NATALIA KOTSYBA

Institute for Interdisciplinary Studies,

Warsaw University

POLUKR (A POLISH-UKRAINIAN PARALLEL CORPUS) AS A TESTBED FOR A PARALLEL CORPORA TOOLBOX

Introduction

There are several major points to consider while working with a pair of medium-density languages like Polish and Ukrainian, among them: availability of texts and cost of developing of language specific processing tools. It is quite obvious that the quantity of available bitexts (original and translated texts) of two medium-sized languages does not allow creating a large and versatile parallel corpus without the participation of a larger community. Whereas fiction is relatively easy to attain, a large amount of parallel texts of other genres is produced and most probably never published by individual translators or translation companies. Language researchers and teachers collect their own problem specific corpora and vocabularies. Most often these projects do not involve the use of the present-day language technologies, as those are still either not sufficiently developed, or not reachable, or, as it is often the case, have not got to people's awareness. Even if such awareness does exist, there is still a long way to a practical realization of using the tools. At the same time, developing various tools for a particular small corpus is quite expensive, which makes their reuse highly welcome. All the mentioned groups might be interested in using a comprehensive parallel corpus of a better quality and bigger size than their self-made projects and one could expect them to be potential co-developers of such an initiative. Therefore, it seems to be a good investment to provide groups of potential users without technical background with means of enriching a common publicly available corpus with their own texts.

This paper is a summary of a multifaceted PolUKR[1] project whose aim is to prepare a methodological and instrumental background for linguists who wish to use and/or create parallel or problem-tailored monolingual corpora but do not have a sufficient technical background for developing their own software.

[1] The project website address is http://www.domeczek.pl/~polukr/index.php?option=welcome.

The rest of the paper is structured as follows: the first section gives an outline of the history of the project, the second one describes the content of the Polish-Ukrainian parallel corpus which is the end product of the project, the third section presents the pipeline of creating a "home-made" parallel corpus, the next section discusses the ways of ensuring format flexibility of a corpus and its compatibility with existing tools, the following section presents the software developed as by-products of the project and at the end there is a short presentation of the current work and plans for further developments.

History of the project

The origin of the corpus is quite informal – it was a follow-up of a corpus linguistics session of the International School for Humanities organized in Warsaw University in 2004 for young researchers in linguistics from the countries of the Central and Eastern Europe. The lack of language resources and processing tools for Slavic languages was especially striking when compared to the developments for Western European ones and this was an impetus for starting creating a corpus for Ukrainian. A Polish-Ukrainian parallel corpus was thought as a provisional solution for a missing comprehensive bilingual dictionary for the two languages.

In November 2004 collecting of texts was started. In April 2005 the corpus project was outlined and in September 2005 its pilot version became available in the Internet. That version contained only 50 small texts (25 pairs), mostly articles that were received from their translators. The texts were aligned at the level of paragraphs, they also contained the basic metainformation: title, author, translator, language of the text. The structural mark-up included only paragraph tags.

Starting from October 2007 the project received a two-year financial support from the Ministry of Science and Higher Education of Poland, which changed its unofficial status and opened new perspectives of development. It was decided to invest primarily into the development of tools and reconsidering the quality of the corpus content and its search possibilities, preparing a methodological and instrumental background for do-it-yourself parallel corpora.

The benefits of the grant period included the following: purchasing modern literature (some of the books, especially the newest ones, are practically unavailable in libraries, besides, the presence of the paper versions is more in the accordance with the copyright law); scanning, digitalizing and editing the texts; developing and bringing to life a common morphosyntactic conceptual scheme; creating user-friendly tools for developing parallel corpora.

As a result, the corpus grew in size from ca. 600 thousand to ca. 3 mln wordforms. The alignment level moved from paragraphs to sentences, all being manually checked. The texts received lemmatization and morphosyntactic information in a unified format; the format of the corpus is a TEI-compatible XML. A standalone search-engine POSHUK was created for searching through the corpus both at the level of a pure text and its meta and morphosyntactic mark-up.

Since February 2011 the new version of the corpus is available for searching online at http://www.domeczek.pl/~polukr/index.php?option=search.

Description of the corpus

Due to the above-mentioned reasons PolUKR is rather quality than size oriented. The corpus contains pairs of texts in Polish and Ukrainian where one of the texts is necessarily original and the other is its translation. All of the texts were written in the XXth or early XXIst century and belong to various genres: fiction, magazine articles, manuals, documents etc. The current size of the corpus is over 3 mln tokens and it is not balanced regarding the genre and original/translation distribution, with a strong prevalence of fiction and more Ukrainian originals than Polish ones.

Genre distribution

- magazine articles: 9 %
- fiction: 85 %
- educational literature, textbooks: 5 %
- official correspondence: 0.3 %
- technical documentation: 0.7 %

There are several levels of mark-up in the corpus: external (meta-information), structural (chapter, paragraph and sentence delimiters) and morphosyntactic (a detailed grammatical characteristics of the word forms).

Meta information includes the title, author or translator if it is a translated text, information whether the text is original or translated, language of the text and language of the original, place and year of publishing, genre, the source of the text, its quality and availability. The texts are lemmatized, i.e. every word form is accompanied by its dictionary form, and the grammatical annotation is presented in the popular international MULTEXT-East format[2]. The next section covers the issue of morphosyntactic description in more detail.

A common morphosyntactic conceptual scheme

The original grammatical information for the Polish language comes from Morfeusz analyzer and TaKIPI tagged that uses its data. The Ukrainian Grammatical Dictionary was used as the original data source and the data were later transformed into the target format and tagged with the help of UGTag morphological analyzer. In both language cases the original data were considerably modified and extended to fit MULTEXT-East format. Both Polish and Ukrainian morphosyntactic tagsets used in the corpus count over 1200 unique tags that are conceptually comparable and are recorded in a common format. Both the Ukrainian and Polish tagsets, together with sample lexicons[3] are part of MULTEXT-East v.4 (released in May 2010). The Polish MULTEXT-East package also includes a one book corpus of George Orwell's "1984". Below is a fragment of a table from [Erjavec 2011] showing counts on MULTEXT-East lexic.

Table 1.

Language	Entries	Words	Lemmas	MSDs
Polish	337,605	174,444	13,601	1213
Ukrainian	318,547	205,348	15,162	1239

Ukrainian morphosyntactic data

Ukrainian Grammatical Dictionary[4] contains ca. 185K lemmas and is organized as a relational database with multiple tables presenting very fine-grained paradigmatic classes. The amount of paradigmatic classes of various frequency of use (from just one

[2] The project site address is http://nl.ijs.si/ME/V4.

[3] Acquiring the lexicon for Polish is due to the courtesy of Marcin Woliński and Zygmunt Saloni.

[4] Acknowledgements go to Igor V. Shevchenko and Volodymyr Shyrokov.

to thousands of lemmas) counts over 1200 unique types. The parameters that make those types include both grammatically related (gender, number, complexity, animacy) and unrelated (stem length, vowel and consonant alternations, etc.) information. It is the basis of a 384 unit tagset that is used by the Ukrainian National Corpus. These grammatical combinations were extended to a MULTEXT-East format with 1239 tags. Many categories were presented in a more detailed way, especially semantic classification of pronouns. Some new classes were created as well, e.g. the active and passive adjectival participle whose detailed classification made over 400 new tags. More details are described in [Kotsyba et al. 2011].

A fragment of morphosyntactic specifications for the Ukrainian adjective follows below, and the full version of the MTE specifications for Ukrainian can be consulted at http://nl.ijs.si/ME/V4/msd/html/msd-uk.html.

MSD (en)	**Features (en)**	**Features (uk)**	**Tokens**	**Examples of usage with lemmata**
Afc-pafn	Adjective Adjective Type=Qualificative Degree=Comparative Number=Plural Case=Accusative Definiteness=Full-Art Animacy=No	Прикметник тип=якісний ступінь=вищий число=множина відмінок=знахідний форма=нестягнена істота=ні	554	азартнішії/азартніший активнішії/активніший актуальнішії/актуальніший амбіціознішії/амбіціозніший ароматнішії/ароматніший багатозначнішії/багатозначніший бадьорішії/бадьоріший боліснішії/болісніший
Afc-pafy	Adjective Adjective Type=Qualificative Degree=Comparative Number=Plural Case=Accusative Definiteness=Full-Art Animacy=Yes	Прикметник тип=якісний ступінь=вищий число=множина відмінок=знахідний форма=нестягнена істота=так	489	азартніших/азартніший активніших/активніший актуальніших/актуальніший амбіціозніших/амбіціозніший ароматніших/ароматніший багатозначніших/багатозначніший бадьоріших/бадьоріший бажаніших/бажаніший балакливіших/балакливіший благородніших/благородніший
Afc-pasn	Adjective Adjective Type=Qualificative Degree=Comparative Number=Plural Case=Accusative Definiteness=Short-Art Animacy=No	Прикметник тип=якісний ступінь=вищий число=множина відмінок=знахідний форма=стягнена істота=ні	554	азартніші/азартніший активніші/активніший актуальніші/актуальніший амбіціозніші/амбіціозніший ароматніші/ароматніший багатозначніші/багатозначніший бадьоріші/бадьоріший бажаніші/бажаніший балакливіші/балакливіший благонадійніші/благонадійніший благородніші/благородніший

Example of an undisambiguated tagged text

```
<?xml version="1.0"?>
<text id="aisecuk." lang="uk">
<body>
<p id="aisecuk.1">
<s id="aisecuk.1.1">
<w lemma="AIESEC" ana="X">AIESEC</w>
<c>(</c>
<w lemma="міжнародний" ana="Ao-fsns">міжнародна</w>
<w lemma="асоціація" ana="Ncfsnn">асоціація</w>
<w_>
<w lemma="студент" ana="Ncmpay">студентів</w>
<w lemma="студент" ana="Ncmpgy">студентів</w>
</w_>
<c>,</c>
<w_>
<w lemma="який" ana="Pq----pna">які</w>
<w lemma="який" ana="Pq---npaa">які</w>
<w lemma="який" ana="Pr----pna">які</w>
<w lemma="який" ana="Pr---npaa">які</w>
</w_>
<w_>
<w lemma="займатися" ana="Vmpip3p">займаються</w>
<w lemma="займатися" ana="Vmpip3p">займаються</w>
</w_>
<w lemma="економія" ana="Ncfsin">економією</w>
<w_>
...
<c>)</c>
```

Polish morphosyntactic data

Over 1000 empirically derived IPIPAN tag configurations were reconceptualized into 1213 tags of the MULTEXT-East mode. The new tagset was designed to cover all common categories in MULTEXT-East tagset and at the same time to keep as much information as possible from the IPIPAN mark-up[5]. Only half of the original tags found one-to-one projections, the rest are either results of contractions or of more detailed representations, a very detailed description can be found in [Kotsyba et al. 2009].

A fragment of morphosyntactic specifications for the Polish verb is below, while the full version of the MTE specifications for Polish can be consulted at http://nl.ijs.si/ME/V4/msd/html/msd-pl.html.

[5] The only discarded category from the original flexemic tagset for Polish was the vocalicity of prepositions (“w/we”).

Tag	Description of the tag	Tokens	Examples of use with lemma ta
Vmpis-pmn	Verb Type=main Aspect=progressive VForm=indicative Tense=past Number=plural Gender=masculine Human=no	606	stanowiły/stanowić, bywały/bywać, zdarzały/zdarzać
Vmpis-pmy	Verb Type=main Aspect=progressive VForm=indicative Tense=past Number=plural Gender=masculine Human=yes	1719	byli/być, mieli/mieć, mogli/móc
Vmpis-pf	Verb Type=main Aspect=progressive VForm=indicative Tense=past Number=plural Gender=feminine	706	musiały/musieć, trwały/trwać, dotyczyły/dotyczyć
Vmpis-pn	Verb Type=main Aspect=progressive VForm=indicative Tense=past Number=plural Gender=neuter	373	były/być, miały/mieć, mogły/móc
Vmpis1sm—y	Verb Type=main Aspect=progressive VForm=indicative Tense=past Person=first Number=singular Gender=masculine Clitic=yes	1302	miałem/mieć, byłem/być, mogłem/móc
Vmpis1sf--y	Verb Type=main Aspect=progressive VForm=indicative Tense=past Person=first Number=singular Gender=feminine Clitic=yes	676	chciałam/chcieć, byłam/być, myślałam/myśleć
Vmpis1sn--y	Verb Type=main Aspect=progressive VForm=indicative Tense=past Person=first Number=singular Gender=neuter Clitic=yes	2	wyłom/wyć, działom/dziać, szłom/iść
Vmpis1pmn-y	Verb Type=main Aspect=progressive VForm=indicative Tense=past Person=first Number=plural Gender=masculine Human=no Clitic=yes	3	wchodziłyśmy/wchodzić, widziałyśmy/widzieć, chodziłyśmy/chodzić

The example above shows one of the considerable discrepancies with the IPIPAN format, i.e. agglutinated verb endings, when they are written together with their stems, are not treated as separate words, they only add information about the person and number. Their presence is indicated by the feature Clitic with the value “yes”.

Example of a disambiguated tagged Polish text in the MTE format.

```
<text id="aisecpl" lang="pl">
<body>
<p id="aisecpl.1">
<s id="aisecpl.1.1">
<w lemma="aiesec" ana="N-mnnsn">AIESEC</w>
<c>(</c>
<w lemma="międzynarodowy" ana="A-pn--sn">międzynarodowe</w>
<w lemma="stowarzyszenie" ana="Ngnn-snen">stowarzyszenie</w>
<w lemma="student" ana="N-myypg">studentów</w>
<w lemma="ekonomia" ana="N-f--sg">ekonomii</w>
<w lemma="i" ana="C">i</w>
<w lemma="zarządzanie" ana="N-n--sg">zarządzania</w>
<c>)</c>
<w lemma="to" ana="Pd--nnnsn--n">to</w>
<w lemma="niepolityczny" ana="A-pf--sn">niepolityczna</w>
<c>,</c>
<w lemma="naukowy" ana="A-pf--sn">naukowa</w>
<w lemma="i" ana="C">i</w>
<w lemma="niekomercyjny" ana="A-pf--sn">niekomercyjna</w>
<w lemma="organizacja" ana="N-f--sn">organizacja</w>
<c>,</c>
<w lemma="w" ana="Spl">w</w>
<w lemma="całość" ana="N-f--sl">całości</w>
<w lemma="zarządzać" ana="Ap-f--sn-ppn">zarządzana</w>
<w lemma="przez" ana="Spa">przez</w>
<w lemma="student" ana="N-myypa">studentów</w>
<c>.</c>
</s>
</p>
```

The corpus creation pipeline

Creating a corpus is a complex task involving numerous minor modules. At the moment all the subtasks that were used for the current PolUKR version are exist as separate programs or just scripts, and certain text manipulation is needed to achieve cooperation between the texts and the programs. The programming languages used are: C, C#, Python, Perl, PHP, Java, XSLT. Ideally, all the tasks should be handled under a common umbrella, a universal flexible corpus manager with integrated modules like: a sentence splitter, a grammatical dictionary and its editor, aligner with the possibility of editing the output, indexers, converters, search engine, but this is quite a complicated project and demands more time and resources.

The pipeline that was used to create the parallel corpus for Polish and Ukrainian includes:

1) preparing the texts in a plain text electronic format, free of linguistic errors and formatting flaws, UTF-8 encoded;

2) recording relevant metainformation for each text into the database;

3) sentence splitting[6] and sentence segmentation correction;

4) manual edition of a higher structural annotation (chapters and their heads);

5) creating and editing alignments with the help of PLUczeK;

6) tagging Polish texts with the help of TaKIPI;

7) tagging Ukrainian texts with the help of UGTag;

8) disambiguation of Ukrainian texts for selected forms of further processing;

9) adjusting TaKIPI format to fit further processing;

10) conversion of the texts into a binary format with the help of the corpus manager.

Some issues connected with the pre-processing of texts are also described in [Kotsyba 2009][7]. More language versions will be present at the PolUKR project site soon.

Language processing tools developed within the project

UGTag, a tagger for Ukrainian language

UGTag[8] is a program for enriching Ukrainian texts with morphosyntactic annotation. A console version together with a dictionary containing over 15 thousand lemmas and 205,348 wordforms with their morphosyntactic descriptions in MULTEXT-East format is available for download at the project's sourceforge site at http://sourceforge.net/projects/ugtag/ since February 2011. The program allows using several input and two output XML formats and an unlimited number of morphosyntactic dictionaries.

[6] A rule-based sentence-splitter written in Python by Oresta Tymchyshyn was used for the PolUKR project.

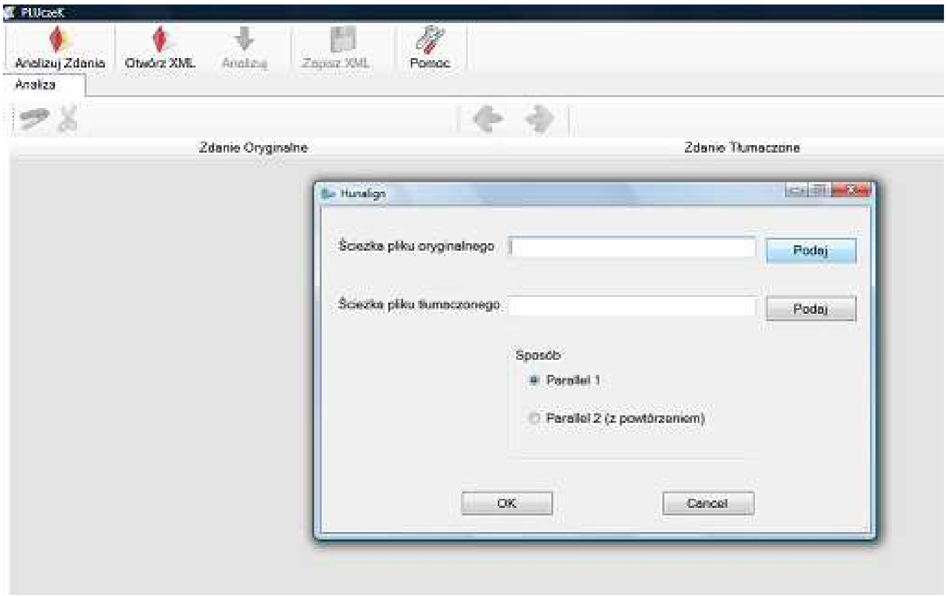
[7] This paper, as well as many other project-related publications, is available at the author's website: http://www.domeczek.pl/~natko/.

[8] A more detailed description is available at the program's site at http://www.domeczek.pl/~polukr/parcor/index.html. The code of the program was written by Andriy Mykulyak.

PLUczeK, a graphic editor and converter for alignments

PLUczeK[9] is a user-friendly alignment editor and converter with GUI used for building parallel corpora, written in C#, based on .NET platform. It works with an independent program Hunalign[10] to pre-align an original and translated texts and allows the user to modify the alignment and record it in the XML format. PLUczeK runs under Windows XP, Vista and Windows 7. It was created for working with the Polish-Ukrainian parallel corpus but can be used for any pair of languages with UTF-8 encoding. The program takes raw UTF-8 encoded texts that are split into sentences and, optionally, paragraphs as input, and produces three XML files as output. Apart from the two texts transformed into XML there is also a stand-off XML alignment file in the XCES format. The program is available for downloading at its sourceforge site at http://sourceforge.net/projects/pluczek/.

PLUczeK: pre-alignment with HunAlign



Ścieżka pliku oryginalnego

Ścieżka pliku tłumaczonego

Sposób

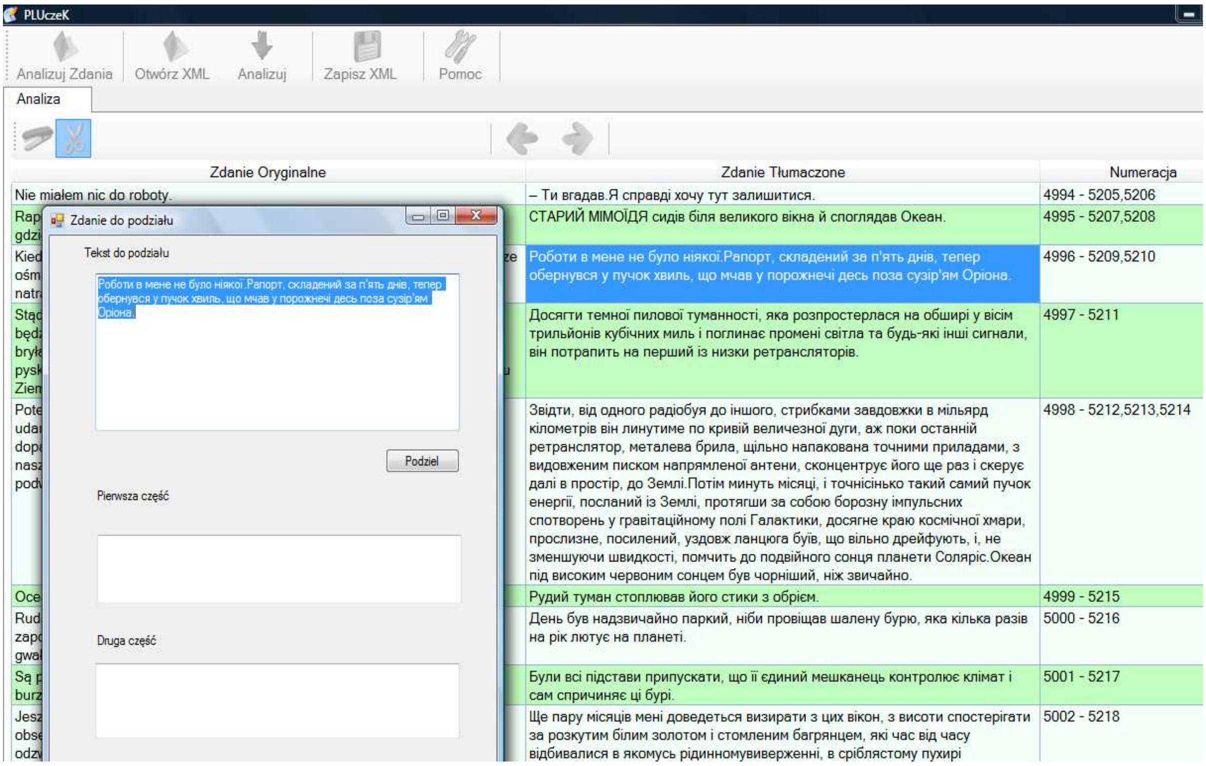
● Parallel 1

○ Parallel 2 (z powtórzeniem)

PLUczeK: editing prealigned texts

[9] A more detailed description as well as user's guidelines are available at the program's site at http://www.domeczek.pl/~polukr/parcor/pluczek_en.html. The code of the program was written by Myroslava Tasinkevych.

[10] The program is available at http://mokk.bme.hu/resources/hunalign.



Роботи в мене не було ніякої.Рапорт, складений за п'ять днів, тепер обернувся у пучок хвиль, що мчав у порожнечі десь поза сузір'ям Оріона.

Zdanie Oryginalne	Zdanie Tłumaczone	Numeracja
Nie miałem nic do roboty.	– Ти вгадав.Я справді хочу тут залишитися.	4994 - 5205,5206
Rap gdzi	СТАРИЙ МІМОЇДЯ сидів біля великого вікна й споглядав Океан.	4995 - 5207,5208
Kied ośm natr ze	Роботи в мене не було ніякої.Рапорт, складений за п'ять днів, тепер обернувся у пучок хвиль, що мчав у порожнечі десь поза сузір'ям Оріона.	4996 - 5209,5210
Stąd będ bryła pysk Zien u	Досягти темної пилової туманності, яка розпростерлася на обширі у вісім трильйонів кубічних миль і поглинає промені світла та будь-які інші сигнали, він потрапить на перший із низки ретрансляторів.	4997 - 5211
Pote udar dop nasz podv	Звідти, від одного радіобуя до іншого, стрибками завдовжки в мільярд кілометрів він линутиме по кривій величезної дуги, аж поки останній ретранслятор, металева брила, щільно напакована точними приладами, з видовженим писком напрямленої антени, сконцентрує його ще раз і скерує далі в простір, до Землі.Потім минуть місяці, і точнісінько такий самий пучок енергії, посланий із Землі, протягши за собою борозну імпульсних спотворень у гравітаційному полі Галактики, досягне краю космічної хмари, прослизне, посилений, уздовж ланцюга буїв, що вільно дрейфують, і, не зменшуючи швидкості, помчить до подвійного сонця планети Соляріс.Океан під високим червоним сонцем був чорніший, ніж звичайно.	4998 - 5212,5213,5214
Oce	Рудий туман стоплював його стики з обрієм.	4999 - 5215
Rud zapo gwał	День був надзвичайно паркий, ніби провіщав шалену бурю, яка кілька разів на рік лютує на планеті.	5000 - 5216
Są p burz	Були всі підстави припускати, що її єдиний мешканець контролює клімат і сам спричиняє ці бурі.	5001 - 5217
Jesz obse odzy	Ще пару місяців мені доведеться визирати з цих вікон, з висоти спостерігати за розкутим білим золотом і стомленим багрянцем, які час від часу відбивалися в якомусь рідинномувиверженні, в сріблястому пухирі	5002 - 5218

Stand-off alignment format:

```
<?xml version="1.0" encoding="utf-8"?>
<cesAlign version="4.1">
    <linkList id="lemsolpluk">
      <linkGrp id="lemsolpl.1" type="body" targType="s" domains="lemsolpl
lemsoluk">
          <link xtargets="lemsolpl.1.1 ; lemsoluk.1.1"/>
          <link xtargets="lemsolpl.1.2 ; lemsoluk.1.2"/>
          <link xtargets="lemsolpl.1.3 ; lemsoluk.1.3"/>
               ...
          <link xtargets="lemsolpl.8.15 ; lemsoluk.9.14"/>
          <link xtargets="lemsolpl.9.1 lemsolpl.9.2 ; lemsoluk.10.1"/>
          <link xtargets="lemsolpl.9.3 ; lemsoluk.10.2"/>
          <link xtargets="lemsolpl.10.1 ; lemsoluk.11.1"/>
          <link xtargets="lemsolpl.10.2 ; lemsoluk.11.2 lemsoluk.11.3"/>
          <link xtargets="lemsolpl.10.3 lemsolpl.10.4 ; lemsoluk.11.4 lemsoluk.11.5"/>
               ...
```

Converters

A number of format converters have been developed to ensure the corpus's flexibility and compatibility with standards and other projects and adjusting to existing searching software. A TAKIPI to MULTEXT-East converter for Polish morphosyntactically annotated texts[11] with a set of conversion tables for tags and lists of selected lemmas was written in Python see [Kotsyba et al. 2009] for a detailed account.

One of the obvious advantages of using XML format in a corpus is the possibility of using XSLT transformations to adjust the format to other tools. Most of the present format flexibility is enabled XSLT transformation sheets, for some types of files Perl scripts are used. Conversion sets were developed for working with Intertext/InterCorp[12], and ParaSol[13] projects.

Searching in the corpus

Local search possibilities

The local distribution of PolUKR comes with a search-engine POSHUK[14] which allows creating complex queries combining information about a word form, lemma and morphosyntax in both language parts of the corpus at the same time.

Another possibility of current local search support is a commercial ParaConc program. Conversion into the ParaConc specific format is enabled by means of Intertext developed within the InterCorp project.

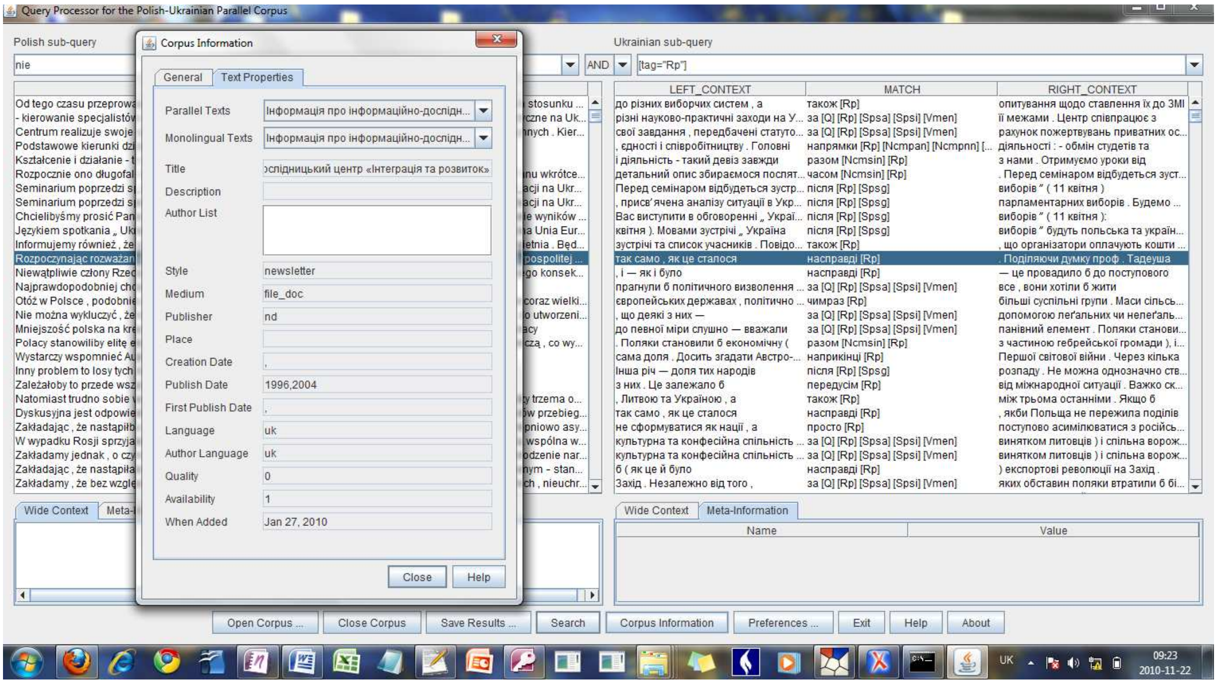
An example of a POSHUK query window is below.

[11] The converter is available at http://www.domeczek.pl/~polukr/mte-conv/. The code of the program was written by Adam Radziszewski.

[12] InterCorp: projekt paralelních korpusů http://www.korpus.cz/intercorp/. I would like to thank Alexander Rosen for his readiness to help and friendly advice.

[13] von Waldenfels, Ruprecht and Meyer, Roland (2004-11): ParaSol - a Parallel Corpus of Slavic and Other Languages. Corpus Resource, University of Bern and Regensburg. Available at: parasol.unibe.ch, www-korpus.uni-r.de/ParaSol. The ParaSol conversion kit for PolUKR was mostly prepared by Ruprecht von Waldenfels.

[14] POSHUK is an abbreviation from Polish, search and Ukrainian, the whole word means *search* in Ukrainian.



Polish sub-query: nie

AND

Ukrainian sub-query: [tag="Rp"]

Od tego czasu przeprowa
- kierowanie specjalistów
Centrum realizuje swoje
Podstawowe kierunki dzi
Kształcenie i działanie - t
Rozpocznie ono długofal
Seminarium poprzedzi s
Seminarium poprzedzi s
Chcielibyśmy prosić Pan
Językiem spotkania „ Uk
Informujemy również , że
Rozpoczynając rozważan
Niewątpliwie człony Rzec
Najprawdopodobniej chc
Otóż w Polsce , podobnie
Nie można wykluczyć , że
Mniejszość polska na kre
Polacy stanowiliby elitę e
Wystarczy wspomnieć Au
Inny problem to losy tych
Zależałoby to przede wsz
Natomiast trudno sobie
Dyskusyjna jest odpowie
Zakładając , że nastąpiłb
W wypadku Rosji sprzyja
Zakładamy jednak , o czy
Zakładając , że nastąpiła
Zakładamy , że bez wzglę

Corpus Information

General | Text Properties

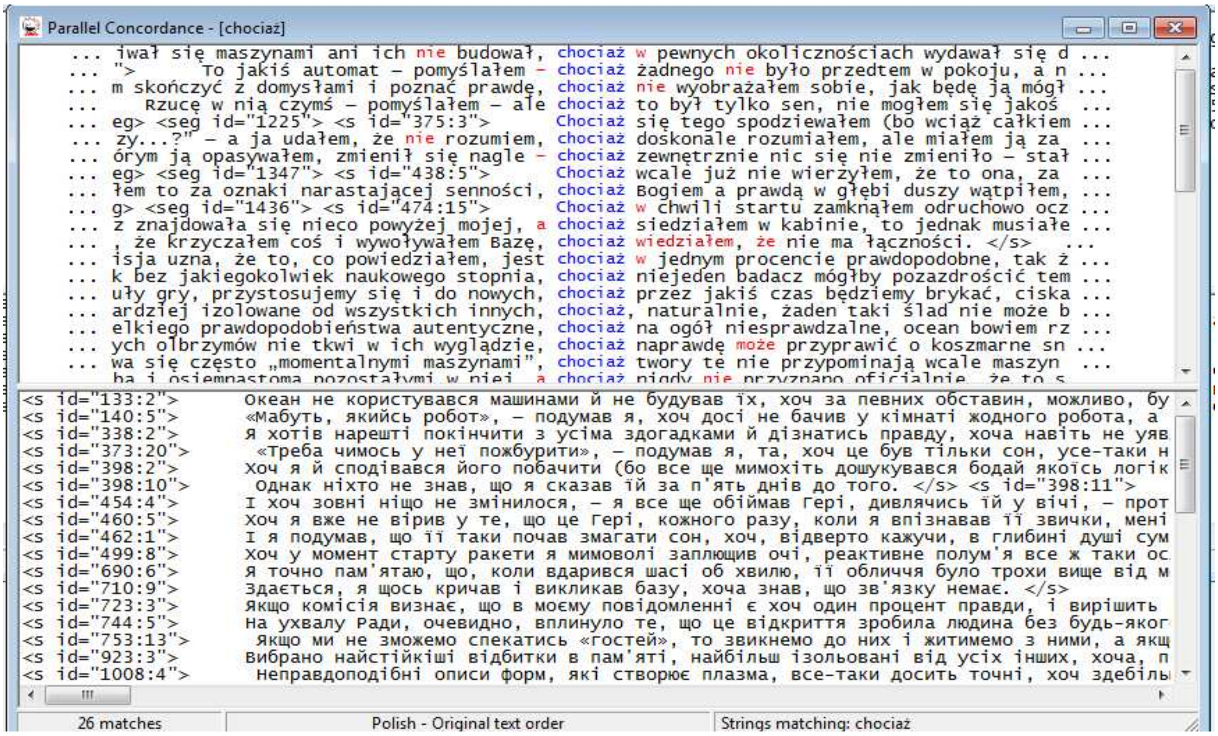
Parallel Texts	Інформація про інформаційно-дослідн...
Monolingual Texts	Інформація про інформаційно-дослідн...
Title	ослідницький центр «Інтеграція та розвиток»
Description	
Author List	
Style	newsletter
Medium	file_doc
Publisher	nd
Place	
Creation Date	,
Publish Date	1996,2004
First Publish Date	,
Language	uk
Author Language	uk
Quality	0
Availability	1
When Added	Jan 27, 2010

LEFT_CONTEXT	MATCH	RIGHT_CONTEXT
до різних виборчих систем , а	також [Rp]	опитування щодо ставлення їх до ЗМІ
різні науково-практичні заходи на У...	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	її межами . Центр співпрацює з
свої завдання , передбачені статуто...	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	рахунок пожертвувань приватних ос...
, єдності і співробітництву . Головні	напрямки [Rp] [Ncmpan] [Ncmpnn] [...	діяльності : - обмін студетів та
і діяльність - такий девіз завжди	разом [Ncmsin] [Rp]	з нами . Отримуємо уроки від
детальний опис збираємося послят...	часом [Ncmsin] [Rp]	. Перед семінаром відбудеться зуст...
Перед семінаром відбудеться зустр...	після [Rp] [Spsg]	виборів ” (11 квітня)
, присв'ячена аналізу ситуації в Укр...	після [Rp] [Spsg]	парламентарних виборів . Будемо ...
Вас виступити в обговоренні „ Украї...	після [Rp] [Spsg]	виборів ” (11 квітня):
квітня). Мовами зустрічі „ Україна	після [Rp] [Spsg]	виборів ” будуть польська та україн...
зустрічі та список учасників . Повідо...	також [Rp]	, що організатори оплачують кошти ...
так само , як це сталося	насправді [Rp]	. Поділяючи думку проф . Тадеуша
, і — як і було	насправді [Rp]	— це провадило б до поступового
прагнули б політичного визволення ...	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	все , вони хотіли б жити
європейських державах , політично ...	чимраз [Rp]	більші суспільні групи . Маси сільсь...
, що деякі з них —	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	допомогою леґальних чи нелеґаль...
до певної міри слушно — вважали	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	панівний елемент . Поляки станови...
. Поляки становили б економічну (	разом [Ncmsin] [Rp]	з частиною гебрейської громади), і...
сама доля . Досить згадати Австро-...	наприкінці [Rp]	Першої світової війни . Через кілька
Інша річ — доля тих народів	після [Rp] [Spsg]	розпаду . Не можна однозначно ств...
з них . Це залежало б	передусім [Rp]	від міжнародної ситуації . Важко ск...
, Литвою та Україною , а	також [Rp]	між трьома останніми . Якщо б
так само , як це сталося	насправді [Rp]	, якби Польща не пережила поділів
не сформуватися як нації , а	просто [Rp]	поступово асимілюватися з російсь...
культурна та конфесійна спільність ...	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	винятком литовців) і спільна ворож...
культурна та конфесійна спільність ...	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	винятком литовців) і спільна ворож...
б (як це й було	насправді [Rp]	) експортові революції на Захід .
Захід . Незалежно від того ,	за [Q] [Rp] [Spsa] [Spsi] [Vmen]	яких обставин поляки втратили б бі...

Wide Context | Meta-Information

Name	Value

Example of using ParaConc for search:



```
... iwał się maszynami ani ich nie budował, chociaż w pewnych okolicznościach wydawał się d ...
... ">       To jakiś automat – pomyślałem – chociaż żadnego nie było przedtem w pokoju, a n ...
... m skończyć z domysłami i poznać prawdę, chociaż nie wyobrażałem sobie, jak będę ją mógł ...
...    Rzucę w nią czymś – pomyślałem – ale chociaż to był tylko sen, nie mogłem się jakoś  ...
... eg> <seg id="1225"> <s id="375:3">        Chociaż się tego spodziewałem (bo wciąż całkiem ...
... zy...?" – a ja udałem, że nie rozumiem, chociaż doskonale rozumiałem, ale miałem ją za  ...
... órym ją opasywałem, zmienił się nagle – chociaż zewnętrznie nic się nie zmieniło – stał ...
... eg> <seg id="1347"> <s id="438:5">        Chociaż wcale już nie wierzyłem, że to ona, za  ...
... łem to za oznaki narastającej senności, chociaż Bogiem a prawdą w głębi duszy wątpiłem, ...
... g> <seg id="1436"> <s id="474:15">       Chociaż w chwili startu zamknąłem odruchowo ocz ...
... z znajdowała się nieco powyżej mojej, a chociaż siedziałem w kabinie, to jednak musiałe ...
... , że krzyczałem coś i wywoływałem Bazę, chociaż wiedziałem, że nie ma łączności. </s>   ...
... isja uzna, że to, co powiedziałem, jest chociaż w jednym procencie prawdopodobne, tak ż ...
... k bez jakiegokolwiek naukowego stopnia, chociaż niejeden badacz mógłby pozazdrościć tem ...
... uły gry, przystosujemy się i do nowych, chociaż przez jakiś czas będziemy brykać, ciska ...
... ardziej izolowane od wszystkich innych, chociaż, naturalnie, żaden taki ślad nie może b ...
... elkiego prawdopodobieństwa autentyczne, chociaż na ogół niesprawdzalne, ocean bowiem rz ...
... ych olbrzymów nie tkwi w ich wyglądzie, chociaż naprawdę może przyprawić o koszmarne sn ...
... wa się często „momentalnymi maszynami", chociaż twory te nie przypominają wcale maszyn  ...
... ba i osiemnastoma pozostałymi w niej, a chociaż nigdy nie przyznano oficjalnie, że to s
```

```
<s id="133:2">      Океан не користувався машинами й не будував їх, хоч за певних обставин, можливо, бу
<s id="140:5">      «Мабуть, якийсь робот», – подумав я, хоч досі не бачив у кімнаті жодного робота, а
<s id="338:2">      Я хотів нарешті покінчити з усіма здогадками й дізнатись правду, хоча навіть не уяв
<s id="373:20">      «Треба чимось у неї пожбурити», – подумав я, та, хоч це був тільки сон, усе-таки н
<s id="398:2">      Хоч я й сподівався його побачити (бо все ще мимохіть дошукувався бодай якоїсь логік
<s id="398:10">      Однак ніхто не знав, що я сказав їй за п'ять днів до того. </s> <s id="398:11">
<s id="454:4">      І хоч зовні ніщо не змінилося, – я все ще обіймав Гері, дивлячись їй у вічі, – прот
<s id="460:5">      Хоч я вже не вірив у те, що це Гері, кожного разу, коли я впізнавав її звички, мені
<s id="462:1">      І я подумав, що її таки почав змагати сон, хоч, відверто кажучи, в глибині душі сум
<s id="499:8">      Хоч у момент старту ракети я мимоволі заплющив очі, реактивне полум'я все ж таки ос
<s id="690:6">      Я точно пам'ятаю, що, коли вдарився шасі об хвилю, її обличчя було трохи вище від м
<s id="710:9">      Здається, я щось кричав і викликав базу, хоча знав, що зв'язку немає. </s>
<s id="723:3">      Якщо комісія визнає, що в моєму повідомленні є хоч один процент правди, і вирішить
<s id="744:5">      На ухвалу Ради, очевидно, вплинуло те, що це відкриття зробила людина без будь-яког
<s id="753:13">      Якщо ми не зможемо спекатись «гостей», то звикнемо до них і житимемо з ними, а якщ
<s id="923:3">      Вибрано найстійкіші відбитки в пам'яті, найбільш ізольовані від усіх інших, хоча, п
<s id="1008:4">      Неправдоподібні описи форм, які створює плазма, все-таки досить точні, хоч здебіль
```

26 matches | Polish - Original text order | Strings matching: chociaż

Web search possibilities

The possibility of searching through the new version of the corpus online is enabled at the moment by means of the Corpus Workbench[15] back-end and the Perl interface for a similar ParaSol project. More options are developed at the moment,

[15] The program site address is http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

including cooperation with the InterCorp, engaging the PARK/Manatee platform, as well as extending POSHUK to be run as a webservice.

Current and future work

At the moment more effort is focused on building on the corpus flexibility and compatibility with standards and other projects, adjusting to existing searching software, so the appearance of more converters is to be expected.

Another module that is currently under development is a rule-based morphosyntactic disambiguation for Ukrainian done within the Constraint Grammar framework. Using this particular framework allows a meaningful and syntax aware handling of disambiguation and opens the door to higher mark-up levels: syntactic and semantic ones.

For most linguistic tasks, like terminology extraction or grammar patterns tracking, versatility of text genres becomes of greater importance. Expanding the size and variability of the corpus, enhancing the quality of the mark-up are also important and it is hoped that a larger community will be involved into these tasks. Currently some of the developed tools are used in at least three individual PhD projects. Raising corpus awareness among traditional philologists, lexicographers, translators, language teachers and learners is crucial for the project development and, more importantly, the use of the corpus by a wide public. This demands preparing various easy-to-follow tutorials and courses, including on-line ones, in the languages of interest. Such resources are currently being developed at the Institute of Interdisciplinary Studies of Warsaw University.

References

Broda B., Piasecki M. and Radziszewski A. 2008: Towards a Set of General Purpose Morphosyntactic Tools for Polish, in “Proceedings of Intelligent Information Systems, Zakopane Poland, 2008”, Institute of Computer Science PAS.

Derzhanski I. Kotsyba N. 2008: The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging, in “Lexicographic Tools and Techniques”, Proceedings of MONDILEX First Open Workshop, 3-4 October 2008, Moscow, p. 68-79.

Derzhanski I., Kotsyba N. 2009: Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian, in the Proceedings of “Metalanguage and Encoding Scheme Design for Digital Lexicography: MONDILEX Third Open Workshop”, 15–16 April 2009, Bratislava.

Erjavec T. 2010: MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora, in “Proceedings of the LREC 2010, Malta, 19-21 May, 2010”.

Kotsyba N., Turska M. 2007: Polish-Ukrainian Parallel Corpus and its Possible Applications, in the Proceedings of the international conference "Practical Applications in Language and Computers", Łódź, 7-9 April 2005. Peter Lang GmbH.

Kotsyba N., Shypnivska O., Turska M. 2008: Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus), in the Proceedings of the international conference "Intelligent Information Systems”, Zakopane, 16-18 June 2008, Warsaw.

Kotsyba N. 2009: The Current State of Work on the Polish-Ukrainian Parallel Corpus (PolUKR), in "Problems of Slavic Lexicography", Proceedings of the international workshop within MONDILEX project, 2-4 February 2009, Kyiv.

Kotsyba N., Mykulyak A., Shevchenko I.V. 2011: UGTag: morphological analyzer and tagger for Ukrainian language, in "Explorations across Languages and Corpora", in the series “Łódź Studies in Language”, (ed. by Stanisław Goźdź-Roszkowski).

Kotsyba N., Radziszewski A., Derzhanski I. 2009: Integrating the Polish language into the MULTEXT-East family: morphosyntactic specifications, converter, lexicon and corpus, in Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop, October 14, 2009, Ljubljana, Slovenia. Ljubljana.

Przepiórkowski A. and Woliński M. 2003: A Flexemic Tagset for Polish, in “The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003”.

Turska M., Kotsyba N. 2006: Polsko-Ukraiński korpus równoległy (PolUKR), in "Materiały LXIII Zjazdu Polskiego Towarzystwa Językoznawczego", Warsaw.

Saloni Z., Gruszczyński W., Woliński M., Wołosz R. 2010: Analizator morfologiczny *Morfeusz*, http://sgjp.pl/morfeusz/.

v. Waldenfels, R. 2006. 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.

Коциба Н. 2009: Морфосинтаксичне тагування польсько-українського паралельного корпусу (PolUKR), in „Proceedings of the International Conference MegaLing'2008. Horizons of Applied Linguistics and Linguistic Technologies, Parthenit, Ukraine, 20-27 September 2008”, Kyiv.

Шевченко И.В., Широков В.А., Рабулець А.Г., 2005: Электронный грамматический словарь украинского языка, in “Труды международной конференции Megaling’2005. Прикладная лингвистика в поиске новых путей. 27 июня–2 июля 2005 года. Меганом, Крым, Украина”, p. 124–129.