TOROT presently contains approximately 250,000 word tokens of fully lemmatised and morphosyntactically annotated 11th–14th century Old East Slavic and 15th–17th-century Middle Russian. The TOROT is a part of a larger family of treebanks of ancient

1 The TOROT was developed as part of the project "Birds and Beasts: Shaping Events in Old Russian" at UiT The Arctic University of Norway. The dictionary resource is a result of its companion project "The Varangian Rus Digital Environment", a cooperation between UiT and two Russian partners: The Higher School of Economics and Moscow State University.

languages, originating in the PROIEL project. The PROIEL project developed an enriched dependency grammar scheme, a scheme for detailed morphological tagging and various other annotation schemes, such as ones for information structure and semantics. Accompanying web annotation, browsing and query tools were also created, now gathered under the Syntacticus umbrella. The PROIEL tools and schemes are developed for and by linguists, and yield data with rich morphological and syntactic information, as well as lemmatisation. The TOROT thus contains a lot of lexicographically interesting information, which we have put to use in the dictionary resource.

The dictionary resource is not a full-fledged dictionary, and provides simple glosses rather than structured and ranked definitions. Glosses are given in English and Russian, and aim to cover all and only the meanings attested in a subset of the TOROT treebank. We currently have approximately 8000 glossed lemmas. In order to make sure that only actually occurring meanings are included in the gloss, the glossers went through all occurrences of every lemma. For particularly high-frequency lemmas, we dispensed with the time-consuming manual process, and semi-automatically selected glosses by aligning the texts with automatically lemmatised modern Russian translations.

	Singular	Dual	Plural
Nom.	варѧгъ (6)	варѧга (1)	варѧзи (11)
Acc.		варѭга (1) варѧга (1)	варѧгы (8) варѧги (7) вариги (1) варѧгы (1)
Gen.			варѧгъ (5)
Loc.			
Dat.			варѧгомъ (4)
Ins.			варѧгы (2) варѧги (1)

Figure 1. Attested paradigm in Syntacticus for *varjagъ* ‘Varangian’

In addition to the mostly manual glosses, the dictionary entries contain generated information based on the treebank data. In addition to a full concordance of all sentences attesting the lemma, we use the detailed morphological annotation to generate paradigms insofar as they are attested for the lemma in question.

The syntactic annotation can also be mined for lexicographically interesting information. In particular, we are able to provide rich valency information for verbs, where we can list and give frequencies of all attested argument structure frames.

Arguments	Non-reflexive	Reflexive
(none)	6	
OBJ (accusative)	9	
OBJ (genitive)	3	
OBJ	1	
OBJ (accusative) OBL (dative)	8	
OBJ (accusative) OBL (participle)	1	
OBJ (genitive) OBL (dative)	3	
OBJ (accusative) OBL (dative) OBL (preposition *за* + genitive)	1	
OBJ (accusative) OBL (preposition *за* + accusative)	1	
OBJ (interrogative adverb *колько*)	1	
OBJ (interrogative adverb *сколько*) OBL (adverb *куды*)	1	
OBL (dative)	7	2
OBL (genitive)	1	
OBL (participle)	1	
OBL (preposition *въ* + accusative)	1	
OBL (preposition *за* + accusative)	2	

Figure 2. Attested argument frames with *otъdati* ‘give away’ in Syntacticus.

The dictionary resource is diachronic in nature, which makes it necessary to give indications of diachronic variation in the use of words. Since the TOROT also contains metadata about the text sources, we can give precise indications of a lemma’s distribution across sources and time periods.

The generated dictionary entries also serve as a useful error detection tool for the treebank, since the systematic representation easily shows up many types of annotation errors, such as morphological misclassification or faulty lemmatisation. The work on the dictionary resource is therefore directly beneficial for the treebank.

The dictionary resource complements the existing Old and Middle Russian dictionaries, and will be useful for students and scholars alike.

••••

Tomaž Erjavec
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Nikola Ljubešić
Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

Darja Fišer
Faculty of Arts, University of Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Training data and tools for processing user-generated content in Slovene, Croatian and Serbian

1 Introduction

The language found on social media, such as tweets, forums, blogs, etc. collectively known as user-generated content (UGC) differs from standard language in many respects, such as non-standard word spellings, frequent use of colloquial expressions, phonetic orthography, as well as omissions of diacritics, as the texts are often written on smartphones, where ASCII letters are quicker to type.

There are two ways of adapting this situation. One is to first standardize the words in UGC, and then use tools trained on standard language for further annotation, and the other is to train the tools with additional, UGC domain data.

In this abstract we present datasets and tools that enable either of these strategies to be used in order to improve automatic annotation of UGC text for three South Slavic languages, namely Slovene, Croatian and Serbian.

2 Annotation workflow

In case of all the languages involved, the annotation proceeded in a similar fashion.

The Slovene datasets were produced first, mostly in the scope of the Janes project (http://nl. ijs.si/janes/). In the first stage annotation guidelines were written and the student annotators were trained on preliminary test data. Then the data was sampled from a large CMC corpus, and imported to the WebAnno [Yimam et al., 2013] platform for manual annotation. The files were distributed among annotators so that each file was annotated by two annotators. Once finished, the disagreements in each file were checked by the curator, who chose the correct annotation. Finally, the annotated files were merged with their source TEI encoding, with specialized scripts developed for this purpose [Erjavec et al., 2016a].

In the second stage the Slovene annotation guidelines were taken on board by the ReLDI project (https://reldi.spur.uzh.ch/), translated to English and annotation campaigns similar to the ones for Slovene were performed for Croatian and Serbian UGC. This approach not only saved time and effort for the resource development, but also produces resources that are harmonized across the three languages.

3 The datasets

The first dataset, Janes-Norm [Erjavec et al., 2016b], contains about 150,000 words of Slovene UGC with correct(ed) tokenization, sentence segmentation and normalization of the words to standard Slovene and — on the basis of the standardized words — automatically assigned morphosyntactic descriptions (MSDs) and lemmas. Technically, one of the most difficult aspects of the annotation and encoding are cases where one non-standard word is mapped to several standardized ones or vice versa.

The second dataset, Janes-Tag is a subset of Janes-Norm, and contains about 55,000 words where the MSDs and lemmas were also manually corrected. Furthermore, the second version of Janes-Tag [Erjavec et al., 2017] was also annotated with named entities (NEs).

The equivalents of Janes-Tag for Croatian and Serbian UGC are ReLDI-NormTagNER-hr and [Ljubešić et al., 2017a] ReLDI-NormTagNER-sr [Ljubešić et al., 2017b], each with about 80,000 words and manually annotated for all six annotation layers, the same as Janes-Tag.

All the datasets are available under CC licenses for download from the CLARIN.SI repository, as well as for exploration and analysis via its online concordancers.

4 Annotation Tools

We have also produced state-of-the-art annotation tools to enable non-standard language processing for the three languages. The first tool is the ReLDI-tokeniser (and sentence segmenter), which is based on manually specified rules and makes use of language-specific resources files, such as lists of abbreviations. The special feature of this tokenizer is that it has two modes: one for the standard language, and one for non-standard language, which is different in two ways: the defined rules are here less strict and there are a few additional rules describing phenomena typical for online communication, such as emoticons.

The second tool, called CSMTiser [Ljubešić et al., 2016] performs word normalization and uses character-level statistical machine translation and was trained on Janes-Norm, ReLDI-NormTagNER- hr and ReLDI-NormTagNER-sr. The character-level accuracy of the normalization procedure on Slovene non-standard Twitter data is 98.5%, while the non-normalized data has a character-level accuracy of 94.8% [Ljubešić et al., 2016].

JANES-tagger [Ljubešić and Erjavec, 2016, Ljubešić et al., 2017] performs MSD tagging and lemmatisation, is based on conditional random fields (CRF), and is trained on standard-language datasets for three languages supplemented with Janes-Tag, ReLDI-NormTagNER-hr, and ReLDI- NormTagNER-sr respectively. The token-level accuracy on Slovene Twitter data before the tagger adaptation was 69% and after the adaptation 86% [Ljubešić et al., 2017]. A similar tagger achieves on standard data accuracy of 94% [Ljubešić and Erjavec, 2016].

Finally, JANES-NER performs NE annotation, is also CRF-based and was trained on the same three datasets as the JANES-tagger. The F1 of the JANES-NER system is 0.69, the "other" class having the lowest F1 of 0.30, followed by organizations with F1 of 0.56, locations having an F1 of 0.80, and the "person" class having the highest F1 of 0.92.

All the above-mentioned tools are available on GitHub, under the CLARIN.SI virtual organisation, at https://github.com/clarinsi.

References

[Erjavec et al., 2016a] Erjavec, T., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., and Zupan, K. (2016a). Annotating CLARIN.SI TEI corpora with WebAnno. In Proceedings of the 2017 CLARIN Annual Conference, pages 1–5. The Association for Computational Linguistics, ACL.

[Erjavec et al., 2016b] Erjavec, T., Fišer, D., Čibej, J., and Arhar Holdt, Š. (2016b). CMC training corpus Janes-Norm 1.2. Slovenian language resource repository CLARIN.SI. http://hdl. handle.net/11356/1084.

[Erjavec et al., 2017] Erjavec, T., Fišer, D., Čibej, J., Arhar Holdt, Š., Ljubešić, N., and Zupan, K. (2017). CMC training corpus Janes-Tag 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1123.

[Gimpel et al., 2011] Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers- Volume 2, pages 42–47. Association for Computational Linguistics.

[Ljubešić and Erjavec, 2016] Ljubešić, N. and Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).

[Ljubešić and Erjavec, 2016] Ljubešić, N. and Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: The case of Slovene. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA).

[Ljubešić et al., 2017] Ljubešić, N., Erjavec, T., and Fišer, D. (2017). Adapting a state-of-the-art tagger for south Slavic languages to non-standard text. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 60–68.

[Ljubešić et al., 2017a] Ljubešić, N., Erjavec, T., Miličević, M., and Samardžić, T. (2017a). Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1170.

[Ljubešić et al., 2017b] Ljubešić, N., Erjavec, T., Miličević, M., and Samardžić, T. (2017b). Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1171.

[Ljubešić et al., 2016] Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. (2016). Normalising Slovene data: historical texts vs. user-generated content. In Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), pages 146–155.

[Yimam et al., 2013] Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible,web-based and visually supported system for distributed annotations. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013), pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

....

Matea Filko
Faculty of Humanities and Social Sciences, Zagreb
matea.filko@ffzg.hr

Krešimir Šojat
Faculty of Humanities and Social Sciences, Zagreb
ksojat@ffzg.hr

Marko Tadić
Faculty of Humanities and Social Sciences, Zagreb
marko.tadic@ffzg.hr

Construction za + infinitive – evidence from the Croatian corpora

In Croatian, prepositions are usually defined as function words which denote relations among entities (cf. Silić & Pranjković 2005). They precede nouns or pronouns in a specific case thus forming prepositional phrases. Namely, the preposition za can be followed by noun or pronoun in genitive, accusative or instrumental case, expressing different range of meanings (temporal, locational etc.). However, evidence from Croatian corpora show that the preposition za is the only preposition that can also be followed by the verb in the infinitive form.

Although many Croatian linguists claim that this construction is ungrammatical in the Croatian language (cf. Rišner 2007), this construction has been used in Croatian for several centuries, most probably under the influence of Italian, German and Latin. In this paper, we analyse the construction *za + infinitive* on the basis of data from the two biggest corpora of Croatian language: Croatian National Corpus (Tadić 2009) and Croatian web corpus hrWaC (Ljubešić & Erjavec 2011). Our main goal is to show that this construction is used in contemporary Croatian and that it occurs in several specific syntactic structures.

In the first step we calculated the frequency of the *za + infinitive* construction in both corpora. The construction appears 3,337 times in the Croatian National Corpus (15.4. per million), a balanced corpus consisting of mainly proofread texts. On the other hand, it appears as many as 147,211 times (105.3 per million) in hrWaC, showing that Croatian speakers tend to use

this construction more frequently in the informal discourse. In the second step we extracted the sample of 1.500 sentences containing the construction *za + infinitive* from each of the analysed corpora in order to detect various syntactic structures in which this construction is used.

The preliminary analysis showed that the construction *za + infinitive* can be used:

– as an attribute:

imati što za prodati 'to have something for sale', *imati što za reći* 'to have something to say', *teški uvjeti za igrati* 'difficult conditions to play', *nešto za raditi* 'something to do'…

It is important to stress that the construction *za + infinitive* can be replaced by the bare infinitive only when it complements relative/indefinite pronouns (*imati što za reći* > *imati što reći*). When it complements a noun, it can only be replaced by the construction *za* + N_{ACC} (*teški uvjeti za igrati* > *teški uvjeti za igru* (*igra* = 'play'); **teški uvjeti igrati*). The same applies when the pronoun is followed by an adjective: *imati što toplo za popiti* ('to have something warm to drink') > *imati što toplo za piće* (*piće* = 'drink'), **imati što toplo popiti*).

– as a verbal complement, where we can distinguish between:

- copula complement

bilo je za očekivati 'it could be expected, lit. it was to expect', *nije za zamjeriti* ‚not to blame smb.', *za krepat od smijeha* ‚laugh one's head off', *vijesti su nam za povratiti* ‚the news were disgusting'

Bare infinitives cannot be used as a nominal part of the predicate. In these cases, it can only be replaced by the different construction with modal verbs: *bilo je za očekivati* > *moglo se očekivati* ('it could be expected'), *nije za zamjeriti* > *ne može se zamjeriti* ('it cannot be blamed').

- part of a complex predicate

ostaje nam za pitati se 'we were left wondering, lit. it remains to us to wonder', *bilo je ugodno za gledati* 'it was pleasant to watch, lit. it was pleasant for watching'

In both cases, the construction can be replaced by the bare infinitive: *ostaje nam pitati se, bilo je ugodno gledati.* Note that in the second case the construction determines the meaning of the nominal part of the predicate, but in this syntactic structure it can't be replaced with the relative clause or with the *za* + N_{ACC} construction.

Although Croatian dictionaries and grammars don't recognize and discuss the *za + infinitive* construction, the evidence from the corpora clearly shows that this construction is frequently used in Croatian and that it occurs in different complex syntactic structures. Although in some cases it can be replaced by the bare infinitive, which is claimed to be a more economical way of expressing the same thing, in this paper we address a question why Croatian speakers nevertheless use this construction. This question requires a more in-depth analysis of all the detected examples. Preliminary results indicate its usage is enabled by the complex semantic structure of the preposition *za* (e.g. purpose, mood, quantity), which in turn contributes to the meaning of the whole *za + infinitive* construction. If a thorough analysis based on the corpus data confirms the difference in meaning between the construction with *za* and the bare infinitive, it could be used as an evidence for the grammaticality of this construction and a basis for a more precise syntactic description of Croatian.

References

Ljubešić, N., Erjavec, T. (2011). hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. In I. Habernal & V. Matousek (Eds.). *Text, Speech and Dialogue 2011. Lecture Notes in Computer Science.* Berlin; Heidelberg: Springer, 395–402.

Rišner, V. (2007). Jezični savjeti Ljudevita Jonkea i suvremena hrvatska norma. *Jezik*, 54, 94–104.

Silić, J., Pranjković, I. (2005). *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta.* Zagreb: Školska knjiga.

Tadić, M. (2009). New version of the Croatian National Corpus. In D. Hlaváčková, A. Horák, K. Osolsobě, P. Rychlý (Eds.). *After Half a Century of Slavonic Natural Language Processing*. Brno: Masaryk University, 199–205.

Olga Goritskaya
Minsk State Linguistic University
goritskaya@gmail.com

Mikita Suprunchuk
Minsk State Linguistic University
suprunchuk@gmail.com

Frequency Dictionary of Belarusian Borrowings in the Belarusian Variety of the Russian Language

There are two official languages in Belarus – Belarusian and Russian. The Russian language functions primarily as a means of communication, while the Belarusian language to a greater extent plays a symbolic role – it represents national identity. According to the last census (2009), 53% of the population consider Belarusian their native language, but only 23% usually speak Belarusian at home. In contrast, only 37% called Russian their native language, but 70% use Russian in their daily lives. The contact of Russian and Belarusian is inevitable in this situation, and it results in Russian-Belarusian mixed speech, code-switching, borrowings and interference in Belarusian and Russian texts, as well as the formation of the Belarusian national variety of the Russian language (Hentschel 2017, Mixnevič 1985, Norman 2008, etc.).

The project aims to create a frequency dictionary of Belarusian borrowings in Russian speech in the Republic of Belarus (preliminary edition). Traditional studies of lexical variation in Belarusian Russian were mostly based on introspection and small amount of data. No comprehensive corpus research of Belarusian borrowings in the Belarusian variety of Russian has been undertaken until now. Our project is designed to fill this gap.

The material for our dictionary was extracted from blogs written in Russian by the residents of Belarus. General Internet-Corpus of Russian (GICR, webcorpora.ru) is the main source of data (cf. Belikov 2013a, Belikov 2013b). GICR is the only Russian corpus, which contains various meta tags, including the information on the author's place of residence, and is big enough to study non-standard variants with low frequency. The corpus is under development now (for instance, it includes texts in Belarusian and other languages, and there is no exact data about the amount of words in texts from different states, cities and other localities), which necessitates the additional processing of search results and imposes limits on the choice of research methods.

Obviously, GICR does not reflect the speech of the whole Belarusian population. For example, the corpus lacks information on users' social characteristics, and the vast majority of the blog authors were born in the period of 1970–1990. Still, we may well suppose that due to the expansive volume, the variety and the unedited nature of its texts, this corpus does reflect general trends in the development of the Belarusian variety of the Russian language.

We extracted the data for the research from the subcorpus of "LiveJournal", which consists of 8.7 billion words. This subcorpus reflects modern Russian speech – most of the texts were created during 2006–2013. It includes not only traditional diary-type blog entries and comments, but also media texts, fiction, etc.

The dictionary is based on our collection of examples gathered while studying Internet communication and publications by other researchers on this topic. Several words and word combinations were taken from "The Language of Russian cities" dictionary (edited by V. Belikov), as well as the corresponding Internet forum (http://forum.lingvolive.com/cat/l26). We also got some data from other sources containing metalanguage reflection on the specificity of Russian in Belarus. Besides, we have studied dictionaries that reflect differences in the lexical systems of the Belarusian and Russian languages (Grabčikov 1980, Vojnič et al. 1985).

In our dictionary, we define absolute frequency of Belarusian borrowings and rank them according to their frequency. The next stage of work is to compare the frequency of Belarusian borrowings and synonymous lexemes in the standard Russian language (see Rieger 2014).

This frequency dictionary will contribute to the research of lexical variation in the Belarusian variety of Russian. Speakers use Belarusian borrowings both unconsciously (because of the influence of the Belarusian language or mixed speech) and deliberately (to express their emotions or attitude). While creating the dictionary we pay special attention to the inter-language paronyms and homonyms, because there is a great amount of contact phenomena concentrated in this zone and speakers are not conscious of a part of them. Besides, some Belarusian borrowings in Russian speech differ in their meaning and usage from the corresponding lexemes of the standard Belarusian and standard Russian languages. We suppose that this dictionary will be useful for studying semantics and pragmatics of the above-mentioned specific lexical units. More-

over, this project reveals which borrowings are key markers of the Belarusian variety of Russian and which Belarusian words are used only occasionally.

References

Belikov, V. et al. (2013a). Corpus as language: from scalability to register variation. In: V. Selegej (Ed.), *Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo: Moskva: RGGU, 12 (19)/1, 84-95.

Belikov, V. et al. (2013b). Big and diverse is beautiful: a large corpus of Russian to study linguistic variation. In: S. Evert, E. Stemle, P. Rayson (Eds.), *Web as Corpus Workshop (WAC-8)*, 24-28. URL: http://corpus.leeds.ac.uk/serge/publications/2013-wac8.pdf.

Davies, M. & Gardner, D. (2010). *A Frequency Dictionary of American English: Word Sketches, Collocates, and Thematic Lists*. London: N.Y.: Routledge.

Grabčikov S. M. (1980). *Mežjazykovye omonimy i paronimy: opyt russko-belorusskogo slovarja*. Minsk. Izdatelstvo BGU.

Hentschel, G. (2017). Eleven questions and answers about Belarusian-Russian Mixed Speech ('Trasjanka'). *Russian Linguistics*, 41/1, 17-42.

Ljaševskaja, O. N. & Šarov, S. A. (2009). *Častotnyj slovar´ sovremennogo russkogo jazyka (na materialax Nacyonal´nogo korpusa russkogo jazyka*). Moskva: Azbukovnik.

Mixnevič, A. E. (ed.) (1985). *Russkij jazyk v Belorussii*. Minsk: Nauka i texnika.

Norman, B. Ju. (2008). Russkij jazyk v Belarusi segodnja. *Die Welt der Slaven*, LIII/2, 289-300.

Rieger, J (2014). *Słownictwo polszczyzny gwarowej na Brasławszczyźnie: oparte głównie na nagraniach i zapisach A. Stelmaczonek-Bartnik, B. Jasinowicz, W. i E. Minksztymów, N. Ananiewej, kartoteki W. Werenicza.* Warszawa: Wydawnictwo Naukowe Sub Lupa.

Vojnič, Y. V. et al. (1985). *Belorussko-russkij paraleksičeskij slovar´-spravočnik*. Minsk: Narodnaja asveta.

••••

Natalia Grabar
CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille.fr

Olga Kanishcheva
Intelligent Computer Systems Dept, National Technical University Kharkiv Polytechnical Institute, Kharkiv, Ukraine
kanichshevaolga@gmail.com

Thierry Hamon
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France; Université Paris 13, Sorbonne Paris Cité, F-93430 Villetaneuse, France
hamon@limsi.fr

Multilingual aligned corpus with Ukrainian as the target language

1. Introduction

Creation of linguistic resources is an important issue for linguistics and NLP. Their availability provides the possibility to design, develop and evaluate methods and tools specific to several contexts and applications. The purpose of this work is to describe multilingual parallel and aligned corpus, in which the target language is Ukrainian, while the current source languages are Polish, French and English.

Ukrainian language is currently provided with little freely available resources. We are mostly interested here by the available corpora, including parallel corpora. Among the existing work, let's notice the national corpus of the Ukrainian language (Дарчук, 2010) available online (http://www.mova.info/corpus.aspx?l1=209) and literary corpus with the work by Ivan Franko (Бук, 2010) built for the research and educational purposes, and a corpus with dialectal texts (Сірук, 2012). Besides, several parallel corpora involving Ukrainian have been proposed, such as Polish-Ukrainian (Kotsyba, 2012) and Bulgarian-Ukrainian (Siruk et al., 2013) corpora.

2. Collection of texts

We use two kinds of texts. The first source is composed of literary corpus in Ukrainian collected from the УкрЛіт and UkrLib websites. The purpose of these websites is to promote literature in Ukrainian. According to the policy of these websites, these works are publicly available and can be used as far as they are cited. For the translated works, we collected publicly available originals from websites like Project Gutenberg (https://www.gutenberg.org/). The second source is composed of medical documents from the MedlinePlus website (https://medlineplus.gov/). These documents contain patient-oriented brochures on several medical topics. These brochures have been created in English and translated in several languages, among which Ukrainian. Here again, Ukrainian is the target language.

Corpus	Occurrence of words	Number of texts
Literature/UK	3,111,656	110
Literature/FR	1,310,732	29
Literature/EN	2,203,350	51
Literature/PL	260,536	30
MedlinePlus/UK	43,184	129
MedlinePlus/EN	46,544	129

In Table above, we indicate the size of the collected corpora (number of texts and number of word occurrences) for each language: Ukrainian (UK), French (FR), Polish (PL), and English (EN).

This dataset contains parallel texts. These source languages have been chosen for their representativity and relation with the Ukrainian language. Polish is also a Slavic language, and is close to Ukrainian. Polish is now quite well researched within the NLP field. We assume that the methods and tools developed for the Polish language can be adapted to Ukrainian provided that there are suitable corpora and resources. English and French languages are well researched from the NLP point of view. We assume, it is possible to take advantage of this research using the transfer methodologies (Yarowsky et al., 2001; Lopez et al., 2002), provided that there are suitable parallel and aligned corpora, and resources.

3. Building of corpus

Documents are converted in the UTF-8 text. Then, the text files are automatically segmented in sentences in each language using strong punctuation. Ideally, such segmentation should provide corpus aligned at the sentence level. Yet, it is necessary to verify the correctness of the segmentation in sentences and the parallelism between the source and target versions of a given document. Indeed, during the translation process, the organization of the sentences and their segmentation can be modified by the translator in order to better convey the meaning. Besides, some sentences can also be omitted. Hence, the manual control and correction during the alignment at the sentence level is necessary. This is a very long and thorough yet necessary process, as it guarantees the quality of the aligned corpora. Only part of the whole set of texts available is aligned.

Corpus	Source	Target
Litterature/FR	507,063	419,479
Literature/EN	502,393	424,730
Literature/PL	260,536	264,200
Medline/EN	46,544	43,184

Table above indicates the size of the currently aligned texts, each of which has undergone manual verification. On the whole, the aligned corpus provides 1,151,593 word occurrences in the target Ukrainian language. As we can see, all medical texts and all literary texts in the Polish/Ukrainian pair has been aligned and verified, while only part of French and English source texts is operational up to now.

4. Conclusion and Future Work

We proposed a description of parallel corpus in which Ukrainian is the target language, while the source languages are Polish, French and English. The corpus mainly contains fiction work but also some texts from the medical field. This corpus is partly aligned at the level of sentences. There are some current exploitations of the corpus for the acquisition of medical terminology (Hamon & Grabar, 2016).

Future exploitations of this corpus may be related to the machine translation, to the acquisition of cross-lingual paraphrases and disambiguation, to various contrastive studies (including stylistics and discourse). An important issue is the creation of tools for the linguistic processing of texts in Ukrainian, like POS-tagging and syntactic parsing.

A subset of the texts is being aligned by two annotators, so that the inter-annotator agreement can be computed. Besides, tools for the automatic alignment of sentences are being investigated, which may allow to enrich the set with the aligned sentences.

References

Дарчук. (2010). Дослідницький корпус української мови: основні засади і перспективи. *ВІСНИК Київського національного університету імені Тараса Шевченка* 21, 45-49.

Бук. (2010). Лінгводидактичний потенціал корпусу текстів Івана Франка у викладанні української мови як іноземної. In: *Theory and Practice of Teaching Ukrainian as a Foreign Language*. 70-74.

Сірук. (2012). Підготовка діалектних текстів для корпусного опрацювання. In: *Комп'ютерна лінгвістика: сучасне та майбутнє*. 43-45.

Hamon, T., Grabar, N. (2016). Adaptation of cross-lingual transfer methods for the building of medical terminology in Ukrainian. *Computational Linguistics and Intelligent Text Processing*, 1-12.

Kotsyba, N. (2012). Polukr (a Polish-Ukrainian parallel corpus) as a testbed for a parallel corpora toolbox. *Philological Studie LXIII*, 181-196.

Siruk, O., Derzhanski, I. (2013). Linguistic corpora as international cultural heritage: The corpus of Bulgarian and Ukrainian parallel texts. *Digital Presentation and Preservation of Cultural and Scientific Heritage 3*, 91-98.

Yarowsky, D., Ngai, G., Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In: *HLT*.

Lopez, A., Nossal, M., Hwa, R., Resnik, P. (2002). Word-level alignment for multilingual resource acquisition. In: *LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Data*. Las Palmas, Spain.

••••

Jane Hacking
University of Utah, United States
j.hacking@utah.edu

Erin Schnur
University of Utah, United States
erin.schnur@utah.edu

Fernando Rubio
University of Utah, United States
fernando.rubio@utah.edu

MuSSeL: Designing and building a corpus of multilingual second language speech

In this presentation, we will describe the ongoing process of designing and compiling the Multilingual Spoken Second Language (MuSSeL) learner corpus. When completed, the corpus will be freely-available on-line, and contain approximately 18,000 spoken texts produced by 1,800 second language (L2) learners of six languages, spanning multiple proficiency levels, ages, and contexts of learning. While extant learner corpora cover a wide range of languages, learner characteristics, and text types, few large-scale, freely-available multilingual corpora of learner language currently exist. Additionally, the majority of existing learner corpora contain written rather than spoken texts and represent the language production of learners at higher proficiency levels (Gablasova, Brezina, & McEnery, 2017; Gilquin & Granger, 2015; Granger, 2002).

Corpus resources for L2 Russian are particularly limited. The Russian Learner Corpus (RLC) http://web-corpora.net/RLC/ and its subcorpus the Russian Learner Corpus of Academic Writing (RULEC) <http://www.web-corpora.net/RLC/rulec are, to our knowledge, the only such freely available resources. Both corpora include samples from heritage speakers of Russian and L2 learners and these are clearly distinguished. The RLC provides access to tagged samples from several thousand learners of Russian, while the RULEC focuses in on a smaller number of learners and offers a rare longitudinal view of writing development. Overall, the vast majority of the samples are of written language although the RLC does contain some transcriptions

of oral samples. A number of studies that draw on these data have been published over the last few years (e.g., Polinsky, Rakhilina, & Vyrenkova, 2016; Rakhilina, Vyrenkova, Mustakimova, Ladygina, & Smirnov, 2016).

The MuSSeL project addresses the overall lack of corpora that focus on the spoken language production of learners across proficiency level and age groups. The corpus draws from professionally-rated oral proficiency exam data and comprises samples from L2 speakers of six languages: Mandarin, French, German, Portuguese, Russian and Spanish. Texts come from learners in three contexts of learning: 3rd (n = 50) and 5th grade students (n = 50) enrolled in Utah's Dual Language Immersion Program, adult classroom learners (n = 100), and adults who have acquired their L2 through in-country immersion (n = 100), totaling 300 speakers of each language. Speech samples are collected during testing using one of two instruments: Adult samples are collected using The American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview by computer (OPIc), which is an adaptive test where the test-taker responds to a series of prompts delivered by computerized avatar, and the response is recorded. Child samples are obtained from the ACTFL Assessment of Performance towards Proficiency in Languages (AAPPL). This test is also delivered by computer. Students are presented with recorded video containing spoken prompts, and the student's response is recorded. Although the instruments for testing children and adults are different, the two tests are developed by the same testing agency and the rating systems are equivalent.

To develop the corpus, student responses are transcribed into both a simple (.txt) version and a version using the Computerized Language Analysis (CLAN) software and annotation conventions (developed for users of the Child Language Exchange Data System, MacWhinney & Snow, 1984). The CLAN software allows researchers to run a variety of built-in analyses, such as calculating mean length of utterance and assessing lexical diversity. The corpus will be freely available online. Texts will be searchable based on speaker attributes (e.g., context of learning, proficiency rating, L2) through a web interface, and downloadable in three formats: text file, chat file (.cha), as well as mp3 file containing the original audio.

We will discuss the specific challenges of corpus design, data collection, and transcription that arose in creating a pilot version of the corpus, and how these challenges have impacted the on-going corpus development process. To demonstrate the functionality of the corpus, and in keeping with the Slavic focus of this conference, we will report pilot study data comparing lexical diversity of L2 Russian and L2 Portuguese in speakers at the ACTFL Intermediate and Advanced proficiency levels. We will also introduce a project that investigates the acquisition of case morphology using Russian data from the corpus.

References

Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: comparing and interpreting corpus frequency information. *Language Learning*, 67(S1), 130–154.

Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 418–436). Cambridge: Cambridge University Press.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: Benjamins.

Polinsky M., Rakhilina, E., & Vyrenkova, A. (2016). Linguistic creativity in heritage speakers. *Glossa*, 43, 1-29.

Rakhilina, K., Kisselev, O., Smolovskaya, E., & Mescheryakova, E. (2015). Russian in the English mirror: (non)grammatical constructions in learner Russian. Corpus Linguistics Conference, University of Lancaster.

Rakhilina, E., Vyrenkova, A., Mustakimova, E., Ladygina, A., & Smirnov, I. (2016). Building a learner corpus for Russian. In Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition (pp. 66-75).

••••

Juho Härme
University of Tampere, Finland
juho.harme@uta.fi

Last year but not yesterday? Explaining differences in the locations of Finnish and Russian time adverbials using comparable corpora

Finnish and Russian are languages traditionally considered to have a free or – in generativistic terms – discourse configurational word order. In both languages the positioning of adverbials in general and time adverbials in particular is especially flexible. If one, however, looks at the distributions of different possible locations for time adverbials, some clear differences can be seen. In this presentation some of these differences are examined in light of the temporal expressions *eilen / вчера ('yesterday') and viime vuonna / в прошлом году* ('last year').

The study is based on two pairs of comparable corpora: first, Finnish and Russian versions of the *Aranea* corpora (Benko 2014) are used as sources of diverse internet-based texts; secondly, the newspaper subcorpus of the *Russian national corpus* is compared to a corpus combining the *Newspaper corpus of the Finnish national library* and the newspaper subcorpus of the *Finnish text collection* of the language bank of Finland.

The study was conducted by first extracting the concordances containing the aforementioned temporal expressions from the corpora and further annotating them using dependency parsers. The annotated concordances were filtered, so that only SVO sentences were taken into account. These occurrences were further filtered to include only the cases where the temporal expressions were considered direct dependents of the finite verb by the parsers; also, expressions like *eilen aamulla* ('yesterday morning'), where the temporal expression is further specified by another adverbial, were filtered out (for more details of these filtering queries cf. https://tinyurl.com/y9sgjyv3 and https://tinyurl.com/yd3psf9e). The final data set contains 7973 Finnish and 7958 Russian sentences. Sentences 1 and 2 are examples of these:

1. Вчера моя жена купила норковую шубу!
   yesterday my wife buy-PRET mink coat

2. Eduskunta näytti eilen todellisen mahtinsa.
   Parliament show-IMP yesterday true power-POSS

The possible locations of the temporal expressions in the SVO sentences were originally divided into four groups (cf. the locations of adverbs in Quirk & Greenbaum 1985): the location before the subject and the verb (P1), the location between the two (P2), the location between the verb and the object (P3) and the location after the object (P4). Since in Finnish P2 is hardly used at all and P3 is not frequent in Russian, the two middle positions where combined. Overall, 25.9 % of the Finnish temporal expressions ended up in P1, 52.99 % in P2/P3 and 21.11 % in P4; the corresponding distribution in Russian is 68.25 %, 27.54 % and 4.21 %. There are, hence, rather clear-cut differences between the languages especially with regards to the sentence-initial position.

In order to find out about the reasons behind the differences, a multinomial bayesian regression model (cf. e.g. Gelman & Hill 2006: 124; Ntzoufras 2009: 300) was constructed with the location of the temporal expression as the dependent variable. The independent variables included, besides the language of the sentence, the type of the source corpus (internet or newspaper), the temporal expression and the type of the subject of the sentence (short/pronominal or long) and the interactions between language and the other variables. The last variable was included as a clue about where the actual sentence was located in the text. Subjects comprised of pronouns or just one noun were considered "short" and multi-noun subjects were labeled "long". Short and pronominal subjects usually indicate that the writer is referring to an agent already active in the discourse, whereas longer subjects are more likely to occur early in the texts, where the referent has not yet been introduced (Lambrecht 1996).

The statistical analysis suggests that, compared to Russian, the Finnish equivalent of yesterday is especially resistant to the sentence-initial position (P1). Also, the concordances from the Finnish newspaper corpus tend to be less likely in P1 as do the Finnish cases with a long rather than a short subject. This points to a basic difference between Finnish and Russian in how temporal adverbials are usually used in structuring a text – especially a news item. In Russian, the temporal adverbial is often used as an anchoring

point for starting the whole text (cf. example above). These are what I call introductory time adverbial constructions. In Finnish, these constructions are a rarity and the preferred way to start a text is through an identifiable agent (such as "the Parliament" in). The fact that the more probable candidate for the sentence-initial position in Finnish is *viime vuonna* suggests that a typical use-case for P1 adverbials is the so-called subtopic strategy (Dik 1989): the writer begins the text by introducing the main topic in the text (e.g. the Parliament), and the main topic is thereafter segmented into various temporal subtopics: two years ago the Parliament did X, last year it did Y and this year it did Z. These kind of sentences are analyzed as instances of a subtopical time adverbial construction. In this presentation I argue that the usage of the introductory and the subtopic constructions are a major factor in explaining the differences in the positions of the temporal expressions examined.

References

BENKO, V. 2014. Aranea: Yet another family of (comparable) web corpora. In: P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.) *Text, Speech and Dialogue: 17th International Conference, Tsd 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings*. pp. 247–256. Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-319-10816-2_31

Dik, S.C. (1989). *The theory of functional grammar. 1. the structure of the clause*. Berlin: Mouton de Gruyter.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Janko, T. (2001). *Kommunikativnyje strategii russkoi retši*. Moskva: Jazyki slavjanskoj kul'tury.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge university press.

Ntzoufras, I. 2009. *Bayesian modeling using winbugs*. New Jersey: John Wiley & Sons.

Quirk, R. and Greenbaum, S. 1985. *A university grammar of english*. London: Longman.

Vilkuna, M. 1989. *Free word order in finnish: Its syntax and discourse functions*. Helsinki: Suomalaisen Kirjallisuuden Seura.

••••

Stefan Heck
University of Tübingen
s.heck@uni-tuebingen.de

Eugen Kravchenko
University of Tübingen
eugen.kravchenko@student.uni-tuebingen.de

Polish być w trakcie + verbal noun – a progressive periphrastic construction?

One well-known characteristic of Slavic verbal aspect is that a wide range of aspectual meanings is clustered into a binary opposition (perfective vs. imperfective). Considering the plurifunctionality of the imperfective in particular (progressivity, habituality, iteration, other modal and pragmatic uses), one might think that the languages would develop new ways of disambiguating this range of meanings. This has happened very noticeably in Romance languages (cf. Dessì Schmid 2014:197–223 on progressive periphrases) and also in Germanic languages (cf. the English progressive or the German colloquial progressive). Slavic languages, however, do not appear prone to developing any new grammatical means of expressing progressivity (Plungjan 2011:297).

The aim of this paper is to present the Polish construction *być w trakcie* + verbal noun (VN). This construction has not been previously studied to our knowledge, though it is noted in Wiemer (in press). It is used to express progressivity and can thus provide a stronger, unambiguous alternative to the polysemous bare imperfective, cf. (1).

1. (…) wyjął butelkę i już był w trakcie otwierania kiedy usłyszałam (…)
   'he took out the bottle and was already opening it when I heard'
   (forum.gazeta.pl)

Its component *w trakcie* 'during' is a prepositional expression used in temporal adverbials, and *być w trakcie* without VN can be used in the sense of 'to be ongoing'. The whole construction is strongly reminiscent of the French progressive construction *être en train de*, a parallel also noted by Wiemer (in press).

We will focus on the use of *w trakcie* together with the copula *być* and a VN. According to Pčelinceva (2016) and Dickey (2000), Polish VNs in *-nie/-cie* are closer to verbs than e.g. their Russian counterparts in that they systematically distinguish verbal aspect and can go with reflexive *się*. Because they are formed regularly from almost all verbs (for restrictions see Fokker 1965:266–268) and because they are semantically very close to their base verbs, we believe the Polish VNs are especially well-suited for use in a periphrastic construction, comparable to the infinitive or gerund in Germanic and Romance language. We will focus our attention mostly on these VNs (and restrict the term verbal noun to them), but also consider other deverbal nouns like *dyskusja, realizacja* etc.

For the first part of our study, we conducted a hand-annotated corpus study using Araneum Polonicum Minus and NKJP, looking for any combinations of *być w trakcie* and a noun or VN in the genitive case (627 items in APMin, 948 in NKJP).

The most surprising find was that *być w trakcie* can occur with perfective VNs (27/286 VNs in APMin, 72/514 in NKJP), which is not mentioned by Wiemer (in press). A perfective verb form should be incompatible with the progressive. This raises the question of whether these really are VNs *sensu stricto*, nominalized verbs with most of their verbal semantics intact, or whether they have gone further on their way to nominalization and in doing so lost the [perfective] component from their meaning (for nouns derived from homonymous VNs, see Fokker 1965:254 or Dickey 2000:237). Pčelinceva (2016) suggests consulting dictionaries to decide whether a given form is a VN or a deverbal (i.e. "less verbal") noun. This does not solve our problem at hand, however, as e.g. *podpisanie* 'signing' is not listed as a separate entry in SJP or SJPD, hence should be a perfective VN, not a common noun; but it still occurs with *być w trakcie*, as does the clearly imperfective *podpisywanie*.

Another variation we have found is that the subject of the copula can be the object of the VN (107/514 in NKJP, 57/286 in APMin) aside from being its subject, cf. (2).

2. Pozostałe części są w trakcie opracowywania. (NKJP)
   'The other parts are being worked out.'

The object of the VN can also surface as a genitive, however:

3. Jesteśmy w trakcie opracowywania kilku nowych projektów… (APMin)
   'We are currently working out a couple of new projects…'

This first part of our study aims to show that the grammaticalization of the construction *być w trakcie* + VN is not very advanced, which is proved by the high degree of variation we find with it. Nevertheless, it appears to be a productive construction that is not confined to literary speech as Wiemer (in press) has claimed, cf. our example (1).

In the second part of our study we compare this Polish construction with Czech and Russian equivalents on InterCorp v10, as well as with French for the sake of comparison with *être en train de.* The Polish is much less frequent, as we would expect. For Czech and Russian, there is no special construction, but rather the use of the polysemous imperfective, often with adverbials to accentuate the progressive meaning.

References

[APMin]: Benko, V. (2015). *Srovnatelné webové korpusy Aranea.* Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz

Dessì Schmid, S. (2014). *Aspektualität – Ein onomasiologisches Modell am Beispiel der romanischen Sprachen.* Berlin/New York: De Gruyter.

Dickey, S. M. (2000). *Parameters of Slavic aspect: A cognitive approach.* Stanford: CSLI Publications.

Fokker, A. A. (1965). Derivation of nouns from verbs in contemporary literary Polish. *Lingua* 13, 240-273.

[NKJP]: *Narodowy korpus języka polskiego.* http://www.nkjp.pl

Pčelinceva, E. Ė. (2016). *Ot glagola k imeni: aspektual'nost' v russkich, ukrainskich i pol'skich imenax dejstvija.* Sankt-Peterburg: Nauka.

Plungjan, V. A. (2011). Tipologičeskie aspekty slavjanskoj aspektologii (nekotorye dopolnenija k teme). *Scando-Slavica* 57, 290-309.

[SJP]: Szymczak, M. (red.) (1999): *Słownik języka polskiego PWN.* Warszawa

[SJPD]: Doroszewski, W. (red.) (1964): *Słownik języka polskiego.* Warszawa.

Wiemer, B. (in press): Slavic aspect: Its rise and inner-Slavic clines as a result of macro-areal diffusion?. In: W. Bisang & A. Malchukov (eds.), *Areal patterns in Grammaticalization.*

Milena Hnátková
Charles University, Czech Republic
milena.hnatkova@ff.cuni.cz

Tomáš Jelínek
Charles University, Czech Republic
tomas.jelinek@ff.cuni.cz

Marie Kopřivová
Charles University, Czech Republic
marie.koprivova@ff.cuni.cz

Vladimír Petkevič
Charles University, Czech Republic
vladimir.petkevic@ff.cuni.cz

Alexandr Rosen
Charles University, Czech Republic
alexandr.rosen@ff.cuni.cz

Hana Skoumalová
Charles University, Czech Republic
hana.skoumalova@ff.cuni.cz

Pavel Vondřička
Charles University, Czech Republic
pavel.vondricka@ff.cuni.cz

Multiword Expressions in Czech: Typology and Lexicon

We propose a typology of multiword expressions (MWE) as a template for entries in a lexicon of Czech MWEs. The goal is motivated by the following considerations:
(i) MWEs play a significant role in any language, and their specifics, often standing in contrast to standard grammatical properties, should be reflected accordingly. Moreover, most MWEs participate in regular morphological and syntactic patterns, which makes them a theoretically and computationally interesting research topic.

(ii) An appropriate analysis and formal description of MWEs may boost the success rate of tasks such as tagging, parsing, word sense disambiguation or semantic annotation. A proper identification of MWEs and their types in any of these tasks may lead to a better analysis of their sentential context as well.
(iii) The lexicon should support recognition and identification of MWEs in running text not only in their standard form, but also in their fragments and variants, including nominalization, passivization or adjectivization, and in more creative modifications of standard MWEs.
(iv) The lexicon is also meant for human users – students and teachers of Czech as L1 and L2, lexicographers, grammarians, translators, even general public.

MWEs are „lexical items that:

(a) can be decomposed into multiple lexemes, and

(b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity“ (Sag et al., 2002). In a language with explicit word boundaries, (a) is easy.

As for idiomaticity (b), useful also in the classification of MWEs, we adopt the proposal of Baldwin et al. (2010). They categorize MWEs according to syntactic structure, fixedness/flexibility and idiomaticity. We extend this taxonomy to reflect the properties of Czech as a free word order language with complex inflectional morphology. Specific extensions concern the following aspects of MWEs:

(i) **definition**

(ii) **style/register**: standard / colloquial / expressive / dialect / other

(iii) **usage type**: proverb / weather lore / comparison / simile / citation / set phrase / grammatical idiom (= compound function word) / term / other phraseme

(iv) **syntactic type**: noun phrase / adjectival phrase / verb phrase / light verb construction / adverbial phrase / prepositional phrase / compound preposition / compound conjunction / compound interjection / clausal unit / compound and sentential unit / other

(v) **use of fragments and variants**: fragments and variants of standard MWEs are identified since the lexicon should make it possible to identify creative modifications of MWEs in running text.

Syntactic information is expressed as a dependency / phrase-structure tree, by valency and for idiosyncratic word order; also internal modifiability of MWE components and possible transformations (de)passivization, nominalization, adjectivization are accounted for.

Moreover, a more detailed specification of some types of idiomaticity is provided:

(i) **lexical:** a word only occurs in MWE

(ii) **morphological** (concerning morphologically specific MWE components)

(iii) **syntactic**: anacoluthon / attraction / aposiopesis / ellipsis / word order / other

(iv) **semantic idiomaticity**, where a degree of meaning compositionality is distinguished:

(iv.1) often compositional: *chodit s holým zadkem* 'not to have two pennies to rub together'

(iv.2) rarely compositional: *kočičí hlavy* 'cobblestones'

(iv.3) never compositional: "*pozdě bycha honit*" 'it is no good crying over spilt milk'

(v) **pragmatic:** MWE is only used in certain situations

(vi) **statistical**: MWE is a fixed but semantically compositional collocation.

Generally, some parts of the description assume the standard rules of Czech as default, stipulating only deviations and irregularities (Hnátková et al., 2017).

A **lexical entry** includes definition of individual **slots** (representing syntagmatic positions of its components) and their possible **fillers** (representing variable realizations of the components), so that descriptions, features or relations may be assigned separately to the MWE as a whole, to its individual components or to their different realizations. Slots may also be marked as optional or mandatory components of the MWE. Fillers may refer to sequences of other slots, allowing for a grouping or construction of tree structures, in order to facilitate assigning descriptions to different partial structures within the MWE.

The lexicon is to contain ca. 7000 entries in the first phase (end of the project, January 2019) and in the next phase (after January 2019) it will be gradually enhanced. At the end of August 2018:

(i) the pilot version of the lexicon includes ca. 100 test entries, still without syntactic structure information

(b) the lexicon manager functionalities are tested (adding, correcting and searching entries).

In September 2018, the number of entries will considerably increase and syntactic structures (dependency and phrase-structure trees) will be automatically added to individual MWE entries. At the conference, not only the typology, but also the lexicon will be presented.

References

Baldwin, T., Kim, S.N. (2010): Multiword expressions. In: Indurkhya, N., Damerau, F.J. (eds.), *Handbook of Natural Language Processing*, 2nd edn., pp. 267–292. CRC Press, Boca Raton.

Hnátková, M., Jelínek, T., Kopřivová, M., Petkevič, V., Rosen, A., Skoumalová, H., Vondřička, P. (2017): Eye of a Needle in a Haystack. Multiword Expressions in Czech: Typology and Lexicon. In: Mitkov, R. (ed.), *Computational and Corpus-Based Phraseology: Second International Conference, Europhras 2017*, London (2017), Proceedings, Springer International Publishing, LNAI 10596, pp. 160–175. doi: 10.1007/978-3-319-69805-2_12, URL: https://doi.org/10.1007/978-3-319-69805-2_12

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D. (2002): Multiword expressions: a pain in the neck for NLP. In: Gelbukh, A. (ed.), CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg. doi:10.1007/3-540-45715-1_1

....

Jakob Horsch
Catholic University Eichstätt-Ingolstadt
Jakob.Horsch@ku.de

A Construction Grammar Account of the Slovak Comparative Correlative Construction

The comparative correlative (CC) construction (cf. Culicover and Jackendoff 1999, Den Dikken 2005) is a structure that in its most simple form consists of two clauses, which shall henceforth be referred to as C1 and C2. CCs have a comparable structure across many languages (cf. Culicover and Jackendoff 1999: 569). In English, the CC typically looks like (1), a line from the popular children's song *The More We Get Together*; in Slovak, CCs typically look like the saying[1] in (2):

(1) [The more we get together,]$_{C1}$ [the happier we'll be.]$_{C2}$

(2) [Čím bližšie Rím,]$_{C1}$ [tým horší kresťan.]$_{C2}$
'The closer Rome, the worse the Christian.'

As a comparison of (1) and (2) shows, CC constructions have a similar structure in both languages: C1 and C2 are introduced by fixed phonological/lexical elements, followed by a comparative element (i.e. in the English C1, the comparative adverb *more* and in C2, the comparative adjective *happier*; and in the case of Slovak, the comparative adjectives *bližšie* (*closer*) in C1 and *horší* (*worse*) in C2), which is then followed by the rest of the clause. One of the most notable differences is the symmetrical *the… the…* encoding of the CC in English as opposed to the asymmetrical *čím… tým…* in Slovak. In the latter, the fixed phonological elements resemble the language's instrumental-case impersonal relative pronoun *čím* and instrumental-case demonstrative pronoun *tým*.

In both languages in terms of semantics, the C2 clause can be described as the effect (apodosis/dependent variable) of a cause (protasis/independent variable) contained in the C1 clause. The construction has asymmetric as well as symmetric properties: First, there is a conditional (asymmetric) relationship between C1 and C2 (being happier is a result of getting together), and second, there is a parallel change in C1 and C2 over a certain time period (as we get together more, our happiness simultaneously increases; cf. Hoffmann 2018).

Owing to the various idiosyncrasies it exhibits, the CC construction has recently attracted increased attention in linguistics (e.g. Borsley 2004a, Den Dikken 2005, Sag 2010, Cappelle 2011, Hoffmann 2018), particularly that of Construction Grammar (CG). However, most of the research done so far has concentrated on the CC construction in English. In Slavic languages, which differ significantly from English in basic parameters such as constituent order, the CC has hardly been explored, with one of the few studies available so far focusing on Polish (Borsley 2004b).

However, as Mirjam Fried notes, the study of Slavic languages holds great potential for the development of Construction Grammar, which in turn is a framework highly suitable for accounting for what she calls "well-studied" but "puzzling" phenomena (2017: 249), i.e. the idiosyncracies commonly observed in CCs (cf. e.g. Borsley, who states that the CC construction "falls outside the scope of syntax proper" (2004b: 59)).

To my knowledge, no research on the CC construction in Slovak has been conducted so far, much less so within the CG framework. Accordingly, this paper will explore the Slovak CC construction, including an empirical analysis of a random sample of corpus data obtained from the Slovak National Corpus (SNC).

The paper presents the Slovak CC construction within a CG framework and is structured as follows: In section 2, I will introduce the Slovak CC using, among others, examples found in the SNC, and point out various idiosyncracies. In section 3, I will show the results of a corpus study using a random sample of 500 hits from the SNC, discussing interesting phenomena such non-iconic clause order and cross-clausal associations. In section 4, I will provide a conclusion, arguing that CG is highly suitable for accommodating CCs and working with idiosyncratic phenomena in general.

References

Borsley, Robert D. (2004a). An approach to English comparative correlatives. In S. Müller (Ed), *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*. Stanford: CSLI Publications, 70–92.

1 1 As found on Slovak newspaper SME's web site (https://zlatyfond.sme.sk/dielo/1444/Zaturecky_Slovenske-prislovia-porekadla-a-uslovia-Dobre-a-zle-mravy-Pravda-a-viera/5, retrieved on 27/2/2018)

Borsley, Robert D. (2004b). On the periphery: Comparative correlatives in Polish and English. *Proceedings of Formal Approaches to Slavic Linguistics,* 12, 59-90.
Cappelle, B. (2011). The *the…the…* construction: Meaning and readings. *Journal of Pragmatics,* 43 (1), 99-117.
Culicover, P. W. and R. Jackendoff (1999). The view from the periphery: The English comparative correlative. *Linguistic Inquiry,* 30, 543–71.
Den Dikken, M. (2005). Comparative correlatives comparatively. *Linguistic Inquiry,* 36, 497–32.
Fried, M. (2017). Construction Grammar in the Service of Slavic Linguistics, and Vice Versa. *Journal of the Slavic Linguistics Society,* 25 (2), 241-276.
Hoffmann, T. (2018). Comparing Comparative Correlatives: The German vs. English construction network. In: H. C. Boas and A. Ziem (Eds), *Constructional Approaches to Argument Structure in German.* Berlin: Mouton de Gruyter.
Sag, I. A. (2010). English Filler-Gap Constructions. *Language*, 86 (3), 486-545.

••••

Laura Janda
UiT The Arctic University of Norway
laura.janda@uit.no

Francis Tyers
Higher School of Economics, Russian Federation
ftyers@hse.ru

Parts Give More Than Wholes: Paradigms from the Perspective of Corpus Data

Native speakers of languages with complex inflectional morphology routinely recognize and produce forms that they have never heard or seen (the "Paradigm Cell Filling Problem", cf. Ackerman et al. 2009). How is this possible? We take a learning perspective on this question and present evidence to show that inflectional morphology is mastered through partially overlapping portions of paradigms in input.

Russian is a good point of reference because among languages that are well-documented and have a large hand-annotated corpus (like SynTagRus, which is the basis for our research), Russian is morphologically relatively complex, in terms of the number of word forms in its paradigms, the number of inflectional classes, and the proportion of irregular and suppletive word forms.

We provide three types of evidence (a-c below) that the inflectional morphology of Russian is based on partial sets of inflected word forms. These parts of paradigms exhibit differ from lexeme to lexeme, yet overlap enough to make it possible to produce unencountered forms both of known and of newly encountered lexemes. Our theoretical perspective is most closely allied with usage-based Cognitive Linguistics (Langacker 2008) and Word and Paradigm Morphology (Blevins 2015, 2016). Our evidence comes from:

a) Comparison of the Percentages of Full Paradigms Attested in Corpora

We compare the percentage of noun lexemes that are attested in all their paradigm forms across five languages with nominal paradigms sizes ranging from 2 to 28 forms, and show that attestation of all forms in a paradigm is rare and there is a consistent relationship (languages with larger paradigms have a lower percentage of fully attested paradigms). For Russian, only 0.06%

of nouns are attested in all 12 forms in the SynTagRus corpus, a portion that, due to Zipf's law, will not change significantly no matter how large the corpus is.

b) Corpus Distribution of Partial Sets of Word Forms

We extract the grammatical profiles (corpus frequency distributions of paradigm forms) for nearly 1000 high-frequency Russian nouns, stratified across 5 declension classes and use these as input for correspondence analysis, measuring the distances between nouns and between paradigm slots. We find that, for any given lexeme, only 1-3 forms are frequent, but this pattern is different for each noun, and as a result the partial sets overlap, collectively populating the space of the paradigm.

c) Computational Experiment

We conduct an experiment modeling the learning of full paradigms for all inflected word classes in Russian (nouns, verbs, and adjectives), as compared with the learning only of the single most frequent form for each lexeme. The task in our experiment is to predict a given form of a previously unencountered lexeme. Both the training and testing data come from a frequency-ordered list of word forms in SynTagRus. Language pedagogy has traditionally relied on presentation of full paradigms, and most computational experiments modelling the learning of inflectional morphology use full paradigms for training (but note a recent pioneering work that departs from this tradition: Malouf 2017). Our experiment shows that learning single forms as opposed to full paradigms is more effective both in terms of the percentage of correctly predicted forms and the edit distance needed to correct errors.

Collectively, these three types of evidence suggest that all paradigms are defective to a greater or lesser extent, since all lexemes have some word forms that are attested rarely or not at all, that inflectional morphology should be modelled in terms of overlapping partial sets of word forms, and that learning is actually enhanced by focusing only on the word forms most likely to be encountered rather than taking entire paradigms as input.

Our results are consistent with a usage-based model of language in which memorization and the learning of patterns coexist. High-frequency forms are likely stored and may also be used as the basis for abstracting schemas for the patterns among word forms.

References

Ackerman, F., J. P. Blevins, and R. Malouf. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In: J. P. Blevins & J. Blevins (Eds.) *Analogy in Grammar: Form and Acquisition*, Oxford: Oxford University Press, 54-82.

Blevins, James P. (2015). Inflectional Paradigms. In M. Baerman (Ed.), *The Oxford Handbook of Inflection*. Oxford: Oxford University Press, 87-111.

Blevins, James P. (2016). *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Langacker, Ronald W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. Morphology.

....

Tomáš Jelínek
Charles University, Czech Republic
tomas.jelinek@ff.cuni.cz

New error annotation of Czech learner corpora

The analysis of texts of non-native speakers has become an important tool for understanding the process of learning a second language and the development of adequate teaching methodologies. In this paper, we propose a new concept of error annotation of texts produced by learners of Czech as a second language which is simpler than previous error annotation systems such as (Rosen 2016, Jelínek et al. 2012, Štindlová et al. 2013), easier to use and complements the previous systems by its focus on morphology. We also describe the procedure of re-annotation of existing learner corpora by the proposed annotation system.

The error annotation in CzeSL uses two levels of manual emendation and error annotation. At the lower level, erroneous word forms are corrected; the result of the higher level of annotation is a correct sentence. To each word correction on both levels, an error label (of about twenty types) is then assigned.

We propose an annotation system that will only use the final, correct emendation (not two levels like CzeSL), significantly simplifying the task of the annotator and facilitating the reproducibility of the error annotation using NLP tools. Our error annotation is based on levels of linguistic description: we identify orthographic errors (ORT), phonological and morphological errors (MPHON), errors of inflection (MORPH), syntactic errors (SYN) and lexical errors together with errors of use (LEX); with optional more detailed sub-labels (e.g. SYN: dep – syntactic error of dependency, ORT: cap – orthographic error of capitalization). In cases where there are two or more possible causes of the error, we select a basic error tag plus one or more tags from the other planes. For example, the sentence *Přijdou mnoho lidi* 'many people will come' with the wrong form of *lidi* (nom.pl) instead of *lidí* (gen.pl) may be an orthographic error (omission of diacritics), morphological error (erroneous case form) or syntactic error (incorrect case choice); the primary error mark is MORPH, with the ORT and SYN flags (the verb *Přijdou 'they will come'* has an incorrect number).

The error annotation can be more accurate due to the fact that the precise locations of errors inside the words are marked. For example, the word *kamara**at**k**y*** 'friends' in the sentence *Mám mnoho kamaratky* 'I have many friends' instead of *kamarádek* has three separate errors (*a*/*á* MPHON:quant + ORT:dia; *t*/*d* MPHON:assim + ORT; *ky*/*ek* MORPH + SYN:dep); each will be marked and error-annotated separately.

In order to get data for machine learning and automatic annotation, we use already annotated CzeSL data, namely the original text (transcribed) and the corrected text (final emendation). In the future, we will use also automatically corrected texts using a combination of rule-based corrections and a stochastic spell-checker and text correction tool (Richter et al., 2012).

The actual annotation of learner texts combines automatic text preprocessing, manual annotation in the Brat environment (Stenetorp et al., 2012) and automatic post-processing of annotated text. We are considering to use a newer annotation environment, WebAnno (Eckart de Castilho et al., 2014), provided that we find a reliable conversion tool from the Brat data format to WebAnno. Preprocessing identifies individual differences between original and corrected text, marks these differences and adds some information about the error type which can help the annotator, but should not influence her decision. The annotator assigns each identified error an error-label and checks for others, unidentified errors. The corrected text cannot be changed in Brat, but can be marked as not properly corrected (to be corrected outside of Brat). Automatic postprocessing assigns morphological tags and lemmas to both original and corrected word forms, for some types of annotator-labeled error tags, sub-labels or flags are added. As a separate information, it records which characters on the part of the original and corrected word form are part of the identified error (eg. in *Prahě/Praze : hě/ze*).

We intend to build a corpus of texts produced by learners of Czech and annotated by the proposed error-annotation system. It will enrich our understanding of interlanguage and lead to better teaching methods of Czech as a second language.

References

Eckart de Castilho, R., Biemann, C., Gurevych, I. and Yimam, S.M. (2014): WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Jelínek, T., Štindlová, B., Rosen, A. and Hana, J. (2012). Combining manual and automatic annotation of a learner corpus. In P. Sojka et al. (eds), *Text, Speech and Dialogue – Proceedings of TSD 2012, no. 7499 in LNCS,* p. 127–134.

Štindlová, B., Škodová, S., Hana, J., & Rosen, A. (2013). A learner corpus of Czech: current state and future directions. In S. Granger et al. (eds), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead, Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve*. Presses Universitaires de Louvain.

Richter M., Straňák P., Rosen A. (2012). Korektor – A System for Contextual Spell-checking and Diacritics Completion. In Kay M., Boitet C.: *Proceedings of the 24th International Conference on Computational Linguistics* (Coling 2012). Mumbai, India, p. 1-12.

Rosen, A. (2016). Building and using corpora of non-native Czech. In B. Brejová (ed), *Proceedings of the 16th ITAT: Slovenskočeský NLP workshop (SloNLP 2016), vol. 1649 of CEUR Workshop Proceedings*, Bratislava, Slovakia, p. 80–87.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou S. and Tsujii, J. (2012). Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, 102–107.

••••

Tomáš Káňa
Masaryk University, Czech Republic
kana@ped.muni.cz

Terminology in and around Diminutives

Terminology is probably the most visible marker of professional/technical language. There are basically two possible ways how to form terms: motivated and not motivated (cf. Schippan 1992, 233). Leaving out the latter, I would like to point out Czech derivatives with (typically) diminutive suffixes functioning as terms (e.g. Czech *sloupek for English column, German (Zeitungs)Kolumne*) and examine if there is any system in the structures of their counterparts in German and English. Potential regularities could ease the orientation in the terminology of several fields in these languages.

The central roles in these issues are played by diminutives. This originally rather non controversial linguistic category has gone through several revisions in the past decades (see Zandvoort, Bauer, Schneider, Nekula) seeing diminutives not only as one-word-derivative, but also as an attributive phrase ("analytical diminutives") *chlapec* (boy) → *malý chlapec* (small/little boy) (Nekula 2010, 306) or Schneider (2003) or as a lexical heteronym denoting a smaller entity than the base ("implied diminutive") such as *house* → *cottage* in English (Zandvoort 1969:303) or *řeka* (river) → *potok* (creek) in Czech (Štícha 1978:116-117). This wide meaning of diminutive seems to be quite controversial because most of these constructs can not or do not convey evaluative meaning – one of the typical features of diminutives (cf. Bauer 1997). Sure, there must be a difference in defining diminutives in different languages according to their respective morphological typology. For this paper, I would like to stick to the definition of Czech diminutives which has been proved on several corpus examples (Káňa 2016, 39-40): A diminutive is a one-word derivative (a word-modification after Dokulil, 1967) with an affix which is typical in coining evaluative words (Káňa 2016, 42). But many Czech words with such affixes do not have an evaluative meaning although they are (most precisely, they were) apparent modifications of a base word such as *bavlnka* (cotton thread) derived from *bavlna* (cotton). A certain part of these derivatives ("pseudodiminutives" after Bilíková 2013) belong to technical or scientific language. I would like to pick up all such terms from the list of diminutive forms of nouns which were identified as the most

frequent in the contemporary written Czech (data of corpus syn, 2010 and InterCorp Czech-German version 7 in Káňa 2016). Further, I present their (non-evaluative) technical meaning and compare it with their German and English counterparts. I assume that most of the German counterparts will be a compound (*stolek* → *Nachttisch*) and the most of the English counterparts will be either a simplex (*desk*) or an attributive phrase (*bedside table*). This assumption is based on the results of counterparts of Czech "real" diminutives with an evaluative meaning such as *dřevíčko*: most significant counterparts in German *Kleinholz* (compound) and in English *stick* (simplex).

As the title of this paper promises, not only diminutive forms will be in the centre of interest here, but also Czech back-formations that are used for coining terms in jargons e.g. *loupák* from common language *loupáček* (a roll with poppy seeds), *limeta* from *limetka* (lime) in food industry.

And finally, a correlation between the morphological language type and the type of coining terms will be examined.

All data for this study were gathered from the multi-lingual corpus InterCorp (various versions) and from the Czech National Corpus (Syn and Syn2015).

References

Bauer, L. (1997). Evaluative Morphology. Search of Universals. *Studies in Language*, 21, No. 3. Amsterdam: Benjamins, 533-575.

Bílková, J. (2013). Pseudodeminutiva v češtině. In: *Gramatika & Korpus 2012: 4. mezinárodní konference*. Hradec Králové: Gaudeamus. [CD-ROM].

Dokulil, M. (1962). *Tvoření slov v češtině 1. Teorie odvozování slov*. Praha: Academia.

Káňa, T. (2016). *Deminutiva na pozadí korpusových dat*. Brno: Masarykova univerzita.

Nekula, M. (2010). Deminutiva a augemntativa v češtině z typologiockého pohledu. In: A. Bičan et. al (Eds.), *Karlík a továrna na lingvistiku*. Brno: Host, 304-315.

Schippan, T. (1992). *Lexikologie der deutschen Gegenwartssprache*. Tübingen: Niemeyer.

Schneider, K. P. (2003). *Diminutives in English*. Tübingen: Max Niemeyer Verlag.

Štícha F. (1978). Substantiva deminutivní formy s lexikalizovaným významem. *Naše řeč*, 61 (1978), č. 3. Praha: Academia, 113-127.

Zandvoort, R. W. (1966, 1969). *A handbook of English Grammar*. London: Longmans.

••••

Witold Kieraś
Institute of Computer Science, Polish Academy of Sciences
wkieras@ipipan.waw.pl

Łukasz Kobyliński
Institute of Computer Science, Polish Academy of Sciences
lkobylinski@ipipan.waw.pl

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
maciej.ogrodniczuk@ipipan.waw.pl

Korpusomat — new functionalities and future development

The purpose of the presentation is to give an overview of available features and further development plans for Korpusomat ('corpus machine', http://korpusomat.pl) — a web application aimed at building automatically indexed and annotated searchable corpora from documents provided by the user. Korpusomat integrates various natural language processing tools and provides an intuitive interface to use them in own projects. The version of the application presented briefly below is now in development and testing stage and will be available in public by mid-2018.

Since its very first version in 2016 Korpusomat has gained a positive response from the Polish corpus linguistics community and was addressed with many feature and enhancement requests. The application is thus in constant development. It has been recently equipped with new search engine, enabling searching various layers of annotation (Brouwer et al., 2017). Each text sent to Korpusomat is automatically annotated with Morfeusz morphological analyzer (Woliński, 2014), with one of the two alternative dictionaries: SGJP (Saloni et al., 2015) or Polimorf (Woliński et al., 2012), and one of the two morphosyntactic disambiguating taggers: Concraft (Waszczuk, 2012) or Toygger (Krasnowska-Kieraś, 2017), each representing different technical approach to the problem. The MTAS search engine allows to query the annotated corpus using Corpus Query Language (CQL) which is familiar to Sketch Engine and National Corpus of Polish users. Query results may be downloaded in CSV format for further off-line processing, i.e. using advanced statistical tools such as R or simply in Excel spreadsheets.

Korpusomat accepts files in many formats: from plain text files and Word DOC(X) files to e-book EPUB and MOBI formats and two layer DJVU files. Each document can be described by metadata, both predefined and user defined. Some metadata fields are automatically completed if the information was provided by the source format. The metadata entries can be later used in searching for restricting the scope of corpus queries and providing basic statistics concerning the corpus. The number of user's corpora and their size are not limited in any way.

Apart from searching the corpora, Korpusomat offers also terminology extraction using TermoPL tool (Marciniak et al., 2016).

The presentation will also give some brief overview of development plans. Future development will focus mainly on integrating other layers of automatic annotation and data extraction, such as named entities, dependency parsing, semantic roles etc. Thus it strongly relies on the state of development of such tools for Polish as well as their performance and accuracy. By presenting Korpusomat to Slavic corpus linguistic community we also hope to gain some feedback and inspiration for further development of the webservice. Also, it is our intention to make Korpusomat a standard, easy to deploy framework for all kinds of static corpora of Polish texts, such as parliamentary corpus, historical corpora and all sorts of special purpose corpora collected as a basis for various linguistic research and projects. It would make any technical updates easier and would allow keeping previously collected corpora processed with the most up-to-date NLP tools.

References

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure, number 136*, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In *Proceedings of 8th Language & Technology Conference*, pages 367–371.

Marciniak, M., Mykowiecka, A., and Rychlik, P. (2016). TermoPL — a flexible tool for terminology extraction. In Calzolari, N., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2278–2284, Portorož, Slovenia. ELRA, European Language Resources Association (ELRA).

Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R., and Skowrońska, D. (2015). Słownik gramatyczny języka polskiego. 3. edition.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India.

Woliński, M. (2014). Morfeusz reloaded. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

Woliński, M., Miłkowski, M., Ogrodniczuk, M., Przepiórkowski, A., and Szałkiewicz, Ł. (2012). PoliMorf: a (not so) new open morphological dictionary for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 860–864, Istanbul, Turkey. ELRA.

••••

Witold Kieraś
Institute of Computer Science, Polish Academy of Sciences
wkieras@ipipan.waw.pl

Marcin Woliński
Institute of Computer Science, Polish Academy of Sciences
wolinski@ipipan.waw.pl

Basic natural language processing toolkit for 19th century Polish

The paper presents a set of tools and resources created for the purpose of automated processing of Polish texts published in the period between 1830 and 1918. The toolkit consists of manually annotated gold-standard corpus, morphological analyser, tagger and automatic transcriber. Also an exemplary corpus of available 19th century texts automatically annotated with the presented tools will be shown.

The most laborious part of the task is manual corpus annotation. Our source data was an over 1 mln tokens large corpus collected in a previous project (Bilińska et al., 2016). The original corpus consists of 1000 samples each of length ca. 1000 words extracted from original first editions of texts published between 1830 and 1918 and was divided evenly into five genres: fiction, drama, short news, essays and popular science. For the purpose of manual annotation 2944 smaller samples (ca. 160 orthographic words from space to space each) were extracted and preprocessed using an automatic rule-based transcriber and a modified ("aged") version of Morfeusz morphological analyser (Woliński, 2014). Subsequently the data was uploaded to a web application called Anotatornia designed and developed for the purpose of manual annotations of historical corpora (Woliński et al., 2017).

To reduce the workload of corpus annotation while maintaining quality, we have introduced a novel annotation mode. Each sample was annotated independently by one human annotator and an automatic tagger. Conflicts between the two were subsequently resolved by an adjudicator ("superanotator"). In the presentation we will show the details of the process of manual annotation of the corpus.

As a result a small manually annotated gold-standard corpus was created. The corpus is 625,000 tokens large and consists of two text layers: original

transliterated text layer and transcribed (orthographically modernized) layer. It serves as a training data for various machine learning tools. So far two stochastic taggers representing different methodologies, namely conditional random fields (Waszczuk, 2012) and bi-LSTM neural networks (Krasnowska-Kieraś, 2017), were trained and evaluated based on the manually annotated corpus. The taggers obtained an overall disambiguating accuracy at 90.8% and 93.95% respectively. Both taggers received slightly lower results comparing to contemporary Polish dataset (92.65% and 95.22% based on 1.2M tokens large gold-standard subcorpus of National Corpus of Polish), however some features of historical Polish seem more difficult to process.

Some machine learning experiments concerning automatic transcription are also scheduled for the near future based on the manually annotated dataset. We are also planning some experiments with dependency parsing of historical texts, this however would require preparation of at least a small sample of syntactically annotated gold standard data for evaluation purposes. The corpus itself will be soon available in public both as XML source files and as a searchable web resource. Searching will be possible through MTAS search engine (Brouwer et al., 2017) allowing to query for both transliterated and transcribed text layers as well as for the morphosyntactic layer. MTAS supports the well known Corpus Query Language (CQL).

Digital libraries nowadays provide more and more historical documents enriched with OCR text layer. The quality of text recognition varies and depends on many factors, however large parts of those archives provide a plain text layer of relatively good quality, especially in the case of second half of the 19 th century. Using the toolkit presented above large collections of such documents can be transformed into searchable annotated corpora. These corpora could represent specific genres, authors or topics providing extensive material for diachronic linguistic research. Last but not least, such corpora could be freely distributed as historical documents provided by digital libraries in most cases are not restricted by copyrights.

References

Bilińska, J., Derwojedowa, M., Kieraś, W., and Kwiecień, M. (2016). Mikrokorpus polszczyzny 1830-1918. *Komunikacja specjalistyczna*, 11:149–161.

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure, number 136*, pages 19–37. Linköping University Electronic Press, Linköpings universitet.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In *Proceedings of 8th Language & Technology Conference*, pages 367–371.

Waszczuk, J. (2012). Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India.

Woliński, M. (2014). Morfeusz reloaded. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

Woliński, M., Kieraś, W., and Komosińska, D. (2017). Anotatornia 2 – an annotation tool geared towards historical corpora. In *Proceedings of 8th Language & Technology Conference*. Poznań, Poland.

••••

Valeria Kolosova
Institute for Linguistic Studies, Russian Academy of Sciences
chakra@eu.spb.ru

Ksenia Zaytseva
Austrian Centre for Digital Humanities, Austrian Academy of Sciences
Ksenia.Zaytseva@oeaw.ac.at

Kira Kovalenko
Institute for Linguistic Studies, Russian Academy of Sciences
Kira.Kovalenko@gmail.com

PhytoLex – the Database of Russian Phytonyms: from Idea to Implementation

Plants have always played an extremely important role in any traditional culture. They served as food, forage, medicine, material for building, clothes, dying, etc. Later some of them were domesticated to make their usage easier. Ancient plant names reflect mythological ideas and language worldview. Among sources, which can tell us about the role of plants in ancient cultures there are archeological findings, anthropological facts, and, of course, texts and inscriptions that allow us knowing plant names and plant knowledge.

In spite of the importance of the plant names investigation, collecting phytonyms is difficult and time consuming, especially for early periods. For some of them, we even do not know the time of their appearing in this or that language and/or their referents. The situation is especially difficult for the Russian language, as the first texts come from the 11th century, which is rather late; at that, most of them being translated from Greek and describing the culture of other people. Though there exist some databases of the early Church Slavonic and Old Russian literature, such as Historical sub-corpus inside the National corpus of the Russian language [1], Corpus Cyrillo-Methodianum Helsingiense – An Electronic Corpus of Old Church Slavonic Texts [2], Old Slavonic Corpus of the University of South California [3], and others, they do not provide semantic search. In fact, all modern sub-corpora of the National corpus of the Russian language have semantic search, while in the Historical part including the Old East Slavic, Birch Bark manuscript, the Old Russian, and Church Slavonic corpora the semantic search is supposed to be provided in the future only for the first one. That means that it is impossible now to have a list of all plant names occurred in the old texts, and that is the reason why most research projects are often based on limited amount of texts or just on the lexicographical materials.

The current project PhytoLex will create favorable conditions for the introduction of new materials into scientific use, for future comparative and typological studies on phytonymy, ethnobotany, folk taxonomy, folk medicine and magic. It will also help overcome fragmentation in Russian studies on folk botany, and provide their compliance with the level and requirements of modern ethnobotanical researches.

Collecting plant names starts from the earliest manuscripts of the Russian literature. The texts chosen for analysis are supposed to cover all the main Old Russian genres from 11th up to 17th centuries, such as religious literature, chronicles, travelogues, lexicographical works (lexicons and phrase-books), herbal books, medicine manuscripts, medical prescriptions and other papers of Apothecary Chancery (Rus. *Aptekarskij Prikaz*).

The sources are being well attributed, including author's name, title, text creation time, as well as time and place of the copy used by a researcher, and information about the book in case if the text was published. For identifying plants, we actively use historical dictionaries and academic books, articles, and theses concerning Old Russian plant names.

To give the full information about plant and its name(s), we create and fill the following description including standard, scientific, and Latin plant names, functions (food, medicine, etc.), metaphorical meaning (if any), word in simplified spelling (close to modern), citation (simplified and as in a source), life form, part of the plant which was mentioned in a text, ways of rendering the foreign phytonym (translation, transliteration, calque, generalisation, etc.) and its foreign etymon.

To make the data more unified, comparable and suitable for analysis, and also to avoid discrepancies, we created a number of controlled vocabularies which, for example, describe functions of plants (decoration, medicine), plant parts (branch, fruit, leaves, root, etc.), literature genres (chronicle, travelogue, herbal book), languages (used or mentioned in texts as plant names sources) and other attributes. In the process of controlled vocabularies creation we are following SKOS [4] standard recommendations and planning to link PhytoLex concepts to external existing thesauri, in particular to the General Multilingual Environmental Thesaurus (GEMET) [5].

The technical implementation of PhytoLex includes data modelling, creation and normalization of controlled vocabularies, development of database and web application for project's data curators and anonymous users on the web, visualization of available geographical data. The project also aims to integrate PhytoLex resources with open access resources like Geonames [6] for georeferencing places mentioned in manuscripts, and Catalogue of Life [7] for scientific name reference.

Overall, the main goal of PhytoLex project is to collect and harmonize data from analogue resources in order to make it available for exploration and analysis, access for further research and reuse.

Acknowledgements

The research is supported by the RFBR (the Russian Foundation for Basic Research), project 17-06-00376 "Russian Phytonyms in the Diachronic Aspect (11-17 cc.)".

References

1. National corpus of the Russian language, http://ruscorpora.ru/. Accessed February 16, 2018.
2. Corpus Cyrillo-Methodianum Helsingiense – An Electronic Corpus of Old Church Slavonic Texts, http://www.helsinki.fi/slaavilaiset/ccmh/. Accessed February 16, 2018.
3. The Historical Syntax of South Slavic, http://www-bcf.usc.edu/~pancheva/HistoricalSyntaxSouthSlavic.html. Accessed February 16, 2018.
4. SKOS Simple Knowledge Organization System Reference, W3C Recommendation, August 18, 2009, https://www.w3.org/TR/skos-reference/. Accessed February 16, 2018.
5. General Multilingual Environmental Thesaurus (GEMET), https://www.eionet.europa.eu/gemet/en/themes/. Accessed February 16, 2018.
6. GeoNames geographical database, http://www.geonames.org/. Accessed February 16, 2018.
7. Catalogue of Life, http://www.catalogueoflife.org/. Accessed February 16, 2018.

Lucie Kopáčková
Czech Language Institute of the Czech Academy of Sciences
kopackova@ujc.cas.cz

Oprahin or Opražin? How to Correctly Form Possessive Adjective from Female First Name or Surname of Foreign Origin in Contemporary Written Czech Language?

How to form possessive adjectives with the suffix *-in* describes every grammar book of Czech language, currently Štícha et al. (2013: 198). Usually, the suffix *-in* is added to the word base without the nominative ending with the consistent consonantic alternation. But how to correctly form these adjectives from some types of female first names of foreign origin shortly describes only Pravdová - Svobodová (2014: 231-233). There are some useful ideas how to form these adjectives, but some first names are missing on the list.

This paper shows how are these adjectives formed by authors of written texts (writers, translators, journalists etc.) in the contemporary Czech language. All presented linguistics data were found in the Czech National Corpus - SYN version 5. The data were obtained by following method. Possessive adjective with the suffix *-in* has its own tag („AU…F.* „); but most of the here presented variants are under the tag „X.* „. It is necessary to search them by a wordform and then all results sort manually, which is time-consuming and laborious process. (see note 1).

The comprehensive analysis showed that from one first name occasionally also from one surname (see note 2) there are often two or three variants. Some adjectives derived from first names with ending *-y (Daisy, Hillary)* or *-ey (Britney)* have as much as four different variants. Also, four variants occurred by names *Sarah* and *Rebecca*. (see note 3). The quantity of variants rises from the ignorance of some irregularities associated with the forming of possessive adjectives from names of foreign origin and probably also from the insecurity about the correct pronunciation of these sometimes exotic sounding names. It is possible to distinguish three basic problems:

1. The predominating uncertainty where and how to add the suffix *-in*

The suffix *-in* follows right after the unchanged name although the nominative ending should be removed (e.g. *Naomi → Naomiin, Maggie → Maggiein*) or the suffix *-in* does not follow the unchanged name because the nominative ending was inaccurately removed (e.g. *Hillary → Hillarin/Hillařin*).

2. The suffix *-in* was somehow modified

The suffix was reduced to *-n* (e.g. *Hillary → Hillaryn, Britney → Britneyn, Kálí→ Kálín*) or extended to *-nin* (e.g. *Daisy → Daisynin, Britney → Britneynin*).

3. The irregularity of consonantic alternation

There is irregular alternation of: *g* with *ž (Meg → Megin, Solange → Solangin*), *k* with *č* (*Brooke → Brookin* : *Broočin*), *r* with *ř (Ginger → Gingeřin : Gingerin*), *ch* with *š* (*Blanche → Blanchin, Uschi → Uschin*), *h* with *ž* (Oprah → *Oprahin : Opražin*) or *cc* [k] with *č* (Rebecca → *Rebečin : Rebeccin : Rebeccin : Rebečkin).*

Note 1

To get the list of adjectives of the type e.g. *Hillaryin* I needed to search the wordform *.*yin.** in the first step. This way I got 230 adjectives. Under the tag „AU…F.* „ you can find only 6 of them. In the second step I needed to search the wordform *Hillaryin.** to get the frequency. The same process was repeated for other possible variants e.g. *Hillařin.*, Hillarin.*, Hillaryn.*.*

By this approach I got lists of adjectives from following first names and surnames with vocalic endings and their possible variants: *-aa (Sanaain), -e (Salomein/Katein), -é (Beyoncéin), -ée (Renéein), -ie (Maggiein, Joliein), -ee (Breein), -oe (Chloein), -ue (Suein), -i (Naomiin), -í (Suguníin), -y (Hillaryin, Perryin), -ey (Britneyin, Winfreyin), -o (Join), -ó (Aikóin), -u (Kijivuin), -ou (Louin), -ú (Icúin).*

Very laborious was also searching for adjectives according to the wordform with all possible consonants from the Czech alphabet positioned in front of the suffix *-in*: from *.*bin.** to *.*žin.**. With this approach I added further adjectives to my list too.

Note 2

The corpus research also revealed a very interesting group of adjectives formed from foreign female surnames; in addition to the above e.g. Edith Piaf *→ Piafin,* Coco Chanel *→ Chanelin,* Agatha Christie *→ Christiin,* Gina Lollobridgida *→ Lollobridgidin.* They are not common. Their frequency is very low, and they appear mostly in newspaper articles. Their uniqueness primarily lies in fact, as I believe, they were never reflected in any research or article till today. From Czech female surnames we don't form possessive adjectives; this is the main cause why the possessive adjectives with the suffix *-in* are many times less numerous than with the suffix *-ův.*

Note 3

Beyoncé *→ Beyoncéin : Beyoncin;* **Jolie** *→ Joliein : Joliin;* **Oprah** *→ Oprahin : Opražin;* **Whoopi** *→ Whoopiin : Whoopin and others*

Audrey *→ Audreyin : Audrein : Audřin,* **Brooke** *→ Brookein : Brookin : Broočin;* **Tori** *→ Toriin : Torin : Tořin and others*

Daisy *→ Daisyin : Daisin : Daisyn : Daisynin;* **Hillary** *→ Hillaryin : Hillarin : Hillaryn : Hillařin;* **Britney** *→ Britneyin : Britnin : Britneyn : Britneynin;* **Sarah** *→ Sařin : Sarin : Sarahin : Saražin and others*

References

Pravdová, M. - Svobodová, I. (Eds.) (2014). *Akademická příručka českého jazyka.* Praha: Academia.

Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny.* Praha: Academia.

Electronic source

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A.: Korpus SYN, verze 5 z 24. 4. 2017. Ústav Českého národního korpusu FF UK, Praha 2017. Dostupný z WWW: http://www.korpus.cz

....

Natalia Kotsyba
Samsung R&D Poland; Institute for Ukrainian, NGO
gnatko@gmail.com

Bohdan Moskalevskyi
Institute for Ukrainian, NGO
msklvsk@icloud.com

An essential infrastructure of Ukrainian language resources and its possible applications

The paper presents an overview of a toolkit of Ukrainian language resources developed by our team and provides some examples of its application for lexicography and linguistic research. In particular, we share the experience of creating a Universal Dependencies (UD) (Nivre et al., 2016) treebank from scratch and stop briefly at organizational issues of managing professional annotators and students, as well as comment on technologies used.

The infrastructure includes:

- a custom-written tool for rapid manual morphological disambiguation with an annotator-friendly web interface supporting "2+1" type of workflow;
- a morphological tagger with rule-based morphological guesser, generating all possible interpretations for a word form, used to feed suggestions to manual annotators;
- a Brat-based (Stenetorp et al., 2012) system for syntax annotation with per-annotator statistics, fine-tuned for speed;
- respective UD-conforming annotation guidelines, with discussions available on Github;
- a gold standard treebank of general Ukrainian comprising 115K tokens, published as UD_Ukrainian;
- a rule-based treebank validator&autofixer exploiting language specifics (e.g. agreement), consisting of more than 250 hand-written rules;
- Enhanced Dependencies (Schuster and Manning, 2016) for UD_Ukrainian, generated by a custom-written script from the basic trees augmented with null nodes for elided predicates and with the distinction of shared/private dependents of a first conjunct;
- a trained UDPipe (Straka et al. 2017) and Stanford's CONLL'17 Shared Task Parser (Dozat et al., 2017) models served as a web visualization and as an API, having universal POS accuracy of 97.5%, full morphological features accuracy of 91.5% and Label Attachment Score of 81.5% (87% for Stanford);
- 3-gig web corpus created with a custom crawler from an online library of classics and major newspapers, and from the general web using Spiderling (Suchomel et al., 2012), filtered through custom post-processing rules, auto-annotated with UDPipe, served via Kontext (Institute of the Czech National Corpus) and Bonito (Rychlý, 2007) interfaces;
- pre-trained word embeddings over the 3-gig corpus with distance and analogy tasks served as a web GUI and an API;
- full Ukrainian localizations for Bonito and Kontext corpus interfaces, involving both rediscovered and newly-invented Ukrainian terminology;
- parallel corpora, manually aligned at sentence level with InterCorp-like (Čermák and Rosen, 2012) workflow: Uk-Polish, Uk-English, Uk-German, Uk-French, Uk-Portuguese, served as Kontext/Bonito, ranging in size significantly from 4M for Polish to 15K for Portuguese, rapidly growing;
- a basic valency dictionary, based on the largest existing so far explanatory dictionary of Ukrainian (СУМ) (Potebnia Institute of Linguistics, 1970–1980), currently containing about 65 thousand entries being projections of ca. 20 thousand lemmas;
- a coreference annotation over the gold standard (early stage, only 7% of the texts were annotated so far);
- a website linking everything together: mova.institute.

The presented infrastructure is aimed at a wide audience of both professional linguists and any users/learners of Ukrainian. A special focus is made on popularizing corpora resources among academia and creative community, also by means of (video) tutorials and blog entries dedicated to specific linguistic phenomena.

All aforementioned resources were developed by Institute for Ukrainian, NGO, a grassroots initiative of linguists and software developers. Each resource is publicly available and is free for non-commercial use.

References

Čermák, F. and Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13 (3), 411–427.

Dozat, T., Qi, P. and Manning, C. D. (2017). Stanford's graph-based neural dependency parser at the CoNLL 2017 Shared Task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies,* 20–30.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, 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. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 1659–1666.

Potebnia Institute of Linguistics. (1970–1980). *Slovnyk ukrainskoi movy* [Dictionary of the Ukrainian Language]. Kyiv: Naukova dumka.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University, 65–70.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* 2371–2378.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Suchomel, V., Pomikálek, J., et al. (2012). Efficient web crawling for large text corpora. In *Proceedings of the 7th Web as Corpus Workshop*, 39–43.

••••

Anna Kryvenko
Kyiv National Linguistic University
annakry@fulbrightmail.org

A reference corpus for discourse dynamics analysis in Ukrainian?

Although there might be no really bad reference corpus for the extraction of "a useful set" of keywords (Scott 2009: 79), the point of departure for this research is the awareness of the relational nature of keywords and, therefore, their potentially high sensitivity to data segmentation, when the purpose is to observe the socio-cultural significance of lexical choices in a specialized discourse from a dynamic perspective. Recent studies have shown that on top of a more general concern about suitability – when a specialized target corpus representing a particular type or genre of discourse is compared to a more general reference corpus supposedly representing all the features of a language (see, e.g., Leech 1998; Aston 2001; Burnard 2002), there are more particular issues of dividing up the data, when specialized corpora are used both as target corpora and as reference corpora (Gabrielatos et al. 2012; Marchi 2018). However, there seems to be not only "remarkably little discussion … about the effects of time segmentation in diachronic discourse analysis using corpora" (Marchi 2018: 174), but also the dearth (to the best of my knowledge) of focus on the effects of dispersion segmentation, i.e. sub-corpora grouped in accordance with dispersion and / or frequency rates of particular linguistic signs in individual texts.

This paper reports on some preliminary findings from my ongoing study, which aims to expose the dynamics of interdiscursivity observed in European integration discourse in national and supranational institutions. The reported fragment of the study focuses on a custom-built corpus featuring texts in the Ukrainian language from the official website of the Verkhovna Rada of Ukraine (the Ukrainian Parliament). This corpus contains over 1.25 million tokens and consists of over 2500 full-size texts explicitly mentioning European integration, which were posted between 2002 and 2017. The genres include parliamentary news, minutes of plenary sittings, hearings and committees meetings, Speaker's addresses, agendas, reports, announcements, etc. The corpus was lemmatized and tagged for POS by the developers of the Large Electronic Dictionary of the Ukrainian Language (VESUM) (Starko 2017).

For the purposes of this study, the corpus was segmented in two different ways. First, it was divided into sixteen sub-corpora, each representing a separate year respectively, and then each sub-corpus was alternately compared with the other fifteen. The described approach draws on the existing practice of using two chronologically distinct sub-corpora of some specialized discourse for comparisons against each other (e.g., Marchi and Taylor 2009; Murakami et al. 2017). Second, the texts in the corpus were grouped into three sub-corpora depending on the number (1, 2–3, or ≥4) of explicit references to European integration in each individual text, with each sub-corpus being alternately compared with the other two. The goal was to identify similar and characteristically different keywords with the timeline and dispersion of recurrence in mind. The software used in this study was AntConc 3.5.7 2018 and the metrics included log-likelihood for keyness (significance) and %DIFF for effect size measure. The results of these searches help to detect continuities, discontinuities and ruptures in patterns of use, which constitute and are constituted by discourse in a wider social context.

The methodological framework of my study rests on the recently established custom to combine quantitative and qualitative approaches to discourse analysis. Primarily, advantage is taken of modern diachronic corpus-assisted discourse studies (MD-CADS), a novel research discipline that pursues changes over recent time in lexical and grammatical patterns of use in corpora capturing a particular sphere of communication, but also it intends to account for extra-linguistic changes that language reflects (Marchi and Taylor 2009; Partington et al. 2013: 265–322). A critical discourse analysis perspective on corpus data (in terms of Baker and McEnery 2015: 5) is utilized here as well, particularly following the Discourse-Historical Approach with its emphasis on the interdependence between discourse and socio-political change (Wodak 2018).

References

Aston, G. (2001). Text categories and corpus users: a response to David Lee. *Language Learning & Technology*, 5 (3):73–76.

Baker, P. & McEnery, T. (2015). Introduction. In Baker, P. & McEnery, T. (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (Palgrave Advances in Language and Linguistics). Basingstoke: Palgrave Macmillan, 1–19.

Burnard, L. (2002). A retrospective look at the British National Corpus. In B. Kettemann and G. Marko (Eds.), *Language and Computers: Studies in Practical Linguistics*. Amsterdam: Rodopi, 51–70.

Gabrielatos, C. et al. (2012). The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics,* 17(2), 151–175.

Leech, G. (1998). Preface to 'Learner English on Computer'. In S. Granger (Ed.), *Learner English on Computer*. London: Addison-Wesley-Longman, xiv–xx.

Marchi, A. & Taylor, C. (2009). Establishing the EU: the representation of Europe in the press in 1993 and 2005. In A. Jucker, M. Hundt & D. Schreier (Eds), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, 201–224.

Marchi, A. (2018). Dividing up the data. In Ch. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review*. London & New York: Routledge, 174–196.

Murakami, A. et al. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora,* 12 (2), 243–277.

Partington A.S., Duguid A. & Taylor C. (2013). *Patterns and Meanings in Discourse. Theory and Practice in Corpus-Assisted Discourse Studies (CADS).* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Scott, M. (2009). In search of a bad reference corpus. In Archer D. (Ed.) *What's in a Word-list? : Investigating Word Frequency and Keyword Extraction* (Digital Research in the Arts and Humanities). London & New York: Routledge, 79–91.

Starko, V. (2017). Computational linguistics projects of the R2U team: present sate and applications, *Ukrainian Language* 3, 86–100 / Starko V. Kompyuterni linhvistychni proekty hurtu R2U: stan ta zastosuvannya, *Ukrayinska mova*, 2017, No. 3. S. 86–100.

Wodak, R. (2018). Discourse and European integration. MIM Working Series, 18 (1) (Willy Brandt Series of Working Papers), 3–29.

••••

Miroslav Kubát
University of Ostrava, Czech Republic
miroslav.kubat@gmail.com

Jan Hůla
University of Ostrava, Czech Republic
jan.hula21@gmail.com

Radek Čech
University of Ostrava, Czech Republic
cechradek@gmail.com

David Číž
University of Ostrava, Czech Republic
davidciz95@gmail.com

Kateřina Pelegrinová
University of Ostrava, Czech Republic
pelegrinovak@gmail.com

Context Specificity of Lemma. Diachronic analysis

The study deals with the application of the neural networks in the linguistic research of word semantics. A recently proposed method of measuring Context Specificity of Lemma (Čech et al. 2018) based on Word Embeddings Word2vec technique (Mikolov et al. 2013) is introduced and illustrated in the analysis of the selected lemmas from various fields (e.g. political discourse or IT). The research is based on the fourth version of SYN series corpora of Czech National Corpus (Hnátková et al. 2014). The results indicate that the method is applicable for detecting the semantic development of a lemma and it could have a potential for linguistic studies. Although neural networks are generally black-box methods, our approach enables the linguistic interpretation of the obtained results. The aim of this contribution is to introduce a method which can detect semantic changes of a lemma from the diachronic viewpoint.

In word embedding methods, each lemma is represented by a vector. The size and the orientation of a vector express the position of a lemma in a semantic multi-dimensional space. Therefore, it is possible to measure similarities among lemmas. If, in an ideal case, there are two lemmas which occur in the identical contexts in the whole corpus, the size and orientation of these two vectors would be identical and, thus, the distance between these two lemmas equals to zero or, reversely, the similarity between them equals one. In the reality, each lemma occurs in different contexts, consequently, they are represented by different vectors which enables us to compute similarities among them.

The method Context Specificity of Lemma (CSL) measures how unique is the context in which the lemma appears in the corpus. Specifically, if the lemma occurs in many different contexts, it will have low context specificity. The context in which the lemma appears is captured with a distributed vector representation which is assigned to every lemma. In this vector representation, it is possible to measure the similarities among lemmas. To be more specific, it means that for each lemma, we can compute its similarity to all other lemmas. Statistics of these similarities (e.g., a mean value) can be used for characterizing the Context Specificity of Lemma. The lower the mean of similarities, the higher the CSL.

Neural networks need huge training data sets to be capable of producing reliable results. We therefore decided to use one of largest Czech corpora - the fourth version of SYN series corpora (Hnátková et al. 2014). The size of the SYN_V4 is 3,626 billion tokens. The SYN corpus is not representative; the dominant component is journalism. Beside journalism there are other two text types: fiction and technical literature. Only journalistic texts were selected for the analysis. The final corpus of our study consists of more than 3 billion tokens (3,045,389,630) and more than one hundred thousand types (102,707). In order to avoid a bias caused by low frequencies, all lemmas with frequency less than 70 were omitted ($f \leq 69$). Since the goal is to analyse diachronic development of the CSL, we divided the data into 19 subcorpora that each represents one year. Only the subcorpus 1990-1996 consists of texts from several years because of the small data sizes.

References

Čech, R., Hůla, J., Kubát, M., Chen, X., Milička, J. (2018). The Development of Context Specificity of Lemma. A Word Embeddings Approach. *Journal of Quantitative Linguistics.*

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the*

54th Annual Meeting of the Association for Computational Linguistics. 1489–1501.
Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavík: ELRA, 160–164
Mikolov, T., Chen, K., Corrado, G.S., Dean, J., Sutskever, I. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of Neural Information Processing Systems (NIPS 26)*, 3111–3119.

....

Moulay Zaidan Lahjouji
Slavisches Seminar, Albert-Ludwigs-Universität Freiburg, Germany
zaidan.lahjouji@slavistik.uni-freiburg.de

The Corpus of Spoken Rusyn – A user-friendly resource for research on Rusyn dialects

This paper aims to report on current developments in the design of and the work with a user friendly and sustainable dialect corpus. It will discuss functional and technical features of a corpus architecture that allows smaller projects with limited work force to build a corpus, which meets the abovementioned expectations. Therefore, some features of *The Corpus of Spoken Rusyn* (http://www.russinisch.de/Varchola2/) are presented. The corpus is a result of a DFG-funded research project on Rusyn dialects.

The goal of the project is to investigate dynamic processes in the East Slavic minority language Rusyn, which is spoken predominantly in South-West Ukraine, East Slovakia, South-East Poland and Northern Hungary. The setting of Rusyn language is divided by several political and linguistic borders as it not only marks a point between East and West Slavic, but also spreads across four national borders as well as the outer border of the EU.

Nowadays, the speakers of Rusyn live to a greater extent in a dynamic environment and under constant and evident pressure by their respective roofing state languages Ukrainian, Polish, Slovak or Hungarian. In this fashion, new divergences within the old Rusyn dialect continuum due to contact with the majority language, that is, so-called border effects, are to be expected (Rabus, 2015; Woolhiser, 2005).

Not only in order to trace these divergences for our project but also to make Rusyn vernacular more accessible for further empirical research, the Corpus of SpokenRusyn(http://www.russinisch.uni-freiburg.de/corpus,Rabus&Šymon, 2015) has been created. The Corpus of Spoken Rusyn is a collection of Rusyn vernacular speech from different regions across the Carpathian Mountains. It consists of several hours of audio recordings with accordingly transcribed speech. The recordings were made in Poland, Slovakia, Ukraine, and Hungary in 2015 (Šymon & Rabus 2015/2016).

The current size of the corpus is more than 100.000 tokens (only informants, including interviewers and notes more than 140.000).[1]

As mentioned above, we endeavour to keep the corpus and its architecture simple, user friendly and easily accessible even beyond the duration limits of our project. The corpus has been built with SpoCo (Waldenfels & Woźniak, 2017) architecture. SpoCo is an easy to use and to maintain system for web-based query for corpora based on standard XML input files.

The great advantage of this system is that it is easy to adopt, also for projects that don't have big resources on their disposal for, e.g., programming a custom server-based web-infrastructure for a corpus. It has been developed for the *VMČ-Corpus* (http://www.vmc.uni-freiburg.de/Mens/, Rabus et Al. 2012)) and the *Ustya River Basin Corpus* (http://parasolcorpus.org/Pushkino, (von Waldenfels et Al. 2014)) and was since then adopted by several other dialect corpora.

The functionality of SpoCo is based on *open CWB* (Evert and Hardie, 2011) but with some extended features. Beside standard input fields (with a graphical keyboard including special Cyrillic characters) for concordance search, our corpus offers a graphical user interface with dropdown lists (for areal, personal and meta informational settings) as well as a CQP Search engine that displays the respectively CQP command in a command line below, corresponding to the information entered in the query and the dropdown lists. This allows users who are not acquainted with CQP search, to enter first simple search queries intuitively and learn to adapt more complex and detailed search commands.

There have been several approaches to morphological tagging of the spoken Rusyn data. The lack of tools for automated tagging of Rusyn, annotated parallel data as well as the orthographically (and morphologically) heterogeneous nature of our transcriptions of spoken data have complicated the process of developing automatic annotation tools. Nevertheless, several training-based efforts e.g. with multi-source approaches on morphosyntactic tagging and the help of the MarMoT (Mueller et al., 2013) have led to respectable results with higher accuracy, as shown in Scherrer & Rabus 2017. Still there is room for improvement and we still endeavour to achieve a higher accuracy rate. The tags can be seen by hovering the mouse over the respectively desired token, found by using query search or the context function of our corpus.

1 Status of 28.08.2018

After executing a query search, users are also able listen to original but anonymized recordings, whereas registered users are allowed to download the segmented WAV-files for further examination, e.g., in PRAAT. Registered users are allowed to edit transcriptions in the corpus, which allows us to find community-based solutions for occurring transcription or tagging issues.

Furthermore, examples of query searches within the corpus will be provided in this paper, in order to make the technical details described above more tangible. The examples will cover a broader field of rather simple lexical searches but also more complex CQP based searches of word forms or grammatical features of the Rusyn varieties. We will intensify the focus on the workflow of our corpus research as well as we will show tendencies, that support the thesis of our project as far as border-effects are concerned.

References

Christ, Oliver (1994). A modular and flexible architecture for an integrated corpus query system. In: *Proceedings of COMPLEX'94: 3rd Conference on Computational Lexicography and Text Research*, 23–32.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In: *Proceedings of the Corpus Linguistics 2011 Conference, Birmingham*, UK. University of Birmingham.

Mueller, T; Schmid,H & Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA, October. Association for Computational Linguistics.

Rabus, A. & A. Šymon (2015): Na novŷch putjach isslidovanja rusyns'kŷch dialektu. Korpus rozhovornoho rusyns'koho jazŷka. In: Koporová, Kvetoslava (Hrsg.): *Rusyn'skŷj literaturnŷj jazŷk na Slovakiji. 20 rokiv kodifikaciji.* Prešov, 40-54.

Rabus, Achim (2015): Current Developments in Carpatho-Rusyn Speech - Preliminary Observations. In: Krafcik P. & V. Padjak (eds.): *Ƶuvilejnyj zbirnyk na čest' profesora Pavla-Roberta Magočija.* Užhorod, 489-496.

Rabus, A. & Scherrer, Y. (2017): Lexicon Induction for Spoken Rusyn - Challenges and Results. In: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, 27-32.

Rabus, A., Savić, S., Waldenfels, R. v. (2012). Towards an electronic corpus of the Velikie Minei Čet'i. In: *Rediscovery: Bulgarian Codex Suprasliensis of the 10th century*. Sofia: Iztok Zapad.

Scherrer, Y & Rabus, A (2017): Multi-source morphosyntactic tagging for spoken Rusyn. In: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 84 – 92.

Schimon, A. & A. Rabus (2016): Wahrnehmungsdialektologische Untersuchungen zum Russinischen in Zakarpattja am Beispiel der Region Chust. In: *Zeitschrift für Slawistik* 61(3), 401-432.

Šymon, A. & A. Rabus (2016): Ysslidovanja rusyns'koho jazŷka yz pohljada vospryymatel'noji dialektologiji. In: *Dynamické procesy v súčasnej slavistike*, S. 71-88. (Nachdruck in Rusyn 5/2016 und 6/2016)

v. Waldenfels, R.; Woźniak, M. (2017). SpoCo – a simple and adaptable web interface for dialect corpora. In: *Journal for Language Technology and Computational Linguistics*, 31(1), 145 – 160.

v. Waldenfels, R.; Daniel, M., Dobrushina, N. (2014): Why Standard Orthography? Building the Ustya River Basin Corpus, an online corpus of a Russian dialect. *Komp'juternaja lingvistika i intellektual'nye technologii: Po materialam ežegodnoj Meždunarodnoj konferencii «Dialog»* (Bekasovo, 4 — 8 ijunja 2014 g.). Vyp. 13 (20). — M.: Izd-vo RGGU, 2014.

Woolhiser, C. (2005). Political borders and dialect divergence/convergence in Europe. In Peter Auer, Frans Hinskens, and Paul Kerswill, editors, *Dialect change*, 236–262. Cambridge Univ. Press, Cambridge

....

Nikola Ljubešić
Jožef Stefan Institute, Slovenia
nikola.ljubesic@ijs.si

Tanja Samardžić
University of Zürich, Switzerland
tanja.samardzic@uzh.ch

Tomaž Erjavec
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Darja Fišer
Faculty of Arts, University of Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Maja Miličević Petrović
Faculty of Philology, University of Belgrade, Serbia
m.milicevic@fil.bg.ac.rs

Simon Krek
Jožef Stefan Institute, Slovenia
simon.krek@ijs.si

"Kad se mnogo malih složi": Collaborative development of gold resources for Slovene, Croatian and Serbian

Introduction

Textual datasets manually annotated with linguistic information are a backbone of the currently dominating paradigm in natural language processing based on machine learning. In this abstract we present a series of collaborations between researchers developing such datasets for Slovene, Croatian and Serbian. Close relatedness of these languages brings an opportunity for a synchronized approach to the development of resources and technologies, to the benefit of all parties. Due to the complex political environment, however, such an approach has not been established. The main synergistic effect

of the collaborations presented here is achieved by drastically lowering the efforts required to produce corpora in additional languages, primarily in the areas of (1) the development of annotation guidelines, (2) setting up the technical requirements for the annotation campaigns and (3) pre-annotation of data with models trained for another, but very close language.

2 Collaborations

2.1 Morphosyntax

One of the first transfers between our languages of interest was a long overdue update of the MULTEXT-East morphosyntactic tagset definitions for Croatian and Serbian. Up to that point, Croatian and Serbian had significantly different tagsets. This introduced artificial obstacles to their cross-linguistic processing, although it was shown that applying Croatian models on Serbian data generates only a minor loss in the quality of annotation [Agić et al., 2013]. New, almost identical tagset definitions were thus proposed, both heavily relying on the recently introduced Slovene tagset.

The collaboration continued in the scope of the Abu-MaTran and ReLDI projects on a simultaneous development of inflectional lexicons for Croatian [Ljubešić et al., 2016a] and Serbian [Ljubešić et al., 2016b] by exploiting paradigm predictions learned on the union of already available data [Ljubešić et al., 2015]. Lexicon entries were also heavily reused between the two resources by encoding the systematic variation in the yat vowel and specifying whether the lexeme is specific for any of the languages [Ljubešić et al., 2016].

Finally, the Serbian SETimes.SR corpus [Samardžić et al., 2017] with a morphosyntactic gold annotation layer was created through manual correction of the labels automatically introduced with a model trained on the parallel Croatian dataset [Agić and Ljubešić, 2014], which generated highly accurate annotations. For performing the annotation corrections, the WebAnno tool [Yimam et al., 2013] hosted by the CLARIN.SI infrastructure was used.

2.2 Dependency syntax

Universal Dependency annotation was added to the SETimes.SR dataset [Samardžić et al., 2017] using a similar approach as above: while adding the Universal Dependencies layer, the Serbian dataset was again preannotated with a model trained on the parallel Croatian dataset [Agić and Ljubešić, 2015], with only 15% of tokens requiring manual interventions.

The UD annotation efforts for Croatian and Serbian on one side and Slovene on the other are currently not coordinated, but this is planned as an additional synergy in the future.

2.3 Basic processing of social media language

In the scope of the Slovene national Janes project, two manually annotated datasets, Janes-Norm [Erjavec et al., 2016] and Janes-Tag [Erjavec et al., 2017] were developed for training and testing the basic annotation layers of Slovene user-generated content (UGC), namely tokenization, sentence splitting, normalization, morphosyntactic tagging, and lemmatization. The work on these datasets comprised writing detailed annotation guidelines, training the annotators, sampling the data and performing multiple annotations and curations in WebAnno [Yimam et al., 2013].

Within the ReLDI project, the Slovene annotation guidelines were translated into Serbian and an annotation campaign similar to the one for Slovene was performed for Croatian and Serbian UGC, leading to two new datasets ReLDI-NormTagNER-hr [Ljubešić et al., 2017a] and ReLDI-NormTag-sr [Ljubešić et al., 2017b]. Given the high complexity of the annotation campaign, reusing the annotation guidelines and the annotation technology proved to drastically lower the efforts necessary to produce the final datasets.

2.4 Named entity recognition

In the scope of producing version 2.0 of the ssj500k training corpus of Slovene [Krek et al., 2017], NE annotation guidelines were written, the existing NE annotations in ssj500k were checked and an additional portion of the corpus was NE annotated.

In the ReLDI project, these guidelines were, with minor extensions, applied to the hr500k corpus of standard Croatian [Ljubešić et al., 2018] and to the SETimes.SR corpus of standard Serbian [Samardžić et al., 2017].

The guidelines were then, within the Janes project, also applied to Slovene UGC in the already mentioned Janes-Tag corpus. Finally, within ReLDI, the above-mentioned Croatian and Serbian UGC datasets (ReLDI-NormTagNER-hr and ReLDI-NormTagNER-sr) were also manually annotated with NEs.

All the annotation projects were performed on the Webanno instance of the CLARIN.SI infrastructure.

2.5 Semantic role labeling

Given the previously observed high synergistic effect, a bilateral Slovene-Croatian project was proposed on collaborative development of semantic role labeling for Croatian and Slovene. Inside this project, joint annotation guidelines were developed and annotation campaigns were run on the ssj500k [Arhar Holdt, 2009] and the hr500k [Ljubešić et al., 2018] datasets, using the same annotation technology.

References

[Agić and Ljubešić, 2014] Agić, Ž. and Ljubešić, N. (2014). The SETimes.HR linguistically an- notated corpus of Croatian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Agić and Ljubešić, 2015] Agić, Ž. and Ljubešić, N. (2015). Universal Dependencies for Croatian (that Work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, volume 1–8.

[Agić et al., 2013] Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria. Association for Computational Linguistics.

[Arhar Holdt, 2009] Arhar Holdt, S. (2009). Učni korpus SSJ in leksikon besednih oblik za slovenščino. *Jezik in slovstvo*, 54(3/4):43–56.

[Erjavec et al., 2016] Erjavec, T., Fišer, D., Čibej, J., and Arhar Holdt, Š. (2016). CMC training corpus Janes-Norm 1.2. Slovenian language resource repository CLARIN.SI. http://hdl. handle.net/11356/1084.

[Erjavec et al., 2017] Erjavec, T., Fišer, D., Čibej, J., Arhar Holdt, Š., Ljubešić, N., and Zupan, K. (2017). CMC training corpus Janes-Tag 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1123.

[Krek et al., 2017] Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., and Kuzman, T. (2017). Training corpus ssj500k 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1165.

[Ljubešić et al., 2018] Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., and Erjavec, T. (2018). Training corpus hr500k 1.0. Slovenian language resource repository CLARIN.SI. http://hdl. handle.net/11356/1183.

[Ljubešić et al., 2017a] Ljubešić, N., Erjavec, T., Miličević, M., and Samardžić, T. (2017a). Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1170.

[Ljubešić et al., 2017b] Ljubešić, N., Erjavec, T., Miličević, M., and Samardžić, T. (2017b). Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource reposi- tory CLARIN.SI. http://hdl.handle.net/11356/1171.

[Ljubešić et al., 2015] Ljubešić, N., Esplà-Gomis, M., Klubička, F., and Preradović, N. M. (2015). Predicting Inflectional Paradigms and Lemmata of Unknown Words for Semi-automatic Ex- pansion of Morphological Lexicons. In *Proceedings of Recent Advances in Natural Language Processing*, pages 379–387.

[Ljubešić et al., 2016a] Ljubešić, N., Klubička, F., and Boras, D. (2016a). Inflectional lexicon hrLex 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1072.

[Ljubešić et al., 2016b] Ljubešić, N., Klubička, F., and Boras, D. (2016b). Inflectional lexicon srLex 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1073.

[Ljubešić et al., 2016] Ljubešić, N., Klubička, F., Željko Agić, and Jazbec, I.-P. (2016). New in- flectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

[Samardžić et al., 2017] Samardžić, T., Starović, M., Agić, Ž., and Ljubešić, N. (2017). Universal dependencies for Serbian in comparison with Croatian and other Slavic languages. In *The 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*.

[Yimam et al., 2013] Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible,web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

••••

David Lukeš
Charles University, Czech Republic
david.lukes@ff.cuni.cz

Zuzana Komrsková
Charles University, Czech Republic
zuzana.komrskova@ff.cuni.cz

Marie Kopřivová
Charles University, Czech Republic
marie.koprivova@ff.cuni.cz

Petra Poukarová
Charles University, Czech Republic
petra.poukarova@ff.cuni.cz

Pronunciation of casual spoken Czech: A quantitative survey

Introduction

Influenced by fairly regular correspondences between phonology and spelling, speakers of Czech will often insist words are pronounced fully even when significantly formally reduced: e.g. *protože* 'because' is heard as [protože] even when pronounced as something more akin to [bže]. And yet, reduced pronunciation variants are ubiquitous, especially in private everyday conversations (Ernestus and Warner 2011), where linguistic self-consciousness is downplayed and the amount of context shared between participants allows them to take especially drastic shortcuts in speech production without hindering understanding (see Lindblom 1990 for an analysis of the tradeoff between the effort spent in speech production and perception). This paper proposes to give a quantitative survey of pronunciation variation in casual Czech in general and formal reduction trends in particular, based on data from the ORTOFON corpus.

From a theoretical perspective, reduced pronunciation variants present a challenge to traditional segment-based representations in phonetics, because in transcribing them, it is often very difficult to determine the constituent segments and their boundaries. The phenomenon is often described in terms of interactions between segments, e.g. as "parallel articulation" (Machač and Zíková 2015), or even as "articulatory prosody" in a conscious attempt to transition to "a paradigm which makes the traditional segment–prosody divide more permeable, and moves away from the generally practiced phoneme orientation" (Kohler and Niebuhr 2011, 1).

Data and methodology

The ORTOFON corpus is the first publicly available (Kopřivová et al. 2017a, 2017b) corpus of casual spoken Czech with a dedicated manual phonetic transcription layer. While earlier hand-annotated data sets obviously exist, access to them is generally restricted. Also, they tend to be smaller in geographic and demographic scope and obtained in formal or semi-formal settings (e.g. interviews). By contrast, ORTOFON focuses on decidedly informal speech, as encountered in private conversations between friends and relatives, and attempts to cover the broadest possible range of regional and sociological backgrounds (Kopřivová et al. 2014).

For the purpose of this study, we have complemented the manual phonetic transcriptions with rule-generated ones which approximate standard pronunciation expected in careful speech.

Results and discussion

It has been shown for multiple languages that reduction likelihood can be predicted based on the frequency and length of a given word form (see e.g. Mitterer 2008 for Dutch and German). Lexical effects are also observed, whereby some lexical items seem to be more disposed towards reduction than others. As these criteria are fairly universal and stem from dynamics that apply across languages, it is not surprising that our preliminary results for Czech point towards the same patterns.

Fig. 1 shows a frequency breakdown of the number of pronunciation variants per word form. Only variants attested at least 5 times were included in order to focus on reliable, repeatedly occurring pronunciations and disregard potential errors. As expected, higher frequency and longer words stand out as those which are particularly inclined towards variation (see the crest of the scatter plot: *normálně*, *vůbec*, *úplně*, *prostě*), partially also because more occurrences of a given type represent more opportunities for varied

pronunciation. However, thanks to the variant frequency threshold, we have hopefully limited the number of highly idiosyncratic variants which would spuriously inflate the count. One word which particularly sticks out with respect to other words with similar frequencies is the aforementioned *protože*; it remains to be seen (by performing multivariate analyses) whether this can be ascribed to other regular predictors like word length, or whether this is perhaps a lexical effect.

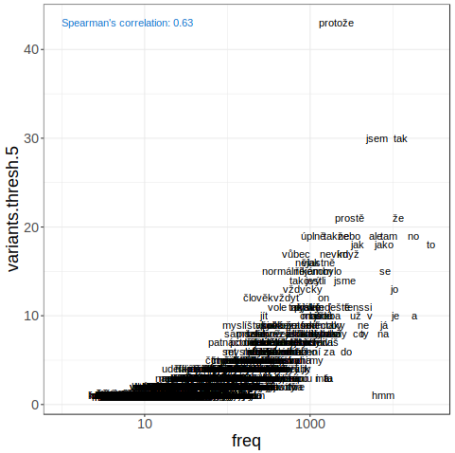

Fig. 1. Number of pronunciation variants per word form vs. its overall frequency in corpus. Only variants occurring at least 5 times were included.

Fig. 2 then shows the average normalized Levenshtein edit distance (Yujian and Bo 2007) between the canonical (rule-derived) pronunciation of a word form and the various actual pronunciations encountered in the wild. This metric aims to be word-length independent by measuring distance in proportion to the length of the entire word. As an average, it also obviates the potential problems arising from comparing items with different absolute frequencies. A positive correlation can be seen here as well: the higher the frequency, the higher the variability. Highly frequent words can be more easily predicted from context, therefore their pronunciation can vary widely (e.g. in response to the surrounding words, in order to make articulation easier) without hampering recognizability. This is particularly true of the very short function words in the upper right corner of the figure.

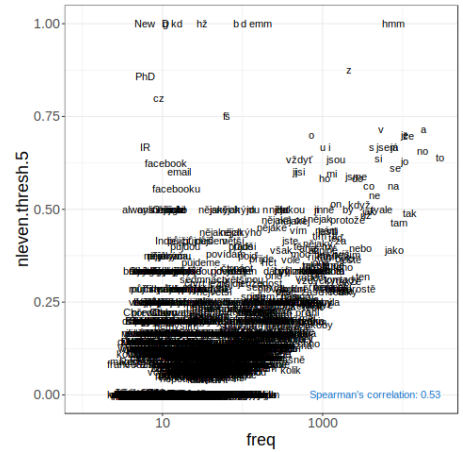

Fig. 2. Average normalized Levenshtein distance between canonical (rule-derived) and actually observed pronunciation of a word form vs. its overall frequency in corpus. Only variants occurring at least 5 times were included.

References

Ernestus, M. & N. Warner. (2011). An Introduction to Reduced Pronunciation Variants. *Journal of Phonetics*, 39, 253–60.

Kohler, K. J. & O. Niebuhr. (2011). On the Role of Articulatory Prosodies in German Message Decoding. *Phonetica*, 68 (1-2), 57–87.

Kopřivová, M., P. Klimešová, H. Goláňová, and D. Lukeš. (2014). Mapping Diatopic and Diachronic Variation in Spoken Czech: The ORTOFON and DIALEKT Corpora. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA), 376–82.

Kopřivová, M., Z. Komrsková, D. Lukeš, P. Poukarová & M. Škarpová. (2017a). ORTOFON V1: Balanced Corpus of Informal Spoken Czech with Multi-Tier Transcription (Transcriptions). http://hdl.handle.net/11234/1-2580.

Kopřivová, M., Z. Komrsková, D. Lukeš, P. Poukarová & M. Škarpová. (2017b). ORTOFON V1: Balanced Corpus of Informal Spoken Czech with Multi-Tier Transcription (Transcriptions & Audio). http://hdl.handle.net/11234/1-2579.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. Dordrecht, Boston, London: Kluwer Academic Publishers, 403–439.

Machač, P., & M. Zíková. (2015). Parallel Articulation: The Phonetic Base and the Phonological Potentiality. *Slovo a Slovesnost*, 76 (1), 3–21.

Mitterer, H. (2008). How Are Words Reduced in Spontaneous Speech? In A. Botinis (Ed.), *Proceedings of the ISCA Tutorial and Research Workshop on Experimental Linguistics*. Athens: University of Athens, 165–68.

Yujian, L., & Bo, L. (2007). A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (6), 1091–1095.

....

Lucie Lukešová (Chlumská)
Charles University, Czech Republic
lucie.chlumska@ff.cuni.cz

Dominika Kováříková
Charles University, Czech Republic
dominika.kovarikova@ff.cuni.cz

Extracting Multi-word Expressions for the Czech Academic Phrase List

Introduction

As John Sinclair suggested with his "idiom principle" almost thirty years ago (Sinclair 1991), our language production is to a great extent phraseological in nature. Morley (2014: 4) claims that much of the language, including the academic domain, is "acquired, stored and retrieved as pre-formulated constructions". Some studies (e.g. Wei & Li 2013: 522) indicate that the proportion of phraseological units in English is even higher than 50 %.

Academic language as a specific genre where the form has to adequately complement the highly accurate content, using formulaic expressions (incl. terms), seems to be a good example of this tendency. It has been a common practice, especially in English speaking countries, to publish academic word lists and phrase lists to help language users acquire fluency in their own academic writing. Since there is no such material available for contemporary Czech, we decided to produce a corpus-based academic phrase list as a resource for Czech university students and Czech Studies scholars both here and abroad.

In our paper, we would like to describe the process of academic phrases extraction and comment on the resulting phrase list.

Academic Czech

The concept of academic language is not particularly rooted in Czech linguistics. Contrary to Hoffmannová et al. (2016: 182), we do not distinguish between scientific/research texts ("vědecké") and academic ("akademické") texts; rather, we have adopted a broader approach to academic Czech cov-

ering a wide range of publications (research articles and monographs, textbooks as well as students' theses) as we believe any of these resources can be helpful in providing a general picture of the formulaic language used in academic environment.

Czech Academic Phrase List

The main objective of our study is to generate a corpus-based list of the most frequent multi-word expressions that are commonly used in academic Czech (in the broad sense) and are not limited to one scientific discipline; i.e. we are searching for a common denominator of all academic texts, regardless of the research field, such as *provést experiment* ('to carry out an experiment'). To reflect the English tradition, we decided to call these multi-word units of all sorts "academic phrases" (cf. Morley 2014), resp. *akademické fráze* in Czech.

Corpus data

For the purpose of this study, we created a subcorpus of the latest available corpus of contemporary written Czech (SYN2015) containing only scientific and academic texts (labelled as SCI). This SCI corpus has app. 10 million words and consists mostly of published research books (esp. monographs) and (university) textbooks.

To identify which phrases are typical only for academic discourse, we used a reference corpus containing Czech newspapers and fiction (almost 70 million words).

Methodology

First of all, we extracted the most frequent n-grams (2-grams, 3-grams and 4-grams) from the SCI subcorpus. Since Czech is a language with a relatively free word-order (for a discussion on the related issue of n-gram extraction in Czech, see e.g. Čermáková & Chlumská 2017), this procedure required taking this variability into account by including all possible n-gram variations (e.g. j*edná se o…* / *se jedná o…*, 'it is…').

Then we applied two filters: first we compared the n-grams frequency in SCI with the frequency in the reference corpus to rule out n-grams used in all written Czech texts. Second, we measured their distribution in individual scientific disciplines to remove specialized multi-word terms from the list as the distribution in academic disciplines proved to be one of the most useful criteria in term identification (Kováříková 2017).

For the core list, only those expressions present in all 24 disciplines of SCI were used and further classified (e.g. using part-of-speech tags and MI-score), resulting in a list of approximately 2,000 items that will be grouped and described in detail.

In the presentation, we will also briefly comment on some interesting groups of words with lower distribution (esp. longer n-grams) as these may also be useful for academic language users.

Preliminary results

Based on our first experiments with the data, we can tentatively distinguish the following types of phrases in the list:

- non-specific multi-word terms with a high distribution across academic disciplines (e.g. *empirický výzkum* 'empirical research', *statistická analýza* 'statistical analysis')
- collocations (e.g. *provést experiment* 'carry out an experiment')
- multi-word prepositions (*s ohledem na* 'with regard to', *v rámci* 'within')
- multi-word linking words (*a i přesto* 'and even though') and discourse markers (*na jedné straně* 'on one hand') (cf. Dobrovoljc 2017)

The final list, loosely inspired by the Academic Phrasebank (Morley 2014), will be made publicly available by September 2018 on the website www.korpus.cz and presented at the conference.

References

Čermáková A. & Chlumská L. (2017). Expressing place in children's literature: Testing the limits of the n-gram method in contrastive linguistics. In T. Egan & H. Dirdal *Cross-linguistic Correspondences: From Lexis to Genre.* Amsterdam/Philadelphia: John Benjamins Publishing Company, 75-95.

Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification: The case of MWDW extraction from the reference corpus of spoken Slovene. *International Journal of Corpus Linguistics*, 22 (4), 551–582.

Hoffmannová, J. et al. (2016). *Stylistika mluvené a psané češtiny.* Praha: Academia.
Kováříková, D. (2017). *Kvantitativní charakteristiky termínů.* Praha: NLN.
Morley, J. (2014). *Academic Phrasebank. A compendium of commonly used phrasal elements in academic English in PDF format.* Available online at (20.2.2018): <http://www.kfs.edu.eg/com/pdf/2082015294739.pdf>.
Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.
Wei, N., & Li, J. (2013). A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, 18 (4), 506–535.

••••

Marek Łaziński
University of Warsaw
m.lazinski@uw.edu.pl

Actional Interpretation of Verbal Aspect in Legal Texts - Corpus Analysis

Verbal aspect can be chosen differently in similar contexts in close related Slavic languages. While Polish penal sanction provision is usually formed with the pronoun *kto* as a subject and with imperfective verbs, Czech provisions use perfective verbs, whereas Russian ones use nominalized forms which are not unequivocally marked for aspect:

(1) Kto zabija (ipf) człowieka podlega karze pozbawienia wolności (…) (Kodeks karny RP 1997, art. 148.1)

'Whoever kills a human being shall be subject to the penalty of the deprivation of liberty'

(2) Kdo jiného úmyslně usmrtí (pf), bude potrestán… (Trestní zákoník ČR 2009, art. 140)

(3) Ubijstvo, to est umyšlennoe pričinenie smerti drugomu čeloveku, nakazyvaetsja… (Ugolovnyj kodeks RF 1996, art. 105.1)

The ipf verb *zabija* in (1) is a typical achievement verb and does not allow for an attempt interpretation 'tries to kill'. However, there are some other verbs in the Polish Penal Code which could raise doubts about whether the committing of a crime is accomplished or merely attempted:

(4) Odpowiada za podżeganie, kto chcąc, aby inna osoba dokonała czynu zabronionego, nakłania ją do tego. (art. 18)

'Whoever, willing that another person should commit a prohibited act, induces the person to do so, shall be liable for instigating.'

In fact, penal codes always use a specific construction to refer to the situation of a crime intended, but not having occured. Although the concept of a punishable attempt is expressed literally in all modern codices, it is possible that the other Slavic languages apart from Polish avoid imperfectives out of a fear of a possible 'attempt' misinterpretation. Such interpretation would be excluded in the case of the achievement *usmrcovat* instead of usmrtit in (2), but it would be possible in the case in the case of imperfective accomplishment verbs *vyrábět* or *opatřovat* instead of *vyrobit* or *opatřit* in the provision below.

(5) Kdo bez povolení vyrobí, sobě nebo jinému opatří nebo přechovává výbušninu… (TzČR, art. 279)

'Who, without permission, obtains for himself or another person, or keeps (in his possession), an explosive…'

Only the activity verb *přechovávát* remains ipf in Czech. In Polish penal provision it is a standard form for all actional classes:

(6) Kto […] wyrabia, przetwarza, gromadzi, posiada, posługuje się lub handluje substancją lub przyrządem wybuchowym… (KkRP, art 171)

'Whoever […] manufactures, processes, accumulates, possesses, uses or trades in an explosive substance or device…'

In order to analyze other possible questionable uses of ipf verbs a corpus comprising the Polish Penal Code was built. In the special part of the Penal Code 1553 ipf verb forms and only 215 pf verb forms.

Not many ipf verbs are potential accomplishments, e.g. *wyrabiać* in (6), *gromadzić* 'to amass/accumulate', *nakłaniać* 'to persuade/induce', *niszczyć* 'to damage/destroy'. In most cases the potential accomplishment is contextually disambiguated as a completed action and not an attempt.

While imperfectives in the present tense denote all kinds of offences and their circumstances, the perfectives in the Polish Penal Code are generally used to specify extenuating and exempting circumstances, such as compensation paid to the victim by the perpetrator, etc:

(5) Kto bierze (ipf) lub przetrzymuje (ipf) zakładnika […] podlega karze […] Nie podlega karze za przestępstwo […], kto odstąpił (past pf) od zamiaru wymuszenia i zwolnił (past pf) zakładnika. (KkRP, art. 252)

'Whoever takes or holds a hostage…, is subject to the penalty… […] Whoever abandoned the intention to extort and released the hostage shall not be subject to the penalty for the offence…'

Assumption for the further research:

The primacy of the imperfective in Polish codes – opposed to pf verbs or verbal nouns in other Slavic languages – can be explained in terms of a tendency for making legal text generalized. The use of perfectives in other languages tries to avoid misunderstandings in the actional interpretation of an attempt or completed action. To determine the factors of aspect choice in legal texts in neighboring languages more deeply a comparative corpus of Polish, Czech and Russian penal codices is planned. A contrastive concordance and frequency analysis let us confirm or falsify the assumption.

References

Dickey S.M. (2000). *Parameters of Slavic Aspect. A Cognitive Approach.* Stanford: CSLI Publications.

Łaziński M. (2015). Verbal aspect and legal interpretation: the use of verbal aspect in Slavic penal codices. In: M. Kitajo (Ed.) *Aspektual'naja semantičeskaja zona: tipologija sistem i scenarii diaxroničeskogo razvitija. Sbornik statej V Meždunarodnoj konferencii Komissii po aspektologii Meždunarodnogo komiteta slavistov*, Kyoto: University of Kioto, 131–136.

Łaziński M., Jóźwiak K. (2017). Verbal Aspect and Legal Interpretation: the Use of Verbal Aspect in the Polish Penal Code. *Polonica XXXVII.* 167-177.

Przetak M. (2013). *Struktura tekstu prawnego na przykładzie kodeksu karnego*, Gdańsk: Wydawnictwo UG.

Jiří Milička
Charles University, Czech Republic
jiri.milicka@ff.cuni.cz

Alžběta Růžičková
Charles University, Czech Republic
ruzickovaalzbeta@seznam.cz

Slovak Vowel Phonotactics: Slavic Origins vs. Hungarian Influences

Introduction

Hungarian language is a typical example of a language with strict vowel phonotactic patterns – the well-known vowel harmony. Contrary, the Czech phonotactic system comprises set of vowel patterns which are not very strict, as described in (Milička – Kalábová, 2018) and the patterns tend to some sort of vowel disharmony: long front vowels tend to be followed by back vowels, long back vowels tend to be followed by short vowels, short front vowels tend to be followed by long front vowels, and short back vowels tend to be followed by short front vowels (see Figure 1).

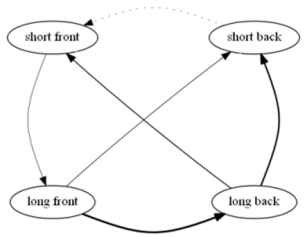

Figure 1.The overrepresented vowel group pairs in Czech.

This discrepancy raisesa question, whether Slovak, which is a close relative of Czechtends to the same vowel phonotactic patterns or whether the patterns are weakened or even altered by the Hungarian influence.

We have followed the same methodology as in (Milička – Kalábová, 2018). The syllabic nuclei of the word types of the Slovak National Corpus were taken into consideration (i.e. all vowels, diphthongs, and syllabic consonants). All vowel bigram frequencies were counted. For example *Slovenka* ('a Slovak woman') occurs 15 times in the Slovak National Corpus, thus the vowel bigrams /o/–/e/, and /e/–/a/ were counted 15 times in our dataset. Consequently we have described all non-random tendencies, i.e. the frequencies found were compared to the random model which resulted to the list of overrepresented and underrepresented vowel pairs (for the methodology details see Milička – Kalábová, 2018).

The phonotactic patterns within the word stems in Czech are different from the patterns on the morphemic seams (i.e. the bigrams in which the first vowel is from the last syllable of the stem and the second vowel is from the first syllable of the ending) therefore we also studied the patterns in the Slovak stems. The stemming algorithms for Slovak and Czech were quite simple and not very reliable so the results on stems and their comparison should be taken *cum granosalis*.

Results

In Slovak, similarly to the Czech language, we observed the tendency towards vowel disharmony, contrary to the Hungarian language; syllables with a front vowel in the nuclear position tend to be followed by back vowel nucleus syllables and vice versa.

As for vowel quantity, „according to the so-called ‚rhythmical law' […] a long vowel, a long liquid, or a diphthong should not be followed by a long segment or diphthong in the next syllable if the two are within the same word" (Hanulíková–Hamann, 2010, pp. 376). This is a codified phonological law in Slovak, and this is also the most prominent pattern we found in our data. Milička – Kalábová 2018 found a somewhat similar tendency in Czech, but only in interaction with vowel quality – there are usually not two neighbouring syllables with long nucleus vowels of the same backness value. However, as our data shows, a very similar tendency can be observed in Hungarian, where a long vowel repels another long vowel in the following syllable, regardless of the vowel quality. This could possibly be a result of areal contact of Slovak and Hungarian.

We analysed both whole words and extracted word stems. In Czech, the results for words and word stems differ from each other. On the other hand,

in Slovak, the results were quite similar for words and word stems – this can be due to the law of rhythmical shortening, "which states that quantity is neutralized in a morphophonemically long syllable after a preceding long syllable" (Short, 1993, pp. 538).

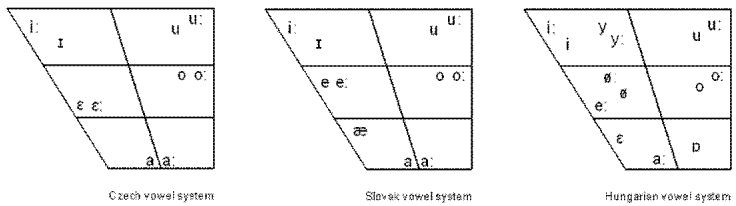

Figure 2. Czech (Skarnitzl – Volín, 2012), Slovak (Pavlík, 2004), and Hungarian (Szende, 1994) vowel system.

To analyse vowel harmony properly, we need to sort vowels into two groups regarding their quality: front vowels and back vowels.

In Czech, /i/ and /e/ are front vowels; /o/ and /u/ are back vowels (Figure 1). According to our analysis, the /a/ phoneme tends to behave as a front vowel, i.e. the phonotactic patterns of /a/ are similar to the phonotactic patterns of the front vowels. To be more precise, the model, that excludes /a/, is more similar to the model that classifies /a/ as a front vowel than to another one classifying it as a back vowel.

The Slovak language uses a different vowel system (Figure 1). The phonemes /i e o u/ fall into the same categories like in Czech, however according to Slovak phonology, /a/ is classified as a back vowel in contrast to its front counterpart /æ/ - the latter vowel's articulation nevertheless tends to shift towards [e] (Short, 1993, pp. 534). Despite the phonological system, the vowel /a/ in Slovak manifests itself as a front vowel, as well as in the Czech language.

In Hungarian, the phoneme /a/ is classified as a back vowel, according to both phonologic classification (Gósy, 1989) and phonotactics.

References

Csaba Oravecz, Tamás Váradi and Bálint Sass. 2014. The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014.* http://www.lrec-conf.org/proceedings/lrec2014/pdf/681_Paper.pdf

Křen, Michal, Tomáš Bartoň, Václav Cvrček, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Renata Novotná, Vladimír Petkevič, Pavel Procházka, VěraSchmiedtová and Hana Skoumalová. 2010. *SYN2010: žánrově vyvážený korpus psané češtiny* [SYN 2010: Genre-Balanced Corpus of Written Czech]. Ústav Českého národního korpusu FF UK, Praha. WWW: http://www.korpus.cz. Accessed 12 Oct 2017.

Leben, William R. 1973. *Suprasegmental Phonology.* Ph.D. thesis, Massachusetts Institute of Technology.

MacKay, David JC. 2003. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press.

McCarthy, John J. 1986. OCP effects: Gemination and Antigemination. *Linguistic inquiry 17.2*: 207–263.

Menzerath, Paul. 1928. Über einige phonetische Probleme. In: *Actes du premier Congres International de Linguistes.* Leiden: Sijthoff

Milička, Jiří and Hana Kalábová. 2018. Vowel Disharmony in Czech Words and Stems. In Fidler, M. – Cvrcek, V. (eds): *Taming the corpus. Frominflection and lexis to interpretation.* Springer.

Nguyen, Noël, and Zsuzsanna Fagyal. 2008. Acoustic Aspects of Vowel Harmony in French. *Journal of Phonetics* 36, no. 1: 1–27.

Ohala, John J. 1994. Towards a Universal, Phonetically-Based, Theory of Vowel Harmony. In *Third International Conference on Spoken Language Processing*, 491–494

Palková, Zdena. 1994. *Fonetika a fonologie češtiny* [Phonetics and Phonology of Czech]. Praha: Karolinum.

Poldauf, Ivan. 1969. *Máme v češtině harmonii samohlásek?* [Do We Have Vowel Harmony in Czech?] Naše řeč 52: 201–209.

Ringen, Catherine O. and Miklós Kontra. 1989. Hungarian Neutral Vowels. *Lingua* 78.2–3: 181–191.

Rounds, Carol. 2001. *Hungarian: An Essential Grammar.* Psychology Press.

Vago, Robert M. 1976. Theoretical Implications of Hungarian Vowel Harmony. *Linguistic inquiry* 7.2: 243–263.

Gósy, Mária. 1989: Vowel harmony: interrelations of speech production, speech perception, and the phonological rules. *Acta Linguistica Hungarica*, 39, 93–118.
Short, David. 1993. Slovak. In *The Slavonic Languages*, 533–592.
Slovenský národný korpus – prim-7.0-public-all. Bratislava. Jazykovedný ústav Ľ. Štúra SAV 2015. Dostupný z WWW: http://korpus.juls.savba.sk.
Hanulíková, Adriana and Silke Hamann. 2010. Slovak. In *Journal of the International Phonetic Association: Illustrations of the IP;* 373–377.
Skarnitzl, Radek and Jan Volín. 2012. Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy,* 18, 7-11.
Pavlík, Radoslav. 2004. Slovenské hlásky a medzinárodná fonetická abeceda. *Jazykovedný časopis*, 55, pp. 87–109.
Szende, Tamás. 1994. Illustrations of the IPA: Hungarian. *Journal of the International Phonetic Association*, 24 (2), 91–94.

••••

Tore Nesset
UiT The Arctic University of Norway
tore.nesset@uit.no

Cascading S-curves: What corpus linguistics tells us about language change

How can corpus data help us understand language change? How are new constructions born? On the basis of a large-scale corpus investigation on Russian numeral constructions, the present paper sheds new light on these questions. First, it is argued that language change can take the shape of cascading s-curves. This is in line with "Piotrowski's law", according to which s-shaped curves are crucial in language change (see Leopold 2005, Blythe and Croft 2012). However, Blythe and Croft do not explain how multiple s-curves interact. The present paper analyzes this interaction in terms of cascades; when one s-curve is about to flatten out at the top, a new s-curve can start.

In Construction Grammar it has been argued that new constructions are born through grammaticalization (Traugott and Trousdale 2013). The second contribution of the present paper is to show that grammaticalization is not the only source of new constructions, insofar as constructions can also be born from so-called rival forms, i.e. forms that compete for the same "functional slot" in a language (Baayen et al. 2013).

Russian numeral constructions are notorious for their syntactic complexity. The present paper focuses on paucal constructions with the numerals *dva* 'two', *tri* 'three' and *četyre* 'four' followed by an adjective and a noun. As shown in (1), these constructions can involve a preposed demonstrative in the nominative plural, an adjectival modifier in the genitive plural, a noun in the genitive singular, as well as a verb in the plural:

(1) *Tol'ko otkuda èti dva zagadočnyx sputnika vzjalis'*?
Only wherefrom these$_{\text{Nom pl}}$ two mysterious$_{\text{Gen pl}}$ companion$_{\text{Gen sg}}$ came$_{\text{Pl}}$
'But where did these two mysterious companions come from?'
(Russian National Corpus)

However, extensive variation is possible. The present study zooms in on the rivalry between the nominative and genitive plural in modifying adjectives.

In order to investigate this variation a database was created with all relevant examples from the Russian corpus (6,581 examples). The data, which span approximately 200 years and represent a variety of genres, were annotated manually for a number of factors that have been argued to be relevant in the literature, but a CART analysis (Classification And Regression Trees, Strobl et al. 2009) where the form of the adjective was the dependent variable, revealed that only three factors have a robust impact: time period, gender of the quantified noun, and the numeral. With regard to time, it is argued that the numerals followed an s-shaped development in the 19th and 20th centuries, whereby adjectives in the nominative plural were gradually replaced by adjectives in the genitive plural. Gender of the quantified noun became relevant in the second half of the 20th century, when constructions with feminine nouns split off and started preferring adjectives in the nominative plural. This development arguably followed an s-curve, which began when the preceding s-curve had almost reached its point of culmination. The numeral itself is also relevant, insofar as constructions with *tri* and *četyre* were generally more innovative than constructions with *dva*.

In addition to showing that language change follows the path of cascading s-curves, the Russian numerals also demonstrate how rival forms can give birth to new constructions. When the feminine nouns split off in the second half of the 20th century, what was one construction became two: one for masculine and neuter nouns with the modifying adjective in the genitive plural (e.g. *dva interesnyx romana* 'two interesting novels'), and one for feminine nouns with the adjective in the nominative plural (e.g. *dve interesnye knigi* 'two interesting books'). This development is arguably not the result of grammaticalization, since the case endings on the adjective have been "fully grammatical" since Old Russian. Instead, it is argued that the new gender-specific constructions arise as a consequence of the rivalry between two fully grammatical forms.

Although the present study only addresses a few of the complexities of Russian numeral phrases, it suffices to show how corpus data have the potential to change the way we think about language change.

References

Baayen. R. H., A. Endresen, L. A. Janda, A. Makarova and T. Nesset (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37 (3), 253–291.

Blythe, R. A. and W. Croft (2012). S-curves and the mechanisms of propagation in language change. *Language* 88 (2), 269–304.

Leopold, E. (2005). Diachronie: Grammatik. In R. Köhler, G. Altmann & R. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook*. Berlin and New York: De Gruyter, 607-633.

Strobl, C., J. Malley and G. Tutz (2009): An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14 (4), 323–348.

Traugott, E. C. and G. Trousdale (2013). *Constructionalization and constructional changes*. Oxford: Oxford University Press.

••••

Jana Nová
Czech Language Institute of the Czech Academy of Sciences
novaj@volny.cz

Vít Michalec
Czech Language Institute of the Czech Academy of Sciences
michalec@ujc.cas.cz

Zdeňka Opavská
Czech Language Institute of the Czech Academy of Sciences
opavska@ujc.cas.cz

Renáta Neprašová
Czech Language Institute of the Czech Academy of Sciences
lexikografie@ujc.cas.cz

Frequency (not) sacred: The headword list of a contemporary Czech monolingual dictionary and corpora

Building the headword list is one of the most important parts of making a dictionary (cf. Atkins et Rundell 2008, Čermák 2010, Filipec 1995, Grundy et Rawlinson 2016). In the Academic Dictionary of Contemporary Czech project (Akademický slovník současné češtiny – ASSC, Kochová et Opavská 2016a,b) building the headword list was one of the initial steps. Building of the ASSC headword list combines qualitative and quantitative criteria, and there are used both frequency and systemic approaches. The raw headword list for the whole planned dictionary (120–150 000 headwords) consists of lexical units having the total frequency of 5 or more in three reference representative corpora SYN 2000 (Čermák et al. 2000), SYN 2005 (Čermák et al. 2005) and SYN 2010 (Křen et al. 2010). Building the raw headword list for the whole dictionary on the same material allows to keep a stable base throughout the whole project, to keep balanced proportions of all alphabetical sections and it also helps not to overestimate newspaper texts, representing the substantial part of large SYN-series corpora.

The preliminary word list for each alphabetical section is generated automatically from the reference corpora mentioned above. Then a member of the ASSC team checks the list word by word, deleting lexicographically unsuitable items (notably spelling errors, lemmatisation mistakes, numerical expressions or foreign words that occur in foreign-language texts), identifying spelling variants etc. After that the raw headword list is prepared for lexicographers who go through it and decide whether to omit, include or add a concrete lemma according to the larger language material: large corpora of the SYN series, the sizeable web corpus Araneum Bohemicum Maximum (Benko 2014), the Newton Media database of written and spoken media texts and the internet. According to word-formative and semantic relations, we add units which had very low frequency in the representative corpora but their frequency in the larger material is sufficient. Some headwords are also added according to their presence in older Czech dictionaries, esp. in *Slovník spisovné češtiny pro školu a veřejnost*. On the other hand we omit units from open word-formative series and specialised terms (where higher frequency is required), or units occurring only in a few unique texts.

However, we must be aware of the fact that our large material sources gradually change (increase). While in 2016 our decisions whether to include or omit a headword used to be mainly based on the SYN v3 (Křen et al. 2014), the largest Czech corpus then, today we use SYN v6 (Křen et al. 2017) where many words have 3–4times higher frequencies; it means we are likely to include words that would not have been put in only 2 years ago. Expecting larger and larger corpora coming during our dictionary project, there is a question how to develop our inclusion rules.

To keep up with the growth of vocabulary we consider adding the newer reference corpus SYN 2015 (Křen et al. 2015) to the raw-headword-list-making process although our survey showed the success rate in identifying suitable dictionary headwords from this corpus appears to be remarkably lower than from the 3 older reference corpora. The ASSC team was also repeatedly advised to use the average reduced frequency (ARF, Savický et Hlaváčová 2002) instead of the absolute frequency when building the headword list. Our survey showed the ARF-based method would be slightly (1-4 %) more successful in identifying suitable headwords, depending on the set threshold ARF or absolute-frequency value. However the main practical difference is in the ARF-based list there are several (though not many) words or word-chains – candidates for inclusion to the dictionary – that do not occur in the absolute-frequency based list, and vice versa. In order not to lose words of this kind, it appears advisable to combine both above mentioned frequency

criteria or to spend more time by manual checking of a longer word list based on lower threshold frequency.

References

Atkins, B. T. S. – Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In P. Sojka – A. Horák – I. Kopeček – K. Pala (Eds.), *Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655.* Springer International Publishing Switzerland, 257-264.

Čermák, F. – Blatná, R. – Hlaváčová, J. – Klímová, J. – Kocek, J. – Kopřivová, M. – Křen, M. – Petkevič, V. – Schmiedtová, V. – Šulc, M. (2000). *SYN2000: žánrově vyvážený korpus psané češtiny.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: http://www.korpus.cz/

Čermák, F. – Doležalová-Spoustová, D. – Hlaváčová, J. – Hnátková, M. – Jelínek, T. – Kocek, J. – Kopřivová, M. – Křen, M. – Novotná, R. – Petkevič, V. – Schmiedtová, V. – Skoumalová, H. – Šulc, M. – Velíšek, Z. (2005). *SYN2005: žánrově vyvážený korpus psané češtiny.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: http://www.korpus.cz/

Čermák, F. (2010). Notes on Compiling a Corpus-Based Dictionary. *Lexikos, 20 (Afrilex-reeks/series*, 20), 559–579.

Filipec, J. (1995). Teorie a praxe jednojazyčného slovníku výkladového. In: F. Čermák – R. Blatná (Eds.), *Manuál lexikografie.* Jinočany: H&H, 14–49.

Grundy, V. – Rawlinson, D. (2016): The Practicalities of Dictionary Production; Planning and Managing Dictionary Projects; Training of Lexicographers. In: P. Durkin (Ed.), *The Oxford Handbook of Lexicography.* Oxford: Oxford University Press, 561–578.

Kochová, P. – Opavská, Z. (2016). Akademický slovník současné češtiny. *Naše řeč*, 99, 57–83.

Kochová, P. – Opavská, Z. (Eds.) (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny.* Praha: Ústav pro jazyk český AV ČR, v. v. i.

Křen, M. – Bartoň, T. – Cvrček, V. – Hnátková, M. – Jelínek, T. – Kocek, J. – Novotná, R. – Petkevič, V. – Procházka, P. – Schmiedtová, V. – Skoumalová, H. (2010). *SYN2010: žánrově vyvážený korpus psané češtiny.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: http://www.korpus.cz/

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A. (2017). *Korpus SYN, verze 6 z 18. 12. 2017.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: http://www.korpus.cz/

Křen, M. – Cvrček, V. – Čapka, T. – Čermáková, A. – Hnátková, M. – Chlumská, L. – Jelínek, T. – Kováříková, D. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Škrabal, M. – Truneček, P. – Vondřička, P. – Zasina, A. (2015). *SYN2015: reprezentativní korpus psané češtiny.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: http://www.korpus.cz/

Křen, M. – Čermák, F. – Hlaváčová, J. – Hnátková, M. – Jelínek, T. – Kocek, J. – Kopřivová, M. – Novotná, R. – Petkevič, V. – Procházka, P. – Schmiedtová, V. – Skoumalová, H. – Šulc, M. (2014). *Korpus SYN, verze 3 z 27. 1. 2014.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: <http://www.korpus.cz/

Savický, P. – Hlaváčová J. (2002). Measures of Word Commonness. *Journal of Quantitative Linguistics* 9, 215–231.

....

Tatiana Perevozchikova
University of Tübingen
tatiana.perevozchikova@uni-tuebingen.de

Pronominal expression of possession in noun phrases in Russian, Czech, and Bulgarian

Competition between reflexive and non-reflexive possessive pronouns has been described for several Slavic languages (e.g. Padučeva 1983, 1985, Timberlake 1980, 2004 for Russian; Čmejrková 2011, Dočekal 2000, Daneš/ Hausenblas 1986 for Czech; Nicolova 1986, Nicolova 2017 for Bulgarian). All studies point that grammatical person determines the choice of possessive pronoun in the first place, i.e. reflexivization is more obligatory in the third person than in the first and in the second person. The optionality of reflexivization in the first and second person has been most often described as conditioned by pragmatic factors (Yokoyama / Klenin 1976). Semantic and syntactic variables have been mentioned in the cited literature but so far not systematically investigated even within one language, let alone comparison between Slavic languages (for an overall contrastive picture of Russian and Czech see Nedoluzhko et al. 2016).

The present study is a part of a bigger project that aims to test the influence of semantic, syntactic, and pragmatic factors on the choice of possessive pronouns in Russian, Czech, and Bulgarian. The presentation will focus on the question of whether the three languages have similar preferences in choosing possessive pronouns in first person singular contexts. Specifically, we ask whether animacy of the pronoun referent and two syntactic factors (the degree of syntactic isolation of the (pro)noun phrase and the presence of other possible controllers in the syntactic structure) constrain the choice of possessive pronouns in Russian, Czech, and Bulgarian in the same way. The literature cited above makes us expect the following:

1) the distribution of reflexive and non-reflexive possessive pronouns is similar in Russian and Czech, whereas Bulgarian has a strong preference for the non-reflexive possessive pronoun at the expense of the reflexive possessive pronoun in the long form and a strong tendency to use only the reflexive possessive pronoun in the short form.

2) non-reflexive possessive pronouns are more likely with animate than with inanimate referents;

3) reflexive possessive pronouns are less likely in case of their (partial) syntactic isolation from the finite predicate and in presence of other possible controllers in the syntactic structure of the clause.

The hypotheses have been tested against two types of data: 1) parallel corpus InterCorp 10 and 2) comparable corpora from the Aranea family (Araneum Bohecum Minus, Araneum Bulgaricum Minus, Araneum Russicum Minus).

The results in general confirm the first hypothesis, although there are differences between parallel and comparable corpora. In Araneum data, we found similar distribution of reflexive and non-reflexive possessive pronouns in Russian and Czech and in short forms of pronouns in Bulgarian. In InterCorp, Russian patterns with Bulgarian in the frequency of reflexive and non-reflexive possessive pronouns as well as in their omission rate. In Czech texts of InterCorp, we found a higher frequency of reflexive possessive pronouns and a lower frequency of non-reflexive possessive pronouns than in Russian and Bulgarian. All three languages also use expressions of external possession.

As for the influence of the three factors investigated, we could confirm that if other possible controllers are present in the clause structure, all three languages prefer to use personal pronouns. Neither animacy nor syntactic isolation alone could predict the use of the possessive pronouns. However, the combination of both factors could partly explain the data, whereby syntactic isolation appears to be a stronger factor for Russian and Czech and animacy seems to be more relevant for the use of short possessive pronouns in Bulgarian.

References

Čmejrková, S. (2011). Posesivní reflexivizace. Zájmeno svůj. Jeho užití a významy. In Štícha, Fr. (Ed.), *Kapitoly z české gramatiky.* Praha, 655–686.

Daneš, F. / Hausenblas, K. (1962). Přivlastňovací zájmena osobní a zvratná ve spisovné češtině. *Slavica Pragensia,* 4, 191-202.

Dočekal M. (2000). Posesivní reflexivum v bohemistice. *Sborník prací FF Brněnské university*, 47–59. Brno.

Nedoluzhko, A. / Schwarz (Khoroshkina), A. / Novák, M. (2016). Possessives in Parallel English-Czech-Russian Texts. *Computational Linguistics and*

Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016", 483-497.
Nicolova, R. (1986). *Bălgarskite mestoimenija*. Sofia.
Nicolova, R. (2017). *Bulgarian Grammar*. Berlin.
Padučeva, E. (1983). Vozvratnoe mestoimenie s kosvennym antecedentom i semantika refleksivnosti. *Semiotika i informatika*, 21, 3-33.
Padučeva, E. (1985). *Vyskazyvanie i jego sootnesennost' s dejstvitelnost'ju (referencialnye aspekty semantiki mestoimenij)*. Moskva.
Timberlake, A. (1980). Reference Conditions on Russian Reflexivization. Language, 56 (4), 777-796.Timberlake, A. (2004). *A Reference Grammar of Russian*. Cambridge.
Yokoyama, Olga. T. / Klenin, E. (1976). The semantics of optional rules: Russian personal and reflexive possessives. In Matejka, L. (Ed.), *Sound, Sign, and Meaning*. Ann Arbor, 249–267.

••••

Alexander Piperski
Russian State University for the Humanities / National Research University Higher School of Economics, Moscow, Russia
apiperski@gmail.com

Aspect-Specific Keywords in Russian

Verbs in Russian, as well as in other Slavic languages, can be either perfective or imperfective, a small group of them being biaspectual. The descriptions of the aspectual system of Russian are numerous, cf. Zalizniak, Mikaelian & Shmelev (2015) for a comprehensive bibliography.

Most approaches to the semantics of Russian aspect take a descriptive and classificatory approach, grouping contexts for imperfective and perfective aspect based on their meanings. In this talk, I propose a different approach in the vein of Firth's (1957: 11) famous saying "You shall know a word by the company it keeps," modified in the following way: "You shall know a grammatical feature by the company it keeps."

To test this approach, words are extracted that co-occur much more frequently with imperfective than with perfective verbs in the past tense, and vice versa. I call them **aspect-specific keywords** rather than collocations, because the notion of collocation normally implies mutual attraction of two words, but no statement about the attraction of specific words to verbs is made here; it is only about some words being encountered more often in the vicinty of one aspect as compared to the other.

Aspect-specific keywords of Russian were extracted from the Araneum Russicum Minus corpus (Benko 2014) comprising approximately 100 million words. The following procedure is used:

1. extract all lemmata occuring at least once near imperfective (query: [tag="Vmis…..e.*"]) and perfective (query: [tag="Vmis…..e.*"]) past tense verb forms within the window of up to 5 tokens to the left or to the right;
2. restrict the list to include only lemmata that occur in the whole corpus at least 100 times;
3. transform the list into relative frequencies (because the two aspects have different frequencies);

4. for each lemma, compute the keyness score *K* for each aspect as compared to the other aspect using Kilgarriff's Add-*N* keyword method (Kilgarriff 2009): $K = (f_{focus} + N) / (f_{reference} + N)$. The constant *N* is set to 1000 ipm; for instance, if a word occurs with a frequency of 4000 ipm in one list and with a frequency of 1000 ipm in the other list, its keyness for the first list will be (4000 + 1000) / (1000 + 1000) = 2.5;
5. sort the list in descending order.

Top 20 keywords for the two aspects are as follows:

Perfective: *vdrug* 'suddenly', *srazu* 'immediately', *itog* 'outcome', *press-služba* 'press service', *nakonec* 'finally', *neožidanno* 'unexpectedly', *snova* 'again', *vskore* 'soon', *prezident* 'president', *RF* 'Russian Federation', *Putin* (proper name), *%*, *rezko* 'sharply', *blagodarja* 'due to', *vyvod* 'conclusion', *pojti* 'go', *tut* 'here', *Medvedev* (proper name), *opjat'* 'again', *glava* 'head, leader';

Imperfective: *vsegda* 'always', *nikogda* 'never', *dolžen* 'must', *ran'še* 'before', *často* 'often', *neodnokratno* 'repeatedly', *inogda* 'sometimes', *postojanno* 'constantly', *možno* 'may', *dolgo* 'for a long time', *vynudit'* 'compel', *každyj* 'each', *by* (subjunctive particle), *ni* 'nor', *ranee* 'before', *protjaženie* 'stretch', *iznačal'no* 'initially', *kogda-to* 'once', *nikto* 'nobody', *čtoby* 'in order to'.

An examination of the keyword list (also beyond the top 20 words) shows that perfective often co-occurs with words indicating immediateness and spontaneity of an action (*vdrug*, *srazu*, *neožidanno*, *rezko*, *vnezapno* 'suddenly') or a single instance of a repeated action (*snova*, *opjat'*, *očerednoj* 'next', *novyj* 'new') as well as with names and titles of public persons, because speakers usually focus on their individual completed actions in the past. Imperfective aspect co-occurs with mood and modality markers (*dolžen*, *možno*, *by*, *nužno* 'necessary'), with negation (*nikogda*, *ni*, *nikto*, *nikakoj* 'none', *nikak* 'in no way'), with words related to repeated and continuous actions (*vsegda*, *často*, *neodnokratno*, *inogda*, *postojanno*) or to long time spans (*protjaženie*, *detstvo* 'childhood', *dolgij* 'long'). Interestingly, keyword list for the imperfective aspect also includes identity-related words, such as *sovetskij* 'Soviet', *evrej* 'Jew', *krest'janin* 'peasant', that occur close to imperfective verbs in the plural.

These findings illustrate the power of keyword identification with respect to specific grammatical features, which can help to describe semantics and the use of these features. A similar approach to Czech, although using different statistical measures, was implemented by Cvrček & Fidler (2017); in the talk, I am also going to compare the results for Czech and Russian.

References

Benko, V. (2014). Yet another family of (comparable) Web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, Speech and Dialogue. 17th International Conference, Brno, Czech Republic, September 8–12, 2014*. Cham: Springer International Publishing Switzerland, 247–256.

Cvrček, V. & Fidler, M. (2017). Probing aspectual context with keyword analysis. In A. Makarova, S. M. Dickey & D. Divjak (Eds.)., *Each venture a new beginning: Studies in honor of Laura A. Janda*. Bloomington, IN: Slavica, 277–296.

Firth, J. R. (1957). *Papers in linguistics*. London: Oxford University Press.

Kilgarriff, A. (2009). Simple maths for keywords. *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool, UK. https://www.sketchengine.co.uk/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf.

Zalizniak, A. A, Mikaelian, I.L. & Shmelev, A. D. 2015. *Russkaja aspektologija: V zaščitu vidovoj pary* [Russian aspectology: Arguing for the existence of aspectual pairs]. (Studia Philologica). Moskva: Yazyki slavyanskoy kultury.

....

Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences;
Institute of Philosophy, University of Warsaw
adamp@ipipan.waw.pl

Agnieszka Patejuk
Institute of Computer Science, Polish Academy of Sciences;
Faculty of Linguistics, Philology and Phonetics, University of Oxford
aep@ipipan.waw.pl

An Enhanced Universal Dependencies Treebank of Polish

The aim of this paper is to present UD_Polish-LFG, the first Universal Dependencies (UD) treebank of Polish making non-trivial use of enhanced dependencies offered by the current version 2 of the UD standard.

The treebank is the result of converting a corpus of Polish sentences annotated with much richer syntactic structures, namely, with syntactic representations adhering to Lexical Functional Grammar (LFG; Bresnan 1982, Dalrymple 2001, Bresnan et al. 2015). In LFG, there are two levels of syntactic representation: the usual constituent structure (c-structure) and a functional structure (f-structure) containing information about grammatical functions, inter alia. For example, the two LFG representations (as visualised via the INESS system, Rosén et al. 2012) for (1) are shown in Figures 1–2, and the resulting UD representation – in Figure 3.

(1)

Także	w	tym	przypadku	fototerapia	może	złagodzić
also	in	this	case	phototherapy.NOM.SG.F	may.3.SG	relieve.INF

lub	znieść	całkowicie	niekorzystne	objawy.	
or	eliminate.INF	completely	unfavourable.ACC.PL.M	symptoms.ACC.PL.M	

'Also in this case, phototherapy may relieve or completely eliminate unfavourable symptoms.'

Any enhanced UD representation consists of two syntactic structures: a basic dependency tree, as shown above the sentence in Figure 3, and an enhanced dependency graph, as shown below the sentence; differences between these structures are shown in red. One trivial difference between them is that, in the enhanced graph, some relations may be subtyped with information about case, as it is understood in UD, where adpositions are treated as extended cases; in the example, the obl relation between *może* 'may' and *także w tym przypadku* 'also in this case' is enhanced to obl:w. Less trivially, enhanced graphs may contain dependencies absent from the basic tree, where only one dependency may target any word. So, in Figure 3, *fototerapia* 'phototherapy' is only the subject (nsubj) of *może* 'may' in the basic tree, but also of *złagodzić* 'relieve' and *znieść* 'eliminate' in the enhanced graph. Similarly, *znieść* is not only a non-initial conjunct (conj), but also a controlled complement (xcomp) of może, just as *złagodzić*, and *niekorzystne objawy* 'unfavourable symptoms' is not only the direct object of *złagodzić*, but also of *znieść*. In the full paper, we discuss the conversion procedure, which required many structural changes from LFG to UD. For example, while prepositions are the heads of prepositional phrases in LFG (see the PP constituent in Figure 1 and substructure with index 66 in the upper part of Figure 2), they are dependents of nouns in UD (see the case dependency from *przypadku* 'case' to *w* 'in' in Figure 3); similarly for numeral phrases and for verbal phrases headed (in LFG, but not in UD) by auxiliaries and copulas. Also, substantial effort was devoted to the right conversion of coordinate structures, headed by the conjunction in LFG but by the first conjunct in UD: various cases of shared dependents (and governors) and various interactions with other phenomena had to be taken into account.

We also compare the resulting enhanced UD treebank with the previous UD treebank, UD_Polish-SZ, available since UD release 1.2. First, UD_Polish-LFG is much larger: it contains 17,246 running sentences (17,190 types; duplicate sentences have different analyses), compared to 8227 running sentences in UD_Polish-SZ (8139 types; duplicate sentences may have the same analyses). In terms of running tokens, the respective numbers are 130,967 (UD_Polish-LFG) vs. 84,316 (UD_Polish-SZ), which implies that UD_Polish-SZ sentences are longer on the average. Second, there are many linguistic differences between the two treebanks, which we view as clear improvements in UD_Polish-LFG, e.g.: 1) direct objects are understood in a standard way (e.g., Gołąb et al. 1968, 132, Urbańczyk 1992, 62), as dependents

becoming subjects under passivisation (rather than as almost any subcategorised nominal phrases), 2) predicative complements are analysed consistently (as xcomp), 3) impersonal *-no/-to* forms are correctly marked as impersonal verbs (rather than as adjectival passive participles), 4) the five genders (Mańczak 1956) are represented directly (rather than assuming that the three masculine genders differ in animacy), 5) three functions of the so-called reflexive marker *się* are distinguished (inherent, impersonal and anaphoric), 6) interrogative and relative uses of pronouns such as *który* 'which' are properly distinguished, 7) abbreviations are assigned appropriate parts of speech (rather than X, which is the UD part of speech used for tokens whose real part of speech is not known), etc.

UD_Polish-LFG is a part of UD release 2.2 (published in July 2018).

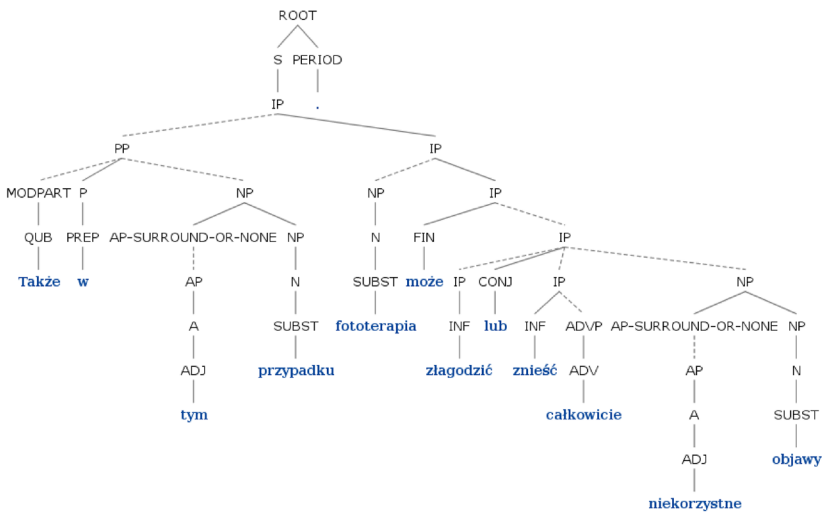

Figure 1: C-structure of (1)

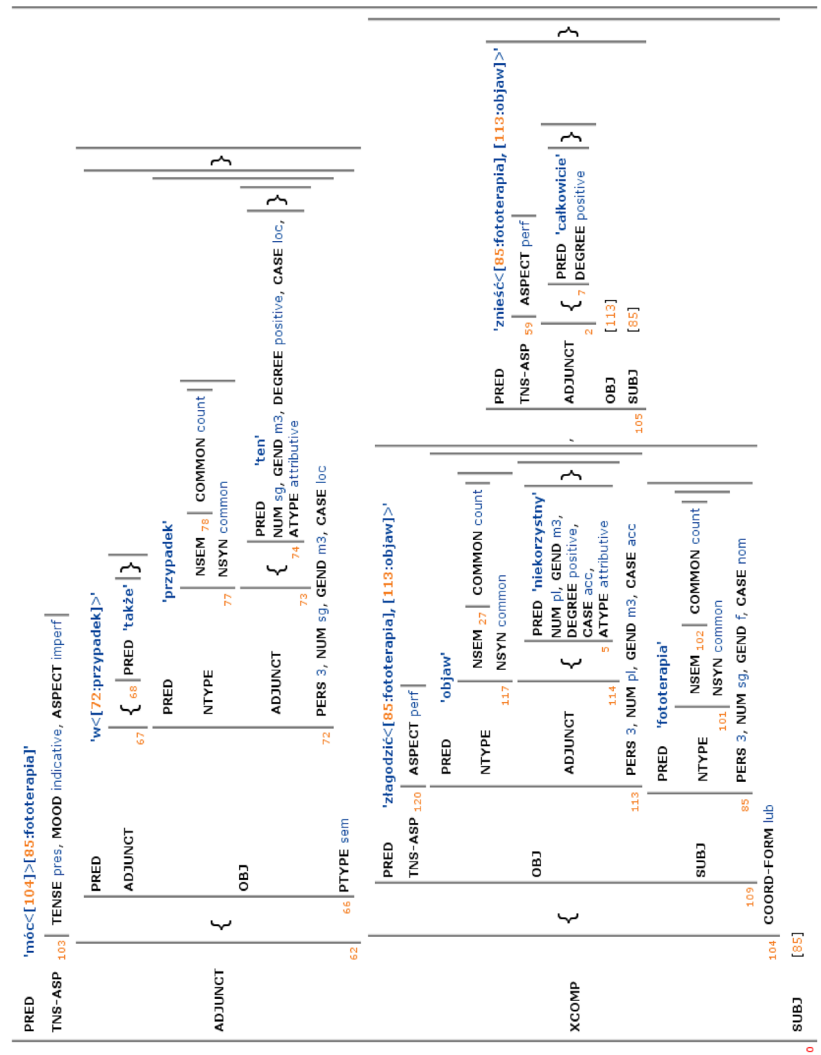

Figure 2: F-structure of (1)

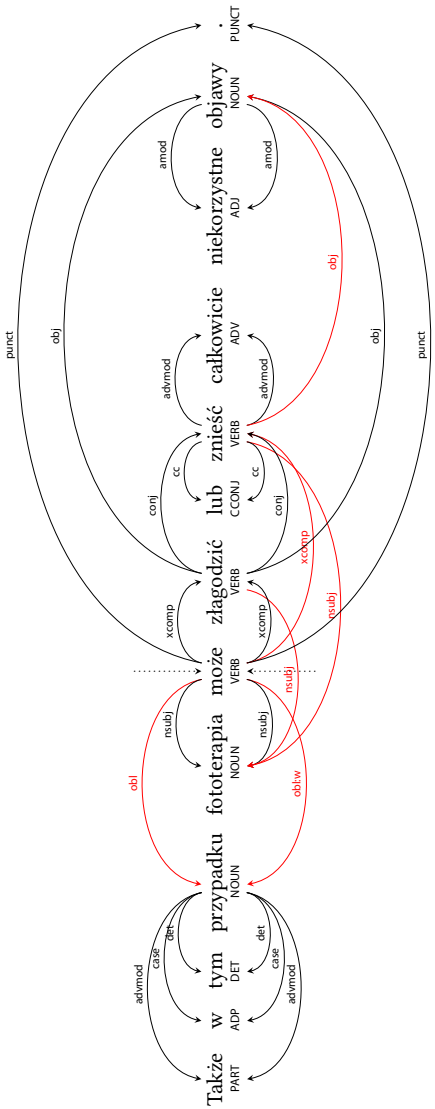

Figure 3: UD representation of (1)

References

Bresnan, J., ed. (1982). *The Mental Representation of Grammatical Relations*. The MIT Press.

Bresnan, J., Asudeh, A., Toivonen, I., and Wechsler, S. (2015). *Lexical-Functional Syntax*. Wiley-Blackwell, 2nd edition.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Academic Press.

Gołąb, Z., Heinz, A., and Polański, K. (1968). *Słownik terminologii językoznawczej*. Wydawnictwo Naukowe PWN.

Mańczak, W. (1956). Ile jest rodzajów w polskim? *Język Polski*, XXXVI(2), 116–121.

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In *LREC 2012 META-RESEARCH Workshop on Advanced Treebanking*, pp. 22–29. ELRA.

Urbańczyk, S., ed. (1992). *Encyklopedia języka polskiego*. Ossolineum.

....

Anna Řehořková
Charles University, Czech Republic
anna.rehorkova@ff.cuni.cz

Czech conditional verb forms in assertive complement clauses

In Modern Czech, there are two options to express doubt about the truthfulness of some subordinate proposition: to use the indicative (1) or the so called conditional mood (2). The second option comprises the auxiliary verb *by* (*bych*, *bys*…) and past participle and is usually translated into English as the past tense form of modal verb *will* + infinitive:

(1) Nemyslím si, že *je*.IND to realizovatelný plán.
'I don't think that this *is* a realizable plan.'

(2) Nemyslím si, že *by.AUX to byl*.PTCP realizovatelný plán.
'I don't think that this *would be* a realizable plan.'

Both options can be used in the same communicative event (cf. Karlík 1982). Historically, though, the interchangeability of these two options was even wider and both were found not only after the dubitative verbs but also after positive verbs of speech or sensory perception. The aim of this paper is to reconstruct the development of these assertive complement *by*-clauses (ACBC) and correct existing assumptions about the extent of their use with the aid of historical corpora.

ACBC are attested in Czech along with the early continuous texts in the 13th century. Searching for similar old evidence in other Slavic languages, Bauer (1960: 146) found such clauses only in Polish. Today, two kinds of *by*-clauses are actually reported for Modern Polish (Tomaszewicz 2010): one with the subjunctive (3) and the other with the so called conditional mood (4). The difference in the placement of the enclitic (*żeby był* vs. *że byłby*) is associated with the difference in modality: whereas the subjunctive can express either the realis or irrealis mood, the other is used only for the irrealis mood.

(3) Nie sądzę, że*by był* trzeźwy, gdyby/jeśli tyle wypił.
'I don't think he *was/is/would be* sober, if/since he had drunk so much.'

(4) Nie sądzę, że *byłby* trzeźwy, gdyby/*jeśli tyle wypił.
'I don't think he *would be*/*was sober, if/*since he had drunk so much.'

The Polish subjunctive thus resembles the Czech ACBC in (2), which can be used in the same situation as the indicative. This is also a feature of the later Latin subjunctive (Harrington – Pucci – Goddard 1997: 48). Given the fact that both Czech and Polish are localized in the West Slavic area, the influence of Latin needs to be assessed.

In Old Church Slavonic, the nearest type of content clauses are indirect questions with the auxiliary form *bi (bimъ, biste…; Trost 1972: 129–134). The bi*-forms, used primarily in unreal conditional clauses or for wish and purpose, are an undeniable source of the so called conditional mood in Slavic languages (along with the pluperfect, according to Sitchinava 2004).

Building on the Slavonic roots of the "conditional" mood, we can consider the influence of Latin on Czech throughout the history. The earliest ACBC after dubitative verbs correspond to the use of the subjunctive in Latin (and differ from what is known about the Old Church Slavonic). In the further development, two tendencies in the usage of this clauses are ascribed to Latin (Bauer 1960: 135, 149, 151, 152): a) the semantic broadening from the association with verbs expressing doubts, negation and other unreal meanings to verbs reporting just unguaranteed information or even someone else's sensory perception; b) the quantitative expansion culminating presumably in the „Humanist" 16th century when the semantic broadening from unreal to real meaning reached the prototypically real complements, as in (5).

(5) znamení […] dal, že *by.AUX pět šífů viděl*.PTCP (1590)
'[he] signalled that [he] *saw* five ships'

However, these findings are not supported by any thorough quantitative study. Do the new contact-induced semantic subtypes of ACBC really indicate the increase in frequency? Our preliminary research based on the corpora of Old and Middle Czech (*Diakorp, Staročeská and Středněčeská textová banka*) suggests that in the timespan of the 14th – 17th century, there

is either no significant shift in the choice between the "conditional" and the indicative mood in assertive complement clauses or only a shift in the opposite direction than expected: towards a lower frequency in the 16th century.

Another question is the straightforwardness of the semantic development in time. E.g. reported speech after a positive verb as in (6)-(8) would be expected with the indicative (*jde*) early, later with the "conditional" mood (*by šel*) but the examples (Bauer ibid.: 149) show a fluctuation:

dixerunt autem ei, quod Iesus Nazarenus transiret (Luc 18: 37)

(6) i pověděchu jemu, ež *jde* Ježíš Nazaretský (late 1300s)
(7) i pověděchu jemu, jež *by* Ježíš nazaretský *šel* (1421)
(8) i pověděchu jemu, že Ježíš nazaretský *jde* (1568)
'so they told him that Jesus of the Nazareth was coming'

The quantitative analysis is undertaken to decide whether there is or is not a consistent tendency to semantic broadening.

References

Bauer, J. (1960). *Vývoj českého souvětí*. Praha: Nakladatelství ČSAV.

Harrington, K. P., Pucci, J., & Elliott, A. G. (Eds.). (1997). *Medieval Latin*. University of Chicago Press.

Karlík, P. (1982). Má čeština konjunktiv? *Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná = Studia minora facultatis philosophicae universitatis Brunensis, Series linguistica, Linguistica Brunensia*, 30, 1–15.

Kučera, K., Řehořková, A., Stluka, M. (2015): *DIAKORP: The Diachronic Corpus*, version 6 (18. 12. 2015). Ústav Českého národního korpusu FF UK. URL: http://www.korpus.cz.

Sitchinava, D. (2004). К проблеме происхождения славянского условного наклонения (On the origins of Slavic conditional). *Ирреалис и ирреальность. Исследования по теории грамматики 3. (Irrealis and irreality. Studies in the Theory of Grammar* 3.) Moscow: Gnosis, 292–312.

Staročeská a středněčeská textová banka (The Bank of Old and Middle Czech Texts), v. 14.10.2016_20:06 (2016). Ústav pro jazyk český AV ČR, v. v. i., Praha. URL: http://vokabular.ujc.cas.cz.

Tomaszewicz, B. (2010). Subjunctive mood in Polish and the Clause Typing Hypothesis. *Formal Studies in Slavic Linguistics*. Cambridge Scholar Publishing: Newcastle.

Trost, K. (1972). *Perfekt und Konditional im altkirchenslavischen*. Wiesbaden: Otto Harrassowitz.

••••

Thomas Samuelsson
Stockholm University, Sweden
thomas.samuelsson@slav.su.se

The Russian adjectives antirossijskij, antirusskij and antisovetskij in Russian media: a corpus study

In this paper, I present a study of three related Russian adjectives: *antirossijskij* 'anti-Russian', *antirusskij* 'anti-Russian' and *antisovetskij* 'anti-Soviet'. The study uses methods of CADS (corpus-assisted discourse studies) (for example Partington, Duguid & Taylor 2013). The aim of this approach is to uncover deeper knowledge, for example hidden meanings that are not obvious. In order to get the most complete results, both quantitative and qualitative methods are applied to linguistic data. Data is also interpreted with the help of corpus-external sources of information.

The negative prefix *anti-* 'anti-' has a polarizing effect on the value-laden, ideological words *rossijskij*, *russkij* and *sovetskij*. The analysis presents how polarized stances are represented with these prefixed words in Russian media in times of polarized politics regarding Russia. A sharpening of policy correlates with Vladimir Putin's return in 2012 to his third presidency, manifested for example in 2011–2013 Russian mass protests, the Ukrainian crisis, and a significant downturn in Russia-West relations. The material is collected from the newspaper corpus "SMI 2000-x gg.", provided by the Russian National Corpus (ruscorpora.ru), and from a compilation of a corpus based on Russian media outlets.

In their unprefixed forms, *rossijskij* and *sovetskij* are demonyms, concerning the Russian Federation and the Soviet Union respectively, while *russkij* is an ethnonym referring to the Russian ethnicity, culture or language. But in usage, *rossijskij* and *russkij* have overlapping semantics. This fuzzy border is for instance used by ethnic nationalists to advance *russkij* at the expense of the non-ethnic *rossijskij* (Kolstø 2016), promoted by the Yeltsin administration as a dissociation from Russia's Soviet past (Blakkisrud 2016), when Russians preferred *sovetskij* (Pain 2016).

The prefix *anti-* is one of the most powerful semantic markers to express opposition in the Russian language (Zelenin 2007: 183). It has repeatedly become activated during times of socio-political polarization in the Russian society, for example during the Russian revolution and the period thereafter (Zelenin 2001), in the course of the demolition of the Soviet totalitarianism in the second half of the 1980s and the beginning of the 1990s (Ferm 1994, Zemskaja 1996), and in recent times (Raciburgskaja 2014).

The search in the newspaper corpus shows that the relative frequency of the prefixed adjective *antirossijskij* increases almost seven times from 2013 to 2014. The less common *antirusskij* shows a boost of twelve times from 2013 to 2014. The still active *antisovetskij* has its relative frequency doubled from 2012 to 2013, but shows no significant change to 2014. The increased activities of *antirossijskij* and *antirusskij* correlate with the dramatic development in Ukraine. An investigation of the contexts of *antirossijskij* and *antirusskij* shows that the changes in relative frequencies can be attributed to a polarized reporting of the events in Ukraine and the responses of the West.

The newspaper corpus (rucorpora.ru) contains data from the following papers: Izvestija, Sovetskij sport, Trud-7, Komsomol'skaja pravda, RIA Novosti, RBK daily and Novyj region 2. The size is about 229 million words from may 2010 to august 2014. Since the subcorpus does not contain data to study newspapers from a broader spectrum of the political field, I have created a new corpus by including news material from some of the largest Russian outlets, communist papers and a nationalist paper. The compiled corpus contains the following papers: Interfax, Lenta, Novaja gazeta, Komsomol'skaja pravda, (KPRF) Pravda, RIA Novosti, Sovetskaja Rossija and Zavtra. It contains around 365 million words from the time period 2000–2018.

The linguistic items in the established newspaper corpus are compared with the same terms in the compiled Russian media corpus, based on media outlets belonging to different ideologies: communism, nationalism and Russian mainstream media. This paper sets out to not only evaluate a single discourse in different corpora, but also to examine additional discourses, such as the nationalistic one and the communist one.

The comparison between the two corpora has revealed differences in frequencies. The preliminary study of the data of the compiled corpus shows variations between the news outlets of different ideologies in frequencies, usages and rhetorical functions.

References

Blakkisrud, H. (2016). Blurring the boundary between civic and ethnic: The Kremlin's new approach to national identity under Putin's third term. In

P. Kolstø & H. Blakkisrus (Eds.), *The New Russian Nationalism, Ethnicity and Authoritarianism 2000–2015*, 249–274. Edinburgh: Edinburgh University Press.
Ferm, L. (1994). *Osobennosti razvitija russkoj leksiki v novejšij period (na materiale gazet)*. Department of Slavic Languages, Uppsala University.
Kolstø, P. (2016). The ethnification of Russian nationalism. In P. Kolstø & H. Blakkisrus (Eds.),*The New Russian Nationalism: Imperialism, Ethnicity and Authoritarianism 2000–2015*, 18–45. Edinburgh: Edinburgh University Press.
Pain, E. (2016). Sovremennyj russkij nacionalizm: dinamika političeskoj roli i soderžanija.*Vestnik obščestvennogo mnenija.* 1–2.
Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)* (Vol. 55). John Benjamins Publishing.
Raciburskaja, L. (2014). Dinamicheskie aspekty internacionalizacii v sovremennom medijnom slovotvorchestve. *Vestnik Volgogradskogo gosudarstvennogo universiteta.* Serija 2: Jazykoznanie, 5.
Zelenin, A. (2001). Slova s pristavkoj anti-, protivo- v ėmigrantskoj publicistike. *Russkaja reč', 3*, 82–86.
Zelenin, A. (2007). *Jazyk russkoj ėmigrantskoi pressy (1919–1939).* St. Petersburg: Zlatoust.
Zemskaja E. 1996. Aktivnye processy sovremennogo slovoproizvodstva. In E. Zemskaja (Ed.), *Russkij jazyk konca XX stoletija (1985–1995)*, 90-141. Moscow: Jazyki russkoj kul'tury.

••••

Ranka Stanković
University of Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs

Miloš Utvić
University of Belgrade, Serbia
misko@matf.bg.ac.rs

Aleksandra Tomašević
University of Belgrade, Serbia
aleksandra.tomasevic@rgf.bg.ac.rs

Ivan Obradović
University of Belgrade, Serbia
ivan.obradovic@rgf.bg.ac.rs

Biljana Lazić
University of Belgrade, Serbia
biljana.lazic@rgf.bg.ac.rs

Development and application of a domain specific corpus for mining engineering

Lexical resources play an important role in management of project documentation in a specific domain. Language corpora and electronic dictionaries are the most important among them. General lexical resources for Serbian have been developed for several decades and have reached a considerable size to date [1]. However, resources covering domain specific terminology still require further development for many fields, including mining engineering.

In this paper we describe how a corpus of engineering documentation in the mining domain is used to enrich Serbian lexical resources, particularly to add terminology specific for the mining domain to the system of Serbian morphological electronic dictionaries [2].

The corpus of mining engineering documentation (RudKor) is developed at the University of Belgrade. This special corpus originated from the ROmeka@RGF digital library [3], firstly as a means of improving the search of the digital library based on linguistic annotation, and then as a resource for various linguistic and terminological research, including extraction of terms

and other tasks in the field of knowledge engineering. The paper compares several possible versions of the professional language corpus that can be developed for the mining domain, that is, the software packages that can be used for creation, management and search of such a corpus. Three different systems were used to create corpus versions for diverse types of usage scenarios: i) IMS Open Corpus Workbench (CWB) and an adaptation of CQPweb, a web-based graphical user interface [4], ii) Unitex [5], used to create a second corpus from the same texts for custom information extraction tasks and iii) NoSketch Engine [6].

Mining is one of the domains that were only recently introduced in system of Serbian morphological dictionaries. The concepts and terminology specific for the mining domain required the introduction of this new domain, and its subdomains. In order to allow the extraction of specific concepts and relations between concepts by creating lexical masks, new semantic markers relevant to the field of mining have also been proposed. For a more precise description of mining terms, a list of sub-domain markers is also defined. They are aimed at marking more specific areas within the mining domain. Sub-domain markers are associated to the domain marker, so for example, the +Mining+Surface marker would indicate the terms belonging to the domain Mining and sub-domain Open-pit exploitation. The inspiration for semantic markers selection for mining was EarthResourceML, a standard for the exchange of XML-based information on mineral phenomena, resources and reserves, mines and mining activities, as well as the production of concentrates, output products, and mining waste.

Integration of dictionaries and corpus is fulfilled through the RESTfull web services. Both CQPweb and NoSketch Engine web-based corpus query interfaces were adapted to use RESTfull web services in order to retrieve synonyms, antonyms, hypernyms and other related terms in lexical resources. Several expansions of query syntax were introduced to support semantic search. E.g. apart from standard positional attribute *lemma, new fictive attributes synlemma, antlemma,... were implemented to expand a value of attribute lemma* (denoted by L) with any lemma in the dictionary (denoted by X) such that there exists a corresponding semantic relation (synonymy for *synlemma, antonymy for antlemma,...) between lemmas L and X. For example, query: [lemma="aktivan"] [synlemma="rudnik"]* retrieves concordances with all inflected forms of lemma aktivan (aktivnom, aktivnog, aktivnim,...) followed by all inflected forms of all synonyms of lemma rudnik /mine/ including e.g. word forms of synonyms površinski kop and kop /open pit, surface mine/: aktivnim površinskim kopovima, aktivnog površinskog kopa, aktivnih rudnika, aktivnog rudnika,...

Available documentation in the field of mining (172 documents) served as a basis for creation of the corpus of texts from the mining domain, and related research work on extraction of mining terminology, text annotation, information extraction, etc. All documents collected were systematized, described by metadata, and stored in the digital library, while the text within the documents was processed using available electronic dictionaries and Unitex local grammars. The Unitex local grammars, often called syntactic graphs, allow description of syntactic patterns that can then be searched in the texts [5, 7]. After the preprocessing of the text, 150,365 sentences and 2,719,086 (100,414 different) words were identified. Around 1900 words (excluding compound terms) specific to mining have been extracted from these texts and included in the system of electronic morphological dictionaries, thus providing for further extraction of domain specific compound terms. For example, the word *mašina* (equipment) can be used as the trigger for the extraction of the terms *pomoćna mašina* (auxiliary equipment), *mašina za bušenje* (drilling equipment), *mašina za transport* (transportation equipment), etc. This paper will present examples of the graphs used for extraction combining part of speech with the semantic and domain markers.

Given that mining is a very complex and multidisciplinary industrial branch, the construction of a special corpus for the mining domain and the enrichment of the system of electronic dictionaries with domain specific terminology can serve as a model for the development of professional language corpora in other engineering fields.

References

1. Krstev, C., (2008). *Processing of Serbian – Automata, Text and Electronic Dictionaries*, Faculty of philology, Belgrade.
2. Tomašević, A., Lazić, B., Vorkapić, D., Škorić, M. & Kolonja, Lj. (2017). The Use of The Omeka Platform for Digital Libraries in the Field of Mining, *Infotheka*, Vol. 17, No. 2, 2017, ISSN: 2217-9461, DOI 10.18485/infotheca.2017.17.2.2.
3. Tomašević, A., Stanković, R., Utvić, M., Obradović, I., Kolonja, B. (2018). Managing mining project documentation using human language technolo-

gy, *The Electronic Library*, ISSN 0264-0473, DOI 10.1108/EL-11-2017-0239. [accepted for publishing]
4. CQP (2017). The IMS Open Corpus Workbench (CWB), CQPweb [Online]. Available: http://cwb.sourceforge.net/cqpweb.php [Accessed 21.12.2017]
5. Unitex (2017). Unitex/GramLab [Online]. Available: http://unitexgramlab.org/ [Accessed 10.08.2018]
6. Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Masaryk University. p. 65-70. ISBN 978-80-210-4471-5.
7. Gross, M. (1997). *The Construction of Local Grammars*. Finite-State Language Processing, The MIT Press, pp. 329-352.

••••

Ilona Starý Kořánová
Charles University, Czech Republic
ilona.koranova@gmail.com

Aspectual homonymy and polysemy in Czech

The Czech aspect is usually described in terms of the perfective – imperfective distinction and aspectual pairs (*psát – napsat*). Aspect is considered a grammatical category of the Czech verb, it is expressed through inflection. Nevertheless, this is not always the case as will be demonstrated through instances of aspectual homonymy and polysemy.

Homonym is an expression that has two meanings/functions, that have no semantic relation. It is accidental similarity between two expressions. Example: *Kvůli mléku dojí krávu. (impf.) Dojí zbytek večeře.* (pf.) A polyseme is an expression with different, but related meanings.

The paper focuses on homonymous and/or polysemous verbal forms of the Infinitive, Present, Imperative, as well as Past and Passive Participles. These forms are defined by the following parameters:

1) Degree of homonymy. The degree of homonymy between two verbs can vary, sometimes only one of the levels of the paradigm is homonymous. In other cases, the level in which the perfective/imperfective forms are homonymous, stretches up to a degree of complete overlap that is a homonymy of the entire paradigm.

Examples: *snít – Narkotika netoleruji. Sním bez drog.* (impf.) and *sníst – Večer sním, na co přijdu.* (pf.) share forms in the Indicative Present only.

okolkovat – Bush okolkuje, odkládá rozhodnutí (impf.) versus *Na lince okolkuje 850 lahví za hodinu* (pf.) are homonymous in the whole paradigm except for Passive Participle since the Passive Participle of the transitive verb *okolkovat* doesn't exist.

2) Dynamics of the axis "aspectual homonymy – aspectual polysemy" of the verbal forms under consideration is another parameter followed.

Example: *Dolétat – Obraz hvězd k nám dolétá s notným zpožděním* (impf.) *Až příští rok dolétá raketoplán, budeme se z oběžné dráhy vracet pouze pomocí padáků* (pf.) The prefix *do-* expresses two different meanings, however a formal and semantic relatedness can nevertheless be observed.

3) The third relevant parameter derives from the fact that aspectual interpretation of a sentence isn't exclusively linked to the verbal form. It is

also dependent on the circumstances in which it is used. Perfective versus imperfective interpretation depends for instance on:

1. **Nature of the subject:** *opadat – nadšení opadá* versus *listí opadá:* If the subject an abstract noun or *voda (water)*, the verb *opadat* is interpreted imperfectively. If the subject is divisible into elements or parts, the verbal form is interpreted as a result of the event and hence perfectively.
2. **The tenses used:** *obejít se – Na dovolené jsme se obešli bez auta.* Events that took place in the past (Preterit) tend to be interpreted as perfective.
3. **Activity or state:** Interpretation of the verbal form as an activity (involves a change) or a state. States "describe situations that do not change over time, e.g. are stative" (Croft, 2012, 34), states fade into qualities and relations, they are perceived imperfectively: *Po oční operaci se už rok obejde bez brýlí.*

The fact that aspectual homonymous and polysemous expressions exist, implies that the aspectual interpretation of a sentence is not given by the morphological make-up of the verb only. Besides that, the aspectual interpretation is co-determined by aspectual markers, by (non)existence of analytical future tense, by compatibility of the particular verbal form with phasal verbs etc., the aspectual interpretation of a sentence is also dependent on the specific situation in which the particular sentence is used.

References

Bermel N., I. Kořánová (2008). From Adverb to Verb: Aspectual Choice in the Teaching of Czech as a Foreign Language. In C. Cravens & M.U. Fidler & S.C. Kresin (Eds.), *Between Texts, Languages, and Cultures*, Slavica Publishers, Indiana University, Bloomington 53-70.

Croft, W. (2012). *Verbs, Aspect and Causal Structure.* Oxford: Oxfrod University Press.

Dostál, A. (1954). *Studie o vidovém systému v staroslověnštině.* Praha: SPN.

Kopečný, F. (1962). *Slovesný vid v češtině.* Praha: Nakladatelství ČSAV.

Petkevič, V. (2010). *Morfologická homonymie v současné češtině.* Praha: NLN.

Starý, Z. (2017) Biaspectuals revisited. *Sali*, 1, 111-123.

Trávníček, F. (1923). *Studie o českém vidu slovesném.* Praha: Česká akademie věd a umění.

Vendler, Z. (1967). Verbs and times. In: *Linguistics and Philosophy.* Ithaca: Cornell University Press, 97-121.

....

Marcin Szczepański
Serbski institut
marcin.szczepanski@serbski-institut.de

Recent challenges and advances in the development of Lower Sorbian corpus resources

The Sorbian languages count as lesser-used languages, minority languages and endangered languages. Nonetheless, they have a relatively rich and alive literary tradition and are objects to active linguistic research. Such specific circumstances determine the process of language resources development and set out the functional requirements to it. This report focuses on some recent advances in the development of the Lower Sorbian corpus resources.

Lower Sorbian text corpus

The earliest attempts to build the Sorbian text corpora date back to 90's. The results were rather unstructured collections of digitised works, irregularly acquired and poorly standardised. The first broader yet not comprehensive unification of the Lower Sorbian text corpus occurred in 2010 and was related to the partial on-line publication of the corpus content along with the query interface provided by Institute of the Czech National Corpus (then Bonito, nowadays KonText) and Sorbian Institute (dolnoserbski.de).

The development efforts intensified around 2015. The goal is to arrange a chain of tasks, which yields a fully operable and versatile text corpus. At this moment the corpus size is 37 million tokens. According to the limited personal capacities of the Sorbian Institute and the lack of a dedicated department or group, many tasks have to be carried out as a more or less official part of different research projects. Setting aside the digitalisation process, metadata maintenance, character encoding, quality assurance, morphological analysis and copyright clearance, the further report focuses on two tasks: the structural annotation and lexical analysis.

The typical corpus-linguistic, quantitative research is not the only purpose of the Lower Sorbian corpus. It aims to be also useful for historical, culture, social and didactic studies. Therefore the word- or even sentence-oriented interaction may not be enough. The longer text units (paragraphs, articles) should be recognised as entities and the content has to be readable

by humans, navigable and bibliographically addressable. This applies especially to the newspapers, which build the corpus core. To achieve that, some kind of structural annotation is necessary and the TEI-P5 standard has been chosen. The internal guidelines define several levels of detail: from a general TEI container (obligatory for all documents) to precise description of the functional layout structure, which should be enough for the text interpretation close to the printed original, and opens the way to the text reproduction (e.g. for a digital library).

Currently the Lower Sorbian corpus does not support queries on lemmas, which is obviously a serious restriction. In case of strongly inflectional Slavic languages an automated lemmatisation is no option. Developing a versatile morphological analyser can be a challenge. However, to a satisfactory extent the goal has been achieved. As the main source of the lexical and grammatical data the German-Lower Sorbian Dictionary (2003-) was used. This digitally-born resource provides rich and precisely encoded information about Lower Sorbian. The Lower Sorbian-German Dictionary (1999, fully processed and digitally encoded in 2012) was queried too. On that basis, an automated tool for generation and recognition of inflected forms has been developed. Even though the system is highly inflection-oriented, it gives satisfactory result for the automated lemmatisation. Its database consists of ca. 76 thousand lexemes and is to be continuously extended (e.g. with the proper names from Muka's dictionary in 2018). It is worth noting, that the mentioned dictionaries, thanks to the rich illustrational language material and digitally accessible information, are already an attractive research source (e.g. for an early form of a phraseological dictionary composed in 2014).

Due to diachronic nature of the Lower Sorbian text corpus and the variety of spellings, the lemmatisation relies upon the normalisation. This task is being carried out as a part of the ESF project „Sorbian knowledge" in its module „Interdisciplinary research corpus". The main goal of this project is to set up an environment for the ongoing processing of the corpus data in automated and interactive ways. The established tools will be also applied to develop, among others, the contemporary Lower and Upper Sorbian reference corpora.

Text and audio corpus of native Lower Sorbian

A special type of corpus is Text and audio corpus of native Lower Sorbian. It is a collection of recordings from the last Lower Sorbian native-speakers. It was completed in 2011-2016 as a part of the international endeavour „Documentation of Endangered Languages". Its outcome is over 100 hours (ca. 700 thousand tokens) of an original speech material. All recordings have been fully transcribed and normalised. Along with that comes a full German translation. For selected parts there is also English translation and phonetic transcription available. The freely accessible long-term repository is a part of The Language Archive. The custom Web-interface at dolnoserbski.de provides extended options for searching in the transcription and translation, as well as a navigable access to the sound and text content.

References

Bartels, H., Thorquindt-Stumpf, K. (2013). Ein neues Ton- und Textarchiv des muttersprachlich-dialektalen Niedersorbischen. *Lětopis*, 60 1, 39-60.

Kaulfürst, F. (2014). Praktyczny przewodnik po korpusie języka dolnołużyckiego. In M. Hebal-Jezierska (Ed.), *Praktyczny przewodnik po korpusach języków słowiańskich*. Warszawa: Uniwersytet Warszawski, 67-75.

Magda Ševčíková
Charles University, Czech Republic
sevcikova@ufal.mff.cuni.cz

Adéla Kalužová
Charles University, Czech Republic
kaluzova@ufal.mff.cuni.cz

Zdeněk Žabokrtský
Charles University, Czech Republic
zabokrtsky@ufal.mff.cuni.cz

A language resource specialized in Czech word-formation: Recent achievements in developing the DeriNet database

The paper reports on recent progress in development of the lexical database DeriNet. DeriNet is a large language resource which has been built at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, since 2013 (Ševčíková and Žabokrtský 2014). It is still one of only few, rather recent resources and tools that focus on derivational morphology of Czech (cf. Deriv by Osolsobě et al. 2009, Morfio by Cvrček and Vondřička 2013, Derivancze by Pala and Šmerk 2015, or the dictionary of affixes by Šimandl et al. 2016).

DeriNet has been designed as a resource specialized in derivation of Czech but, recently, the structure of the database has been modified in order to allow for capturing compounding and combined word-formation processes, too. The ambition is to cover a major part of the word-formation system of Czech in all its complexity.

DeriNet contains more than 1 million lexemes in four part-of-speech categories (nouns, adjectives, verbs, and adverbs). The set of lexemes in DeriNet is based on the large-scale morphological dictionary of Czech called MorfFlex CZ (Hajič and Hlaváčová 2013). DeriNet is thus considerably larger than most comparable resources developed for Czech (cf. the resource Derivancze) and for other languages, e.g. Word Formation Latin (Litta et al. 2014), Démonette for French (Hathout and Namer 2014), DErivBase for German (Zeller et al. 2013), DerivBase.Hr for Croatian (Šnajder, 2014), or CELEX for English, German, and Dutch (Baayen, 1995).

In DeriNet, the approach to derivation is based on the linguistic account of Czech word-formation provided by Dokulil (1962), most importantly on the notions of a word-formation line (a chain of words that were derived subsequently in several steps, ex. (1)) and of a word-formation nest (a set of word-formation lines that share one or more items; the derivational nest with the root noun *dřevo 'wood'* contains also rows listed in (2)). A base word and a word derived from it were connected by a link represented as an oriented edge. In the original structure of the database, at most one base word was allowed for each lexeme. Word-formation nests were thus modelled as rooted trees where the root node corresponds to the unmotivated or underived word.

(1) *dřevo 'wood' → dřevák 'wooden shoe' → dřeváček 'small wooden shoe'*
(2a) *dřevo 'wood' → dřevěný 'wooden' → dřevěnice 'wooden cottage'*
(2b) *dřevo 'wood' → dřevař 'woodcutter' → dřevařův 'woodcutter's'*
(2c) *dřevo 'wood' → dřevař 'woodcutter' → dřevařský 'related to woodcutters'*

Pairs of base words and derivatives were captured by automatic and semi-automatic methods. Whereas the core part of (regular, high frequency) derivational relations was created by rules based on substitution of affixes (or either longer or shorter strings), less frequent and irregular patterns (esp. with morphophonemic alternations) had to be identified by manually compiled lists and manual annotation. Existing data resources were used for specific groups of derivatives; for instance, the Vallex dictionary (Lopatková et al. 2017) proved useful when searching for verbs with the same root morpheme (cf. Ševčíková et al. 2016 and 2017 for details). The current version of the database, DeriNet 1.5, contains more than 1 million lexemes connected with approx. 785 000 derivational links.

In addition to derivation, a pilot annotation has been carried out recently that focused on identification of compounds in the database. Words which were coined primarily by compounding or combined processes, such as *modrooký 'blue-eyed'*, were marked explicitly as compounds in DeriNet. Words which were derived from compounds (e.g. *dřevorubecký 'related to lumber-*

jacks' from *dřevorubec 'lumberjack'*) are assumed to inherit the "compoundness" feature from their bases. Loan words which are compound in their original language, e.g. *kovboj 'cowboy'*, were not marked as compounds, unless both (all) stems also can be considered Czech words.

So far, compounds have been identified using various methods, most prominently searching for frequent compound parts, and marked by adding "C" to their part-of-speech tag. This preliminarily annotation is available for approximately 30 000 compounds in DeriNet 1.5. In the very next future, the identification of compounds will be continued and compounds will be connected with their multiple parents.

DeriNet 1.5 was released under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA) in the Lindat/Clarin repository (Vidra et al. 2017). The DeriNet data can be also searched by two online tools, namely DeriNet Search and DeriNet Viewer (see http://ufal.mff.cuni.cz/derinet/search and http://ufal.mff.cuni.cz/derinet/viewer). In the paper, the potential of the resource for both linguistic research and experiments in Natural Language Processing will be exemplified by recent case studies based on the data.

Acknowledgement:

This work was supported by the Grant No. GA16-18177S of the Czech Science Foundation. It has been using language resources developed, stored, and distributed by the INDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

Baayen, R. H. et al. (1995). *The CELEX lexical database* (release 2). Data/software. Philadelphia, PA: LDC.

Cvrček, V. – Vondřička, P. (2013). Nástroj pro slovotvornou analýzu jazykového korpusu. In J. Klímová (ed.), *Gramatika a korpus 2012*. Gaudeamus: Hradec Králové.

Dokulil, M. (1962). *Tvoření slov v češtině 1: Teorie odvozování slov*. Praha: Nakladatelství ČSAV.

Hajič, J. – Hlaváčová, J. (2013). *MorfFlex CZ*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

Hathout, N. – Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11, pp. 125–168.

Litta, E. et al. (2016). Formatio formosa est. Building a Word Formation Based Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016)*, pp. 185–189.

Lopatková, M. et al. (2017). *Valenční slovník českých sloves VALLEX*. Praha: Karolinum.

Osolsobě, K. (2009). Deriv – nástroj pro automatické vyhledávání slovotvorných vztahů. Slovotvorný stroj pro češtinu – sen nebo skutečnost? In: *Přednášky a besedy z XLII. běhu LŠSS*. Brno: FF MU, s. 132–137.

Pala, K. – Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In: *Proceedings of Text, Speech and Dialogue 2015*. Berlin: Springer, pp. 515–523.

Ševčíková, M. – Žabokrtský, Z. (2014). Word-Formation Network for Czech. In: *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC 2014)*. Paris: ELRA, pp. 1087–1093.

Ševčíková, M. et al. (2016). Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině. *Časopis pro moderní filologii*, 98, pp. 62–76.

Ševčíková, M. et al. (2017). Identification of aspectual pairs of verbs derived by suffixation in the lexical databse DeriNet. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*. Milan: EDUCatt, pp. 105–116.

Šimandl, J. a kol. (2016). *Slovník afixů užívaných v češtině*. Praha: Karolinum.

Šnajder, J. (2014). DerivBase.Hr: A High-Coverage Derivational Morphology Resource for Croatian. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavík: ELRA, pp. 3371–3377.

Vidra, J. et al. (2017). *DeriNet 1.5*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2422.

Zeller, B. et al. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics 2013*, pp. 1201–1211.

••••

Svatava Škodová
Charles University, Czech Republic
svatava.skodova@ff.cuni.cz

Sebrat se a …. a construction between coordination and subordination in contemporary Czech

This text investigates the syntactic behaviour of the word SEBRAT SE[1] in constructions where a coordinate form displays subordinate properties, resulting in constructions that cannot be clearly categorized as either coordination or subordination and that can be characterised as peudocoordination (Ross, 2015).

The pattern I am going to investigate is:

V1 *sebrat se* – *a/and* – V2

i.e. ***Seber se a*** *hledej znovu! /* ***Go and*** *look for him again.*
Seber se a *padej!* / ***Go on,*** get moving!
Sebrala se a *hned se vrátila. /* ***She picked herself up****. She came straight back .*
Sebral se a *zmizel, jen co se narodilo třetí dítě. /* ***He up and*** *left soon's the third kid came along .*

My research of this construction in Czech is based on examples gathered from the Czech National Corpus.

The aim of this paper is to introduce the formal and semantic features of pseudocoordination in Czech, which is an inflective language and constructions of compound character are rare in its grammatical system (for Russian see Kiparski, 1971; Škodová, 2009).

I consider pseudocoordination (PseCoor) to refer to the use of the coordinator '*and*' in constructions that behave unlike prototypical coordination (ProCoor), defined as a transitional state between coordination and subordination (Haspelmath, 2005). The resulting constructions still display some properties

1 There is no straighforward translation of the word SEBRAT SE into English. Very often it is not translated at all. Sometimes, the present continuous tense is used to express the meaning; sometimes, the verb GO (usually in imperativ) is used.

of coordination and cannot be definitively identified as either coordination or subordination. Thus, syntactic analysis of pseudocoordination is challenging and important, and thorough description is required.

Both ProCoor and PseCoor construction types have in common a binary coordinative structure, using the coordinator *a* (*and*). The main claim is that even though these two types share the same surface structure (*sebrat se*) V1*and*V2, they do not represent the same phenomenon of coordination and it is necessary to distinguish them, as proposed.

I have proposed a two-part analysis. Firstly, PseCoor is analysed as a complex predicate formed on the level of syntax (Hilpert, 2008). This analysis immediately accounts for a number of properties of PseCoor which allows the comparison with ProCoor. Secondly, PseCoor is analysed as a means of aktionsart, more precisely as a variety of coordination of substages in the event structure. This also accounts for a number of characteristics of PseCoor, this time on the level of semantics.

In this way, I also presented the criteria for the distinction between ProCoor on one side and PseCoor on the other side in Czech. I argued that ProCoor is a biclausal structure coordinating two separate events while PseCoor coordinates two verbs into one complex predicate and the coordinator *a/and* serves for coordination in the frame of substages of a single event (comp. Ross, 2015).

On the semantic level I characterised PseCoor as a complex event, the substages of which are coded into two conjuncts of the coordinative construction. It appears that the verb in the first conjunct denotes an event that expresses the preparation phase for the activity denoted by the verb in the second conjunct. The pseudo-coordinative verb in the first conjunct lexicalises a manner component in the internal event structure. The verb *sebrat se* in the first conjunct goes through the process of desemantisation and, instead of the meaning of taking, expresses dynamic aspects of the second event and the decision of the actor of the event to finish the second event.

References

HASPELMATH, Martin. (2005). Nominal and Verbal Conjunction. In Martin HASPELMATH, Matthew S. DRYER, David GIL and Bernard COMRIE (eds.) *World atlas of language structures*. Oxford: Oxford University Press, 262-265.

HILPERT, Martin and Christian KOOPS. (2008). A quantitative approach to the development of complex predicates. *Diachronica* 25 (2), 240-259.
KIPARSKY, Valentin [Кипарский, Валентин]. (1971). «Взял и … л». In Viktora Ivanoviča Borkovskogo (ed.) *Проблемы истории и диалектологии славянских языков.* Moscow: Izdatel'stvo "Nauka.", 134-139.
ROSS, Daniel, Ryan GRUNOW, Kelsey LAC, George JABBOUR and Jack DEMPSEY. (2015). Serial Verb Constructions: a distributional and typological perspective. *Presented at the Illinois Language and Linguistics Society (ILLS) 7*, University of Illinois at Urbana- Champaign, Urbana, Illinois. http://hdl.handle.net/2142/88844
ŠKODOVÁ, Svatava. (2009). *Pseudokoordinace v syntaxi češtiny.* Liberec: Bor.

••••

Petar Vuković
University of Zagreb, Croatia
petar.vukovic@ffzg.hr

The second future tense in contemporary Croatian: A corpus-driven study in grammatical semantics

The second future tense in contemporary Croatian is a periphrastic form composed of the l-participle and the auxiliary biti in perfective present (budem, budeš, bude, budemo, budete, budu). Its use is limited to several types of subordinate clauses joined to main clauses expressing future actions. In the main clauses, verbs are usually in the first future tense, although imperative, conditional, and present tense forms are also possible. In the normative grammars, the second future tense was traditionally referred to as "the future anterior", since it was believed to denote a future action preceding the future action of the main clause. However, this is neither complete nor accurate description of its grammatical meaning. It is also important to note that the second future is used mostly with imperfective verbs, while perfective verbs typically take present forms to express the same grammatical meaning(s) in these syntactic contexts. Despite that, the second future tense of perfective verbs is not entirely ungrammatical and some linguist even claim that it is used to mark resultative meaning (Silić & Pranjković 2005). Despite that, prescriptive grammars often warn against this use (Raguž 1997).

The aim of the paper is to contribute to the functional and grammatical description of the second future tense in contemporary Standard Croatian. The paper is based on an analysis of a random sample of 10% of all examples of the second future tense forms found in the Croatian National Corpus 2.5, first morphologically annotated corpus of contemporary Standard Croatian with over 100 million words.

The analysis demonstrates that the second future tense has at least two different grammatical meanings, which are expressed in different syntactic contexts. In temporal clauses it refers to a future action in general, which can precede, be simultaneous with or follow the future action of the main clause. The first future tense cannot be used in temporal clauses. On the other hand, in conditional, relative, local, modal, quantitative and comparative clauses, the second future tense refers to a future action preceding or simultaneous

with the future action from the main clause. The first future tense can be used in these types of subordinate clauses and it refers to a future action following the future action of the main clause. There are several more types of subordinate clauses in which the second future tense can be used, but the Croatian National Corpus does not contain enough examples of them, so the analysis was not possible.

In approximately 90% of the analysed examples, the second future tense is used with imperfective verbs, but the analysis does not support the claim by Silić and Pranjković that the function of the second future tense with perfective verbs is to mark resultative meaning. If examples with perfective verbs have something in common, it is the fact that most of them are connected to syntactic parallelisms and/or spoken discourse.

References

Grickat, I. [И. Грицкат] (1956/57). О неким особинама футура II (футура ефзактног). *Наш језик*, 8, 89-103.

Katičić, R. (1986). *Sintaksa hrvatskoga književnog jezika*. Zagreb.

Katičić, R. (1992). Kategorija gotovosti u vremenskom značenju glagolskih oblika. In: *Novi jezikoslovni ogledi*. Zagreb, 172-183.

Kravar, M. (1959/60). Futur II. u našem glagolskom sistemu. *Radovi Filozofskog fakulteta u Zadru*, 1, 30-50.

Milošević, K. (1970). *Futur II i sinonimski oblici u savremenom srpskohrvatskom književnom jeziku*. Sarajevo.

Raguž, D. (1997). *Praktična hrvatska gramatika*. Zagreb.

Silić, J. & I. Pranjković (2005). *Gramatika hrvatskoga jezika*. Zagreb.

Svane, G. [Г. Сване] (1959). О конструкцији будем + партицип на –л у српскохрватском језику. *Scando-Slavica*, 5, 30-51.

Vuković, J. (1967). Futur drugi i ekvivalentni glagolski oblici po upotrebi u srpskohrvatskom jeziku. In: *Sintaksa glagola*. Sarajevo, 246-274.

••••

Adrian Jan Zasina
Charles University, Czech Republic
adrian.zasina@ff.cuni.cz

Evaluating a corpus-driven approach in L2 classroom on the example of Czech

In the recent decade, the involvement of corpus methods in language teaching has become increasingly popular in the Czech Republic. Quite understandably, the attention focused first on English teaching in the Czech context (cf. Thomas 2006) and then step by step extended to Czech as a foreign language. Today there are a number of papers on using corpora in L2 Czech teaching (Lukšija 2010, Osolsobě 2010, Vališová 2012, Konečná & Zasina 2014), however, there is still a lack of studies aiming to validate the utility of corpus-driven exercises and their real influence on students' progress.

This paper aims to fill this gap and comment on whether a corpus-driven approach in the L2 classroom has a positive impact on students' language development, based on a thorough analysis of learners' errors and a semester-long teaching experiment (inspired by a similar study by Leray & Tyne 2016 on L1 French).

Data and methodology

As a basis for the study, two corpus resources were used. First, I analysed the CzeSL-SGT (Šebesta et al. 2014) learner corpus which includes automatic students' error annotation in addition to the standard morphological tagging. The corpus contains texts of non-native Czech speakers with several different L1 and has the total of 1,147,477 tokens (incl. punctuation). I focused only on the subcorpus of Slavic L1 speakers amounting to 769,126 tokens (incl. punctuation) to identify students' most frequent errors (e. g. spelling, declension) in a homogeneous group. Second, I used the representative corpus of contemporary written Czech SYN2015 (Křen et al. 2015; Křen et al. 2016) as a main source of data for creating tailor-made teaching materials reflecting the most problematic areas of study in L2 Czech.

Experiment and evaluation

The next step consists of an evaluation of the proposed materials and methods in a real university class. The experiment took place in the spring

semester 2018 at the Jagiellonian University and included two comparable groups of university students of Czech Studies (15 students in total) who enrolled for the course of grammar and lexis exercises. In the first group, the corpus-driven approach was employed, whereas the second group (a control group) learnt the same subject based on traditional textbooks and methods.

To ensure the comparability of the groups and to minimise any secondary factors influencing the acquisition of foreign language, I decided to work with two homogeneous groups in terms of mother tongue (Polish) and age (19–24). The goal of the investigation was to examine whether the corpus-driven approach improves the effectiveness of language teaching and whether the proposed procedure for explaining a problematic language phenomenon using corpora (i.e. a. identifying the problem, b. solving the problem using corpus methods, c. interpreting the results) is in fact functional in practise. Both groups were subjected to an entry test and a final test. The entry test results in both groups helped to reveal any major error areas that were covered during the semester such as: declension, long vowels, grammatical gender, stylistic variants, collocability, vocalisation of prepositions, past participle, and differentiation of hard and soft adjectives. The findings based on the final test showed that both groups significantly improved their Czech; the group with corpus approach was better by 23.04 percentage points (36.98%) and the control group was better by 21.86 percentage points (35.94%). It has exposed that both methods were efficient, but a corpus-driven approach brought a slightly better result. Previous study by Leray & Tyne (2016) shows improvement in favour of the corpus method as well.

Using corpus methods in SLA of Czech is a rather new approach in Czech didactics and more research needs to be done in this area. The present study proved that corpus exercises help students acquire linguistic knowledge as well as traditional methods, and they might be used in the classroom as a supplement of traditional language learning. The presented research will hopefully cast a new light on using such methods in practise and provide useful information both for researchers and teachers of Czech as a foreign language.

References

Konečná, H., & Zasina, A. J. (2014). Studium českého jazyka a internet [Studing Czech and the Internet]. In E. Rusinová (Ed.), *Přednášky a besedy ze XLVII. běhu LŠSS*. Brno: Masarykova univerzita. 104–112.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. (2015). *SYN2015: a representative corpus of written Czech*. Praha: Ústav Českého národního korpusu FF UK. Available at: http://www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., & Zasina, A. J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: ELRA. 2522–2528.

Leray, M., & Tyne, H. (2016). Homophonie et maîtrise du français écrit: apport de l'apprentissage sur corpus [Homophony and proficiency of writing in French: contribution of data-driven learning]. *Linguistik online*, 78(4), 131–150.

Lukšija, M. (2010). *Korpus jako zdroj dat při prezentaci předložek do/na s místním směrovým významem ve výuce češtiny pro cizince* [Corpus as a data source in presenting of prepositions *do/na* with local directional meaning in L2 Czech teaching]. BA thesis. Brno: Masarykova univerzita. Available at: http://is.muni.cz/th/217240/ff_b/

Osolsobě, K. (2010). Jak se učit česky s korpusem [How to study Czech using a corpus]. In E. Rusinová (Ed.), *Přednášky a besedy z XLIII. běhu LŠSS*. Brno: Masarykova univerzita. 112–119.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M., & Rosen, A. (2014). *CzeSL-SGT: corpus of non-native speakers with automatical annotation*, version 2 from 28 Jul 2014. Praha: Ústav Českého národního korpusu FF UK. Available at: http://www.korpus.cz.

Thomas, J. (2006). Using Corpora in Language Teaching and Learning. *Teaching English with Technology, A Journal for Teachers of English*, roč. 2005, 6 (1).

Vališová, P. (2012). Data-driven learning a výuka češtiny jako cizího jazyka [Data-driven learning and L2 Czech teaching]. *CASALC Review* 2, 22–39.

••••

Adrian Jan Zasina
Charles University, Czech Republic
adrian.zasina@ff.cuni.cz

Michal Škrabal
Charles University, Czech Republic
michal.skrabal@ff.cuni.cz

Morfio.pl – the possibilities for the application of Czech corpus tools to other languages

In our paper, we focus on the possibility of the application of the Morfio tool (Cvrček & Vondřička 2011a, 2011b) to other languages besides Czech. The Morfio tool was successfully applied to the Polish part of the InterCorp parallel corpus last year (Zasina 2017). In this pilot study the competitive ending *-a*/*-u* in the genitive singular form of Polish masculines was analysed. The Morfio tool has proven to be a suitable complement to the Polish provenance data and tools such as the Polish National Corpus (NKJP) and the corpus browsers PELCRA and Poliqarp. With none of these was it possible to search for a pair of two words in the Polish corpus which have a common part and differ only in the given inflectional suffixes. In other words, they did not provide a ready estimate of the productivity of word-forming models. Yet, lists generated in this way have wide potential uses: a pedagogical, but also a lexicographic, translatological, etc., one.

To use the Morfio tool, it was necessary to specify a query *subst:sg:gen.** (for the Polish tagset see Szałkiewicz & Przepiórkowski 2012). For a fully-fledged and user-friendly Polish version of the tool, an inventory of Polish alternations (for both vowels and consonants) and phonemic groups have been added. In this respect, we rely on prestigious linguistic publications (Kowalik 1999, Ostaszewska & Tambor 2000). In addition, there is nothing to prevent other language versions for further (not only) Slavic languages. The competition of the genitive endings *-a* and *-u* should be studied on wider, general Slavic material (see similar research for Czech: Šimandl 2003, Bermel & Knittl 2012a, 2012b, Bermel, Knittl & Russell 2014; for other West Slavic languages: Žigo 2012; Bígl 2013). It turns out that some languages show different tendencies for the choice of ending in words with the same Proto-Slavic origin (Stieber 2005). We appreciate the Morfio tool as a universal application potentially usable for contrastive research of word formation. The question of the relevant language mutation is secondary, the availability of corpus data is a crucial issue.

Owing to the Morfio tool, we obtained 534 word pairs (with minimal frequency 4) which had to be manually revised. The pairs which included the given name/surname vs. toponym (such as *Jordana* × *Jordanu* 'river', *Waszyngtona* × *Waszyngtonu* 'city; state', *Harvarda* × *Harvardu* 'university') were discarded from the list as well as pairs which do not have the same etymology (*muła* 'mule' × *mułu* 'silt'; *popa* 'orthodox priest' × *popu* 'music genre'; *posta* 'contribution to the Internet discussion' × *postu* 'fasting') and incorrectly annotated words.

The final list of the duplicate forms of the genitive is divided into three groups. The first one contains variants that refer to the same denotation (*bilarda* || *bilardu* 'billiards'; *filara* || *filaru* 'pier'; *wraka* || *wraku* 'wreck'). These are either equal to one another, or one of them is preferred in certain collocations (*od rana do wieczora*, rather than **od rana do wieczoru* 'from morning till evening').

The second group includes examples where the endings *-a* and *-u* change the meaning but the words nonetheless have the same etymological origin (*browara* 'beer' × *browaru* 'brewery'; *mostka* 'sternum' × *mostku* 'bridge'; *skręta* 'joint' × *skrętu* 'turn').

The third group consists of pairs where the variance of the genitive is questionable. These are instances with insufficient or unclear evidence in which one of the variants appeared exclusively or predominantly in the film subtitles, eventually it might concern a typo, mistake or annotation error as well. The list thus suggests potential duplicate forms in Polish. It contains both pairs with the same semantics as variants differing the meaning (*doła* 'depression' × *dołu* 'pit'; *fleta* || *fletu* 'flute'; *kanta* || *kantu* 'edge').

It turned out that this classification can be further developed. Firstly, only those pairs consisting of a name/surname and toponym were excluded from the list, while common nouns with different etymology now form the fourth group. Moreover, from the second group we have divided a subgroup of nouns that have the same etymological basis, but they differ in the category of animateness: one member of a pair is inanimate, the other one animate (including proper names, e.g. *Urana* 'Uranus' × *uranu* 'uranium'). We have applied this new, more detailed classification to new language data: the NKJP_1M corpus (Degórski & Przepiórkowski 2012), the one-million sample

of the NKJP corpus. We were interested in how the results will be affected by both the small size of this subcorpus (cf. the size of the Polish part of InterCorp v9: 83.8 million running words) and its balance and representativeness (features missing for InterCorp). The resulting list is, of course, much smaller, including hapaxes, yet all five groups of masculines are represented. Besides, 26 new pairs not found previously have emerged (e.g. *cyklona* || *cyklonu* 'cyclone', *SMS-a* 'sports school' × *SMS-u* 'SMS', *świerka* || *świerku* 'spruce', *zamka* 'lock' × *zamku* 'castle') although the size of data has decreased substantially.

References

Bermel, N. & Knittl, L. (2012a). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8, 241–275.

Bermel, N. & Knittl, L. (2012b). Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments. *Russian linguistics* 36 (1), 91–119.

Bermel, N., Knittl, L. & Russell, J. (2014). Absolutní a proporcionální frekvence v ČNK ve světle výzkumu morfosyntaktické variace v češtině. *Naše řeč* 97 (4–5), 216–227.

Bígl, R. (2013). *Vývoj lužickosrbského skloňování a stupňování*. Praha: Karolinum.

Cvrček, V. & Vondřička, P. (2013a). Morfio. Praha: Ústav Českého národního korpusu FF UK. Dostupné z http://morfio.korpus.cz.

Cvrček, V. & Vondřička, P. (2013b). Nástroj pro slovotvornou analýzu jazykového korpusu. *Gramatika a korpus 2012*. Hradec Králové: Gaudeamus.

Degórski, Ł. & Przepiórkowski, A. (2012). Ręcznie znakowany milionowy podkorpus NKJP. In A. Przepiórkowski et al. (Eds.), *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 51–58.

Kowalik, K. (1999). Morfonoligia. In R. Grzegorczykowa, R. Laskowski & H. Wróbel (Eds.), *Gramatyka Współczesnego Języka Polskiego. Morfologia*. Warszawa: Wydawnictwo Naukowe PWN, 87–123.

Ostaszewska, D. & Tambor, J. (2000). *Fonetyka i fonologia współczesnego języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Stefańczyk, W. T. (2015). Kupiłem grejpfrut czy grejpfruta? O błędach językowych w glottodydaktyce polonistycznej. *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców* 22, 99–105.

Stieber, Z. (2005). *Zarys gramatyki porównawczej języków słowiańskich*. Warszawa: Państwowe Wydawnictwo Naukowe.

Szałkiewicz, Ł. & Przepiórkowski, A. (2012). Anotacja morfoskładniowa. In A. Przepiórkowski, M. Bańko, R. L. Górski & B. Lewandowska-Tomaszczyk (Eds.), *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 59–96.

Šimandl, J. (2003). Od čtvrtku do pátka. *Naše řeč* 86 (3), 161–164.

Zasina, A. J. (2017). Konkurence koncovek *-a* a *-u* v genitivu singuláru neživotných maskulin v polštině. In M. Stluka & M. Škrabal (Eds.), *Liſka a czban – Sborník příspěvků k 70. narozeninám prof. Karla Kučery*, 90–98. Praha, Czech Republic: NLN.

Žigo, P. (2012). Variantnosť v retrospektíve. In K. Buzássyová, B. Chocholová & N. Janočková (Eds.), *Slovo v Slovníku. Aspekty lexikálnej sémantiky – gramtika – štylistika (pragmatika). Na počesť Alexandry Jarošovej*. Bratislava: VEDA, 27–40.

Шарабуряк, А. А. (2012). Rozszerzenie klasy gramatycznej rzeczowników żywotnych we współczesnym języku polskim. *Наукові записки Національного університету* 26, 362–364.

••••

Jan Patrick Zeller
Universität Hamburg, Germany
jan.patrick.zeller@uni-hamburg.de

Syntagmatic corpus analyses of mixed speech: code-shifting in Belarusian trasyanka and Ukrainian suržyk

Due to the influence of the Labovian approach (cf. for example Labov 2006 [1966]), studies on sociolinguistic variation usually have a paradigmatic point of view. Abstracting from where in the conversation a given variable occurs, they typically investigate the influence of sociodemographic characteristics of speakers like age, sex, social network, or social class, on the realization of this variable, and the influence of the speech situation ("style"), the latter being understood as a rather stable factor, undergoing no or only negligible changes during one conversation. From this emerged a "view of variation (involving isolated, loose elements) as being very different from code-mixing (involving stretches of items from different systems)" (Muysken 2000, 126).

The purely paradigmatic approach has been criticized, arguing that it shows only half of the picture of linguistic variation since also "sequences of variants produced by individuals display systematic patterns" (Tamminga et al. 2016, 300; cf. also Sankoff & Laberge 1978, Gries 2016). Both in cases of sociolinguistic "language-internal" variation and in cases of contact between closely related varieties like dialect and standard, it is often the case that speakers not only switch between the respective styles or varieties discretely. They can also make their speech gradually approximate the standard or the dialect (Auer 1986). This "code-shifting" or "style-shifting" can be functional in the same way as code-switching can, and can be in connection with aspects like the interlocutor, the speech situation or the topic of conversation. Using a "semi-syntagmatic" approach, dividing one conversation into different segments according to the topics of conversation, Schilling-Estes (2004) for instance shows how the proportions of sociolinguistic variants differ in the course of one conversation. This calls for the integration of the syntagmatic axis in the investigation of linguistic variation.

Since sociolinguistic studies on Slavic languages following the variationist paradigm are rare in general, it is no wonder that the syntagmatic aspect of linguistic variation has not been addressed in Slavic linguistics as well. In this talk, I will deal with variation in two contact situations between closely related languages: Belarusian-Russian Mixed Speech ("Trasyanka") and Ukrainian-Russian Mixed Speech ("Suržyk"). In Ukraine and Belarus, these mixed forms of speech, i.e. speech containing features of Belarusian / Ukrainian autochthonous dialects and Russian are widespread phenomena. Paradigmatic studies on these phenomena have shown that the proportions of Belarusian / Ukrainian and Russian features at different linguistic levels, as well as the degree of variation and stabilization of these forms of speech, are connected with social characteristics of the speakers. In this paper, I will address the syntagmatic aspect based on two corpora: the Oldenburg Corpus of Belarusian-Russian Mixed Speech and the Oldenburg Corpus of Ukrainian-Russian Mixed Speech. Each of the two corpora contains close to 400 000 words, both are tagged for grammatical values, for correspondence of tokens and utterances with standard Belarusian / Ukrainian and Russian on both phonic and deeper structural levels, and for sociodemographic characteristics of the speakers. I will show that many speakers not only switch between mixed speech on the one hand and Ukrainian/Belarusian and Russian on the other hand. They can also gradually vary the proportion of Belarusian/ Ukrainian and Russian elements in their speech according to aspects like the topic of conversation and the speaker constellation. This includes variation on the phonetic-phonological and on the lexical-morphological level. To neglect this syntagmatic variation would mean to ignore an important aspect of linguistic variation in general and of the speakers' competence in contact situations of closely related languages.

References

Auer, P. (1986). Konversationelle Standard/Dialekt-Kontinua (Code-Shifting). *Deutsche Sprache*, (1986), 97-124.

Gries, St. Th. (2016). Frequencies of (co-)occurrence vs. variationist corpus approaches towards alternations: variability due to random effects and autocorrelation. In P. Baker &;J. Egbert (Eds.), *Triangulating methodological approaches in corpus linguistic research*. New York, 108–123.

Labov, W. (2006 [1966]). *The social stratification of English in New York City*. Cambridge.

Muysken, P. (2000). *Bilingual speech. A typology of code mixing*. Cambridge.
Sankoff, D. & Laberge, S. (1978). Statistical dependence among successive occurrences of a variable in discourse. In D. Sankoff (Ed.), *Linguistic variation: Models and methods.* New York, 119-126.
Schilling-Estes, N. (2004). Constructing ethnicity in interaction. *Journal of Sociolinguistics,* 8, 163-195.
Tamminga, M., MacKenzie, L. & Embick, D. (2016). The dynamics of variation in individuals. *Linguistic Variation*, 16 (2), 300-336.

••••