Recent Developments in ParaSol: Breadth for Depth and XSLT based web concordancing with CWB

Ruprecht von Waldenfels

Institut für slavische Sprachen und Literaturen, Universität Bern

Abstract. The article describes the Slavic parallel corpus ParaSol, developed in Bern and Regensburg. The paper gives an account of recent developments, focussing on conceptual decisions concerning corpus make up and the user interface.

1 Introduction

ParaSol is a multilingual Slavic parallel corpus comprising original and translated prose texts developed in collaboration of the University of Bern in Switzerland and the University of Regensburg in Germany; the acronym stands for *Parallel Corpus of Slavic and Other Languages*. Having initially been developed in Regensburg under the name *Regensburg Parallel Corpus*, it is now headed in Bern University and hosted on servers in both places. Web site development and text aquisition is shared between the two institutes[1].

The following principles, some of them new or modified, guide the development of ParaSol:

- *original and translated prose in many Slavic and some non-Slavic languages (breadth for depth)*
- *variation of (preferably Slavic) source languages*
- *automatic preprocessing and alignment*
- *linguistic annotation such as lemmatization and POS tagging*
- *public availability through a web concordancer*
- *crowdsourcing: users are encouraged to collaborate*

The present article focusses on text strategies and on new developments concerning the web interface.

[1] The following people are part of the project: Ruprecht von Waldenfels (head of the project and the Bern team; overall corpus architecture, corpus maintainance, interface design and text aquisition); Roland Meyer (head of the Regensburg team, CWB integration and interface design); Veronika Wald, Dmitrij Nikolenko (text aquisition, Regensburg); Vivian Kellenberger, Michael Reinhard, Karin Zurbuchen (text aquisition, Bern); Andreas Zeman (text aquisition and interface design, Bern).

1.1 Similar projects

ParaSol is most comparable to two other current projects:

- InterCorp [6] , developed in the Czech republic under the auspices of the Czech National Corpus, a corpus built in a cooperation of numerous departments across the country and focusing on language pairs composed of Czech and a second language, one of currently 21 European languages (see http://www.korpus.cz/intercorp). Alignment is done manually between these pairs (with some supplementary automatic alignment with Czech as a pivot done for texts that are present in more than one pair). Where possible, linguistic annotation (lemmatization and POS-tagging) is included. Access is provided, after registration, via a web interface.
- The Amsterdam Slavic Parallel Aligned Corpus (ASPAC), developed in Amsterdam by Adrie Barentsen. This corpus focuses on all Slavic languages and also includes some other European languages. Alignment is done manually; all language versions are aligned in a tabular fashion so as to reflect equivalence to the original (see below). As a consequence, all difficulties resulting from omissions or additions in translation, or varying source documents, are resolved explicitely. No lemmatization or POS-tagging is performed on the text. While it is not searchable through a web interface, access to the corpus is available for research on personal request (see home.medewerker.uva.nl/a.a.barentsen/page3.html).

2 Corpus composition

2.1 Breadth for Depth

Like in ASPAC, and in contrast to InterCorp, both of which ParaSol cooperates closely with, the focus of the ParaSol corpus has developed to be on breadth, rather than depth, of coverage. In other words, the corpus composition strategy at this points stresses the inclusion of more language versions of a given text, rather than more texts for a given language pair. The augmentation of specific pairs of languages (as, e.g., in the past the Polish-Russian, Slovak-Bulgarian and German-Slovak pairs) was at the center of attention in the earliest phase of ParaSol (then RPC), since its rechristening as ParaSol in 2009, preference has been given to the inclusion of texts in many Slavic languages.

This development reflects a typical position of Slavic studies outside the Slavic speech communities: rather than being focussed on any of the particular national languages, our interest potentially involves all Slavic languages and, to a considerable extent, a comparative perspective on Slavic; see [8] for an approach where translation variants in diverse languages is crucial.

Moreover, our students typically study more than one Slavic language in a variety of combinations. Since ParaSol is used in pedagogical applications, especially in early stages where students do not yet have full command of these languages, having a wide range of language versions per text is an important

asset. For these reasons, ParaSol has been focussing on texts that are present in many Slavic languages rather than continueing a focus on depth, that is, the addition of texts of any specific language pair.

It is therefore no coincidence that ParaSol is in this way most similar to the Amsterdam Slavic Parallel Aligned Corpus, also developed outside of the Slavic countries. Both corpora differ in this from the Czech project InterCorp, which focusses on pairs of languages with Czech.

In distinction to the projects mentioned above, ParaSol strives to balance source languages as far as practical in order to be able to deal with translation effects. As of summer 2011, 8 novels from 7 source languages[2] are available in translation into almost all Slavic literary languages: J.K. Rowling's *Harry Potter and the sorcerer's stone* (English), Milan Kundera's *Nesnesitelná lehkost bytí* (Czech), Mikhail Bulgakov's *Master i Margarita* (Russian), Nikolaj Ostrovskij's *Kak zakaljalas' stal* (Russian), Ivo Andrić's *Na Drini ćuprija* (Serbian); Umberto Eco's *Il nome della rosa* (Italian), Patrick Sueskind's *Das Parfüm* (German), Stanisław Lem's *Solaris* (Polish). The reader is referred to the project web sites (see below) for a current list of texts included in the corpus.

In addition to a complete coverage of each text in all (major) Slavic languages, we also strive to include German, French and Italian, the national languages of Switzerland and Germany. These are the most frequent non-Slavic native languages of our students. Moreover, we try to include Modern Greek for research interest, as this is a language especially interesting for comparison being both a member of the Balkan Sprachbund, like Macedonian and Bulgarian, and an aspect language. The Baltic languages, most closely related to Slavic, are also represented. Aside from that, we take an opportunistic stance to including other languages.

The corpus project initially grew out of the recognition that in contrastive work, researchers often compile their own small parallel corpora. ParaSol is conceived as a corpus architecture that can accomodate such projects. We continue to encourage users to contribute and make use of its facilities, adhering to a wiki spirit of crowdsourcing in corpus compilation.

3 Design decisions and web interface

3.1 Annotation

As far as possible, texts in the corpus are lemmatized and POS tagged; where such tools are not publically available, this is done in cooperation with institutions that develop these tools in the context of the national corpora (see web site for a list of cooperations and [7] for more details).

[2] Thanks are due to Emmerih Kelih who has contributed Nikolaj Ostrovski's *Kak zakaljalos' stal'* in eleven Slavic languages [2] and a large number of translations in the *Bulgakov* subcorpus.

3.2 Alignment

A conceptual decision was taken to rely on pairwise alignments, rather than on a table-like alignment architecture that would involve transitive alignment properties. To assess the differences, consider the example in figure 1, where a corresponding text segment is divided into two sentences in only two of three languages.

DE		RU		PL
DE.1 Lass mich.	↔	RU.1 Pusti.		PL.1 Puść, nie chcie,
DE.2 Ich will nicht, dass Du mich berührst.	↔	RU.2 Ne xoču, čtoby ty ko mne prikasalsja.	↔	zebyś mnie dotykał!

Fig. 1. An alignment example with differences in segmentation across languages.

Let us suppose a user is interested in Russian *pusti* and Polish *puść*, cognate items both translated as *let!*. With pairwise alignment, each language version is aligned to each other language version independently. This means that if one chooses to base one's search on the Russian text, RU.1 *Pusti.* will be aligned to German DE.1 *Lass mich.* and Polish PL.1 *Puść, nie chcie, zebyś mnie dotykał!*. If, however, the search is based on Polish, the segment PL.1 *Puść, nie chcie, zebyś mnie dotykał!* is aligned to the German PL.1-2 *Lass mich. Ich will nicht, dass Du mich berührst.* and Russian RU.1-2 *Pusti. Ne khoču, čtoby ty ko mne prikasalsja.* On a pairwise basis, alignment is thus maximally precise, but differs depending on which language the search is based on.

In table-like alignment, in contrast, rows such as the one in the example above are, like in a table, considered a single segment aligned across *all* versions. The more fine grained equivalence relations between Russian and German are disregarded. Therefore, any query will output the same segments regardless of which language variant the search is based on.

The decision for pairwise alignment makes the approach more robust: if any one of the language pair based alignment relations breaks down for some reason, e.g., because the text is abridged or censored, this does not result in degradation of the alignment quality of any other pair. Since ParaSol exclusively relies on automatic alignment, robustness is very important.

Alignment was initially done with bsa [3]; we have now moved to hunalign [5] (see Rosen [4] for a comparison of aligners). Where possible, alignment is done on files containing word forms replaced with lemmas in order to reduce the search space during alignment[7].

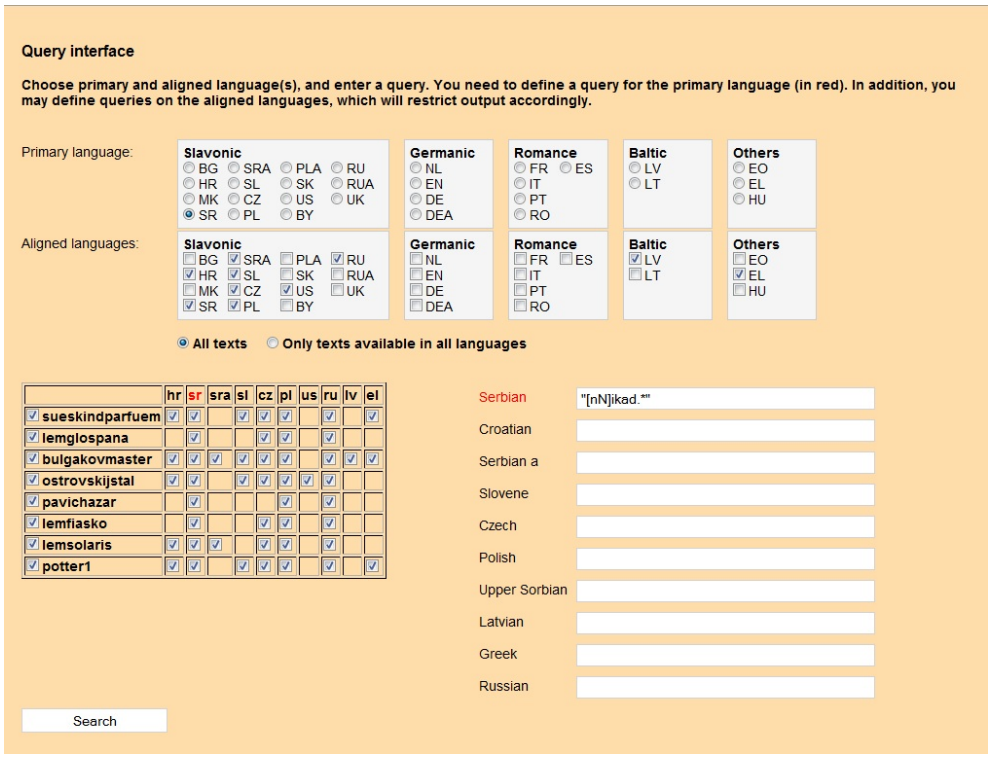

Fig. 2. Query for *[Nn]ikad.** in Serbian, with a variety of aligned languages, not all present in all texts.

3.3 Query interface

The design of the interface[3]reflects the conceptual decision for pairwise alignment. The user first chooses some primary language, and then selects a set of aligned languages. As the user selects and deselects languages, the list of corpus files on the lower left side of the interface and input fields for the query strings on the lower right side appear and adapt to reflect the user's choices. This is implemented in javascript and partly backed by entries in an SQL-data base. There is an option to either restrict the set of texts to those texts that are available in all languages, or to perform the query on all texts which are present in the primary language (see the screenshot in figure 2).

Input fields for the query strings accept standard CQP syntax and directly channels queries to CWB[1], which now fully supports unicode encoded corpora. Annotation varies from language to language, but typically, three levels are supported: word form, lemma, and morphosyntactic tag. Each query opens a new result window (a feature inspired by the RNC).

3.4 XML/XSLT based concordance

The interface, originally developed essentially as a wrapper for the HTML output module of CWB, now utilizes client-based XSLT for the display of the XML encoded result returned by CWB. As of the moment of writing, however, CWB does not yet support XML output (although this is a planned feature, Stefan Evert, p.c.). Instead, the SGML output module is used, which, however, is faulty in respect to entity resolution. Regular expressions in the php code are used to derive valid XML from this faulty SGML representation. The resulting XML text is transferred to the client together with an XSLT style sheet that transforms it to HTML.

While the transformation from SGML to XML slows down output considerably, the transition to an XML based output system is justified by a number of advantages. First, this decision basically amounts to dividing content generation (the XML file) from output display (the HTML file resulting from the XSLT transformation), thus adding to the modularity of the system. The question of output generation is for the time being solved in a provisional way with transformations from SGML; this will have to be reviewed as soon as an XML module is ready. Since content generation and display are separate issues now, this temporary solution does not stand in the way of further development of the display module. Also, since XSLT is a language without side effects, directly geared to manipulating structured data, using a XSLT style sheet is much simpler, and at the same time more flexible and more robust than php code.

As an example, consider queries where not all texts are available in all languages the user is interested in, as in the query of the screen shot in figure 2. In order to format the resulting table, a server based php solution would have to

[3] The web interface has been developed by Roland Meyer, Regensburg; Andreas Zeman, Bern; Ruprecht von Waldenfels, Bern

keep track of which corpus is available in which language and check for consistency with the actual result table returned by CWB - for a variety of reasons, this can fail, and strategies to deal with this have to be employed. In contrast, a client side XSLT solution works locally on the resulting XML file alone. As long as this is a valid XML file, all necessary display decisions such as widths of the columns or the column labels can be taken on the data alone; since this involves much less assumptions and variables, this much more robust and at the same time easier to implement.

As a whole, moving to XML and XSLT technology has in our case resulted in much more rapid and flexible evolution of the concordance window (as shown in the screenshot in figure 2). Lemmas and morphological tags are now shown as tool tips, and basic statistics are computed on the basis of the result file on the client side.

The style sheet is much simpler and, owing to the fact that XSLT has no side effects, more robust than a server side construction of a HTML file. Moreover, since this is a modular solution, we can very easily offer more output formats now by simply adding an option to use different style sheets, which may ultimately even may be user developed or user supplied.

80159 Ni u kom slučaju i nikada !	Ни в каком случае и никогда !	Nekad un nekādā gadī jumā !	V nobenem primeru in nikoli !	Ni u kom slučaju i nikada !	Nigdy , w żadnym wypadku ! |	- Už nikdy , v žádném případě nevezmu ve vašem bufetu nic do úst !	Ni u kom slučaju i nikada !	Σὲ καμιὰ , περίπτωση καί ποτέ !
80203 Sir ni u kom slučaju i nikada ne može da bude zelene boje , to vas je neko prevario .	Брынза не бывает зеленого цвета , это вас кто - то обманул .	Brinzai taču nav jābūt zaļā krāsā , jūs esat maldināts .	Ovčji sir ni zelene barve , nekdo vas je moral potegniti .	Ne postoji ovčiji sir zelene boje , vas je to neko prevario .	Zielona bryndza nie istnieje , ktoś Musiał pana oszukać . |	Můj milý , brynza nesmí být nazelenalá , to vás někdo ošidil .	Ovčji sir nije zelene boje , to vas je netko obmanuo .	Δὲν υπάρχει φέτα πράσινου χρώματος , κάποιος σάς ξεγέλασε .

40 hits in corpus *lemsolaris*.

	ru	pl	cz	hr	sra
1899 Nikad pre ga nisam video .	Раньше я никогда не видел Снаута .	Nigdy go jeszcze nie widział | em . |	Osobně jsem se s ním ještě nikdy nesetkal .	Nikada ga još nisam vidio .	Nikad ga još nisam bio video .
9867 Ovaj aneks prvog solarističkog godišnjaka bio mi je poznat , to jest znao sam za njegovo postojanje , ali ga nikad nisam imao u rukama , pošto je predstavljao čisto istorijsku vrednost .	О приложении к первому тому " Соляристического ежегодника " я знал , то есть слышал , что оно существует , но никогда не держал его в руках , поскольку оно представляло собой только историческую ценность .	Ów aneks do pierwszego solary - stycznego rocznika znał | em , to znaczy , wiedział | em o jego istnieniu , ale nie miał | em | miałem go nigdy w ręce , przedstawiał bowiem czysto historyczną wartość . |	Onen dodatek k prvnímu ročníku Solaristické ročenky jsem znal , či spíše věděl jsem o jeho existenci , ale neměl jsem jej nikdy v rukou , měl totiž už jen historickou cenu .	Taj sam dodatak prvome solarističkom godišnjaku poznavao , to jest , znao sam za njegovo postojanje , ali ga nikada nisam imao u ruci , jer je predstavljao čisto povijesnu vrijednost .	Ovaj aneks uz prvi solaristički godišnjak znao sam , to jest bilo mi je poznato da postoji , ali nikad ga nisam imao u ruci , jer je predstavljao čisto istorijsku vrednost .
9891 A za nekog pak Ravincera , niti za njegov Mali apokrif nikad nisam ni čuo .	Однако я понятия не имел ни о Равинцере , ни о " Малом Апокрифе " .	Natomiast o jakimś Ra - vintzerze ani o jego „ Małym Apokryfie ” nigdy nawet nie słyszał | em . |	Zato však o žádném Ravintzerovi ani o jeho " Malém apokryfu " jsem v životě neslyšel .	Međutim , za nekakvoga Ravintzera i njegov Mali apokrif , nisam nikada čak ni čuo .	Međutim o nekom Ravinceru i o njegovom ' Malom apokrifu ' nikad nisam bio čak ni čuo .
12592 Nikad nisam bio na Stanici , ali sam šest nedelja stanovao u njenoj vernoj kopiji koja se nalazila u Institutu , na Zemlji .	Я никогда не был на Станции , но во время подготовки прожил шесть недель в ее точной копии , находящейся в Институте	Nie był | em nigdy na Stacji , ale przez sześć tygodni mieszkał | em — w ramach wstępnego treningu — w jej dokładnej kopii ,	Nebyl jsem nikdy na stanici , ale při výcviku jsem bydlil šest týdnů v její přesné kopii , kterou mají v Institutu na Zemi .	Nikada nisam bio na Postaji , ali sam šest mjeseci - u okviru pripremnih vježbi - boravio u njezinoj točnoj kopiji , koja se nalazila u	Nisam nikad bio na Stanici , ali šest sedmica sam proveo - u okviru uvodne pripreme - u njenoj tačnoj kopiji koja se nalazila u

Fig. 3. Query result for *[Nn]ikad.** in Serbian, with differing number of aligned languages.

4 Summary

The present article has given a short overview of the ParaSol, a Parallel Corpus of Slavic Languages, focussing on two recent developments: a change in the corpus composition strategy with an aim to include more language versions of a given text, rather than more texts for a given language pair (breadth for depth) as well as a move to XML/XSLT technology for the web concordancer.

Bibliography

[1] Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*.

[2] Kelih, E. (2009). Slawisches parallel-textkorpus: Projektvorstellung von „kak zakaljalas' stal' (kzs)". In Kelih, E., Levickij, V., and Altmann, G., editors, *Metody analizu teksta/Methods of Text Analysis*, pages 106–124, Černivci. ČNU.

[3] Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*, pages 135–144.

[4] Rosen, A. (2005). In search of the best method for sentence alignment in parallel texts. In Garabík, R., editor, *Computer Treatment of Slavic and East European Languages: Third International Serminar, Bratislava 10-12 November 2005*, pages 174–185, Bratislava.

[5] Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, pages 590–596.

[6] Vavřín, M. and Rosen, A. (2008). Intercorp: A multilingual parallel corpus project. In *Proceedings of the International Conference Corpus Linguistics - 2008*, pages 97–104. St. Petersburg State University.

[7] von Waldenfels, R. (2006). Compiling a parallel corpus of slavic languages. text strategies, tools and the question of lemmatization in alignment. In Brehmer, B., Ždanova, V., and Zimny, R., editors, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pages 123–138. München.

[8] von Waldenfels, R. (t.a.). Aspect in the imperative across slavic - a corpus driven pilot study. *Oslo Studies in Language. Special Issue ed. by A. Grønn & D. Haug.*