Natalia Kotsyba
Polska Akademia Nauk

Polsko-Ukraiński Korpus Równoległy PolUKR i jego następca PolUKR-2

Polish-Ukrainian Parallel Corpus PolUKR and its successor PolUKR-2

Abstract

The paper discusses the present stage of development of one of the aspects of an ongoing project aiming at creating electronic resources for the Ukrainian language. Parallel corpora make an important part of this project. The *Polish-Ukrainian Parallel Corpus* (*PolUKR*) was developed in 2004-2010, first in the Institute of Slavic Studies of the Polish Academy of Sciences, later at the faculty "Artes Liberales" of the University of Warsaw. The first two versions of *PolUKR* are available for search online at http://domeczek.pl/~polukr.

PolUKR consists of texts written originally either in Polish or Ukrainian, i.e., it does not contain any texts translated from a third language, but only immediate translations of its own texts. It had been aligned at the level of sentences automatically, afterwards the alignments were edited manually. Both the Polish and Ukrainian sentences had been supplied with the morphosyntactic layer of annotation. The characteristic feature of *PolUKR* is its purpose-built morphosyntactic categorical apparatus, common for the two corpus languages, and its morphosyntactic tagsets based on it. The tagsets are also used in the multilingual European project MULTEXT-East (1996-2010), version 4 "MONDILEX", available at http://nl.ijs.si/ME/V4/.

While the pilot versions of *PolUKR* concentrated rather on developing corpus-making technologies, in both their technical and theoretical linguistic aspects, the new version, presently developed in cooperation with the National University of Lviv and Lviv Polytechnical University in Ukraine, aims at: 1) first of all, extending the size of the corpus up to 30 million words (as previously, with the biggest possible attention to original Polish or Ukrainian texts, but without a strict limitation on this feature); 2) optimalization of the morphosyntactic description for the Ukrainian language, i.e., disambiguation of ambiguous interpretations and extension of the grammatical dictionary for new, unknown words. Work on the shallow syntax for

Ukrainian is also planned. *PolUKR-2* will be used as a basic corpus resource for creating a great Ukrainian-Polish dictionary with ca. 80 thousand entries.

Słowa kluczowe: korpus równoległy, język polski, język ukraiński, tagset morfoskładniowy, MULTEXT-East, PolUKR
Keywords: parallel corpus, Polish, Ukrainian, morphosyntactic tagset, MULTEXT-East, PolUKR

1. Wprowadzenie

Polsko-Ukraiński korpus równoległy (PolUKR) był rozwijany[1] w latach 2004-2011 w Instytucie Slawistyki PAN oraz na wydziale „Artes Liberales" Uniwersytetu Warszawskiego najpierw jako projekt eksperymentalny. Dwuletni grant NCN 2007-2009 oraz roczny udział jego autorów w projekcie europejskim MONDILEX (2008-2009) umożliwiły istotny postęp projektu pod względem jakości. Wersja pilotażowa oraz fragment pierwszej wersji PolUKRu są dostępne do przeszukiwania pod adresem: http://domeczek.pl/~polukr.

PolUKR był wzorowany na korpusie IPIPAN (Przepiórkowski, 2004), co oznaczało docelowo podobną strukturę, format i poziom anotacji oraz dostępność do przeszukiwania w Internecie. Podobnie do tekstów korpusu IPIPAN teksty w PolUKR-ze są podzielone na zdania, zaopatrzone w informacje morfoskładniowe i zapisane w formacie XML (XCES) zgodnym z TEI. PolUKR zawiera tylko autentyczne teksty napisane w języku polskim albo ukraińskim i ich bezpośrednie tłumaczenia. Żeby zapewnić możliwie wysoką jakość materiału korpusowego, wyrównania tekstów były dokonywane na poziomie zdań, przy czym wyniki wyrównań automatycznych zostały sprawdzone ręcznie. W celu umożliwienia wiarygodnych badań w zakresie gramatyki konfrontatywnej opracowano i zastosowano w korpusie wspólny tagset morfoskładniowy ze spójną anotacją dla obu języków. Wśród zadań, wykraczających poza możliwości czasowe i finansowe dotychczasowej realizacji projektu znalazło się ujednoznacznienie morfoskładniowe dla języka ukraińskiego oraz dopracowanie wersji internetowej wyszukiwarki POSHUK. Brak dostępnych zasobów językowych (głównie dla języka ukraińskiego) oraz odpowiedniego oprogramowania do opracowania tekstów korpusu na czas opracowania pierwszej wersji korpusu spowodował skierowanie istotnej części wysiłków autorów i dostępnych zasobów finansowych na ich uzupełnienie. Z tym wiąże się względnie mała objętość wersji korpusu dostępnej do przeszukiwania: do ostatecznego stadium opracowania doszło ok. 600 tys. słów z zebranych pierwotnie 3 milionów.

1 Historia powstania projektu oraz podstawowe programy, stworzone w trakcie prac nad nim i udostępnione publicznie, zostały opisane w (Kotsyba, 2012).

2. Pozyskiwanie tekstów

Podczas pierwszych prac eksperymentalnych teksty były pozyskiwane bezpośrednio od tłumaczy albo z bibliotek internetowych. Preferowane było pierwsze źródło, ponieważ dostarczało materiały przeważnie bardzo dobrej jakości oraz jednocześnie pozyskiwana była zgoda na publiczne wykorzystanie tekstu za pośrednictwem wyszukiwarki. Biblioteki internetowe dziesięć lat temu oferowały o wiele skromniejsze zasoby niż obecnie nie tylko ilościowo, lecz także pod względem jakości: obecne w nich teksty były pozyskiwane drogą skanowania i zastosowania automatycznego OCR dla źródeł papierowych; bardzo często tak uzyskane pliki nie podlegały dalszej redakcji. Oprogramowanie, z którego korzystano w tamtych czasach (najczęściej to był program FineReader), pozostawiało dużo błędów[2]. Skutkiem tego był znacznie wydłużony czas redakcji tekstów, czasami też odrzucano teksty najgorszej jakości, ponieważ ich poprawianie było mniej opłacalne od ponownego przepisywania.

Oportunistyczne podejście do pozyskania tekstów miało wpływ na ogólną reprezentatywność i zawartość korpusu. Z literatury pięknej znalazły się w nim współczesne teksty postmodernistyczne, które zawierają specyficzne słownictwo (w tekstach ukraińskich są to liczne wtrącenia takie, jak surżyk oraz transliterowane zdania w językach obcych, głównie w języku rosyjskim), nietypową budowę zdań (w utworach współczesnych ukraińskich autorów Oksany Zabużko i Jurija Andruchowycza nie są rzadkością ponadstronicowe zdania). Teksty te nie odzwierciedlają typowego języka, dlatego nie są zbyt praktyczne dla badań ogólnojęzykowych. Nietypowe teksty, jak i teksty z błędami po OCR, czasami stwarzają więcej problemów przy przetwarzaniu przez programy, np. surżyk i liczne neologizmy czy okazjonalizmy nie są opisane w słownikach gramatycznych, długie rozbudowane zdania bywają trudne do podziału, a tym samym także do sparsowania.

Otrzymane wsparcie grantowe, które umożliwiło m.in. zakup papierowych wersji utworów literatury pięknej, odzwierciedlającej język literacki, opłacenie usług skanowania i wyczytywania tekstów po zastosowaniu OCR, pozwoliło na dobór tekstów, który był lepiej nakierowany na przyszłe korpusowe potrzeby badawcze. Największy nacisk jednak, jak już wspomniano wyżej, był położony na tworzenie brakującego oprogramowania oraz opracowanie i wdrożenie zasad analizy lingwistycznej.

2 Np. FineReader 6.0 nie rozpoznawał dużej ukraińskiej litery Ї, która pojawia się na początku np. takich wyrazów o wysokiej frekwencji jak formy zaimków „jej, ją, ich", co skutkowało później błędami przy podziale na zdania; numery stron trzeba było usuwać ręcznie; program też zostawiał łącznik w miejscu podziału słów na sylaby. Większość z tych wad została usunięta w późniejszych wersjach FineReadera.

3. Oprogramowanie korpusowe[3]

3.1 Podział na zdania

Do podziału tekstów na zdania został utworzony program SentSplit, który bazuje na ręcznie opisanych regułach. Jest to edytowalny skrypt w języku Python, który umożliwia dodawanie skrótów używanych z kropką dla obu języków w miarę ich znajdowania[4]. Ze względu na swoją specyfikę regułową SentSplit ma pewne wymagania względem formatowania i zawartości tekstów wejściowych, co stanowi dodatkowy czynnik wspomagający kontrolowanie ich jakości. Jednocześnie wyniki podziału na zdania uzyskane przez program trzeba często poprawiać – jednak w sytuacjach, nieprzewidzianych przez reguły, program podaje komunikat o błędzie. Najczęściej problemy tego rodzaju są spowodowane błędami formatowania albo obecnością niealfanumerycznych znaków, które pozostają po błędach w OCR-ze.

3.2 Wyrównanie

Wyrównanie lub inaczej zrównoleglanie (ang. alignment) tekstów w wersji eksperymentalnej dokonywane było na poziomie akapitów, przy czym program do przeszukiwania tekstów „zakładał", że podział na akapity był identyczny w obu wariantach językowych. Bliskość struktury tłumaczonych i oryginalnych, krótkich publicystycznych tekstów, które weszły do pilotażowej wersji korpusu, praktycznie nie wymagała ingerencji w autorski podział na akapity. Natomiast przy większych tekstach rozbieżności znacząco rosły. Ponadto akapity były jednostkami tekstu, które tworzyły kontekst przy wyszukiwaniu, co nie było specjalnie wygodne przy akapitach większych rozmiarów. W pierwszej wersji PolUKRu wyrównanie zostało dokonane już na poziomie zdań za pomocą ogólnie dostępnego programu Hunalign (Varga et al., 2005). Wyniki działania tego programu zawierały błędy, które były poprawiane ręcznie przez redaktorów. W tym celu został stworzony program do edycji wyrównań PLUczeK[5]. Wszystkie wyrównania tekstów w PolUKR-ze zostały poprawione za pomocą tego edytora. Dodatkowym plusem jego działania było konwertowanie tekstów wyjściowych do standardowego formatu XML (XCES).

3.3 Morfoskładnia

Informacje morfoskładniowe dla języka polskiego zostały wprowadzone do tekstów za pomocą jednej z pierwszych wersji tagera TaKIPI, opracowanego na

3 http://www.domeczek.pl/~polukr/index.php?option=software.

4 SentSplit opiera się na dość uniwersalnych regułach budowy zdań, dlatego może być stosowany dla innych języków, m.in. był pomyślnie sprawdzany także na tekstach angielskich, niemieckich, francuskich, bułgarskich i rosyjskich. Autorką programu jest Oresta Tymczyszyn.

5 Program PLUczeK jest dostępny pod adresem: http://www.domeczek.pl/~polukr/parcor/pluczek.html.

Politechnice Wrocławskiej (Piasecki, 2007). Informacje te następnie były modyfikowane i konwertowane do docelowego formatu za pomocą specjalnie stworzonego konwertera KIPI2MTE[6], zob. (Kotsyba et al., 2009). Anotacja tekstów ukraińskich została dodana za pomocą programu UGTag (Kotsyba et al., 2011), który wykorzystuje dane Ukraińskiego Słownika Gramatycznego autorstwa Igora Szewczenki (Шевченко et al., 2005) zmodyfikowane na potrzeby opracowanego wspólnego tagsetu.

W ramach prac nad ujednoliceniem opisów morfoskładniowych obu języków został najpierw stworzony wspólny tagset polsko-ukraiński, wzorowany na bardziej czytelnym i intuicyjnym sposobie zapisu tagów Korpusu IPIPAN (Kotsyba et al., 2008). W PolUKR-ze jednak ostatecznie znalazł zastosowanie inny, częściowo wzorowany na poprzednim, wspólny tagset, opracowany w ramach wielojęzycznego (17 języków) projektu europejskiego MULTEXT-East (MTE), wersja 4 „MONDILEX", dostępny pod adresem http://nl.ijs.si/ME/V4/ (Erjavec, 2012) razem z przykładowym, oznakowanym za jego pomocą, korpusem i leksykonem[7].

Potrzeba wspólnego tagsetu wynikła z jednego z pierwotnie stawianych celów tworzenia korpusu, mianowicie, zastosowania go do gramatycznych i semantycznych badań konfrontatywnych. Podobne rozumienie terminologii morfoskładniowej w porównywanych językach jest przydatne także w szeregu zastosowań maszynowych. Na przykład przy automatycznym generowaniu słowników dwujęzycznych porównywane są charakterystyki morfoskładniowe wyrazów. Wobec tego, samo podobieństwo formalne przy różnym rozumieniu terminów albo różne nazywanie podobnych zjawisk prowadzi do powstawania błędów, których przy uspójnionym opisie można uniknąć.

Samo ustalanie tagsetu już należy do badań konfrontatywnych. Nawet kwestie pozornie nieskomplikowane, jak np. definicja i zakres rzeczownika, mogą dostarczyć problemów użytkownikom korpusu równoległego, w którym użyto różnych tagsetów[8]. Do informacji morfoskładniowej obu języków korpusu wykorzystano źródła o konceptualnie odmiennych podejściach do ekstrakcji informacji oraz jej organizacji i zapisu, a także różny stopień granulacji tych informacji. W każdym przypadku takiego zróżnicowania trzeba było podjąć decyzję dotyczącą docelowości kodowania informacji. Z jednej strony należało liczyć się z ewentualną stratą informacji (czego zamierzaliśmy unikać), z drugiej

6 Konwerter jest dostępny na stronie http://www.domeczek.pl/~polukr/mte-conv/, zob. też http://clip.ipipan.waw.pl/LRT.

7 Leksykon polski jest zmodyfikowanym i przekonwertowanym fragmentem słownika gramatycznego autorstwa Marcina Wolińskiego, Zygmunta Saloniego, Jana Tokarskiego i in. Zob. notkę: http://nl.ijs.si/ME/V4/msd/html/msd-pl.introduction.html.

8 Problem znacznie się powiększa z rozszerzeniem o kolejne języki (Derzhanski, Kotsyba, 2009; Rosen, 2010).

strony pojawiła się konieczność uzupełnienia brakujących informacji w drugim języku. Często zastosowanie takiej brzytwy Ockhama uzasadniane było nie tyle potrzebami teoretyczno-lingwistycznymi, ile praktycznymi możliwościami. Dla porównania podajemy wybrane statystyki, dotyczące dwóch początkowych tagsetów: tylko 6 kategorii gramatycznych[9] było tożsamych formalnie; 21 kategorii było specyficznych dla języka ukraińskiego, 23 kategorie były specyficzne dla języka polskiego, przy czym suma kategorii w obu tagsetach stanowiła 50 jednostek. Jako przykład źródła takich różnic można przytoczyć to, że ukraiński tagset traktował przymiotniki i przysłówki stopnia wyższego i najwyższego jako osobne "techniczne" części mowy, podczas gdy w polskim były one opisywane pod wspólnymi fleksemami. Oba tagsety zawierały kategorię predykatywu, ale jej traktowanie istotnie się różniło, co czyniło formalne podobieństwo kategorii praktycznie bezużytecznym[10].

IPIC tag	MTE tag	MTE extended	Tokens	Example
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgy-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=yes Syntactic_Type=nominal	44	*niej*
ppron3:sg:gen:f:ter:nakc:praep	Pp-3f--sgasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=genitive Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--say-n	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=yes Syntactic_Type=nominal	11	*nią*
ppron3:sg:acc:f:ter:nakc:praep	Pp-3f--saasn	Pronoun Type=personal Person=third Gender=feminine Number=singular Case=accusative Clitic=agglutinant Definiteness=short-art Syntactic_Type=nominal		*ń*

Rysunek 1. Mapowanie tagów Korpusu IPIPAN (IPIC) do tagsetu MTE-PL.

W porównaniu z tagsetem Korpusu IPI PAN stworzony tagset MTE-PL ma rozbudowany system znaczników zaimkowych, bardziej zbliżone do tradycyjnych kategorie części mowy, ruchome końcówki i wskaźniki modalne typu „by" traktowane są w nim wspólnie z podstawą. Z 1298 oryginalnych tagów 101 otrzymało więcej niż jedną projekcję na tagset MTE: 60 tagów przymiotnikowych otrzymało po 13 projekcji w MTE; 18 tagów substantywów – po 2–7 MTE;

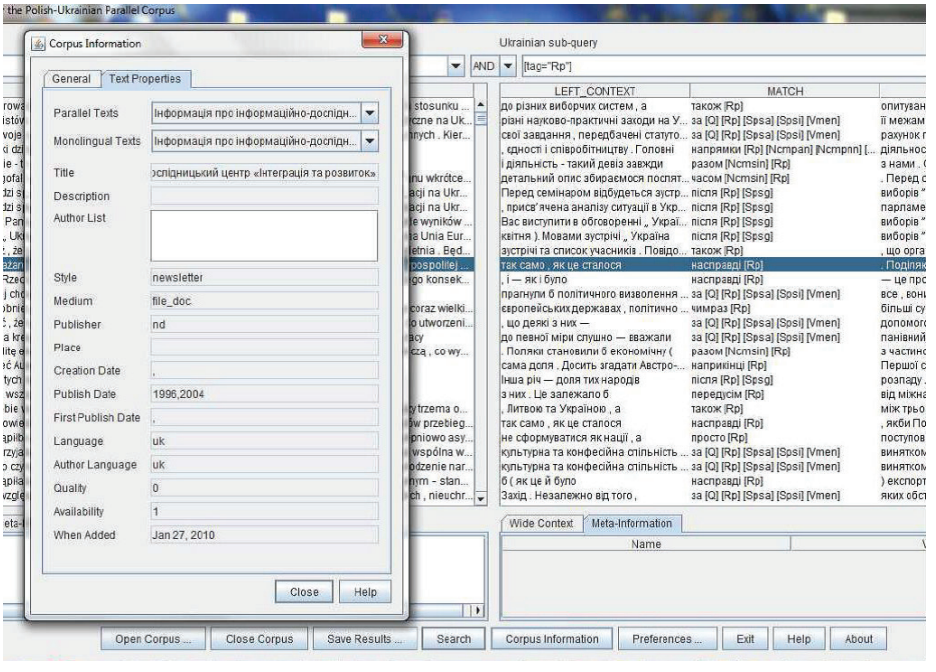
9 W polskim są to odpowiednio fleksemy (Przepiórkowski, Woliński, 2003), w ukraińskim – części mowy.

10 Predykatyw był jednym z największych źródeł problemów przy pracy nad wspólnym tagsetem ze względu na różne rozumienie tego terminu. Dla porównania: Korpus IPIPAN zawiera 26 predykatywów (wyliczenia własne za pomocą wyszukiwarki Poliqarp), Ukraiński Słownik Gramatyczny zawiera 176 predykatywów (Derzhanski, Kotsyba, 2008).

qubliki zostały podzielone na 7 kategorii z 27 tagami; predykatywy zostały podzielone na 3 kategorie i 4 tagi (Kotsyba et al., 2009).

3.4 Wyszukiwarki dedykowane

Dla wersji pilotażowej korpusu sporządzona została prosta wyszukiwarka internetowa, która obsługiwała wówczas 35 par krótkich tekstów publicystycznych, wyrównanych na poziomie akapitów. Możliwe było wyszukiwanie za pomocą wyrażeń regularnych, co uzupełniało częściowo brak lematyzacji, ale język zapytań był niezbyt przyjazny dla użytkownika. Do pierwszej pełnej wersji korpusu stworzona została w języku Java stacjonarna wyszukiwarka POSHUK, w której zaimplementowano wyszukiwanie według metainformacji oraz prostych tagów, zob. rys. 2. Wyszukiwarka ta później nie była rozwijana. Wśród wyszukiwarek webowych, z którymi eksperymentowano, żeby umożliwić dostęp do korpusu przez Internet, warto wymienić Park[11] – jeden z pierwszych programów tego typu projektu Intercorp[12]. Obecnie dostępny korpus korzysta z CWB[13] i ParaVozu (Meyer et al., 2014).



Rysunek 2. Interfejs wyszukiwarki POSHUK.

11 http://parcor.ibi.uw.edu.pl/Park/. Od kwietnia 2015 roku autorzy Intercorpu całkowicie zrezygnowali z tego programu na rzecz nowej, dużo szybszej i zintegrowanej z korpusem jednojęzycznym wyszukiwarki Kontext: https://kontext.korpus.cz/.

12 http://ucnk.ff.cuni.cz/intercorp/.

13 http://cwb.sourceforge.net/.

4. PolUKR-2

Następca tytułowego projektu, PolUKR-2, ma na celu istotne rozbudowanie ilościowe korpusu tak, aby umożliwić na szeroką skalę badania leksykologiczne i gramatyczne, a także wspomóc prace nad powstającym dużym słownikiem ukraińsko-polskim[14]. Planowana jest objętość od 10 do 30 milionów wyrazów w zależności od możliwości finansowych. Dotychczas opracowano kolejne 6,5 milionów wyrazów. Dobrane zostały głównie oryginalne teksty polskie, napisane w XIX bądź XX wieku, i ich tłumaczenia ukraińskie. Wyrównanie nowych tekstów zostało dokonane w ramach warsztatów tłumaczeniowych studentów filologii polskiej (Uniwersytet Narodowy im. Iwana Franki we Lwowie). Prace trwały przez dwa semestry (2013/2014).

Jeszcze jeden projekt badawczo-dydaktyczny skierowany na rozwiązanie problemu ujednoznaczniania został zorganizowany w ramach zajęć praktycznych w Katedrze Lingwistyki Stosowanej Politechniki Lwowskiej. Prace z ręcznego ujednoznaczniania tekstów trwały jeden semestr (jesień 2013 r.)[15]. Jednocześnie prowadzone były prace nad tworzeniem reguł ujednoznaczniania w ramach formalizmu Constraint Grammar (Karlsson, 1990). Opracowano m.in. reguły ujednoznaczniania wybranych przyimków.

5. Podsumowanie

W ciągu ostatnich dziesięciu lat od początku prac nad projektem sytuacja w lingwistyce korpusowej zmieniła się na korzyść: jest więcej dostępnych tekstów lepszej jakości w postaci elektronicznej, pojawiły się kolejne tłumaczenia. Ponadto podobne projekty korpusowe są rozwijane przez wiele ośrodków, co daje możliwość wymiany tekstów. Jakość działania programów służących do opracowania tekstów (np. FineReader 10.0) też znacznie się poprawiła. Pojawiły się nowe dostępne wyszukiwarki i wyrównywarki. Tendencje te są dowodem, że wysiłki zainwestowane w teorię lingwistyczną i rozwój technologii były trafnym posunięciem, o wiele lepszym niż ekspansja ilościowa – koszt opracowania tej samej ilości tekstów obecnie jest dużo niższy, co znaczy, że nadszedł właściwy czas, żeby zająć się powiększeniem korpusu. Najbardziej pozytywną zmianą jest jednak rosnące zainteresowanie korpusami ze strony językoznawców, leksykografów, tłumaczy, co zwiększa motywację do kontynuacji prac nad projektem.

14 http://clip.ipipan.waw.pl/UkrPolDict.

15 Ilość przerobionych tekstów wciąż nie wystarcza dla danych treningowych tagera, ale planowane są kolejne prace w tym zakresie.

Literatura

Derzhanski, Ivan, Kotsyba, Natalia (2008): The Category of Predicatives in the Light of Consistent Morphosyntactic Tagging. W: *Lexicographic Tools and Techniques, Proceedings of MONDILEX First Open Workshop, Moscow, Russia, 3-4 October 2008,* 68–79. http://domeczek.pl/:natko/papers/ID_NK_tagSlav.pdf, (01-03-2016).

Derzhanski, Ivan, Kotsyba, Natalia (2009): Towards a Consistent Morphological Tagset for Slavic Languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. W: Radovan Garabík (red.)*: Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop, Bratislava, Slovakia, 15–16 April 2009*, 9–26.

Erjavec, Tomaž (2012): MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation* 46(1), 131–142.

Karlsson, Fred (1990): Constraint Grammar as a Framework for Parsing Unrestricted Text. W: Hans Karlgren (red.): *Proceedings of the 13th International Conference of Computational Linguistics, Volume 3.* Stroudsbug, PA: Association for Computational Linguistics, 168–173.

Kotsyba, Natalia (2012): PolUKR (a Polish-Ukrainian Parallel Corpus) as a Testbed for a Parallel Corpora Toolbox. *Prace Filologiczne* LXIII, 181–196.

Kotsyba, Natalia, Shypnivska, Olha, Turska, Magdalena (2008): Linguistic principles of organizing a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). W: Mieczysław Kłopotek (red.): *Proceedings of the International Conference on Intelligent Information Systems, 16-18 June 2008, Zakopane, Poland*, 475–484.

Kotsyba, Natalia, Radziszewski, Adam, Derzhanski, Ivan (2009): Integrating the Polish language into the MULTEXT-East family: morphosyntactic specifications, converter, lexicon and corpus. W: Tomaž Erjavec (red.): *Proceedings of Research Infrastructure for Digital Lexicography: MONDILEX Fifth Open Workshop, October 14, 2009, Ljubljana, Slovenia*, 37–55.

Kotsyba, Natalia, Mykulyak, Andriy, Shevchenko Ihor V. (2011): UGTag: morphological analyzer and tagger for Ukrainian language. W: Stanisław Goźdź-Roszkowski (red.): *Explorations across Languages and Corpora,* Frankfurt am Main: Peter Lang, 69–82.

Meyer, Roland, von Waldenfels, Ruprecht, Woźniak, Michał, Zeman, Andreas (2006-2015): *ParaVoz – a simple web interface for querying parallel corpora.* Second Version. Bern, Regensburg, Berlin, Kraków. https://bitbucket.org/rvwfels/paravoz, (17 October 2015).

Piasecki, Maciej (2007): Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly* 11(1-2), 151–167.

Przepiórkowski, Adam, Woliński, Marcin (2003): A Flexemic Tagset for Polish. W: *The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*, Budapest: Association for Computational Linguistics, 33–40.

Przepiórkowski, Adam (2004): *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version.* http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/, (03-03-2016).

Rosen, Alexandr (2010): Mediating between Incompatible Tagsets. W: Lars Ahrenberg, Jörg Tiedemann and Martin Volk (red.) *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora AEPC 2010 December 2, 2010, Tartu, Estonia,* 53–62.

Varga, Daniel, Németh, Péter, Halácsy, Péter, Kornai, András, Trón, Viktor, Nagy, Viktor (2005): Parallel corpora for medium density languages. W: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov (red.) *Proceedings of the International Conference on Recent Advances on Natural Language Processing*, 590–596.

Шевченко, ИГорь, ШирОков, ВОлодимир, Рабулець, АЛександр (2005): Электронный грамматический словарь украинского языка. W: Труды международной конференции «*Megaling'2005. Прикладная лингвистика в поиске новых путей». 27 июня – 2 июля 2005 года. Меганом, Крым, Украина*, 124–129.