Natalia Kotsyba

Institute of Slavic Studies
PAS (Warsaw)

The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR).

Objectives of creating the corpus

PolUKR[1], a Polish-Ukrainian parallel corpus was launched as a pilot corpus project in the Institute of Slavic Studies of the Polish Academy of Sciences in 2004. The corpus is intended for use as a tool for both human and machine users, and language material for compiling bilingual Polish<>Ukrainian dictionaries and a contrastive grammar for Polish and Ukrainian. It can also be used as a translation database and language learning materials.

Acquisition and preprocessing of parallel texts

Currently the corpus contains ab. 2 mln tokens (ab. 500K tokens in 70 parallel texts are publicly available for search through the web interface at http://corpus.domeczek.pl) that represent mostly modern Ukrainian and Polish literature (the 2nd part of the XXth century). Part of them was received from the translators[2], then the corresponding original was sought for and prepared accordingly. Another group was downloaded from existing digital libraries[3]. The quality of the texts was often unsatisfactory, as in most cases electronic texts were acquired through scanning the paper editions that were later submitted to the automatic Optical Character Recognition (OCR) procedure and needed further corrections. A large group of the texts was originally in the hard copy format, they were scanned, cleaned from images, page numbers and other unnecessary information, then OCRed with the help of the FineReader 9.0 program, checked for mistakes that appeared as a consequence of a poor OCR, recorded as MSWord documents and converted into simple UTF-8 encoded xml files that contain the information about the division into paragraphs extracted from doc files with the help of the AutoReplace function.

The text metadata are recorded into a MySQL database placed on the server. They include (if available): author, title, language, year of creation, publication place, year and publishing house , genre, translator, year of translation, source and original format of the text, etc. This information

[1] The current, extended version of the corpus is partially supported by MNiSW (Polish Ministry of Science and Higher Education) grant N N 104 0403 33 in 2007-2009.

[2] We would like to thank Katarzyna Kotyńska, Anna Łazar, Ola Hnatiuk and Helena Krasowska for sharing their texts.

[3] Some of the libraries used can be found at: http://lib.ru, http://www.ae-lib.org.ua/, http://www.4shared.com/dir/3997557/7fe59813/ebooki.html, http://exlibris.org.ua/, http://ukrcenter.com/library/default.asp, http://www.share.net.ua/, http://lib.proza.com.ua.

may be used to restrict the scope of the search e.g. one can choose only the texts created after a specific date or by a specific author.

Structural annotation

The texts are segmented into chunks that can be of two types: paragraphs and sentences. Sentences are always parts of paragraphs. Such structure of the document is encoded in a corresponding Document Type Definition file.

Morphological annotation

For adding the morphosyntactic information for the Polish texts we use the freely available TaKIPI toolset developed by Marcin Woliński, Adam Przepiórkowski, Adam Radziszewski and Maciej Piasecki, that includes a text chunker, a lemmatizer, a morphological tagger and a disambiguator.

Morphological tags are stored as value lists containing morphological class and grammatical categories adequate for a given class, e.g., the grammatical characteristics of *jedziecie* (*you*$_{pl}$ *go*) will be *fin:pl:sec:imperf* (finite verb form, plural, second person, imperfective aspect). If an ambiguity occurs for a given segment, several tags are listed. After the disambiguation procedure the most verisimilar "candidate" is given the disambiguation value "1".

An example of a tagged chunk "Dokąd jedziecie?"

```
<chunk type="p" xlink:href="#p5">
<chunk type="s">
   <tok>
     <orth>dokąd</orth>
     <lex disamb="1">
      <base>dokąd</base>
      <ctag>qub</ctag>
     </lex>
   </tok>
   <tok>
     <orth>jedziecie</orth>
     <lex disamb="1">
       <base>jechać</base>
       <ctag>fin:pl:sec:imperf</ctag>
     </lex>
   </tok>
   <tok>
     <orth>?</orth>
     <lex disamb="1">
       <base>?</base>
       <ctag>interp</ctag>
     </lex>
   </tok>
</chunk>
</chunk>
```

For the Ukrainian language we use the UGS (Ukrainian Grammatical Dictionary) developed at the ULIF NASU by Igor Shevchenko and Oleksandr Rabulets, that enables lemmatization and morphological annotation of texts, although its does not support disambiguation at the moment.

A common morphosyntactic tagset for Polish and Ukrainian was developed by us for the corpus needs based on the mentioned resources, see [Kotsyba et al. 2008, Коциба 2009] for details. Language specific categories and values are preserved, as our intention was not to lose any information. All the details will not be seen at the GUI search-level, but will be accessible for advanced users through self-defined regex-based corpus queries. The basic changes we had to introduce include a higher POS granulation for Ukrainian and regrouping some word classes for Polish to fit a more traditional understanding of the parts of speech. These quasi-changes are realized with the help of the mechanism of aliases and effect only the GUI search level. Reorganizing of information about the degree for Ukrainian adjectives and adverbs from the lexical to grammatical level has also been done to keep to the standards both in traditional grammars and current commonly accepted NLP treatment of the degree as a grammatical category. The special treatment of predicatives that was followed by us as well is described in detail in [Derzhanski, Kotsyba 2008].

The above format was also used for the Ukrainian language while converting the original annotated files.

Alignment

Presently the parallel texts are aligned at the paragraph level dynamically, i.e. paragraphs are enumerated during the searching procedure and paragraphs with the same order number that the ones where the searched fragment is found are shown along with the KWICs. The difference in the paragraph division had to be removed manually, so that their order numbers and content where equal. This situation is provisional – the paragraph level of the alignment is unsatisfactory as most paragraphs are too lengthy to easily spot the searched equivalent. The intended alignment level are sentences and, eventually, words.

One of the freely available programs that aligns parallel texts at the sentence level is the language independent HunAlign. The result of the alignment is recorded either as an intertwined text or as sets of corresponding sentences, so called link groups, represented by sentence numbers or other identifiers. Additional numeric information about the accuracy of alignment can be included as well. The program foresees the use of a corresponding bilingual dictionary to ensure a higher accuracy of the alignment. Such a dictionary can also be generated by the program itself from the currently fed in bitexts, if not available otherwise. The results of aligning Polish and Ukrainian texts without a dictionary were far from satisfactory. For the purpose of a more accurate alignment we have developed a bilingual dictionary structured according to the HunAlign demands. It is recorded as a plain text where each entry takes a separate line: the original word or expression, @-sign, the equivalent word or expression. Since many words and expressions have several equivalents due to polysemy, the same entries on the left side can be repeated with different equivalents. The alignment dictionary was generated automatically from the database version of the Polish-Ukrainian dictionary that is currently developed as a joint project of ULIF NASU and ISS PAS, and contains 31088 entries.

Fragment of the dictionary:

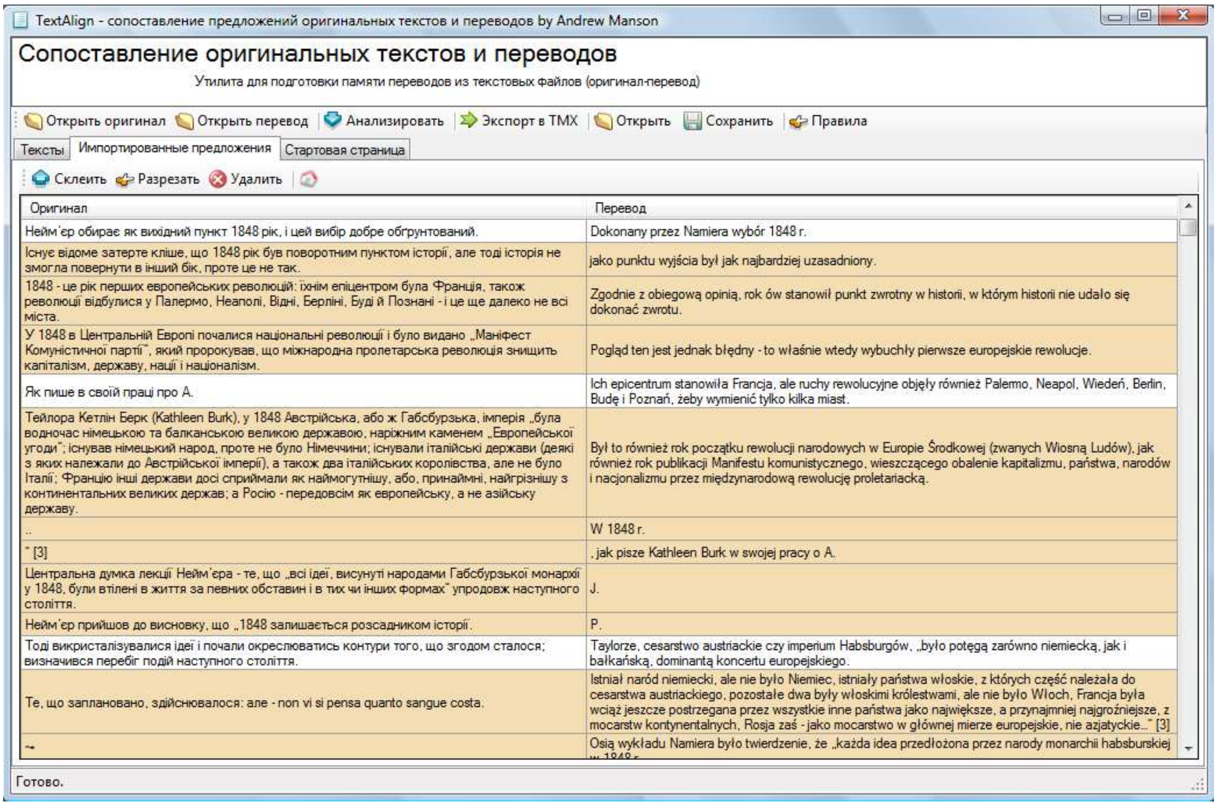
```
białokrusz @ окис свинцю
blałolicy @ білолиций
blałoramienny @ білоплечий
białoruszczyzna @ білорущина
```

```
białoruszczyzna @ все, що білоруське
bibułka @ папіросний папір
bibułka @ цигарковий папір
bibułomania @ манія збирати старі рукописи
biczować @ батожити
biczowanie @ батоження
biczyk @ батіжок
biczykowaty @ подібний до батіжка
bić @ бити
biję @ б'ю
biec @ бігти
```

Since both Polish and Ukrainian are highly inflected languages, basic dictionary forms are not enough. Either we need lemmatized texts, or a dictionary with all possible forms generated. The first option seems to be easier to realize, but for this we need to adjust the alignment algorithm and to work with already annotated texts.

Another option for aligning is the TextAlign, a user friendly software with GUI and editing possibilities. The only possible input format there is RTF (rich text format), the output is a TMX file with an intertwined parallel text. The main problem with the unequal number of sentences in parallel texts that effected the quality of the results produced by the fully automatic and hardly controllable Hunalign is compensated by the possibility of an easy and quick alignment edition in the TextAlign. However, the sentence segmentation algorithm in the TextAlign is too simple for satisfactory results.

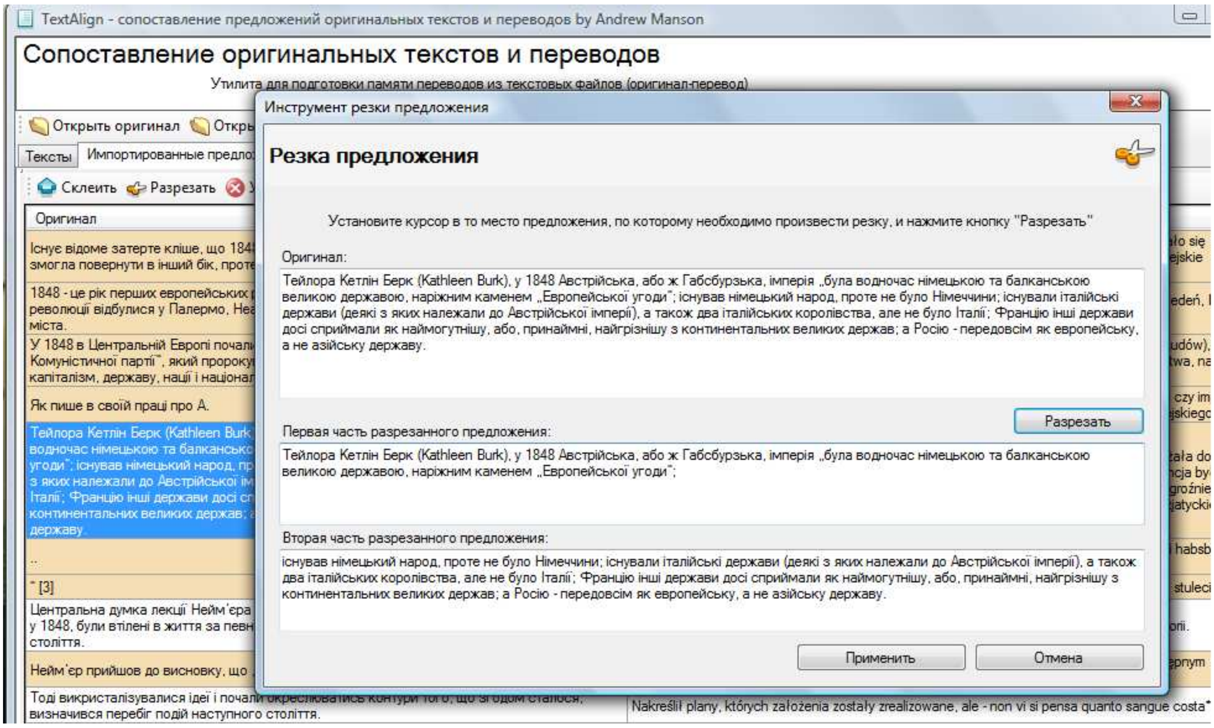
Example of alignment results by TextAlign, pre-editing phase


TextAlign - сопоставление предложений оригинальных текстов и переводов by Andrew Manson
Сопоставление оригинальных текстов и переводов
Утилита для подготовки памяти переводов из текстовых файлов (оригинал-перевод)
Открыть оригинал
Открыть перевод
Анализировать
Экспорт в TMX
Открыть
Сохранить
Правила
Тексты
Импортированные предложения
Стартовая страница
Склеить
Разрезать
Удалить
Готово.

Оригинал	Перевод
Нейм'єр обирає як вихідний пункт 1848 рік, і цей вибір добре обґрунтований.	Dokonany przez Namiera wybór 1848 r.
Існує відоме затерте кліше, що 1848 рік був поворотним пунктом історії, але тоді історія не змогла повернути в інший бік, проте це не так.	jako punktu wyjścia był jak najbardziej uzasadniony.
1848 - це рік перших європейських революцій: їхнім епіцентром була Франція, також революції відбулися у Палермо, Неаполі, Відні, Берліні, Буді й Познані - і це ще далеко не всі міста.	Zgodnie z obiegową opinią, rok ów stanowił punkt zwrotny w historii, w którym historii nie udało się dokonać zwrotu.
У 1848 в Центральній Европі почалися національні революції і було видано „Маніфест Комуністичної партії", який пророкував, що міжнародна пролетарська революція знищить капіталізм, державу, нації і націоналізм.	Pogląd ten jest jednak błędny - to właśnie wtedy wybuchły pierwsze europejskie rewolucje.
Як пише в своїй праці про А.	Ich epicentrum stanowiła Francja, ale ruchy rewolucyjne objęły również Palermo, Neapol, Wiedeń, Berlin, Budę i Poznań, żeby wymienić tylko kilka miast.
Тейлора Кетлін Берк (Kathleen Burk), у 1848 Австрійська, або ж Габсбурзька, імперія „була водночас німецькою та балканською великою державою, наріжним каменем „Европейської угоди"; існував німецький народ, проте не було Німеччини; існували італійські держави (деякі з яких належали до Австрійської імперії), а також два італійських королівства, але не було Італії; Францію інші держави досі сприймали як наймогутнішу, або, принаймні, найгрізнішу з континентальних великих держав; а Росію - передовсім як европейську, а не азійську державу.	Był to również rok początku rewolucji narodowych w Europie Środkowej (zwanych Wiosną Ludów), jak również rok publikacji Manifestu komunistycznego, wieszczącego obalenie kapitalizmu, państwa, narodów i nacjonalizmu przez międzynarodową rewolucję proletariacką.
..	W 1848 r.
" [3]	, jak pisze Kathleen Burk w swojej pracy o A.
Центральна думка лекції Нейм'єра - те, що „всі ідеї, висунуті народами Габсбурзької монархії у 1848, були втілені в життя за певних обставин і в тих чи інших формах" упродовж наступного століття.	J.
Нейм'єр прийшов до висновку, що „1848 залишається розсадником історії.	P.
Тоді викристалізувалися ідеї і почали окреслюватись контури того, що згодом сталося; визначився перебіг подій наступного століття.	Taylorze, cesarstwo austriackie czy imperium Habsburgów, „było potęgą zarówno niemiecką, jak i bałkańską, dominantą koncertu europejskiego.
Те, що заплановано, здійснювалося: але - non vi si pensa quanto sangue costa.	Istniał naród niemiecki, ale nie było Niemiec, istniały państwa włoskie, z których część należała do cesarstwa austriackiego, pozostałe dwa były włoskimi królestwami, ale nie było Włoch, Francja była wciąż jeszcze postrzegana przez wszystkie inne państwa jako największe, a przynajmniej najgroźniejsze, z mocarstw kontynentalnych, Rosja zaś - jako mocarstwo w głównej mierze europejskie, nie azjatyckie..." [3]
--	Osią wykładu Namiera było twierdzenie, że „każda idea przedłożona przez narody monarchii habsburskiej w 1848 r.

It can be seen from the example above that sentence borders are defined basing on punctuation marks without considering common abbreviations ended with full stops, which can generate wrong sentence segmentation.

Example of a manual splitting procedure with the help of TextAlign.



TextAlign - сопоставление предложений оригинальных текстов и переводов by Andrew Manson
Сопоставление оригинальных текстов и переводов
Утилита для подготовки памяти переводов из текстовых файлов (оригинал-перевод)
Открыть оригинал
Тексты
Склеить
Разрезать
Оригинал
Инструмент резки предложения
Резка предложения
Разрезать
Применить
Отмена

Установите курсор в то место предложения, по которому необходимо произвести резку, и нажмите кнопку "Разрезать"

Оригинал:

Тейлора Кетлін Берк (Kathleen Burk), у 1848 Австрійська, або ж Габсбурзька, імперія „була водночас німецькою та балканською великою державою, наріжним каменем „Европейської угоди"; існував німецький народ, проте не було Німеччини; існували італійські держави (деякі з яких належали до Австрійської імперії), а також два італійських королівства, але не було Італії; Францію інші держави досі сприймали як наймогутнішу, або, принаймні, найгрізнішу з континентальних великих держав; а Росію - передовсім як европейську, а не азійську державу.

Первая часть разрезанного предложения:

Тейлора Кетлін Берк (Kathleen Burk), у 1848 Австрійська, або ж Габсбурзька, імперія „була водночас німецькою та балканською великою державою, наріжним каменем „Европейської угоди";

Вторая часть разрезанного предложения:

існував німецький народ, проте не було Німеччини; існували італійські держави (деякі з яких належали до Австрійської імперії), а також два італійських королівства, але не було Італії; Францію інші держави досі сприймали як наймогутнішу, або, принаймні, найгрізнішу з континентальних великих держав; а Росію - передовсім як европейську, а не азійську державу.

At the moment we are developing a PLUczeK program that will combine the features of the HunAlign and the TextAlign. It will include an editable plugging-in module of text-segmentation at the paragraph and sentence levels, which has to ensure language independence of the program. The sentence segmentation module is rule based, it presupposes the use of such heuristics as common abbreviation to function as a stop list, combinations and sequences of abbreviations and punctuation marks, forms of the reported speech presentation (that can also be different across languages), cf. also [Rudolf, 2004]. The program will work with both plain texts and morphologically annotated xml files, addressing either the information about the actual form of the token, or its lemma, as well as using grammatical information for sentence segmentation (a verb or a preposition cannot be a proper name, hence, written with a capital letter they signal about the beginning of a sentence, etc.). The program will also have a GUI interface and enable editing of the segmentation.

We have chosen the XCES format for alignment records. The information about corresponding sentences is stored in a separate file. An example fragment of an alignment file is below (sentences 1 i 2 of the second link group are translated as one sentence).

```
<cesAlign>
 <cesHeader>
...
...
  <translations xml:base="http://corpus.domeczek.pl/corpus">
    <translation trans.loc="exampleAna.ua.xml" lang="ua" xml:lang="ua"
    n="1" />
    <translation trans.loc="exampleAna.pl.xml" lang="pl" xml:lang="pl"
   n="2" />
  </translations>
 </profileDesc>
</cesHeader>

 <linkList>
   <linkGrp id="p1" targType="s">
     <link>
       <align xlink:href="#p1s1" />
       <align xlink:href="#p1s1" />
     </link>
     <link>
       <align xlink:href="#p1s2" />
       <align xlink:href="#p1s2" />
     </link>
   </linkGrp>
   <linkGrp id="p2" targType="s">
     <link>
       <align xlink:href="#xpointer(id('p2s1')/range-to(id('p2s2')))"
   />
       <align xlink:href="#p2s1" />
     </link>
     <link>
       <align xlink:href="#p2s3" />
       <align xlink:href="#p2s2" />
     </link>
   </linkGrp>
 </linkList>
 </cesAlign>
```

Even sentence alignment cannot reach a 100% accuracy due to objective reasons. In the table below, fragments that are parts of one sentence are highlighted with the same shade.

Dokonany przez Namiera wybór 1848 r. jako punktu wyjścia był jak najbardziej uzasadniony.	Нейм'єр обирає як вихідний пункт 1848 рік, і цей вибір добре обґрунтований.
Zgodnie z obiegową opinią, rok ów stanowił punkt zwrotny w historii, w którym historii nie udało się dokonać zwrotu.	Існує відоме затерте кліше, що 1848 рік був поворотним пунктом історії, але тоді історія не змогла повернути в інший бік,
Pogląd ten jest jednak błędny –	проте це не так.
to właśnie wtedy wybuchły pierwsze europejskie rewolucje.	1848 – це рік перших європейських революцій:
Ich epicentrum stanowiła Francja, ale ruchy rewolucyjne objęły również Palermo, Neapol, Wiedeń, Berlin, Budę i Poznań, żeby wymienić tylko kilka miast.	їхнім епіцентром була Франція, також революції відбулися у Палермо, Неаполі, Відні, Берліні, Буді й Познані – і це ще далеко не всі міста.

This means that mistakes are practically unavoidable, especially with large amounts of texts, but its is still possible to keep the general quality of the corpus sufficient for working with it and receiving objective results.

Conclusions and further work

The current state of PolUKR enables already searching for translation equivalent and can be used as a translation memory database both by human translators and researchers and machines. But the corpus can be enhanced in a number of ways, like finer alignment level, enriching with further annotation of different types, including also semantic and referential information. Automatic word-level alignment can be of significant help while compiling bilingual dictionaries. The search engine has to be adjusted to enable searching for the new information as well.

Literature

Broda Bartosz, Piasecki Maciej & Radziszewski Adam. Towards a Set of General Purpose Morphosyntactic Tools for Polish. Proceedings of Intelligent Information Systems, Zakopane Poland, 2008. Institute of Computer Science PAS, 2008.

Ivan Derzhanski and Natalia Kotsyba. *The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging of Slavic Languages*. Proceedings of the International Workshop within MONDILEX project, Moscow, 2-4 October 2008.

Hunalign - sentence level aligner: http://mokk.bme.hu/resources/hunalign.

Natalia Kotsyba, Olha Shypnivska and Magdalena Turska. *Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus)*. Proceedings of Intelligent Information Systems, Zakopane, Poland, 2008. Institute of Computer Science PAS, 2008.

Adam Przepiórkowski and Marcin Woliński. *A Flexemic Tagset for Polish*. In: *The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003.* http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf

Michał Rudolf. *Metody automatycznej analizy korpusu tekstów polskich. Pozyskiwanie, wzbogacanie i przetwarzanie informacji lingwistycznych.* Warszawa, 2004.

TextAlign in MT2007 (Memory Translation Computer Aided Tool): http://mt2007-cat.ru/index.html.

Magdalena Turska and Natalia Kotsyba. *Polsko-Ukraiński korpus równoległy (PolUKR).* „Materiały LXIII Zjazdu Polskiego Towarzystwa Językoznawczego", Warszawa.

Magdalena Turska and Natalia Kotsyba. *Polish-Ukrainian Parallel Corpus and its Possible Applications*, Proceedings of the International Conference "Practical Applications in Language and Computers, 7-9 April, Łódź", Peter Lang GmbH, 2007.

v. Waldenfels, R. *Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment*. In: Brehmer, B., Zdanova, V., Zimny, R. (Hrsg.); Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 123-138, 2006.

Коциба Наталія. *Принципи морфосинтактичного таґування польсько-українського паралельного корпусу (PolUKR)*. Proceedings of the International Conference "MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies, Parthenit – Crimea, Ukraine, September 2008", 2009 (in preparation).

Широков В.А, О.В.Бугаков, Т.О.Грязнухіна, О.М.Костишин, М.Ю.Кригін, Т.П.Любченко, О.Г.Рабулець, О.О.Сидоренко, Н.М.Сидорчук, І.В.Шевченко, О.О.Шипнівська, К.М.Якименко. *Корпусна лінгвістика*. Київ: Довіра, 2005.

Abstract

The article describes the present state of work on PolUKR, the Polish-Ukrainian parallel corpus, developed in the Institute of Slavic Studies of the Polish Academy of Sciences since 2004. Presented are the ways of bitexts' acquisition, their structure and pre-processing stages; the solutions concerning the common morphosyntactic annotation pattern for Polish and Ukrainian, as well as annotation methods; the alignment format and the software used or developed for the corpus needs.

Recommendations

One of the objectives of the current project is to develop a scheme for creating a parallel corpus for any pair of Slavic languages. At the moment a researcher who deals with Slavic parallel corpora envisages several major problems that need to be attended to. One of the still unresolved issues is a common morphological annotation tagset for Slavic languages that should ensure uniform search through both parts of a corpus at the same time. Technical bilingual dictionaries for sentence alignment as well as a user friendly alignment editor are necessary to enable controllable high-quality alignment. A free, platform independent search engine for parallel corpora is also needed.