Brehmer, B., Ždanova, V., Zimny, R. (Hrsg.) 2006. Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 123-138. (Die Welt der Slaven. Sammelbände/Сборники. 27.)

COMPILING A PARALLEL CORPUS OF SLAVIC LANGUAGES
Text strategies, Tools and the Question of Lemmatization in Alignment[1]

1. Introduction

This article describes the Regensburg Parallel Corpus of Slavic Languages (RPC), which is currently being developed at the Institute of Slavistics at Regensburg University. The corpus is intended as a research and teaching tool for linguists working on Slavic languages. In the first part of this paper, I will outline the purpose and architecture of the corpus and give a summary of its current state.

A crucial part of assembling a parallel corpus is assuring that text segments of one language are aligned to the corresponding segments in other languages. In the second part of the paper, I will investigate how the lemmatization of parallel texts influences their alignment quality and discuss this issue on the basis of experiments with two aligners, *hunalign* and *bsa*.

2.1. The Regensburg Parallel Corpus of Slavic Languages

Examining translated texts has always been important in comparative linguistics. Yet, in contrast to language-specific research where the availability of large monolingual corpora has had a huge impact on research methods, parallel corpora have remained much less important in comparative linguistics; presumably, the main reason for this is that there are so few of them available. But some do exist, and their number has been growing in recent years.

Multilingual Slavic corpora projects that need to be mentioned here are, to the best of my knowledge, first of all Adrian Barentsen's *Amsterdam Slavic Parallel Aligned Corpus* (ASPAC)[2], a large manually aligned corpus of belletristic texts in Slavic and non-Slavic languages. This corpus is probably most similar to RPC, with the difference that it relies on transitive alignments, is not linguistically annotated and has no public interface. Other multilingual corpora include: the *Acquis Communautaire* (Erjavec et al. 2005), a corpus of translations of European Community legislation including many Slavic languages; *Opus*, a corpus of public domain texts in many languages publically available on the net[3] (Tiedemann/ Nygaard 2004); and a number of further projects relating to fewer languages or including only a limited number of texts. Last but not least, in the Czech Republic the ambitious parallel corpus project "Intercorp"[4] is underway, which will include Czech and a wide range of other languages, some manually aligned data and a balanced selection of text types (Rosen 2005).

[1] The corpus is the result of work by Roland Meyer and myself. Thanks are due to various other people who have contributed to it, directly or indirectly: by getting material, scanning and correcting; by valuable comments, discussions and support. These include but are not restricted to Adrian Barentsen, Monika Banášová, Dagmar Divjak, Radovan Garabik, Christine Grillborzer, Björn Hansen, Tomáš Káňa, Natalia Kotsyba, Michail Michailov, Adam Przepiórkowski, Alexander Rosen and Christian Wolff. I am indebted to Dagmar Divjak, Roland Meyer and Alexander Rosen for reading and commenting on a draft of this paper.

[2] There is no published account on this corpus; for an example of work with it, see Barentsen (2006)

[3] See http://logos.uio.no/opus/.

[4] See https://trnka.ff.cuni.cz/ucnk/intercorp/.

There are several difficulties specific to building multilingual parallel corpora. The first difficulty concerns obtaining the material: one needs texts that have been translated into many languages; fine literature is an obvious possibility here. Another equally important issue is alignment, that is, the process of defining and encoding which segments of texts in different languages are translations of each other. Finally, one would like to have annotated data, that is, tagged corpora, which facilitates conducting complex queries and statistical analyses. Working on several languages multiplies the difficulties involved in annotation, since for every language, individual solutions for automatic annotation have to be supplied.

The RPC is a parallel corpus that tries to address these problems from a pragmatic perspective. It differs from existing corpora in some important respects. Intended to serve as a research and instruction tool for linguists and students of Slavistics doing comparative work, it is envisioned as a corpus of all Slavic languages with primarily belletristic texts that have been aligned to each other and are linguistically annotated.

Given the high workload typical of compiling a parallel corpus, the leading idea behind the RPC is to utilize existing tools in a flexible modular system, to minimize human intervention and to rely on cooperation between institutions and researchers to enlarge it and facilitate sophisticated linguistic annotation. Researchers or students often compile their own corpus, be it electronic or not, for specific language pairs and with specific research questions in mind. Combining these efforts in a multilingual Slavic corpus effectively eases the task and is beneficial to all researchers engaged in similar work. An important aspect of the Regensburg Parallel Corpus is that it not only provides a corpus, but also an infrastructure for sharing the work involved in corpus building. The RPC is designed to be easy to augment both in terms of texts as well as in terms of languages and to be flexible according to the needs of researchers cooperating in enlarging the corpus according to their own demands.

One can think of a multilingual parallel corpus as a set of subcorpora that contain two or more languages each. Many researchers are interested in a specific subcorpus for a small set of languages; when enlarging such a corpus, adding texts that are translated into many Slavic languages makes the work beneficial also to researchers working on other language combinations. At the moment, for example, Stanisław Lem's novel *Solaris* is available in RPC in Polish, Russian, German, Belorussian and Serbian. Enlarging the Polish-Czech subcorpus of RPC can therefore be achieved by simply adding the Czech version of *Solaris*; at the same time, this will augment the Polish-Czech, Russian-Czech, German-Czech, Belorussian-Czech and Serbian-Czech subcorpora.

The focus of the RPC lies on post-war fiction, although other periods and genres can be considered for inclusion; at present, it additionally contains some legal and journalistic texts. A second compositional strategy emphasizes the inclusion of originals (and their translations) in, if possible, all Slavic languages in order to at least partly balance the problems connected with interference and 'translationese'; texts with multiple translations are preferred for the same reason. This ties up with the above mentioned strategy to add texts that have been translated into many Slavic languages[5].

The current make-up of the corpus is given in Tables 1 and 2. At the moment, the largest subcorpora are the Polish-Russian and German-Slovak portions (see Banášová (forthcoming) for an example of work with RPC).

[5] The best candidates for books translated into all Slavic languages are, aside from biblical and marxist material, international bestsellers; unfortunately, they all seem to be translated from one language: English.

BoellClown		DE					RU			SK		
BoellFrau		DE								SK		
BulgakovMaster	BY	DE				PL	RU		SB			
EUConst		DE								SK		
EndeMomo		DE					RU					
GralsWelt		DE								SK		
KafkaErz		DE								SK		
LemAstronauci						PL	RU					
LemFiasko						PL	RU					
LemGlosPana						PL	RU					
LemKatar						PL	RU					
LemKongres		DE				PL	RU					
LemPamWannie						PL	RU					
LemPokoj						PL	RU					
LemPowGwiazd						PL	RU					
LemSolaris	BY	DE				PL	RU				SX	
LemWizjaLokalna						PL	RU					
NabokPnin		DE	DEa							SK		
PavicHazar						PL	RU				SX	
Potter1		DE		EN	HR	PL	RU	RUa	SB	SK		UK
Potter2		DE		EN		PL	RU	RUa				UK
Potter3						PL	RU	RUa				
Potter4						PL	RU					
Potter5						PL	RU					
SloOestHK		DE								SK		
StrugLebedi		DE				PL	RU					
StrugPiknik		DE				PL	RU			SK		

Table 1: Texts currently included in RPC

Abbreviations (alphabetic order, lower-case letters after the language abbreviations signify multiple translations): BoellFrau: *Heinrich Böll, Frauen vor Flusslandschaft*; BoellClown: *Heinrich Böll, Bekenntnisse eines Clowns*; BulgakovMaster: *Михаил Булгаков, Мастер и Маргарита* ; EUConst: *Constitution of the European Union*; EndeMomo: *Michael Ende, Momo*; GralsWelt: *Grals Welt (a journal published in several languages)*; KafkaErz: *Franz Kafka, Erzählungen*; LemAstronauci: *Stanisław Lem, Astronauci*; LemFiasko: *Stanisław Lem, Fiasko*; LemGlosPana: *Stanisław Lem, Głos Pana*; LemKatar: *Stanisław Lem, Katar*; LemKongres: *Stanisław Lem, Kongres futurologiczny*; LemPamWannie: *Stanisław Lem, Pamiętnik znale ziony w wannie*; LemPokoj: *Stanisław Lem, Pokój na Ziemi*; LemPowGwiazd: *Stanisław Lem, Powrót z gwiazd*; LemSolaris: *Stanisław Lem, Solaris*; LemWizjaLokalna: *Stanisław Lem, Wizja lokalna*; NabokPnin: *Vladimir Nabokov, Pnin*; PavicHazar: *Milorad Pavić, Hazarski rečnik*. Potter1-5: *The five first volumes of J.K. Rowling's Harry Potter series*; SloOestHK: *Bekanntmachungen der Slovakisch-Österreichischen Handelskammer*; StrugLebedi: *Аркадий и Борис Стругацкие, Гадкие лебеди* ; StrugPiknik: *Аркадий и Борис Стругацкие, Пикник на обочине*.

language abbr.	in full	tokens
BY	Belorussian	208 177
DE	German	1 154 356
EN	English	208 986
HR	Croatian	90 581
PL	Polish	1 956 713
RU	Russian	2 458 435
SB	Serbian (Cyrillic script)	244 277
SK	Slovak	620 370
SX	Serbian (Latin script)	154 199
UK	Ukrainian	179 630

Table 2: Language abbreviations and subcorpus size

2.2. System architecture

2.2.1. Input Module

Inclusion of new texts and languages is kept as simple as possible; this enables researchers to contribute to the corpus according to their needs. After digitizing the text, which usually involves scanning, optical character recognition and spell checking, a header with bibliographic information is prepended and chapter divisions are annotated. The resulting text is saved as simple unicode text. All preprocessing can be done with a simple text editor or word processor.

This text file is then processed by the input module of the corpus system performing tokenization, sentence splitting and conversion to XML. This process is kept as simple and language-independent as possible: there is only one option used in tokenization for languages that include the apostrophe as a letter, such as Ukrainian; simple heuristics are employed by the sentence splitter to identify problematic issues (e.g. abbreviations).

Lemma tags are, if possible, added at this stage: the module produces a common list of word forms in all new texts per language; this list can be input into a language specific lemmatizer[6]. After processing, the resulting word-lemma list is used to supply lemma tags for each token during the generation of the XML files. The lemmatization at this stage is minimally complex: no disambiguation of homonyms is attempted, and a stemmer can be used where no lemmatizer is available. By restricting the processing to word form-lemma lists and disregarding all further information a lemmatizer might output, integrating lemmatizers is kept semiautomatic: the user only has to feed the automatically compiled word form lists into the different lemmatizers, make sure the second word on each line of the output files are lemmata and copy them to a place the input module will find them.

There are pragmatic reasons why annotation is carried out this way: first, lemmatization provides a basic, yet very useful annotation; it enables to query for lexemes rather than for word forms. Secondly, open source lemmatizers are reasonably easy to find on the web for many languages; if no lemmatizer is available, it can be substituted by an algorithmic stemmer. Finally, lemmatization effectively improves automatic alignment by constraining the search space (see the second part of this article).

[6] The following lemmatizers are used at this stage: Morphy (Lezius 2000); Stempelator (Weiss 2005), RMorph (Gelbukh/Sidorov 2003); the online ASUsilc BCS tagger (Sipka 2006).

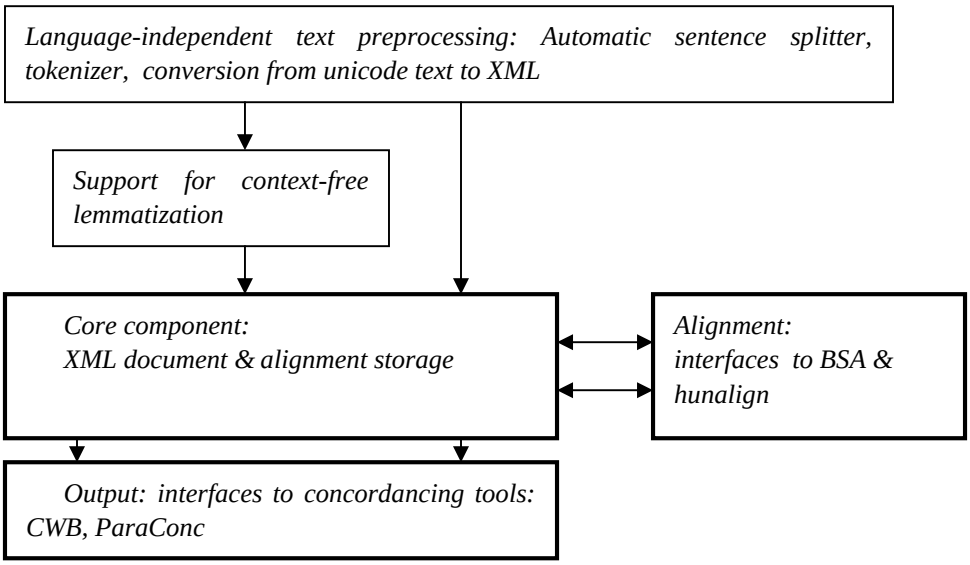

Figure 1: RPC structure

The input model thus ensures *essential* annotation that can be included with little effort. The need of fully-fledged interfaces to more sophisticated tools is avoided at this stage; richer annotation (e.g. including morphological tags from the lemmatizers used) would be more complicated to automate for this range of languages and is therefore carried out without central support directly on the XML files in the text repository.

2.2.2. Text repository and alignment module

XML (in a modified version of TEI-lite) was chosen as the central storage format because it is transparent, easy to exchange and can accommodate different levels of annotation. The files are kept in a flat directory structure; file names are used to identify texts and language versions.

The alignment information is kept as stand-off XML, that is, as individual files for every text/language pair with pointers specifying which segments of a text are aligned to which portion of the other. Multiple translations into one language are also aligned to each other and treated analogously. Rather than aligning all texts into segments that are equivalent in all languages, we compute pairwise alignments which are then combined in searching. Alignment in RPC is therefore not transitive: partition of the data is always in respect to one source language that is queried, with the output in other languages determined by the alignment relations to the source language rather than to other target languages.

When a new file is included, alignments to all existing language versions of this text are computed automatically. This is done by scripts that output the necessary format and start the aligner. At this moment two aligners are supported, *bsa* (Moore 2002) and *hunalign* (Varga et al. 2005). Keeping the corpus texts and the alignments apart facilitates later adaptation to new, better aligners. No manual correction is supported.

Once the texts are stored in the repository, additional annotation can be included by manipulating the XML-files directly. The XML specification provides five tags, named <tag1>…<tag5>. These are pseudotags: dependent on the tagger they refer to different types of information. We are indebted to Radovan Garabik of the Slovenský národný korpus in Bratislava for the tagging of the Slovak texts (Garabik 2005) and to Adam

Przepiórkowski from IPI PAN in Warsaw for the tagging of Polish texts (Piasecki/ Godlewski (forthcoming)). Russian tagging was performed locally at our institute by Roland Meyer (Betsch/Meyer 2003). Accordingly, the pseudo tags have different content in these languages and corpus users must acquaint themselves with these tag sets in order to utilize the annotation.

2.2.3. Output modules

Two output modules are implemented. The first one generates files that can be processed by ParaConc (Barlow 1999), a commercially available parallel concordancing program that is stored on local computers (see Figure 1). This program does not fully support rich annotation; one cannot query for combinations of morphosyntactic tags or lemmas, but it is very convenient for visualizing larger text portions and simple searches. For copyright reasons, the corpus can be viewed through ParaConc at our institute only as whole texts have to be loaded locally. Since the RPC relies on pairwise alignments, only the equivalences in two languages can be viewed at a time using ParaConc, which supports only transitive multiple alignments.

Worldwide access is provided by a second output module that produces the input files for IMS Corpus Workbench (Christ 1994), a powerful corpus management program running on a web server and providing online access to the corpus through a web interface[7]. After choosing source and target languages and a set of texts, the user can run complex queries on the source language texts. The resulting hits for the source language as well as all their aligned target language correspondences are returned. Hits can be constrained by formulating an additional condition on the target languages (see pictures 2 and 3). Since the web interface outputs only citations, that is, non-coherent text chunks of a few sentences, it can be queried by registered users outside the university, too.

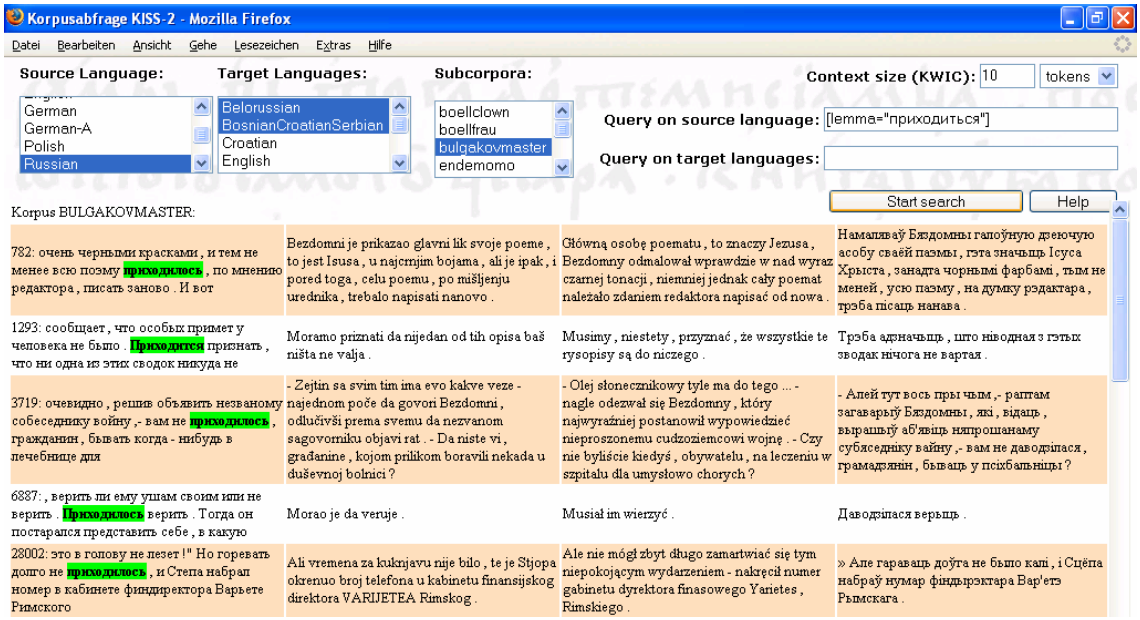

Figure 2: A multilingual query in the web interface to CWB

[7] The integration of RPC into CWB and the web interface were programmed by Roland Meyer.

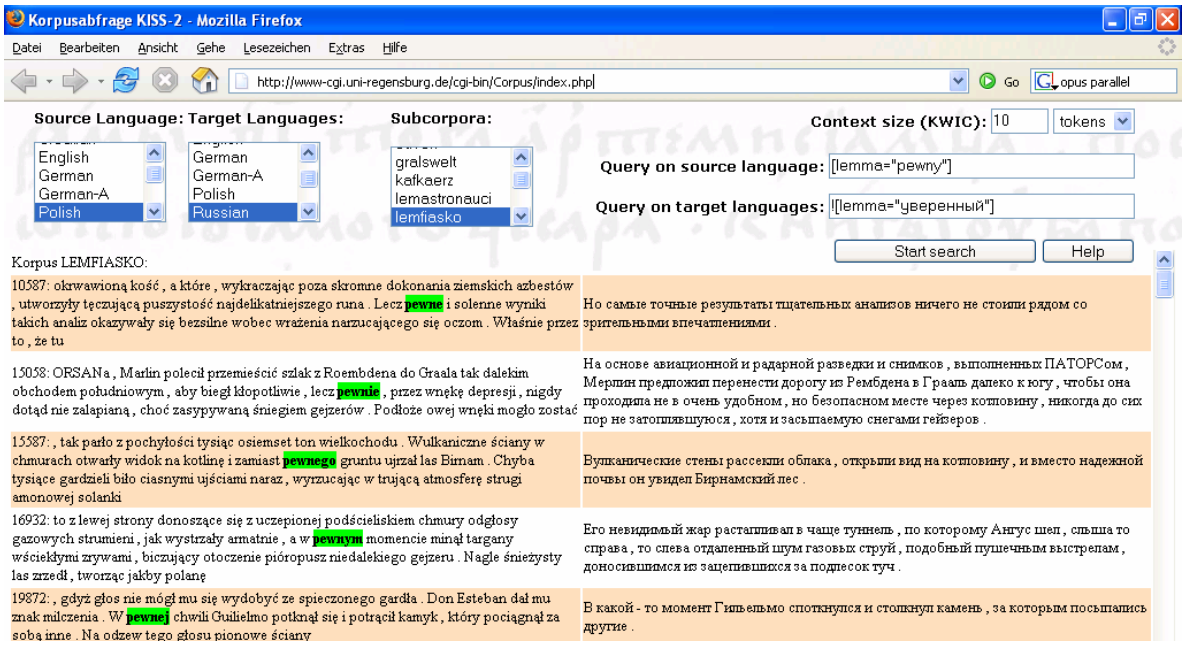

Figure 3: Querying CWB for Polish *pewny* not translated as Russian *уверенный*

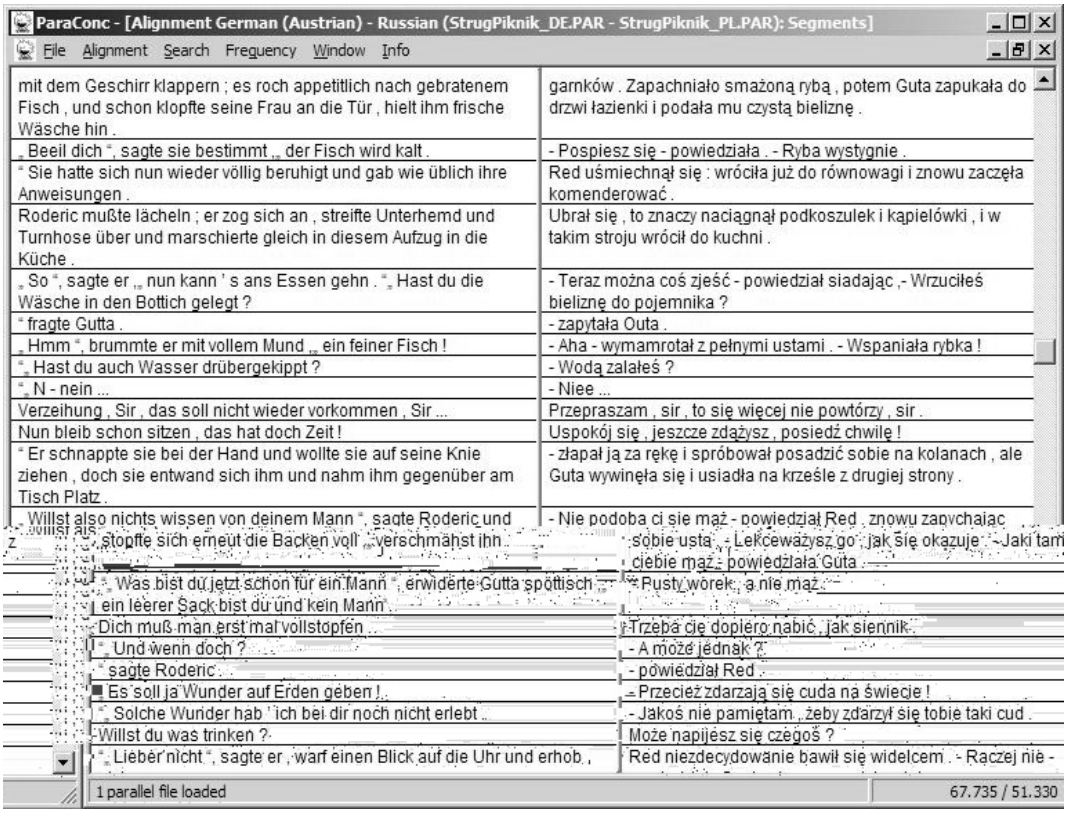

Figure 4: RPC in ParaConc

3. Using lemmatization to improve alignment

3.1. Introduction

Aligning texts is probably the most crucial part in building a parallel corpus. A parallel corpus can be effectively used only if queries return the right correspondences. Alignment can be defined in the following way (cf. Véronis/Langlais 2000) :

Presume we have two texts, text A and text B with segments (sentences) $a_1 \ldots a_n$ and $b_1 \ldots b_n$. A bead is then defined as a pair of segment sets of each language that correspond to each other: according to the cardinality of these sets, one can speak of 1-1 beads, as in ($\{a_1\},\{b_1\}$), 1-2 beads, as in ($\{a_2\}$, $\{b_2, b_3\}$), 0-1 beads ($\{\},\{b_4\}$) and so on. As an example, take the following two fragments which were manually aligned as a 1-2 bead:

Polish
(A1) Puść, nie chcę, żebyś mnie dotykał!

Russian
(B1) Пусти.
(B2) Не хочу, чтобы ты ко мне прикасался.

Since the Polish sentence relates to two Russian sentences, the correct bead is ({A1}, {B1,B2}). An alignment of a given text pair, then, is the full set of beads ($Bead_1 \ldots Bead_n$).

For RPC, we looked out for an aligner that was free to use, preferably open source, independent of language specific resources such as seed lexica or stop word lists (which are hard to obtain for smaller languages), and without need for manual preprocessing such as paragraph markup. The initial aligner used in RPC was a modified version of *bsa* (Moore 2002); when *hunalign* (Varga et al. 2005) appeared, support for this second aligner was added.

Bob Moore's *bilingual sentence aligner* (bsa) utilizes sentence length – in a modification of the classical algorithm by Gale and Church (1991) – to compute a preliminary alignment. The best 1-1 beads of this alignment are then used to build a statistical translation model. Finally, the full corpus is reprocessed assessing alignment candidates in relation to both translation model and sentence length correspondences.

Hunalign is similar in spirit. It also goes through three stages: first it computes correspondences on the basis of sentence length, then it builds a probabilistic dictionary and combines this lexical resource and the sentence-length approach in the final alignment.

While *hunalign* gives arbitrary beads, bsa outputs only 1-1 beads above a certain confidence threshold. It does compute 2-1, 1-2, 0-1 and 1-0 beads in an intermediate stage, however, and since in contrast to tasks such as statistical machine translation training, in our case it is important to align the full set of sentences in the corpus, these intermediate results are used rather than the final output. Since the 0-1 and 1-0 beads generally seem to be unreliable, RPC considers only the 1-1, 1-2 and 2-1 beads in these results and uses a simple heuristic to fill the gaps resulting where no beads were output: if there are text portions unassigned between two beads in both languages, they are assumed to be aligned to each other; if there is such text in one language only, it is assumed to belong to the next aligned segment. This adapted version of *bsa* therefore does not handle omissions; all material is assumed to have correspondences in the other language.

Text	tokens (pl)	tokens (ru)	sentences (pl)	sentences (ru)	sample beads
BulgakovMaster	148 342	145 213	10 660	10 279	145
LemAstronauci	104 111	108 202	7 478	7 867	87
LemFiasko	110 141	115 611	7 677	7 827	117
LemKongres	39 462	39 690	3 000	3 182	42
LemPowGwiazd	85 528	89 334	8 502	8 643	113
LemSolaris	69 918	62 697	5 886	5 794	71
LemWizjaLokalna	103 103	109 655	4 939	5 478	59
Potter1	86 577	90 823	7 848	7 300	103
Potter2	95 754	101 250	8 347	7 954	91
StrugLebedi	77 261	76 607	7 124	7 226	94
StrugPiknik	62 058	61 799	5 280	5 414	62

Table 3: The Polish-Russian subcorpus

3.2. Research questions

Since both *bsa* and *hunalign* use statistics to find word correspondences, it seems plausible that replacing word forms with their lexemes (or even stems) should improve the quality of alignment[8]. Such a replacement can be seen as treating the grammatical information encoded in flectional morphology as noise; it reduces the number of word types and therefore leads to more robust statistics and less combinations that need to be evaluated. *Bsa*, for instance, cuts off a certain portion of the word types encountered and uses only the 5000 most frequently used word types. Reducing the number of types in the text therefore increases the amount of text actually used in computing the translation model. From this perspective, it is clearly expected that aligning lemmata rather than word forms should improve alignment.

On the other hand, aligning lemmata might also have negative effects: in translation there is much symmetry of combined lexical and grammatical information so that, especially in languages as closely related as the Slavic languages, word forms will in fact often correspond to word forms; replacing word forms by lemmata neglects this information.

Aside from this principal issue, there is also a practical problem: what can be done if there is no lemmatizer available for a particular language? Belorussian, for which we do not have any computational linguistic tools available, is a case in point: does one get acceptable results by aligning Belorussian texts to the lemmatized versions of texts in other languages or does alignment quality actually deteriorate in such a case? Intuitively, there is no reason to assume why lemmatization of only one of the two languages should improve results, unless one is much more analytic than the other, which could, for instance, play a role in aligning English-Russian data.

Experiments were therefore conducted to gain insights into the following questions:

1. Does lemmatization improve alignment quality?
2. What are the effects of lemmatizing only one of two languages?
3. Which aligner fares better on the corpus, *bsa* or *hunalign*?

[8] Suffice it to say that this is hardly an original thought; lemmatization has been used to preprocess data in statistical machine translation, as well as in parallel text alignment. However, to my knowledge it has not been investigated in detail before.

Text	tokens (de)	tokens (ru)	sentences (de)	sentences (ru)	sample beads
BoellClown	88 025	89 550	4 620	5 133	65
EndeMomo	79 099	66 402	5 849	6 230	94
LemKongres	45 142	39 690	3 587	3 182	53
LemSolaris	81 762	62 697	5 922	5 794	83
StrugLebedi	91 333	76 607	7 969	7 226	126
StrugPiknik	78 697	61 799	5 202	5 414	67
BoellClown	88 025	89 550	4 620	5 133	65

Table 4: The German-Russian subcorpus

3.3. Experimental setup

Rather than using a full prealigned corpus, two representative[9] random samples from a Russian-Polish subcorpus (1000 sentences) and from a German-Russian subcorpus (500 sentences) were taken. Taking random samples tries to do justice to the heterogeneity of the texts that occur in a belletristic corpus with as little manual processing as the RPC. Texts translated from both languages as well as from a third language (English and Russian, respectively) were included in these subcorpora; for details see Table 3 and 4. Note that the number of sentences chosen per text is not even, as a random sample of each subcorpus as a whole was taken.

The sentences in the sample were manually aligned to their equivalents in the other language; since they were part of larger beads, the number of sentences effectively increased in the process. These beads were then compared to the automatic aligners' output with and without prior lemmatization.

3.4. Alignment metrics

The usual way to measure alignment quality in respect to a manual gold standard is in terms of *recall* (what proportion of correct correspondences were included by the aligner), *precision* (what proportion of the automatically determined correspondences were correct) and *f-measure*, the harmonic mean of these two, $F = 2 \cdot (\text{recall} \cdot \text{precision})/(\text{recall} + \text{precision})$. Sentence level metrics were used in order to take partially correct alignments into consideration (see Véronis/Langlais 2000, for discussion of different evaluation methods). To give an example, consider the following possible alignments of examples B1 and B2 quoted above and repeated here:

Polish
(A1) Puść, nie chcę, żebyś mnie dotykał!

Russian
(B1) Пусти.
(B2) Не хочу, чтобы ты ко мне прикасался.

Alignment 1: ({},{B1}) ({A1},{B2})
Alignment 2: ({},{B1, B2}) ({A1},{B3})
Alignment 3: ({A0, A1, A2},{B0,B1,B2,B3})

Alignment 1 takes B1 to have been added in the target text, and A1 to correspond to B2. This is, of course, not as bad as alignment 2, which presumes both B1 and B2 to have

[9] Representativity was measured on sentence length in absence of a better measure.

been added and aligns A1 with segment B3 (here not shown but clearly a wrong choice). Alignment 3, finally, takes a very large bead that includes, among others, A1 as well as B1 and B2. Alignment 3 is less than perfect, but not 'false' from a user's point of view: the right sentences are aligned to each other, and that they are part of a segment of up to four sentences does not necessarily pose a problem. None of alignments 1-3 are identical to the manual alignment ({A1}, {B1,B2}) and therefore all would yield zero alignment and recall. In fact, however, they differ in quality: some of them do provide a good approximation.

In order to account for such partially correct alignments in sentence level metrics, the beads are not evaluated directly: instead, both manual and automatically assigned beads are transformed by computing the cartesian product of their corresponding segments to arrive at a larger set of beads that explicitly reflects partial correspondences. Recall and precision is then measured in respect to these transformed beads. The manual alignment now becomes ({A1},{B1}), ({A1},{B2}) Alignment 1 does not change under this transformation and gets assigned 0.5 recall and precision, which adequately mirrors its partial usefulness.

Random sampling is problematic, however, in regard to large beads such as Alignment 3. It is transformed to the set ({A0}, {B0}), ({A0}, {B1}), ({A0}, {B2}), ({A0}, {B3}), ({A1}, {B0}), ({A1}, {B1}), ({A1}, {B2}), ({A1}, {B3}), ({A2}, {B0}), ({A2}, {B1}), ({A2}, {B2}), ({A2}, {B3}). Many of these correspondences might actually be correct, but since the neighboring segments were not part of the sample, they were not checked manually. Consequently, they cannot be recognized as correct and are therefore considered as false. The recall value for this alignment 3 is 2/2 = 1, but precision drops to 2/12 = 0.167. Large beads, which might still be quite useful for the linguist, if they are correct in the sense that they connect a set of translational equivalents of three or four sentences, are thus heavily penalized by this setup. The metric was therefore modified to take into account only correspondences with at least one segment covered in the manual alignment. In this case, since only {A1},{B1}and {B2} are mentioned in the manual alignment, the transformed alignment 3 is restricted to ({A0}, {B1}), ({A0}, {B2}), ({A1}, {B0}), ({A1}, {B1}), ({A1}, {B2}), ({A1}, {B3}); precision rises to a more plausible 0.33. In this way, the distortion inherent to random sampling is reduced; very large beads will, however, still result in lower precision than if all relevant alignments were part of the manually checked sample.

3.5. Results

Results are given in Tables 5-8. Note that due to random sampling and modified sentence level metrics, the absolute figures resulting from the experiment are not directly comparable to the figures given in other studies. They are, however, in the same general range as those quoted by Rosen (2005) and Singh and Husain (2005) and can be compared among themselves.

It should be kept in mind that it is actually unclear whether the figures given *per text* are in fact characteristic of the full alignment of this text; but, since the aligners always operate on the whole text, the alignment of each of its sentences is at least strongly influenced by its full alignment. Any comparison between texts and, indeed, between subcorpora therefore has to be considered as tentative and relating foremost to the specific data set tested; the comparison between different alignment setups, however, is straightforward since it relates to the very same set of data.

German-Russian			
	recall	precision	f-measure
bsa	0,90	0,70	0,79
bsa, lemmatized	**0,92**	0,80	0,86
hunalign	0,80	0,90	0,85
hunalign, lemmatized	0,85	**0,93**	**0,89**
Polish-Russian			
	recall	precision	f-measure
bsa	0,92	0,90	0,91
bsa, lemmatized	**0,94**	0,93	**0,93**
hunalign	0,85	0,93	0,89
hunalign, lemmatized	0,89	**0,96**	0,92
Polish-Russian with only one language lemmatized			
	recall	precision	f-measure
bsa, PL lemmatized	**0,93**	0,91	**0,92**
bsa, RU lemmatized	**0,93**	0,84	0,89
hunalign, PL lemmatized	0,85	**0,92**	0,88
hunalign, RU lemmatized	0,84	0,91	0,87

Table 5: Overall results

The figures clearly show that on average (and with only one exception on a per-text basis) performance in terms of f-measure improves with lemmatization. This applies to both aligners and both language pairs. Although the absolute differences seem small, they reflect considerable improvements: For instance, the 0.02 difference in f-measure between lemmatized and non-lemmatized *bsa* in the Polish-Russian subcorpus with f-measure values at 0.91 and 0.93, respectively, reflects an error rate reduction of 0.9 to 0.7, that is, over 20 %.

Only in one case does the non-lemmatized variant outperform the lemmatized variant: *bsa* fails in aligning Lem's *Solaris*, presumably because it is abridged in some places[10], with f-measure down to 0.63; the non-lemmatized variant scores 0.71. This is due to a low precision of 0.49: *bsa* aligns huge chunks in the vicinity of the abridged data, apparently because there are not enough high certainty anchor points available.

Generally speaking, the hypothesis that lemmatizing only one text is not helpful was confirmed: f-measure was generally worse when only one language version of the texts was lemmatized in comparison to aligning the unaltered texts.

There was a surprising result, though: as Table 5 shows, lemmatizing only the Polish text in fact always performed better than if only Russian was lemmatized, and, with *bsa*, even improved quality in comparison to the alignment of non-lemmatized text. I have no explanation for this, although it might be surmised that differences in quality of the lemmatizers play some role here; it underscores the results from the other experiments that differences in text lead to differences in alignment that are hard to predict yet.

There is a notable difference between the language pairs: German-Russian, perhaps not surprisingly, aligns worse than Polish-Russian in terms of f-measure. This might be due to the general similarity of Slavic languages, but could also be attributed to stronger influence of interference leading to more literal translations. Texts translated from a third language were among the worst performing in both language pairs.

[10] Interestingly, the parts that were not translated into Russian mainly concern theological discussions.

	bsa unchanged			**bsa lemmatized**			**hunalign unchanged**			**hunalign lemmatized**		
	rec.	prec.	f-m.	recall	prec.	f-m.	recall	prec.	f-m.	recall	prec.	f-m.
BoellClown	0,86	0,62	0,72	0,88	0,78	0,83	0,73	0,86	0,79	0,90	0,98	**0,94**
EndeMomo	0,94	0,85	0,89	0,95	0,91	0,93	0,94	0,98	0,96	0,96	0,99	**0,98**
LemKongres	0,85	0,54	0,66	0,84	0,75	**0,79**	0,64	0,81	0,71	0,64	0,81	0,71
LemSolaris	0,93	0,51	0,66	0,95	0,66	0,78	0,92	0,94	0,93	0,93	0,95	**0,94**
StrugLebedi	0,91	0,87	0,89	0,96	0,88	**0,92**	0,81	0,92	0,86	0,83	0,91	0,87
StrugPiknik	0,91	0,59	0,72	0,89	0,74	0,81	0,66	0,80	0,72	0,79	0,91	**0,85**
allSents	0,90	0,70	0,79	0,92	0,80	0,86	0,80	0,90	0,85	0,85	0,93	**0,89**

Table 6: Results for the German-Russian subcorpus (highest f-measure highlighted)

	bsa unchanged			**bsa lemmatized**			**hunalign unchanged**			**hunalign lemmatized**		
	rec.	prec.	f-m.	recall	prec.	f-m.	recall	prec.	f-m.	recall	prec.	f-m.
BulgakovMaster	0,92	0,95	0,93	0,95	1,00	**0,98**	0,84	0,94	0,89	0,90	0,98	0,94
LemAstronauci	0,96	0,95	0,96	1,00	0,97	**0,99**	0,95	0,94	0,95	0,95	0,97	0,96
LemFiasko	0,97	1,00	0,98	0,99	1,00	**1,00**	0,97	0,99	0,98	0,98	0,98	0,98
LemKongres	0,90	0,92	0,91	0,94	0,96	**0,95**	0,83	0,88	0,85	0,87	0,94	0,90
LemPowGwiazd	0,94	0,96	0,95	0,94	0,98	**0,96**	0,85	0,89	0,87	0,93	0,99	**0,96**
LemSolaris	0,88	0,60	0,71	0,90	0,49	0,63	0,69	0,81	0,75	0,85	0,97	**0,91**
LemWizjaLokalna	0,97	0,93	0,95	0,97	0,99	**0,98**	0,97	0,97	0,97	0,96	0,92	0,94
Potter1	0,87	0,73	0,79	0,88	0,86	**0,87**	0,76	0,91	0,83	0,81	0,93	0,86
Potter2	0,79	0,89	0,84	0,86	0,93	**0,89**	0,68	0,92	0,78	0,71	0,95	0,82
StrugLebedi	1,00	0,95	0,97	1,00	0,95	**0,97**	0,93	0,97	0,95	0,93	0,97	0,95
StrugPiknik	0,90	0,91	0,91	0,93	0,96	**0,94**	0,86	0,92	0,89	0,87	0,93	0,90
allsents	0,92	0,90	0,91	0,94	0,93	**0,93**	0,85	0,93	0,89	0,89	0,96	0,92

Table 7: Results for the Polish-Russian subcorpus (highest f-measure highlighted)

	bsa, Russian lemmatized			**bsa, Polish lemmatized**			**hunalign, Russian lemmatized**			**hunalign, Polish lemmatized**		
	recall	prec	fm.	recall	prec	fm.	recall	prec	fm.	recall	prec	fm.
BulgakovMaster	0,93	0,91	0,92	0,95	0,98	**0,97**	0,86	0,96	0,91	0,82	0,93	0,87
LemAstronauci	0,95	0,91	0,93	0,97	0,95	**0,96**	0,93	0,94	0,94	0,92	0,91	0,92
LemFiasko	1,00	0,99	**1,00**	0,97	1,00	0,98	0,96	0,98	0,97	0,94	0,94	0,94
LemKongres	0,90	0,92	0,91	0,94	0,94	**0,94**	0,81	0,86	0,83	0,87	0,92	0,89
LemPowGwiazd	0,94	0,89	0,91	0,92	0,98	**0,95**	0,78	0,80	0,79	0,90	0,94	0,92
LemSolaris	0,89	0,46	0,61	0,86	0,63	0,73	0,73	0,86	0,79	0,82	0,94	**0,87**
LemWizjaLokalna	0,97	0,80	0,88	0,97	0,94	0,96	0,99	0,99	**0,99**	0,94	0,94	0,94
Potter1	0,89	0,71	0,79	0,87	0,86	**0,86**	0,73	0,87	0,79	0,74	0,92	0,82
Potter2	0,85	0,74	0,79	0,85	0,72	**0,78**	0,65	0,89	0,75	0,65	0,87	0,75
StrugLebedi	1,00	0,95	**0,97**	1,00	0,95	**0,97**	0,90	0,95	0,92	0,90	0,95	0,93
StrugPiknik	0,90	0,89	0,90	0,92	0,96	**0,94**	0,86	0,92	0,89	0,86	0,88	0,87
allSents	0,93	0,84	0,89	0,93	0,91	**0,92**	0,84	0,91	0,87	0,85	0,92	0,88

Table 8: Results for the Polish-Russian subcorpus with only one language lemmatized

Which of the aligners is better *in general*, however, cannot be decided on the basis of this experiment. On average, *bsa* fares slightly better than *hunalign* on Polish-Russian, while *hunalign* scores higher f-measure on the German-Russian subcorpus. Results differ depending on which text is aligned; while lemmatized *bsa* performs best on most Polish-Russian texts, its results were less stable than *hunalign's*, which did not seem to vary this much. Note that the relevant figure for low f-measure in aligning *Solaris* is due to precision, the value which is potentially biased by the metric used. Maybe one could say that *bsa* is sensitive to these omissions, but this needs to be investigated in detail in order to draw definite conclusions.

Taking into account the primitive heuristics used in filling out the blanks *bsa* leaves during alignment, it is surprising that it does not do worse than *hunalign*, where this was not necessary[11]. On the other hand, the outcome that *bsa* often performed better than *hunalign* might be explained by the fact that *bsa* builds its translation model on the full subcorpus of all texts to be aligned, while *hunalign* constructs its probabilistic dictionaries on individual texts. I has been shown that *bsa* results consistently improve with the amount of data aligned (Singh/Husain 2005), and the same is to be expected for *hunalign* since it relies on similar unsupervised learning techniques. There is thus space for improvement of both aligners.

As stated above, all conclusions in respect to individual texts have to be taken with caution. Alternatively, the division into texts could be regarded as an arbitrary segmentation of the test set that gives clues about how the two aligners cope with data of different characteristics. As seen from the results on these data segments as well as on average, the differences and similarities in f-measure do not generally reflect analogous differences and similarities in recall and precision of the aligners. An important result is thus that these two aligners *differ* with respect to each other and have different weaknesses and strengths, which becomes important in a committee approach such as investigated in Rosen (2005).

3.6. Conclusions

Summarizing the results, several points can be made:

1. There is a clear answer to whether or not lemmatization is helpful for alignment: it is, so the approach taken in RPC, i.e. semiautomatically performing even less-than-perfect lemmatization at inclusion stage, is well justified.

2. If only one text is lemmatized, results generally degrade but might improve, possibly depending on which of the two languages is lemmatized. Since this question only becomes important if one does not have a choice as to which language to lemmatize, it is generally safer to use unaltered text in both languages.

3. *Hunalign* and *bsa* work with comparable accuracy, but with different strengths and weaknesses, one of them being their ability to handle omissions. For the time being, the decision for or against the use of one of the two aligners in RPC cannot be justified on empirical grounds; since no manual intervention in alignment is provided, it seems that continuing to use *bsa* as default aligner is not an unwise choice.

In general, more research is needed. More work has to be done on textual characteristics important for alignment; how to diagnose them, and on what grounds to choose between aligners, would likewise be important here (see Singh/Husain 2005 for research on this topic and similar conclusions). In systems with more human

[11] It seems probable, though, that *hunalign* has to use heuristic of an essentially similar kind internally where its computations do not give high confidence results.

intervention than RPC, some automatic or semiautomatic environment that facilitates this choice *post hoc* could help in maximizing alignment quality.

Most promising, however, seems to be a committee voting solution for alignment, where several aligners with different properties would be used in combination. *Hunalign* and *bsa* are good candidates for membership in such a committee.

References

Banášová, M. (forthcoming). Možnosti prekladu nemeckých modálnych slovies do slovenčiny v romá ne H. Bölla. In: *Zborník XV. medzinárodného kolokvia mladých jazykovedcov VARIA XV*. Banská Bystrica.

Barlow, M. 1999. MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics* 4 (1), 319-327.

Barentsen, B. 2006. O pol'skich i russkich sootvetstvijach anglijskogo sojuza *till/until*. In: Bobrowski, I., Kowalik, K. (eds.): *Od fonemu do tekstu. Prace dedykowane Profesorowi Romanowi Laskowskiemu.* Kraków, 65-80.

Betsch, M., Meyer, R. 2003. Automatic annotation of Russian texts: evaluation of different tagging methods. In: Kosta, P. et al. (eds.): *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages (FDSL IV) held at Potsdam University, November 28-30, 2001.* Frankfurt/Main, 231-242.

Christ, O. 1994. *The IMS Corpus Workbench Technical Manual.* Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R. 2005. Massive multilingual corpus compilation; Acquis Communautaire and totale. In: *2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05)*. Poznań.
Available at: http://www.jrc.cec.eu.int/langtech/

Gale, W.A., Church, K.W. 1991. A program for aligning sentences in bilingual corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.* Morristown, 177-184.

Garabik, R. 2005. Levenshtein Edit Operations as a base for a morphology analyzer. In: Garabik, R. (ed.): *Computer Treatment of Slavic and East European Languages. Proceedings of Slovko 2005.* Bratislava, 50-58.

Gelbukh, A., Sidorov, G. 2003. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. In: Gelbukh, A. (ed.): *Computational Linguistics and Intelligent Text Processing (CICLing-2003).* Berlin et al., 215-220.
Available at: www.cic.ipn.mx/~sidorov/GelbukhSidorovMorphCICLING2003.ps

Lezius, W. 2000. Morphy – German morphology, part-of-speech tagging and applications. In: Ulrich, H. et al. (eds.): *Proceedings of the 9th EURALEX International Congress*. Stuttgart, 619-623.

Moore, R.C. 2002. Fast and accurate sentence alignment of bilingual corpora. In: *AMTA '02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users.* London, 135-144.
Available at: http://research.microsoft.com/research/downloads/default.aspx

Piasecki, M., Godlewski, G. (forthcoming). Reductionistic, tree and rule based tagger for Polish. In: Kłopotek M. et al. (eds.): *Intelligent Information Processing and Web Mining. Proceedings of the International IIS: IIPWM'06 Conference held in Ustron, Poland, June 19-22, 2006*. Berlin.

Rosen, A. 2005. In search of the best method for sentence alignment in parallel texts. In: Garabik, R. (ed.): *Computer Treatment of Slavic and East European Languages. Proceedings of Slovko 2005.* Bratislava, 174-185.
Available at: http://utkl.ff.cuni.cz/~rosen/public/mybibl.html

Singh, A.K., Husain, S. 2005. Comparison, selection and use of sentence alignment algorithms for new language pairs. In: *Proceedings of the ACL Workshop on Building and Using Parallel Texts, June 2005.* Ann Arbor, 99-106.
Available at: http://www.aclweb.org/anthology/W/W05/W05-0816

Sipka, D. 2006. Computer-assisted early inclusion of authentic Slavic materials. *Porta Linguarum. Revista Internacional de Didáctica de las Lenguas Extranjeras* 6, 33-41.

Tiedemann, J., Nygaard, L. 2004. The OPUS corpus – parallel & free. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04).* Lisbon.
Available at: http://logos.uio.no/opus/

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., Trón V. 2005. Parallel corpora for medium density languages. In: *Proceedings of RANLP'2005*. Borovets, 590-596.

Véronis, J., Langlais, P. 2000. Evaluation of parallel text alignment systems: the arcade project. In: Véronis, J. (ed.): *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht, 369-388.

Weiss, D. 2005. *Stempelator: A Hybrid Stemmer for the Polish Language.* Institute of Computing Science, Poznan University of Technology. (Research Report RA-002/05.)

Резюме

Создание параллельного корпуса текстов - еще более сложная задача, чем создание одноязычного корпуса, поскольку для него требуется не только тексты , переведенные на несколько языков, но и соотнесение эквивалентных фрагментов текста на разных языках, т.н. алигнирование (анг. alignment). К тому же сложности лингвистической разметки умножаются с возрастанием числа языков в корпусе.

Регенсбургский паралле льный корпус – это параллельный корпус преимущественно художественных текстов на славянских языках, специфика которого заключается в том, что его система направлена на упрощение работы пользователя при вводе в корпус как новых текстов, так и дополнительных языков. В первой части статьи обсуждаются архитектура и состав текстов этого корпуса. Во второй части описаны эксперименты , направленные на исследование проблем , связанных с автоматическим алигнированием при помощи двух программ: *hunalign* и *bsa*. В статье выдвигается гипотеза о том , что предварительная лемматизация текстов является адекватным способом улучшения качества алигнирования. Кроме того, вследствие ряда различий используемых между этими программами их комбинация представляется многообещающей для решения поставленных задач.

Regensburg Ruprecht von Waldenfels
ruprecht.waldenfels@sprachlit.uni-regensburg.de