ParaViz, an online visualization tool for studying variation in meaning based on parallel texts

Ruprecht von Waldenfels
Dept. of Slavic languages and literatures
University of California, Berkeley
ruprecht.waldenfels@gmail.com

Michał Woźniak
Institute of Polish
Polish Academy of Sciences, Cracow
michal.wozniak@ijp-pan.krakow.pl

Abstract

ParaViz is a modular corpus query and analysis tool for use with a word aligned, linguistically annotated multilingual corpus of parallel translated texts. Representing an addition to classic query-based corpus tools, ParaViz makes it easy to assess differences in the meanings of cognate or otherwise comparable items in different languages based on their distribution in parallel texts. Translations are thus essentially used as semantic annotations, allowing for a bottom-up analysis of semantics in a network of texts in many languages.

The tool takes as input a user-supplied operationalization of the variables under comparison. It then provides the user with two perspectives on the distribution of these variables in the parallel corpus: on the one hand, a close-up perspective of word-aligned corpus examples, color-coded in respect to the user-provided parameters; on the other hand, a bird's view perspective with visualizations that provide overviews of the aggregated differences in use. Data sets with the categorized data is made available for download so it can be further analyzed.

Initially developed as an offline version with a specific research topic in mind, the tool has been adapted as an online tool and will be available for use with the ParaSol corpus (Waldenfels 2011). We feel the publication of such tools in a format that makes it accessible for the research community at large is an important part of addressing the issues of research result replication and sustainability of research efforts in digital humanities in general.

1 Introduction

The article reports on work on ParaViz, a complex query and visualization system for word aligned, linguistically annotated parallel corpora used to investigate cross-linguistic similarity of linguistic variables. ParaViz builds on a simple insight: similarities in the distribution of linguistic items in parallel texts, i.e., multiple translations of the same text in different languages, reflect their functional and semantic similarity across languages in a distributional model of semantics (for such models, see the overview in Sahlgren 2008). By comparing such distributions in a word aligned corpus, notions of comparative semantics can be achieved bottom-up; see Cysouw and Wälchli (2007), Dahl (2014) for related approaches, Waldenfels (2015b) for more background and the description of an earlier version of ParaViz.

ParaViz in its offline version is a functional and powerful research instrument developed in a concrete research project that aims to investigate language convergence and divergence based on a parallel corpus (von Waldenfels, 2014). Perhaps typically for such a project, its production version today resembles a patchwork of different technologies and involves many semi-automated steps. It is designed to be used with a locally available parallel corpus – data that is not

```
<parameter id="NounSuffixes">
  <type id="O" name="OST">
    <criteria><lng>ru</lng>
      <regexp level="lem">ость$</regexp><regexp level="tag">^N.*</regexp>
    </criteria>
    <criteria><lng>sl</lng>
      <regexp level="lem">ost$</regexp><regexp level="tag">^S.*</regexp>
    </criteria>
   <criteria><lng>pl</lng>
     <regexp level="lem">ość$</regexp><regexp level="tag">^subst.*</regexp>
    </criteria>
  </type>
  <type id="S" name="STVO">
    <criteria><lng>ru</lng>
      <regexp level="lem">ство$</regexp><regexp level="tag">^N.*</regexp>
    </criteria>
    <criteria><lng>sl</lng>
      <regexp level="lem">stvo$</regexp><regexp level="tag">^S.*</regexp>
    </criteria>
    <criteria><lng>pl</lng>
      <regexp level="lem">[cs]two$</regexp><regexp level="tag">^subst.*</regexp>
    </criteria>
  </type>
</parameter>
```

Figure 1: A sample parameter file, defining classes of cognate suffixes that are defined for each language slightly differently.

unproblematic to share from both a practical and a legal point of view[1]. These facts make it difficult for other researchers to use the methodology and tools that were developed for the original project and make it virtually impossible to replicate its results, both of which would be highly desirable.

The aim in developing an online version of ParaViz is to reuse the existing system to create a web service which takes care of all technicalities and provides users with an easy way to conduct their own research with this corpus. In the design of the system, we aim to significantly lower the threshold for researchers that want to do comparable research, empowering them to use our methods and, replicate, test and expand on our results.

Section two introduces the functions of the tool in some more detail and in the context of ongoing research. Section three presents a description of design choices and the user interface. Section four concludes with an outlook to further developments.

2 Multilingual query and visualization based on parameter files

ParaViz allows the user to define items for comparison using a standardized parameter file in XML format. In this file, the user can define cross-linguistic types the use of which is compared on the basis of word and sentence alignment encoded in the corpus. The types are defined as regular expressions over tokens and their linguistic annotation, which at the moment of writing involves morphosyntactic tagging and lemmatization; in the future, semantic annotation may be added.

As an example, figure 2 shows a set of parameters that defines the cognate suffix classes OST and STVO, both forming abstract nouns, in three Slavic languages. These suffixes are used in all Slavic languages for the derivation of abstract nouns, e.g., Russian *molod-ost'* 'youth', Polish *rad-ość* 'happiness', Serbian *mogućn-ost* 'novelty' all represent instances of the same cognate suffix 'OST'. It is represented in slightly different forms and with slightly different usage profiles across the Slavic languages. In the original project, over ten such cognate suffix classes are defined.

The system then provides the user with two representations of the data that result from

[1]For the relevance of this point in the Swiss legal context, see `www.swisscorpora.ch`.

6174 Словно нам было известно бог знает сколько представителей данного вида , в то время как **представитель** был только один — правда , весом 17 миллиардов тонн .	Немовби нам було відомо хтозна - скільки представників цього виду , тимчасом як **насправді** існував тільки один - щоправда , вагою в сімнадцять більйонів тонн .	Zupełnie jak gdyby śmy znali Bóg wie ile egzemplarzy gatunku , podczas gdy w **rzeczywistości** wciąż był tylko jeden , co prawda wagi siedemnastu bilionów ton .	Jako kdybychom znali bůhvíkolik exemplářů tohoto druhu . Ve **skutečnosti** je znám pořád jen jeden , i když - a to je co říci - o váze sedmnácti bilionů tun .	Pod prstami mi šušťali farebné diagramy , kresby , rozbory , spektrogramy , demonštrujúce typ a tempo **premeny** podstaty a jej chemické reakcie .	Kot da bi poznali bogve koliko primerkov te vrste , medtem ko je v **resnici** še vedno bil samo eden , resda pa je tehtal sedemnajst bilijonov ton .	Kao da poznajemo bogzna koliko primjeraka vrste , dok je u **stvarnosti** još uvijek bio tek jedan , istina težak sedamsto bilijuna tona .	Baš kao da smo poznavali bog te pita koliko primeraka vrste , dok je u **stvarnosti** neprestano postojao samo jedan , istini za volju težak sedamnaest biliona tona .
6490 А может , импульсы , где - то далеко , за тысячи миль от **исследователей** , порождающие его огромные образования ?	Може , імпульси , які десь далеко , за тисячі миль від **місця** **дослідження** , спричинюють його велетенські утворення ?	Może impulsy , powodujące powstawanie jego olbrzymich tworów , gdzie ś gdzieś , o tysiące mil od **badaczy** ?	Anebo snad impulsy , které vyvolávaly **vznik** jeho olbřímích výtvorů někde tisíce mil od místa výzkumů ?	Možno impulzy , ktoré spôsobujú vznik jeho obrovitých foriem kdesi na tisíce míľ od **pozorovateľa** ?	Morda impulzi , ki so sprožali nastajanje njegovih orjaških tvorb nekje tisoče milj stran od **raziskovalcev** ?	Možda impulsi koji su uzrokovali nastajanje njegovih divovskih tvorevina negdje na tisuće milja od **istraživača** ?	Možda impulsi koji su uzrokovali nastanak njegovih džinovskih struktura , hiljadama milja daleko od **istraživača** ?

Figure 2: A word aligned corpus sample with color coding according to user-supplied parameter file.

applying these parameters as queries in the corpus. First, it produces random samples of relevant corpus examples with the respective aligned word forms given in bold and in different colors according to the criteria in the user supplied parameter file. This allows users qualitative insight into the data and lets them gauge the error rate of the operationalization; see figure 2.

Second, the co-occurrence patterns found in the corpus are visualized as NeighborNets: figure 3 shows two such graphs. The left graph represents similarities and differences in the distribution of OST across 14 Slavic versions of the same text, partly in multiple translations. It shows that the use of the suffix in duplicate translations, e.g., Polish and Polish/2, is very similar, and between different languages, it mostly follows the accepted division of languages into East, West and South Slavic. However, there is an important exception: Russian and Bulgarian cluster together, showing that use of this suffix in these two languages is very similar due to extensive language contact in the history of these languages.

The right graph gives an overview of the similarity of cognate suffix classes across all Slavic languages: here, all the suffixes are compared in relation to how often they are used in equivalent word forms across different Slavic versions of the same text. Here, we see that STVO and OST, together with STVIE, NIE and CIJA, form a distinct branch of similar items. This is because their distribution reflects an obvious semantic similarity: all these suffixes are used to derive abstract nouns, with different languages using them for different items and following different semantic models. While this observation may seem trivial in hindsight, it is significant that here it is arrived at based solely on corpus data, rather than secondary data, and extends also to resource-low languages such as Macedonian, in respect to which secondary sources may be difficult to come by.

Comparing the distribution of these morphemes in the corpus thus affords insight into the meanings of these suffixes. In general, many other items can be operationalized in a similar fashion, e.g. reflexive pronouns, pronominal forms, the use of tense, aspect, indefinite pronouns, prepositions, case forms or names; for representative case studies, see von Waldenfels (2014; 2015a). Translation is thus extremely valuable in providing a knowledge-richt type of annotation that links all the texts on a semantic level, making quick and rather comprehensive assessments of a wide range of issues possible. In general, our approach provides data-driven notions of relative semantic similarity, rather than semantic substance. Comparison of distribution provides a way to structure the data in non-arbitrary ways, rather like the semantic maps approach in linguistic typology (Haspelmath, 2003). More modes of analysis, such as the clustering of language-specific

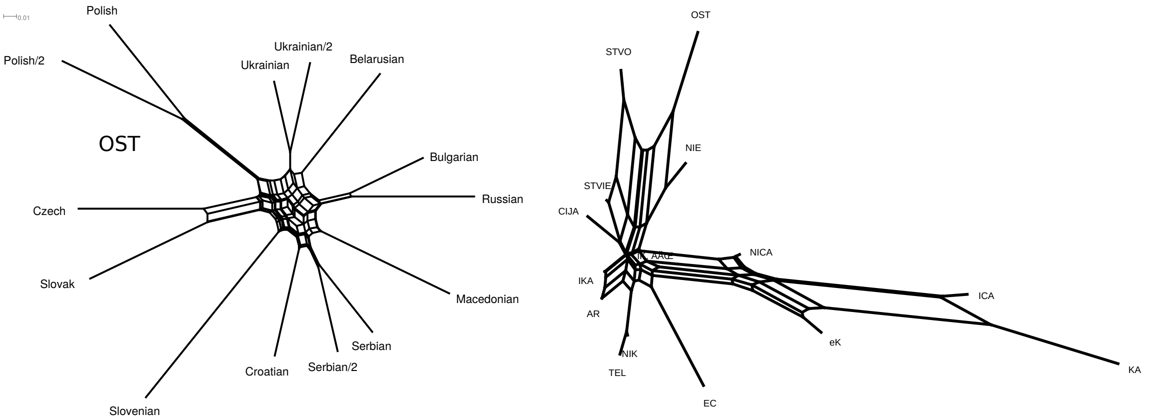


Figure 3: NeighborNets (Bryant and Moulton, 2004) represening similarity of use of nouns derived with the suffix class OST in different Slavic versions of the same text (left); similarity of suffix classes across Slavic (right).

members of these suffix classes, are not implemented yet; for these and more details on the method, see (von Waldenfels, 2015b).

ParaViz uses its own built-in parallel corpus - there is no possibility (at least at this stage) to use custom corpora. Currently the system uses ParaSol, a multilingual parallel corpus primarily geared towards linguistic contrastive and typological research[2] (von Waldenfels, 2011). ParaSol focuses on Slavic, but also includes Romance, Germanic, Finno-Ugric, Greek, Armenian and other languages. Most languages are lemmatized and POS-tagged; a subset of the corpus is word aligned using UPLUG (Tiedemann, 2003). We hope that with time, the system will be used by researchers interested in very different language combinations.

3 The application

While ParaViz in its off-line version is a functional and powerful research instrument, it was developed in a patchwork-like way for a specific research project. It involves many manual steps and uses many technologies (including Perl, XSLT, Java, Python, R and specialized visualization software), running on a Linux operating system. This makes it virtually useless to anyone but its creators. The aim of the online version is thus to take care of all technicalities and provide users with an easy way of conducting their own research with parallel corpora.

ParaViz is developed with Django, a state-of-the-art Python powered web framework which is database oriented. The online version exhibits an extremely simple layout and has many performance issues, but the main functions are implemented and working properly. Generally, we do not plan to implement all functions of the offline system, but rather aim to provide users with all the relevant data files so that they can be used with other tools.

Registered users get assigned their own project space on the server where they may then create their own experiments. Experiment are the basic data objects: they can be created, run, examined, changed or deleted by the user. An experiment represents a linguistic problem that is investigated; in the above example, this would be the use of cognate noun suffixes in different Slavic languages. Each experiment consists of two input objects: a parameter file (see section two above) and a set of options that defines a configuration of texts, languages, and a number of parameters geared to managing lexical and selection effects during the experiment. At this stage parameter files have to be prepared and uploaded by the user; in the future, a graphical tool for the creation of such files may be added. Both parameters and options are saved, modified, and copied independently of experiments, for which they are reused.

The main page of the user interface (figure 4) is minimalistic, listing experiments, parameter

[2]http://www.parasolcorpus.org

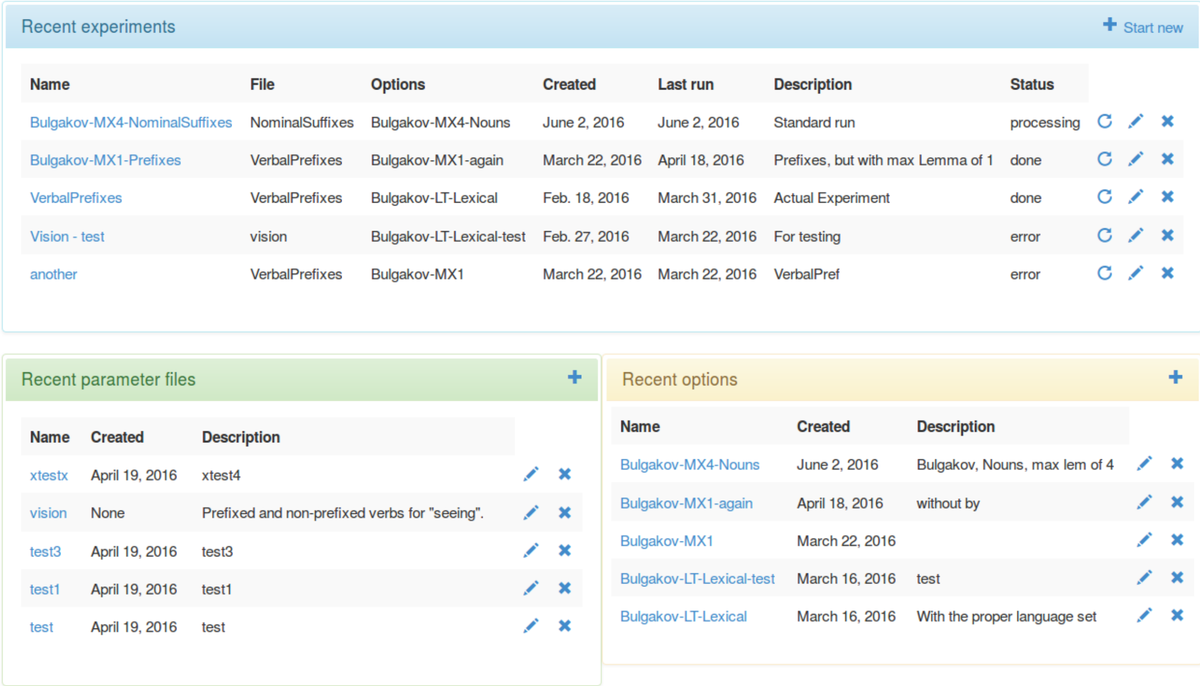


Figure 4: Main user interface

files and option sets. Each list can be expanded to show all of its objects and each object can be viewed in detail or removed. Users can also create new experiments or option objects on the basis of existing ones.

When users hit the “Run” button, the system checks if there are enough resources (both CPU and memory usage) and either starts to execute scripts or adds the experiment to the queue of experiments to be run as soon as resources are available. The experiment status changes from ‘created’ to ‘waiting’, ‘processing’, and finally, either ‘done’ or ‘error’. The processing time is typically a few minutes, depending on the experimental settings.

After the experiment has been processed, the user is provided with an overview of its results as shown in figure 5. The system first gives a graph with the clustering of types (here: nominal suffixes) and their basic frequencies broken down by language and type. Then, for each type, it offers a graph of its use across languages, and a matrix of languages that provides the number of eqivalent word forms where two versions agree in using this type, and a second matrix with the number of equivalent word forms where only one of two versions use it - that is, of the data that is visualized in the graph to the left. Clicking on these numbers will open a new window with a random sample of color-coded corpus examples (see above figure 2) illustrating this case. Finally, files with classifications representing the use of the variables in the corpus, as well as the nexus files used to generate the networks, can be downloaded alongside their pdf versions for publication.

Given the right operationalization, thus, the tool provides users with both qualitative data allowing them to assess operationalization and the language data itself, as well as an aggregate perspective based on the same data, allowing them to proceed quickly in the analysis of the comparative issue they are engaged with.

4 Summary and outlook

In our paper, we have presented a tool that uses a semantically rich multilingual resource, translated texts, for the comparison of the use of multilingual categories. The tool was originally developed as an offline set of scripts and procedures developed for a specific project. Here, we aim to make it available to the research community at large in order to make the methodological

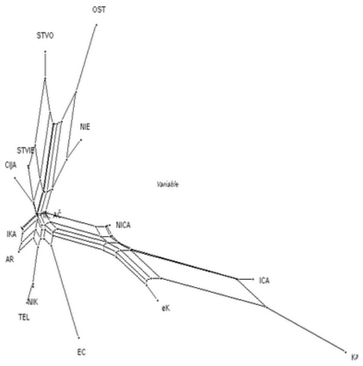
Bulgakov-MX4-NominalSuffixes

Name	Description	Options	Parameter file	Created	Last run	Status
Bulgakov-MX4-NominalSuffixes	Standard run	Bulgakov-MX4-Nouns	NominalSuffixes	June 2, 2016	June 2, 2016	done

Run again Change

Results: Overall

All types (clustered)



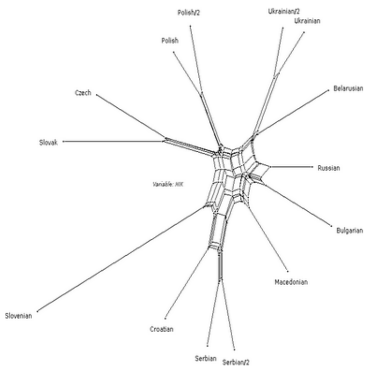
Frequency table

	AČ	TEL	AR	eK	IK	ICA	IKA	NIK	NICA	EC	OST	STVO	STVIE	NIE	CIJA	KA
ru	75	211	84	541	414	161	70	302	161	355	469	193	117	1110	102	1984
by	43	5	1314	600	130	152	68	306	121	307	287	120	0	750	135	2449
uka	128	16	253	877	283	142	21	433	127	491	296	95	0	684	123	2200
uk	118	13	340	763	261	119	20	409	120	447	311	101	0	597	125	2120
pl	63	45	377	1197	119	88	34	183	96	392	685	127	0	1152	271	1692
pla	60	37	372	1138	115	71	32	375	88	424	670	144	0	286	333	1738
cz	57	260	139	835	357	245	20	457	111	492	552	34	58	662	701	2002
sk	80	438	606	587	376	188	23	458	96	618	583	114	0	481	189	2132
sl	90	129	441	1229	97	838	32	865	485	973	588	163	0	845	89	1211
hr	212	70	420	937	128	869	83	666	457	582	456	127	0	611	129	999
sr	132	32	619	1114	227	898	79	667	351	676	567	103	0	687	88	1065
sra	130	54	584	1153	189	893	77	663	370	563	553	100	0	794	106	997
mk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bg	143	287	289	164	118	327	45	402	161	394	427	149	114	648	148	2030

Types (jump to)

- AR
- AČ
- CIJA
- EC
- ICA
- IK
- IKA
- KA
- NICA
- NIE
- NIK
- OST
- STVIE
- STVO

Type:OST



download | matrix

OVERLAP

	bg	cz	hr	mk	pl	pla	ru	sk	sl	sr	sra	uk	uka
bg	-	202	268	189	231	220	320	227	253	262	270	183	177
cz	-	-	213	156	234	228	205	378	248	203	217	156	145
hr	-	-	-	217	243	236	311	259	316	309	332	185	178
mk	-	-	-	-	163	162	218	187	218	215	233	124	105
pl	-	-	-	-	-	435	264	286	269	251	256	218	198
pla	-	-	-	-	-	-	252	269	253	239	239	214	197
ru	-	-	-	-	-	-	-	247	284	288	292	212	215
sk	-	-	-	-	-	-	-	-	289	252	254	186	164
sl	-	-	-	-	-	-	-	-	-	310	325	212	193
sr	-	-	-	-	-	-	-	-	-	-	387	188	181
sra	-	-	-	-	-	-	-	-	-	-	-	199	183
uk	-	-	-	-	-	-	-	-	-	-	-	-	212
	-	-	-	-	-	-	-	-	-	-	-	-	-

CONTRAST

	bg	cz	hr	mk	pl	pla	ru	sk	sl	sr	sra	uk	uka
bg	-	225	159	238	196	207	107	200	174	165	157	244	250
cz	358	-	339	396	318	324	347	174	304	349	335	396	407
hr	186	237	-	239	213	220	145	197	140	147	124	271	278
mk	232	259	202	-	260	261	205	236	205	208	190	299	318
pl	440	443	431	517	-	250	421	399	416	434	429	467	487
pla	441	443	434	517	243	-	418	401	417	431	431	456	473
ru	145	256	160	251	213	215	-	222	185	181	177	257	254
sk	352	198	316	390	301	309	329	-	294	331	329	397	419
sl	321	331	260	352	316	324	297	290	-	278	263	376	395
sr	270	321	225	320	284	289	242	285	238	-	147	346	353
sra	251	301	192	288	272	282	227	274	210	138	-	326	342
uk	127	158	123	185	104	100	100	132	111	126	117	-	99
uka	114	150	118	194	105	99	81	136	115	117	116	82	-

Figure 5: Result page: at the top, experiment description and overview of results with basic statistics. Below, one of a number of rows with per-type results providing NeighborNet graphs and matrices of corpus examples where translations into two languages do or do not agree in using the suffix in question. The matrices give the number of cases; clicking on the number will open a new window with a color-coded random sample of these cases in the corpus.

results of our research available to other scholars and open our results to replication, aims that have become, in our view, both more relevant and more readily attainable with the development of digital humanities.

The basic functions of this tool are grounded in a distributional model of semantics that utilizes translation as semantic annotation providing a data-driven method to derive comparative models of meaning of a large range of possible linguistic variables. While the original research was done only on Slavic languages, the tool is language independent and the corpus data it is used on involves many Romance, Germanic, Finno-Ugric and other languages.

Rather than building an offline version that would cater to a computationally literate community only, we have opted to prepare an online version built around the existing scripts. Focusing on functionality and transparency, we have devised a simple interface that enables the researcher to perform basic comparisons of a wide range of user-definable variables based on their use in the parallel corpus and download the relevant categorizations for further use. In the future, we plan to add a number of further analytic functions and, if time allows, provide a graphical tool for the construction of the parameters that form the basis of the experiments.

Acknowledgment

Funding by the Swiss National Science Foundation, grant 151230 *Convergence and divergence of Slavic from a usage based, parallel corpus driven perspective*, is gratefully acknowledged.

References

David Bryant and Vincent Moulton. 2004. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265.

Michael Cysouw and Bernhard Wälchli, editors. 2007. *Parallel Texts: Using translational equivalents in linguistic typology. Special Issue of STUF 60/2.*

Östen Dahl. 2014. The perfect map: Investigating the cross-linguistic distribution of tame categories in a parallel corpus. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages*, pages 268–289. De Gruyter Mouton, Berlin, New York.

Martin Haspelmath. 2003. The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In M. Tomasello, editor, *The new psychology of language: Cognitive and functional approaches to language structure. Vol. 2*, pages 211–42. Laurence Erlbaum Associates, Mahwah, NJ.

Magnus Sahlgren. 2008. The distributional hypothesis. *Rivista di Linguistica*, 20(1):33–53.

Jörg Tiedemann. 2003. *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing.* Ph.D. thesis, Uppsala University, Uppsala, Sweden. Anna Sågvall Hein, Åke Viberg (eds): Studia Linguistica Upsaliensia.

Ruprecht von Waldenfels. 2011. Recent developments in parasol: Breadth for depth and xslt based web concordancing with cwb. In Daniela Majchráková and Radovan Garabík, editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20–21 October 2011*, pages 156–162, Bratislava. Tribun EU.

Ruprecht von Waldenfels. 2014. Explorations into variation across slavic: taking a bottom-up approach. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, within and across Languages*, pages 290–323. De Gruyter Mouton, Berlin, New York.

Ruprecht von Waldenfels. 2015a. Inner-slavic contact from a corpus driven perspective. In Emmerich Kelih, Stefan Michael Newerkla, and Jürgen Fuchsbauer, editors, *Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven*, pages 237–263. Peter Lang, Frankfurt.

Ruprecht von Waldenfels. 2015b. The paraviz tool: Exploring cross-linguistic differences in functional domains based on a parallel corpus. In Gintaré Grigonyté, Simon Clematide, Andrius Utka, and Martin Volk, editors, *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015, May 11–13, 2015, Vilnius, Lithuania.*