Magdalena Turska
Natalia Kotsyba
Institute of Slavic Studies
Polish Academy of Sciences

PolUKR - POLISH-UKRAINIAN PARALLEL CORPUS (A PROJECT)

1. Introduction: available Polish and Ukrainian language resources

Nowadays there seems to be ever increasing need of digitalized and preprocessed language resources, especially for scholarly work but also for quotidian purposes of the wide audience. Such resources would be primarily on-line dictionaries and encyclopedias but also language corpora, either mono- or multilingual. For many users, especially language learners, bilingual parallel corpora prove very useful in their studies. These could also serve as learners' and translators' resources and provide basic material for linguistic research.

Although dictionaries and encyclopedias are quite common throughout the internet, corpora for less popular languages are still relatively rare and small, while parallel corpora based on them are still rarer, even to the point of non-existence.

Situation for Polish-Ukrainian language pair does not seem to look particularly optimistic in that field[1]. Existing printed bilingual dictionaries tend to be either quite old, thus not reflecting the current state of the language and often difficult to reach or small in volume (up to 30 000 entries), not very useful in professional work. Another disadvantage is the common method of preparation of Polish-Ukrainian and Ukrainian-Polish dictionaries via Russian language that causes many inadequacies and errors in the resulting dictionary. As for on-line resources there are attempts to create bilingual dictionaries but for now no serious results were achieved (cf. www.slovnyk.org, and the magazine 'Ridna Mova', which covered up to now only letter 'A' of the Polish-Ukrainian dictionary). As for the explanatory dictionaries, there are digitalized versions of most popular Polish and Ukrainian dictionaries though these share the vices of printed editions.

Currently there exist two large electronic corpora of the Polish language: IPI PAN Corpus (www.korpus.pl) with 300 million segments and PELCRA Corpus (korpus.ia.uni.lodz.pl) with 85 million tokens. For Ukrainian language no corpus is yet available though the work is well advanced at Ukrainian Academy of Science.

At the moment there is no parallel corpus for Polish and Ukrainian languages, although it seems that with a reasonably sized parallel corpus one might achieve immediate goals, such as providing tools for students and translators (especially a parallel concordancer) and basis for linguistic work (with tools for corpora annotation and advanced search). Not only that, but starting from a parallel corpus it is possible to create high quality bilingual dictionaries that will reflect latest changes in the languages immediately, which is especially important considering the present speed of information growth.

2. Corpus representation

We have decided to base the structure of our parallel Polish-Ukrainian corpus on the existing IPI PAN corpus of Polish texts. The main tasks faced while creating such a parallel corpus are the necessity of extending the tagset to cover the Ukrainian language specific morphological information as well as adding information about the alignment of parallel documents.

The corpus consists of separate documents stored in CES file format. Every document consists of several files: the header part that includes all the metadata, the actual content (the body part) and additional part

[1] We present more detailed discussion of this question in [Kotsyba, Turska 2006]

containing morphological annotation.

For every document of the parallel corpus there exists a corresponding parallel document (co-text), containing either its translation or the original version itself. In addition, there is one file storing alignment information for every bi-text (a pair of texts).

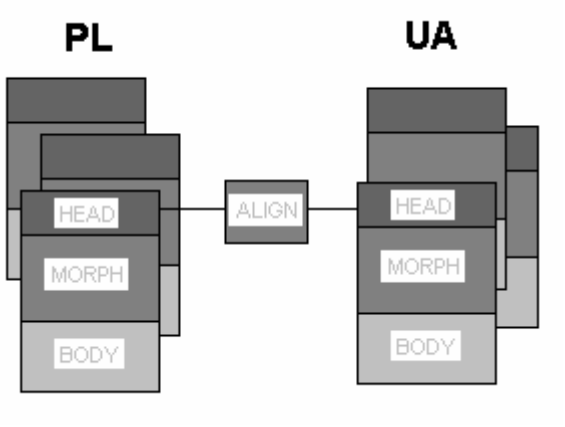

Illustration 1. Parallel corpus structure

3. Metainformation. The HEADER part

With every text in parallel corpus we associate metadata concerning author or translator, title, origin and publisher and many other. This information may be used to restrict the scope of the search e.g. choose only the texts created after a specific date or by a specific author. We believe this could prove very useful in various kinds of research, including dialectal studies, language interference, sociolinguistics etc., although in many cases some or even most of the metainformation is not available.

The information we strive to obtain is as follows:

Author & Translator	**Original Text & Translation**
- birth year - country and region of origin - mother tongue - other languages - education - location	- date of creation - date of first publication - author(s)/translator(s) - publishing medium - genre - first publisher - intended reader

4. Morphological annotation. The MORPH part

To transform a plain text into the morphologically annotated corpus bi-text additional tools are required, especially morphological analyzers for Polish and Ukrainian languages. We intend to use Morfeusz analyzer for the Polish part[2] and the software created on the basis of the grammatical dictionary of Ukrainian at ULIF (Ukrainian Linguistic Informational Fund at Ukrainian Academy of Sciences) [3] for the Ukrainian texts.

Morphological tags are stored as value lists containing morphological class and grammatical categories adequate for given class, e.g.: *jechać* (*to go*) it will be *fin:pl:sec:imperf*. If ambiguity occurs for a given segment, several tags are listed.

```
<chunk type="p" xlink:href="#p5">
<chunk type="s">
   <tok>
```

[2] courtesy of the author Marcin Woliński [Woliński, 2003],
[3] [Shirokov et al. 2005]

```
      <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>
```

Illustration 2. Morphological annotation

Our tagset for morphological information encoding is based on the IPI PAN corpus[4], yet it is still necessary to expand it to cover Ukrainian specific morphological classes. We tried to agree our extensions with the solutions developed at ULIF. The initial comparison of the available tagsets developed for Ukrainian and Polish allows at the moment only to make some preliminary generalizations: The approaches used to describe morphological systems in [Woliński 2003] and [Shyrokov et al. 2005] respectively differ considerably and it will take a great deal of effort to map the tagsets. One possible way seems to be to increase the common tagset by adding the categories that exist in one language description but not in the other. There will be an initially common group, of course, but *prima facie* it looks discouragingly small – only 6 categories coincide formally. As many as 21 of the categories from this formal union of the tagsets (counting 50 entries) are unique for the Ukrainian morphological system and 23 are unique for the Polish one.

Hence, the preliminary analysis shows that disregarding the intuitively assumed similarity of the two grammatical systems due to the kinship of the languages there are more differences in their formal descriptions than one could expect. The main reason, however, lies rather in different conceptualization of the grammatical categories. For instance, the ULIF scheme treats comparative and superlative adjectives as categories of the same order as positive adjectives (which is similar to some existing tagsets for other languages like the one used for the British National Corpus), while Woliński does that according to the traditional grammatical intuitions, namely, he introduces degree as the inflectional property of adjectives and adverbs. On the other hand, both tagsets include the category of "predicative", however, the scope of its grammatical meaning is different in the two cases. At the moment it would be too risky to make any claims about the common tagset, but it is clear that one cannot base it exclusively on the formal characteristics of the existing classifications. Some theoretical review is still necessary.

5. Alignment. The ALIGN part

The possibility of presenting parallel documents in the aligned form is indispensable for the parallel corpora resources to function properly.

AIESEC (międzynarodowe stowarzyszenie studentów ekonomii i zarządzania) to niepolityczna, naukowa i niekomercyjna organizacja, w całości zarządzana przez studentów.	AIESEC (міжнародна асоціація студентів, які займаються економією й управлінням) - неполітична, освітня і некомерційна організація, повністю керована студентами.

[4] [Przepiórkowski, Woliński 2003]

Działamy w ponad 950 szkołach wyższych. Do naszej organizacji należy 60.000 studentów z 84 krajów świata. Naszym celem jest rozwój jednostek i krajów pod względem gospodarczym, kulturalnym i moralnym. Szczególną uwagę przywiązujemy przy tym do międzynarodowego wzajemnego zrozumienia, jedności i współpracy.	Діємо в понад 950 вищих учбових закладах і об'єднуємо 60 тис. студентів з 84 країн світу. Наша мета - розвиток людей і країн в економічному, культурному і моральному аспектах. Зокрема, особливу увагу ми приділяємо міжнародному взаємозрозумінню, єдності і співробітництву.

Illustration 3. Text alignment on paragraph level

Text alignment on the paragraph level is used comparatively rarely, mostly due to their size which tends to be unnecessary large for grasping the context of the searched item. More 'user friendly' attitude would be aligning the corpus on the sentence level. However, singular sentences of the original text may be sometimes translated as two or more ones and vice versa (see the second paragraph of the example below).

AIESEC (międzynarodowe stowarzyszenie studentów ekonomii i zarządzania) to niepolityczna, naukowa i niekomercyjna organizacja, w całości zarządzana przez studentów. Działamy w ponad 950 szkołach wyższych. Do naszej organizacji należy 60.000 studentów z 84 krajów świata. Naszym celem jest rozwój jednostek i krajów pod względem gospodarczym, kulturalnym i moralnym.	AIESEC (міжнародна асоціація студентів, які займаються економією й управлінням) - неполітична, освітня і некомерційна організація, повністю керована студентами. Діємо в понад 950 вищих учбових закладах і об'єднуємо 60 тис. студентів з 84 країн світу. Наша мета - розвиток людей і країн в економічному, культурному і моральному аспектах.

Illustration 4. The same bi-text aligned on the sentence level

```
<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>
```

Illustration 5. Fragment of the alignment information file (sentences 1 i 2 of the second paragraph translated as one sentence)

6. How to acquire bi-texts?

To acquire a substantial number of parallel texts in our case it is essential to organize the efficient cooperation of several institutions. The common sources are: publishing houses, translators, regulations and other legal documents, multi-lingual proceedings protocols, technical documentation and manuals, bilingual magazines and web sites. In our work we intend to gather data from all of these, trying to keep the corpus relatively 'balanced'.

As a means to gain translators' cooperation we propose to create a friendly, on-line working environment with the access to discussion forums and bulletin boards but above all to the growing digital language resources database with the possibility for the users to contribute to them. This, we hope, could be the meeting point for the scholars, translators and students community. As a desired side-effect the material database could steadily grow in size and cover many genres of modern texts.

7. Virtual lexicographical laboratory (VLL) - towards a high quality dictionary

With the available digital language resources it will be possible to create the Polish-Ukrainian bilingual dictionary using the following algorithm[5].

1. generating the entry list based on items found in corpora
2. distinguishing the senses that differ in translation after concordances analysis according to a coherent classification
3. joining the senses between languages

We are convinced that the outlined above corpus-based approach can help create a dictionary that will cover a wide range of lexicon, reflect the current state of both languages, clearly show the distinction between different word senses (especially when used on-line and linked with references to the parallel corpus) and have a relatively small number of errors.

An important step towards this goal is to create an environment for geographically distributed research teams in the field of linguistic. Such, as we call it, Virtual Lexicographical Laboratory should provide (besides above mentioned functionality) means to prepare mono- and multilingual dictionaries, freely annotate the corpus according to the users research goals and also work on semantic networks.
For this environment to work it is essential for it to be:
- on-line, to provide constant, simultaneous access for researchers everywhere
- user-friendly, to let even those not accustomed to computers cooperate and contribute to the research
- flexible, to change with changing users needs
- standardized, to allow information exchange with other databases and research teams.

The latter two postulates are strictly intermingled, as only thorough standardization allows to replace and add some new modules, especially prepared by other parties, in the working system so that it does not disturb the other parts and the users. To meet this end we intend to comply with Unicode standard for language encoding, XML/CES for bi-texts representation and modular architecture allowing to use any specific tools (e.g.

[5] We discuss this in detail in [Kotsyba, Turska 2006].

morphological analyzers) with the help of a wrapper translating input and output formats. This will guarantee the VLL to be suitable for use with other languages and prepared for further development.

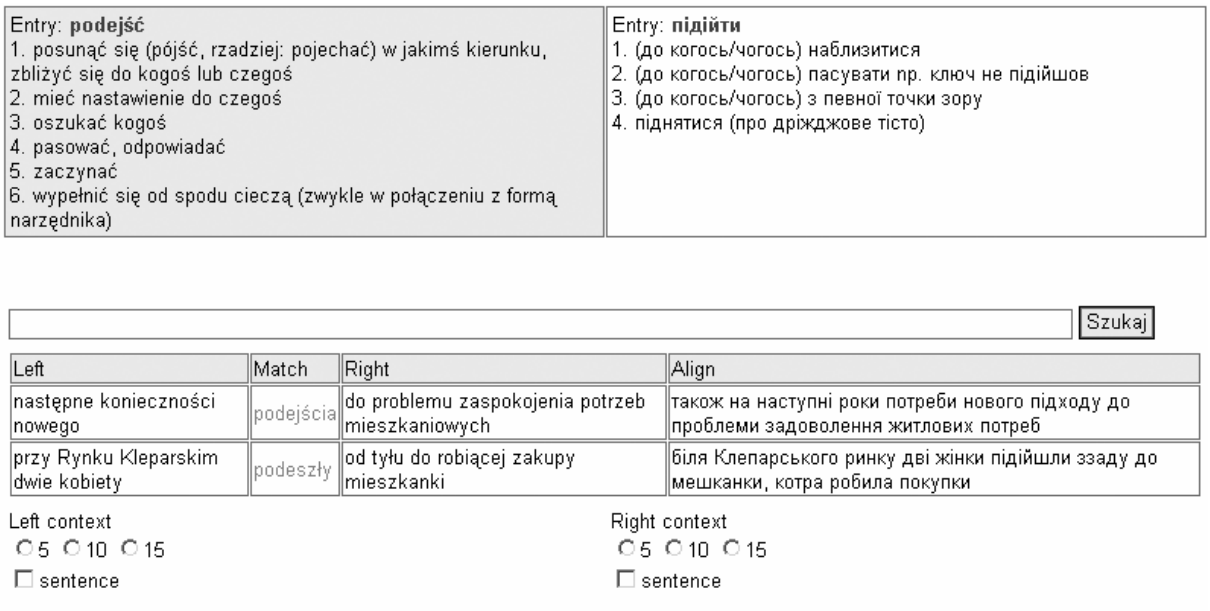

Illustration 6. VLL interface sketch

8. Summary

At the moment, the preliminary version of PolUKR is freely accessible through the web site http://corpus.domeczek.pl. It contains ca 50 parallel documents in XCES format. New documents are added as they are obtained. It is possible to search the corpus in a simple method based on the regular expressions mechanism. The search engine is being developed and after the morphological annotation process is completed it will be possible to use much more powerful query language, similar to the IPI PAN solution[6]. The search results are presented in the form of parallel concordances aligned on the paragraph level.

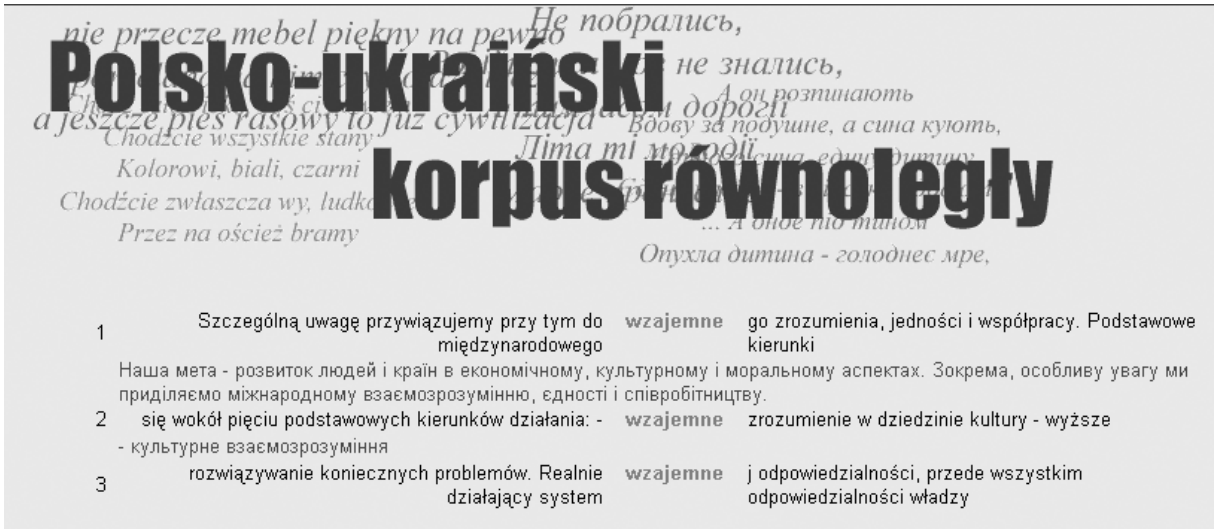

Illustration 7. Search results screen

Further work on our corpus may be divided in two main paths: organizational with its main goal to assure the cooperation of institutions and individuals in possession of translated Polish and Ukrainian texts and the technical one, that needs to provide means and tools to be used with the corpus. Priorities for the technical

[6] See about Poliqarp guery language in [Przepiórkowski 2004].

path are alignment on the sentence level and morphological annotation of the corpus together with the adequate search engine. The next stage will be the work on Virtual Laboratory software.

We believe that such a parallel corpus with an efficient search engine, a powerful yet easy to grasp query language and user friendly parallel concordancer could become one of most useful resources for all Polish and Ukrainian scholars, translators and students as well as be a contribution for future developments of parallel corpora for other languages.

REFERENCES

Corpus Encoding Standard: http://www.cs.vassar.edu/CES/CES1-0.html

Corpus Typology: http://webdeptos.uma.es/filifa/personal/amoreno/teaching/cl/corpus-typ.html

EAGLES Guidelines: http://www.ilc.pi.cnr.it/EAGLES96/browse.html

Korpus IPI PAN: http://korpus.pl

Korpus referencyjny języka polskiego PELCRA: http://korpus.ia.uni.lodz.pl/

Kotsyba, N., Turska, M. (2006). „Leksykografia polsko-ukraińska – stan obecny i perspektywy". In *Semantyka i konfrontacja językowa*, t. III. Warszawa, SOW.

PolUKR – Polsko-Ukraiński Korpus Równoległy: http://corpus.domeczek.pl

Przepiórkowski A. (2004). *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*: http://dach.ipipan.waw.pl/~adamp/Papers/2004-corpus/

Przepiórkowski A., Woliński M. (2003). *A Flexemic Tagset for Polish* http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf

Turska, M., Kotsyba, N. „Polsko-Ukraiński korpus równoległy (PolUKR)". In *Materiały LXIII Zjazdu Polskiego Towarzystwa Językoznawczego*, Warszawa (in print).

TEI Guidelines for Electronic Text Encoding and Interchange: http://etext.virginia.edu/TEI.html

Ukrainian Linguistic Portal: www.ulif.org.ua

Woliński, M. (2003). „System znaczników morfosyntaktycznych w korpusie IPI PAN". In *Polonica XXII-XXIII*, p. 39-55.

Широков В.А (1998): *Інформаційна теорія лексикографічних систем*. - Київ: Довіра.

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 (2005). *Корпусна лінгвістика*. Київ: Довіра.

KEYWORDS

corpus linguistics, parallel corpus, morphological annotation, alignment, parallel concordancer, Polish, Ukrainian, XML, XCES