The ParaViz Tool: Exploring Cross-linguistic Differences in Functional Domains Based on a Parallel corpus

Ruprecht von Waldenfels
Institute of Polish,
Polish Academy of Sciences, Cracow
ruprecht.waldenfels@gmail.com

Abstract

ParaViz is a modular corpus query and analysis tool in development for use with a word-aligned, linguistically annotated multilingual parallel corpus. Representing an addition to classic query-based corpus tools, it allows to assess the cross-linguistic variation in the functional domain of items or structures that are defined as cognate or otherwise equivalent by the user. ParaViz provides the user with two perspectives on such data: on the one hand, a close-up perspective with word-aligned corpus examples that are classified and color-coded according to the user's criteria; on the other hand, a bird's view perspective with word lists and NeighborNet visualizations that offer an overview of the aggregated differences in use. Together they enable researchers to quickly find and explore convergent and divergent functional patterns of equivalent formants in different languages.

1 Introduction

ParaViz is a modular corpus query and analysis tool in development for use with a word-aligned, linguistically annotated multilingual corpus. It is deployed with ParaSol, a small multilingual parallel corpus primarily geared towards linguistic contrastive and typological research of Slavic (Waldenfels 2011), but may in general be used with any massively parallel word-aligned corpus such as Opus (Tiedemann, 2012) or InterCorp (Čermák and Rosen, 2012). ParaViz functions as a stand-alone component at the moment of writing and is planned to be implemented as a web application.

ParaViz adds a new type of functionality to parallel corpus querying going beyond what is described in Volk et al. (2014). It supplements and builds on ParaVoz (Meyer, Waldenfels, Zeman 2014), a corpus query interface which allows querying parallel corpora through a traditional web interface on the basis of complex queries involving token sequences across aligned languages, including negative queries on aligned segments.

Section 2 of this paper gives an introduction to functional comparison using parallel texts as implemented in ParaViz. In section 3, I describe the implementation, and conclude in section 4.

2 Functional Comparison Based on Parallel Texts

The function of a linguistic item as understood here includes its semantic, pragmatic, or other characteristics. As a rule, functional characteristics of linguistic elements are harder to describe than their formal characteristics and involve complex analysis of corpus examples. This makes the comparison of such functions particularily difficult, since it presupposes a consistent, comparable analysis of these functions across many languages, a task that quickly becomes very complex with a growing number of languages.

To see this, consider a comparison of the German, Dutch, and English perfect. In all three languages, it is quite straightforward to describe cognate perfect constructions that consist of an auxiliary ('have' or 'be') and a past participle. However, in order to compare the functional profiles of such constructions, we first need to describe the functions of each construction, taking care to do this in a consistent, comparable manner that is indeed relevant for the comparison and does not miss important contrasts. This is a very time-consuming and difficult task for even a medium number of languages.

The fact that aligned parallel corpora involve translationally equivalent texts in many languages can be harnessed to quickly attain insights on functional differences and similarities based on

formal definitions alone (see the papers in Cysouw and Wälchli (2007) and Dahl (2014) for related approaches). The basic notion is straightforward: if two items in two languages are often used as translations of each other, we assume that their functional potential is similar. This notion can be used to do quite complex comparisons of functional domains. In application to the above example, the fact that these perfect constructions differ in their function quickly emerges from simply observing that their distribution is very different in the parallel corpus.

For a different, rather lexical example, let us assume users want to compare the functional domain of color terms in different languages. In order to do that, users define which words represent the lexical categories red, blue and yellow across many languages, and the system then compares the use of these words (and thus, the lexical categories) across languages in translationally equivalent expressions. For example, such a comparison would show that German *blau*, English *blue*, French *bleu* are often used in the same segments, but that the distribution of Russian *sinij* stands apart, since this item denotes a dark hue of blue. Moreover, the German representative of the lexical concept 'blue' is used in contexts that it isn't used in English and French since it also refers to a state of drunkenness (*er ist blau* 'he is drunk'), while in English, *blue* also denotes a melancholic state of mind (as in *I'm feeling blue*). If these uses are attested in the parallel corpus, the differences in the denotation of hues as well as in non-literal uses are readily apparent in the distribution of these terms in translationally equivalent segments across the languages in question.

In this way, the functions of variables of different types ranging from grammatical categories such as tense, aspect, or case to lexical categories such as words for the color red or derivational suffixes can be easily compared. Further applications and a more detailed description of the approach are found in von Waldenfels (2014).

3 The ParaViz system

ParaViz is meant to simplify the type of comparison outlined in section 2 by offering a standardized way to easily conduct such functional comparisons for a wide range of variables. This is done by offering the user a mechanism to define variables in a formal way and outputing the results of a classification of the corpus data based on these definitions. This section describes the standalone application as it is functional at the moment of writing; in the future, this system is planned to be implemented as a web service.

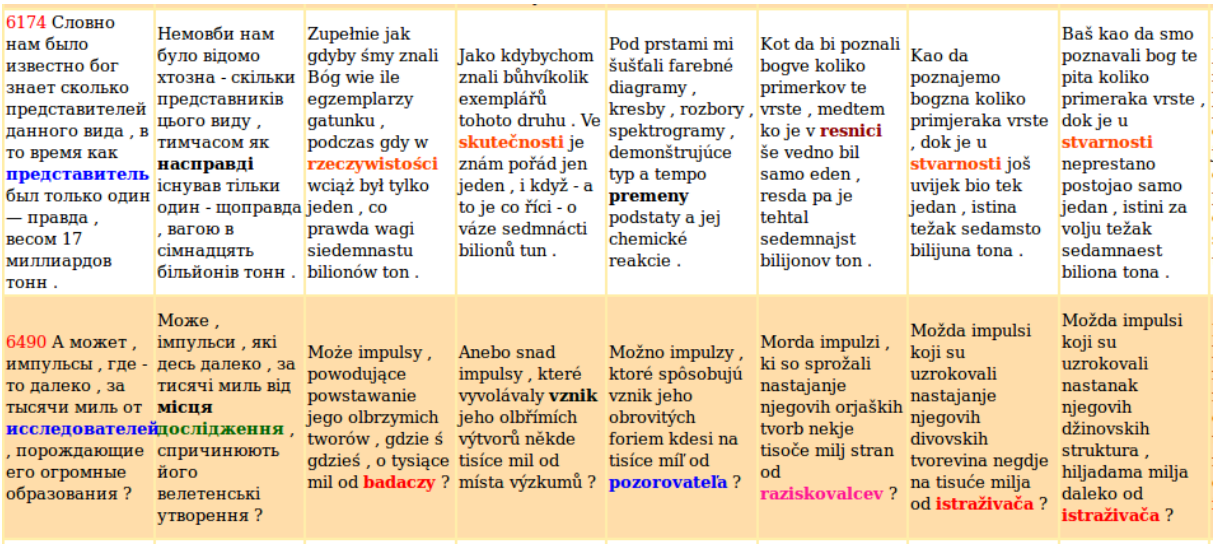
ParaViz is used with ParaSol, a small multilingual parallel corpus primarily geared towards linguistic contrastive and typological research[1]. ParaSol focusses on Slavic, but also includes Romance, Germanic, Finno-Ugric, Greek, Armenian and other languages. The word forms in most languages are lemmatized and POS-tagged; a subset of the corpus is word-aligned using UPLUG (Tiedemann, 2003).

3.1 Operationalization of Variables

As a first step in the process, users define the sets of elements which they want to compare across languages. This is done by the operationalization of variables in parameter files in XML format. In such a paramter file, variables are defined as constraints over word-aligned word forms and their annotation. At the moment, the parameter files allow the definition of such variables as regular expressions over tokens, their lemmas and POS tags, as well as over tokens, lemmas and POS tags directly adjacent. The following example defines the suffix classes OST and STVO, both denoting abstract nouns, in two Slavic languages:

```
<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>
</type>
<type id="S" name="STVO">
 <criteria><lng>ru</lng>
  <regexp level="lem">ство$</regexp>
  <regexp level="tag">^N.*</regexp>
 </criteria>
...
 <criteria><lng>pl</lng>
  <regexp level="lem">[cs]two$</regexp>
  <regexp level="tag">^subst.*</regexp>
```

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



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

```
</criteria>
</type>
```

The user then defines a filter condition for one of the languages that is taken as the primary language. For example, for the comparison of nominal suffixes above, the user may choose to classify only nouns. In other cases, the user may want to restrict the classification to some list of lemmata; for example, the user may be interested in a particular lexical or grammatical domain. The primary language will usually be the language of the original, but in general, any language may be chosen.

The tokens in the primary language that satisfy the filter condition as well as well their word-aligned equivalents in the other languages are classified in the next stop. This classification assigns a type to each token in question based on the criteria defined in the parameter file.

The system then does a corpus search and classification of the corpus data, which is offered to the user in two ways; first, as random samples of the corpus hits in context; second, in an aggregate form.

3.2 Qualitative Perspective: Color-coded Corpus Samples

As a first result, the user is given random samples of corpus data conforming to the definitions. This enables the user to review and refine the operationalization of his or her parameters. The word-aligned forms in these corpus results are color-coded to reflect the types previously defined in the parameter file; an example output is given in figure 1. This perspective affords a qualitative assessment and allows the user to explore the data.

This part of the output is based on ParaVoz, a modular corpus query interface for CorpusWorkbench[2] (CWB; see Evert & Hardie (2011)) published as open source (Meyer et al. 2014). It is designed as an easy to use, easy to install and easy to maintain flexible corpus interface for a parallel corpus hosted by CWB. ParaVoz (and CWB) uses CQP, a query language also used with a number of other corpus engines such as the NoSketchEngine (Rychlý, 2007) or Poliqarp[3]. ParaVoz does not use the CWB output directly, but, having configured CWB to SGML mode, reformats its output into a convenience XML format using regular expressions.

The output module of ParaVoz then uses XSLT to transform the XML result document. Using XML and XSLT at this stage allows rapid adaption to diverse types of corpus data. In this case, word alignment visualization is realized by linking ids that are encoded as token annotations. These annotions are compared, and if a token in one of the target languages is aligned to the target tokens in the source language, it is shown in bold. Using the same script, each target form is classified according to the user-defined definitions and color-coded accordingly.

[2]http://cwb.sourceforge.net
[3]http://poliqarp.sourceforge.net

```
'Bulgarian'  -b---z-aa--aanan----------o-------c
'Belarusian' -z---o-----a-----zz----b-------b----
'Czech'      --------yyyyy--n------d-------z-----
'Croatian'   -b----ii-rrr---------z----------o--
'Macedonian' -------------ynnz-------------------
'Polish'     y---nn-------nb----------yobb-------
'Russian'    -----o-bbaaa--------b---------------
'Slovak'     -o-b----y-aay----------------zz-a-y-
'Slovenian'  -----o-rrrrr--b-----------d---z-r-z-
'Serbian'    b-----i---rr-------z----z-------zd--
'Ukrainian'  --n--z--aaaa---n----z---z-zz---p----
```

Figure 2: Strings representing the word-aligned tokens as classified into types according to the user-supplied parameter file.

3.3 Aggregate Perspective

In a second perspective, the system outputs visualizations of the aggregate differences in distribution of the variables across different languages. As descibed above, tokens in the primary language and their equivalents are classified with respect to the user defined operationalization in terms of word, lemma, tag, and other possible levels, just as it is done for the random samples. The examples are thus converted to strings as shown in figure 2. If the strings are seen as a table, each column represents a set of word-aligned word forms, with each word form represented as a single letter reflecting the type it was classified as.

The functional similarity or dissimilarity between the distribution of the variables in multilingual versions of the same text is then computed by determining for each pair of texts the overlap in the occurrences of this variable in aligned tokens, i.e.:

$$dist(V_{lng1}, V_{lng2}) = 1 - \frac{V_{lng1} \cap V_{lng2}}{V_{lng1} \cup V_{lng2}}$$

In other words, the system computes strings of word-aligned corpus positions that are labeled according to the classification in the parameter file and computes the hamming distance between these strings. This computation is used to arrive at distance matrices describing the similarity or dissimilarity of the distribution of the variable between texts. These matrices are visualized in NeighborNets (Huson and Bryant, 2006), a clustering algorithm that was chosen since it preserves much of the ambiguity we find in this data.

Using different filters and definitions of the features that are being compared, the system can then be used to output different visualizations of the differences in distribution of the variables

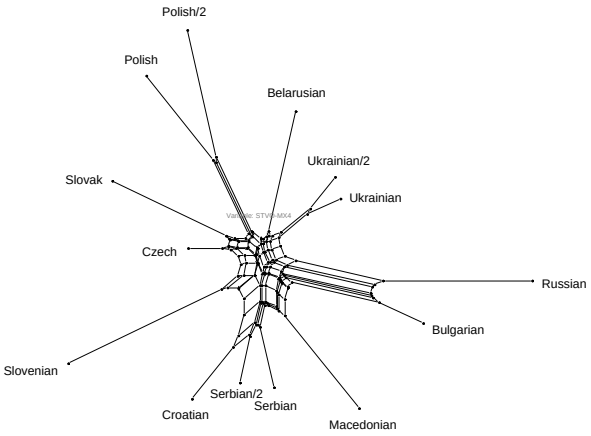


Figure 3: Similarity of use of nouns derived with the suffix class OST in multiple versions of the same text in different Slavic languages

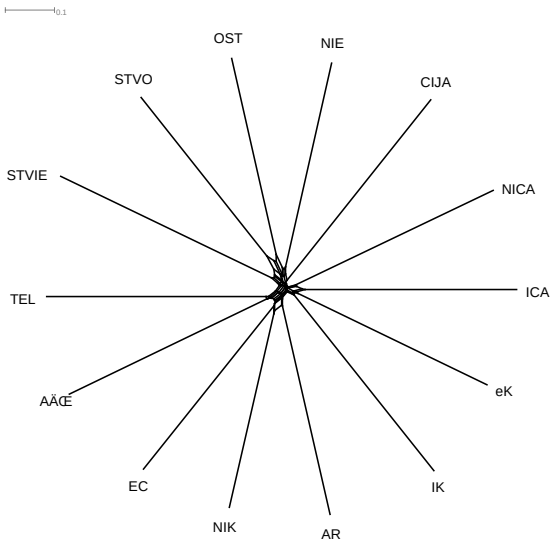


Figure 4: Suffix classes across Slavic: functional similarity.

in question. This concerns (a) the distances of texts/languages to each other with respect to some variable (see fig. 3); (b) the distances of the variables to each other taken as cross-linguistic types; this is calculated for each pair of variables by determining the proportion of examples where both are used in equivalent word forms (see fig. 4); (c) the distance of the language specific instantiations of the variables to each other; this allows to see whether formants of different classes overlap in their domain (see fig. 5). In addition, it outputs documents with word lists of equivalent tokens in different languages and their classification (not shown here).

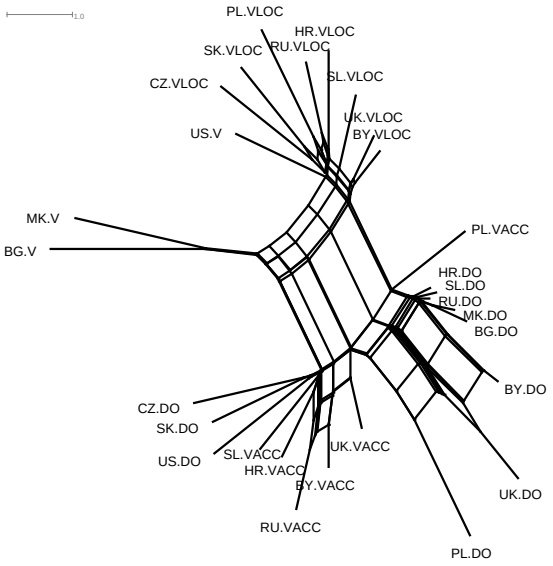


Figure 5: Functional similarity of Slavic prepositions (language specific representations)

4 Concluding Remarks

I have presented a component that supplements a query-based interface to a word-aligned multilingual parallel corpus with a new comparative corpus evaluation component which is available offline and is planned to be implemented as a web service. This corpus evaluation component will provide users with the possibility to upload their own parameter files which provide complex definitions of comparable items in different languages based on their formal characteristics. The corpus is then evaluated in respect to the functional similarity of the items in question. Crucially, the component aims to give both an aggregate and a detailed view of the data, so that the user keeps the possibility to interpret the aggregate picture, and refine the parametrization as necessary for his or her needs.

Acknowledgments

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

References

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

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.

Stefan Evert and Andrew Hardie. 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.

Daniel H. Huson and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23:254–267.

Roland Meyer, Ruprecht von Waldenfels, and Andreas Zeman. 2006-2014. Paravoz - a simple web interface for querying parallel corpora. https://bitbucket.org/rvwfels/paravoz.

Pavel Rychlý. 2007. Manatee/bonito - a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masaryk University.

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.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Martin Volk, Johannes Graën, and Elena Callegaro. 2014. Innovations in parallel corpus search tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.