BULLETIN DE LA SOCIÉTÉ POLONAISE DE LINGUISTIQUE, fasc. LXV, 2009
ISSN 0032–3802

ADAM PRZEPIÓRKOWSKI
Instytut Podstaw Informatyki PAN
Uniwersytet Warszawski
Warszawa

RAFAŁ L. GÓRSKI
Instytut Języka Polskiego PAN
Kraków

BARBARA LEWANDOWSKA-TOMASZCZYK
Uniwersytet Łódzki
Łódź

MAREK ŁAZIŃSKI
Uniwersytet Warszawski
Warszawa

Narodowy Korpus Języka Polskiego

Po kilku dekadach, w których jakąkolwiek wartość przypisywano w językoznawstwie jedynie danym, które pochodzą z introspekcji, daje się dostrzec wyraźny zwrot ku empiryzmowi. Bez wątpienia główną tego przyczyną jest radykalna zmiana zarówno jakości danych empirycznych, jak i łatwości ich pozyskiwania. Dzięki korpusom elektronicznym językoznawca nie jest zdany na konieczność czynienia uogólnień na podstawie jednostkowych przykładów, przeciwnie – może ich pozyskać niejednokrotnie setki czy tysiące, co zajmuje mu często nie więcej niż kilka sekund. Tym samym, dzięki znacznie szerszej podstawie eksperymentalnej, pewne zjawiska – zdawałoby się dobrze znane – ukazują się w nowym świetle. Ta łatwość pozyskiwania danych rzutuje również na przetwarzanie języka naturalnego. W wielu wypadkach lepsze efekty dają metody oparte nie na wiedzy językoznawczej, ale na statystyce.

Jak się wydaje, językoznawstwo polskie (tak w sensie przedmiotowym jak i podmiotowym) wspomniany zwrot ku empiryzmowi ma wciąż przed sobą. Niewątpliwie ważną przyczyną tego stanu rzeczy jest brak korpusu spełniającego wszystkie wymogi współczesnego językoznawstwa. Jeśli opisywany projekt zakończy się sukcesem, możemy mieć nadzieję na radykalną zmianę tej sytuacji. W chwili gdy piszemy te słowa, w Internecie pod adresem www.nkjp.pl jest dostępna bezpłatnie

demonstracyjna wersja Narodowego Korpusu Języka Polskiego o rozmiarze ponad 450 milionów słów. Powstał on drogą scalenia istniejących dotąd korpusów języka polskiego.

Początkowo prace nad tworzeniem korpusów w Polsce były rozproszone. Pierwszy elektroniczny korpus języka polskiego to korpus, który stanowił podstawę dla Słownika Frekwencyjnego Współczesnej Polszczyzny. Kolejny to korpus Instytutu Języka Polskiego PAN w Krakowie, który niestety nigdy nie stał się publicznie dostępny. Nieco później rozpoczęto tworzenie korpusu Wydawnictwa Naukowego PWN[1]. Kolejne inicjatywy to korpus zespołu PELCRA[2] z Zakładu Językoznawstwa Komputerowego i Korpusowego Uniwersytetu Łódzkiego i wreszcie korpus Instytutu Podstaw Informatyki PAN[3] w Warszawie. Każdy z nich ma inne mocne strony. Korpus PWN jest najlepiej zrównoważony i urozmaicony. Z kolei atutem korpusu PELCRA jest duży, liczący ponad 600 000 słów komponent mówiony – transkrypty spontanicznych rozmów. Korpus IPI PAN zaś jest bardzo duży (z górą 200 000 słów) i anotowany morfosyntaktycznie.

Projekt opisywany w niniejszym artykule stanowi połączenie sił wymienionych czterech zespołów. Jest to sytuacja w pewnym sensie wyjątkowa – zespół NKJP znajduje się w tej szczęśliwej sytuacji, że może bazować na dotychczasowych doświadczeniach, również tych negatywnych, w przeciwieństwie do większości dużych korpusów narodowych, które powstawały od podstaw.

Projekt zakłada stworzenie „megakorpusu" o wielkości 1 miliarda słów. Oczywiście korpus tej wielkości nie może być zrównoważony – nie ma dostatecznej liczby tekstów w języku polskim, które reprezentowałyby w odpowiedniej ilości wszystkie style funkcjonalne. W związku z tym będzie to tzw. korpus oportunistyczny – złożą się nań wszystkie teksty, jakie uda się pozyskać, bez oglądania się na zróżnicowanie tematyczne czy stylistyczne. W tym wypadku wartością samą w sobie jest jego wielkość. Bez wątpienia dane pochodzące z tak wielkiego korpusu mają swoją wartość, nawet jeśli nie jest on reprezentatywny i tym samym pochodzące z niego dane mogą być jednostronne, reprezentatywne dla pewnej dominującej grupy tekstów. Prawdopodobnie trzon takiego korpusu będą stanowiły najłatwiejsze do pozyskania teksty, a więc wszelkie teksty urzędowe, w tym protokoły parlamentarne, teksty ustaw, a także teksty prasowe. Zapewne jednak dla większości badaczy istotniejszy będzie korpus zrównoważony, który ma liczyć 300 milionów słów. Jest to bardzo dużo, dość powiedzieć, że to trzykrotnie więcej niż swoisty standard wyznaczony przez British National Corpus. Dla porównania: jest to równowartość ok. 50 000 arkuszy wydawniczych. By uzmysłowić sobie, jak obfity materiał może taki korpus przynieść, dość powiedzieć, że w 600 razy mniejszym korpusie Słownika Frekwencyjnego słowo *biały* występuje 87 razy, słowo *łańcuch* 26 razy[4].

---

[1] http://korpus.pwn.pl/

[2] http://korpus.ia.uni.lodz.pl/

[3] www.korpus.pl, opisany w pracach Przepiórkowskiego (2004 i 2005)

[4] Przestrzegamy tu oczywiście przed naiwną ekstrapolacją tych liczb – w korpusie liczącym 300 mln słów niekoniecznie znajdziemy odpowiednio 52 000 i 16 600 wystąpień.

Zatem celem projektu jest dostarczenie środowisku językoznawczemu, a także informatykom zainteresowanym przetwarzaniem języka naturalnego czy wreszcie szerokim kręgom osób zainteresowanych językiem polskim, wielkiego korpusu, który spełnia wszelkie wymogi współczesnej nauki. Celem nie mniej istotnym, choć formalnie rzecz biorąc pobocznym, jest wypracowanie szeregu narzędzi dla przetwarzania języka naturalnego.

Korpus stanowi również źródło danych ilościowych. Po to, by dane te miały jakąkolwiek wartość, budowa korpusu powinna odzwierciedlać jakąś rzeczywistość. Można więc przyjąć, że powinna odzwierciedlać ofertę wydawniczą lub percepcję języka. Inną możliwością jest przypisanie jednakowej reprezentacji wszystkim a priori wyróżnionym typom tekstów[5].

Reprezentatywność korpusu jest w omawianym projekcie rozumiana jako reprezentacja percepcji języka przez polską społeczność językową. W praktyce chodzi o odzwierciedlenie struktury czytelnictwa w Polsce – powszechnie czytane typy tekstów mają większy udział w korpusie niż te, które są czytane przez statystycznie niewielką liczbę osób. Główny powód przyjęcia tej koncepcji jest raczej praktycznej natury. Otóż, gdyby przyjąć, że korpus ma odzwierciedlić populację tekstów drukowanych, to okazałoby się, że znikomą jego część będą stanowiły książki, bo taka jest proporcja pomiędzy tekstami ukazującymi się w książkach i prasie.

Od korpusu oczekuje się jednak nie tylko reprezentatywności, ale także zrównoważenia, a więc takiej budowy, która nie daje żadnemu typowi tekstów wyraźnej przewagi. Dość arbitralnie można przyjąć, że taką granicą jest 50%, tzn. że żaden z typów tekstów nie może stanowić więcej niż połowę korpusu. Istnieją wreszcie teksty, których zakresu recepcji nie sposób ustalić, a jest ona zapewne marginalna, ale których nie sposób nie uwzględnić w korpusie, który ma reprezentować całą polszczyznę. Mamy tu na myśli m.in. ustawy, ulotki, instrukcje. Kolejnym problemem jest ustalenie zakresu czytelnictwa takiego medium, jakim jest Internet. Tu trzeba od razu przyznać, że udział tych tekstów w korpusie został ustalony arbitralnie. Reprezentatywność korpusu ustalono na podstawie badań czytelnictwa prowadzonych przez Instytut Książki i Czytelnictwa Biblioteki Narodowej, a także organizacje powołane do kontroli nakładu prasy. Budowa korpusu ma mieć następującą postać:

[5] Nie ma tu miejsca na szczegółowe omawianie wszystkich koncepcji reprezentatywności. Zainteresowany czytelnik może znaleźć szerszą informację w artykule (Górski 2008).

Prasa		publicystyka i krótkie wiadomości prasowe	49%
Książki	fikcjonalne	literatura piękna	16,0%
	niefikcjonalne	książka publicystyczna	1,0%
		literatura faktu	5,5%
		teksty informacyjno-poradnikowe	5,5%
		teksty naukowo-dydaktyczne	2,0%
		niesklasyfikowane teksty książek niefikcjonalnych	1,0%
Inne		w tym: internet (blogi, czaty, fora)	7,0%
		teksty użytkowe, ustawy, protokoły	3,0%
		teksty mówione	10,0%

Zapewne ideałem byłby losowy dobór tekstów, trzeba sobie jednak od razu powiedzieć, że jest to postulat niemożliwy do spełnienia. Jesteśmy tu bowiem ograniczeni przede wszystkim dwoma czynnikami, mianowicie zgodą właścicieli praw autorskich i dostępnością tekstu w wersji elektronicznej.

Warto zwrócić uwagę na komponent mówiony korpusu. Obejmować ma on 10%. Oczywiście nagranie, a następnie transkrypcja takiej ilości tekstów nie jest możliwa, jeśli ma być tylko jednym z zadań tak rozbudowanego projektu. Tak więc jedynie 1% całego korpusu będą stanowiły transkrypcje spontanicznych „podsłuchanych" rozmów[6]. Zauważmy wszakże, iż stanowi to 3 miliony słów, co zapewni bardzo szeroką egzemplifikację i znacznie przekracza dotychczas publikowane zbiory naturalnych dialogów. Większość wszakże tego komponentu będą stanowiły gotowe zapisy tekstów mówionych, takie jak protokoły parlamentarne czy zapisy wywiadów radiowych.

Kolejny podkorpus, liczący 1 mln słów, dokładnie odzwierciedlający budową korpus zrównoważony, zostanie wygenerowany drogą losowania niewielkich fragmentów z tekstów, które wchodzą w skład korpusu. Będzie on ręcznie anotowany językowo. Grupa lingwistów dokona podziału na zdania, ujednoznaczni wszystkie homonimiczne formy w tekście i przypisze wybranym wyrazom znaczenia właściwe w ich kontekście. Każde zdanie będzie interpretowane niezależnie przez dwie osoby, które w razie rozbieżności będą musiały sprawdzić swoją interpretację, co pozwoli wyeliminować pomyłki. Najczęściej bowiem rozbieżności wynikają ze zwykłego błędu, choć niekiedy są wynikiem odmiennych interpretacji. Ten korpus tworzymy dla dwu celów: po pierwsze dla wielu zastosowań ściśle językoznawczych istotny jest korpus, w którym użytkownik może być pewny interpretacji fleksyjnej wszystkich jednostek – jeśli zada pytanie np. o mianownik, to korpus zwróci wszystkie mianowniki i tylko mianowniki. Musimy się wtedy oczywiście pogodzić z niewielkimi rozmiarami takiego korpusu, niemniej dla wielu badań np. o charak-

[6] Oczywiście uczestnicy rozmowy zawsze muszą wyrazić zgodę i podpisać stosowną licencję.

terze gramatycznym będzie on wystarczający. Drugi cel to „trenowanie" narzędzi informatycznych, o czym piszemy poniżej.

Oprócz informacji morfosyntaktycznych podkorpus ten będzie zawierał także ręcznie wprowadzone oznakowanie nazw własnych, ręczną dezambiguację (ujednoznacznienie) znaczeń czasowników, rzeczowników, przymiotników i przysłówków, identyfikację podstawowych typów fraz składniowych oraz – w mniejszym zakresie – informacje o głębokich (semantycznych) strukturach argumentów czasowników. Wszystkie te informacje są niezbędne lub przydatne w takich zastosowaniach, jak ekstrakcja informacji czy tłumaczenie maszynowe, istotne jest więc stworzenie danych treningowych i testowych dla narzędzi informatycznych znajdujących automatycznie takie informacje lingwistyczne. Zgodnie z naszą wiedzą nie istnieją obecnie korpusy języka polskiego zawierające informacje tego typu.

Anotacja

Korpus ma być bogato anotowany na różnych poziomach. Każdy tekst będzie opatrzony szczegółowymi danymi bibliograficznymi, co pozwoli dany cytat połączyć z autorem[7] i tytułem dzieła. Zamieścimy także szczegółowe dane dotyczące medium tekstu (książka, prasa, internet), stylu funkcjonalnego, a także klasyfikacji tematycznej. Jeśli chodzi o tę ostatnią to zamierzamy opatrywać teksty kodem Uniwersalnej Klasyfikacji Dziesiętnej oraz klasyfikacją tematyczną Biblioteki Narodowej. Teksty mówione z kolei będą zawierały szczegółowe informacje o uczestnikach rozmowy (ich wiek, miejsce zamieszkania, wykształcenie itp.). Obok informacji o charakterze bibliograficznym i socjolingwistycznym zamieszczone będą także informacje dotyczące sposobu przetworzenia tekstu do postaci elektronicznej, statusu tekstu itp. Dane te, raczej nieistotne dla przeciętnego użytkownika korpusu, mają dużą wagę dla utrzymującego go zespołu, z kolei klasyfikacja tematyczna pozwoli nam łatwo zorientować się w zróżnicowaniu korpusu.

Kolejny poziom anotacji odzwierciedla strukturę tekstu – przede wszystkim podział na akapity, ale także rozdziały, tomy itp. Trzeba tu jednak od razu zaznaczyć, że dokonywanie dokładnej anotacji tekstu jest bardzo pracochłonne, a równocześnie dla znakomitej większości zastosowań – zbędne. W związku z powyższym tylko pewien podkorpus będzie szczegółowo anotowany pod względem struktury tekstu.

Następny poziom to znakowanie morfosyntaktyczne (fleksyjne). Każde słowo w korpusie będzie opatrzone charakterystyką fleksyjną. Dzięki temu użytkownik może poszukiwać w korpusie tak jednostek leksykalnych, jak i kategorii fleksyjnych bądź ich kombinacji. Korpus będzie też anotowany pod względem składniowym. W tym ostatnim wypadku chodzi jednakże jedynie o analizę powierzchniową, a więc nie o pełną analizę składniową zdania, a jedynie wyróżnienie poszczególnych fraz. Ostatnim wreszcie poziomem anotacji będzie znakowanie semantyczne – dla dość ograniczonej liczby słów zostaną wydzielone podstawowe znaczenia. Na-

[7] Należy oczywiście pamiętać, że autorzy wielu tekstów prasowych i użytkowych są anonimowi.

stępnie słowom tym zostaną przypisane przez odpowiedni program znaczenia właściwe w ich kontekście[8].

Jak widać, tekst będzie obudowany szeregiem informacji, które nie stanowią jego integralnej części. Podobnie zresztą w edycji filologicznej obok właściwego tekstu mamy do czynienia z komentarzem i aparatem krytycznym. Istotne jest wszakże z jednej strony powiązanie tekstu z komentarzem, a z drugiej – ścisłe oddzielenie tych dwu. Tym, co w korpusie pozwala powiązać tekst z owymi informacjami dotyczącymi danych bibliograficznych, struktury tekstu, jak i wspomnianymi opisami gramatycznymi, jest język znaczników XML, a dokładnie standard TEI P5. Posługiwanie się zestandaryzowanym schematem anotacji jest o tyle istotne, że korpus może być przetwarzany przy pomocy różnych narzędzi informatycznych, niekoniecznie stworzonych specjalnie dla niego[9].

Znakomita większość użytkowników korpusu będzie korzystała z niego za pośrednictwem jednej z dwu wyszukiwarek: Poliqarp, stworzonej w Instytucie Podstaw Informatyki PAN, lub PELCRA[10]. Pierwsza za cenę dość skomplikowanej składni zapytań pozwala zadawać bardziej wyrafinowane pytania, druga z kolei ma składnię prostszą, przyjazną użytkownikowi, choć i nieco bardziej ograniczone możliwości. Oba te programy zapewniają dostęp do korpusu przez Internet. Podstawowym ich zadaniem jest tworzenie konkordancji, czyli listy słów kluczowych wraz z ich najbliższym kontekstem. Konkordancje mogą być sortowane alfabetycznie wg słowa kluczowego lub kontekstu. Inną ważną ich funkcją jest wyszukiwanie kolokacji. Istotnym poszerzeniem funkcjonalności wyszukiwarek jest możliwość ograniczenia przeszukiwania do konkretnego stylu funkcjonalnego, autora bądź ram chronologicznych. PELCRA oferuje także dostęp programistyczny[11].

Do pewnych nietypowych zastosowań niezbędny będzie dostęp do wersji źródłowej korpusu, ze względu jednak na ochronę interesów autorów i wydawców będzie można z niego korzystać jedynie za pośrednictwem instytucji współtworzących korpus.

Jest oczywiste, że anotacja językowa tak dużego korpusu może być przeprowadzona jedynie automatyczne. W tym celu tworzymy szereg narzędzi informatycznych. Duża ich część będzie „trenowana" na wspomnianym wyżej znakowanym ręcznie podkorpusie. W praktyce chodzi o to, by zebrać dane dotyczące prawdopodobieństwa wystąpienia pewnych sekwencji jednostek. Przykładowo: kropka może być znakiem końca zdania, skrótu, może też w liczbach oddzielać jednostki od części dziesiątych. Podobnie ciąg *myślą* to forma od czasownika *myśleć* lub od rze-

---

[8] Będzie to jednak zapewne rozwiązanie prototypowe, które należy widzieć w kategoriach eksperymentu, a nie gotowego narzędzia.

[9] Sami twórcy korpusu przekonali się o roli standaryzacji w czasie scalania trzech korpusów wyjściowych. Szczęśliwie anotacja ich była oparta na XML, jednak duże rozbieżności pomiędzy korpusami nie ułatwiały scalenia.

[10] Przeszukiwarka Poliqarp została opisana w pracach (Przepiórkowski et al. 2004; Przepiórkowski 2004; Janus, Przepiórkowski 2007a, 2007b), zaś przeszukiwarka PELCRA w pracy (Pęzik et al. 2004; Waliński, Pęzik 2007).

[11] Za wcześnie jeszcze na podawanie szczegółów – obie wyszukiwarki są intensywnie rozwijane.

czownika *myśl*. Ponieważ zależy nam na tym, żeby odszukać właściwą w danym kontekście interpretację tych ciągów, musimy dokonać ujednoznacznienia. Odbywa się to następująco: najpierw rozpoznawane są formy fleksyjne w tekście, przy czym formom homonimicznym przypisuje się wszystkie możliwe interpretacje. Kolejny program przypisuje formom właściwą w danym kontekście interpretację w ten sposób, że na podstawie wspomnianego wyżej ręcznie znakowanego korpusu ustala, jaka jest najbardziej prawdopodobna sekwencja form gramatycznych. By posłużyć się przykładem: *myślą* w ciągu *lotną myślą* będzie raczej rzeczownikiem, skoro występuje bezpośrednio po przymiotniku, zaś w ciągu *chłopcy myślą* raczej czasownikiem, skoro występuje bezpośrednio po rzeczowniku w mianowniku. Nie stoi za tym żadna wiedza językoznawcza, a jedynie odtwarzanie obliczonego na podstawie ręcznie anotowanego korpusu prawdopodobieństwa wystąpienia jakiejś sekwencji form gramatycznych. Oczywiście w rzeczywistości stosuje się tu znacznie subtelniejszą statystykę niż ukazana w powyższych przykładach, niemniej ilustrują one ogólną ideę. Precyzja tego rodzaju ujednoznaczniania jest dość wysoka – wyraźnie powyżej 90% jednostek w tekście jest interpretowanych prawidłowo; trzeba jednak pamiętać, że większość tekstu zajmują jednostki jednoznaczne, tak więc w rzeczywistości, tam gdzie chodzi o subtelny problem (np. regularna homonimia wewnątrz paradygmatu fleksyjnego rzeczownika), ujednoznacznianie nie może być aż tak precyzyjne. Warto tu dodać, że zadanie utrudnia zarówno swobodny szyk polszczyzny, jak i fakt, że liczba możliwych kombinacji kategorii fleksyjnych dochodzi do 1500[12].

Jednym z zadań projektu są tzw. słowa tygodnia, czyli udostępniana w Internecie wraz z komentarzem lista słów, dla których w tekstach wybranych czasopism zanotowano wyraźnie wyższą niż zazwyczaj frekwencję (por. Łaziński i Szewczyk 2006). Będzie to zapewne przyciągało uwagę szerszej publiczności.

Trudno przewidzieć wszystkie zastosowania korpusu. Przypomnijmy – korpus jest jedynie narzędziem i od badacza zależy, jakie pytanie mu zada. Jednym z pierwszych zastosowań jest leksykografia. NKJP jest podstawowym źródłem dla *Wielkiego Słownika Języka Polskiego* PAN.

Badania prowadzone na korpusach innych języków pokazały, że narzędzie to pozwala dostrzec wiele faktów, które bez niego pozostałyby niewidoczne. Korpusy znajdują zastosowanie w leksykologii, gramatyce, pragmatyce, analizie dyskursu czy rekonstrukcji językowego obrazu świata. Służą do tworzenia materiałów dydaktycznych. Można się też spodziewać, że NKJP będzie ważnym elementem dydaktyki uniwersyteckiej – korpus dostarcza materiału do szeregu prac, które z jednej strony nie przekraczają możliwości studenta, a z drugiej dają satysfakcję odkrycia czegoś nowego.

Należy wreszcie podkreślić, że narzędzia informatyczne stworzone na potrzeby korpusu znajdą zastosowania praktyczne. Programy rozpoznające formy gramatycz-

---

[12] Ze względu na językoznawczy profil BPTJ ograniczamy omówienie technicznych aspektów dezambiguacji do minimum. Narzędzia stosowane w korpusie do tego celu opisane są w pracach (Piasecki, Godlewski 2006; Dębowski 2004)

ne w tekście czy dokonujące powierzchniowej analizy składniowej są nieodzownymi składnikami programów do wyszukiwania informacji w tekście, tłumaczenia automatycznego czy inteligentnych wyszukiwarek internetowych. Założeniem projektu jest to, że wszystkie narzędzia powstałe w jego ramach będą powszechnie dostępne.

The National Corpus Polish

Summary

The paper describes an ongoing project which aims at compiling a large corpus of modern Polish publicly accessible via web. Four major Polish corpus developers are involved, that is Institute of Computer Science, Institute of Polish Language (both at the Polish Academy of Sciences), Chair of English Language and Applied Lingustics, Łódź University and PWN Scientific Publishers.

The intended corpus will be 1 billion running words large, with a balanced subcorpus containing 300 million words. The corpus will be structurally and linguistically annotated. The access will be possible with two different interfaces. A second aim – maybe even as important as compiling the corpus – is to develop a series of tools for natural language processing designed especially for Polish.

Key words: linguistic corpus, corpus linguistics, natural language processing

Bibliografia

Dębowski Łukasz (2004): Trigram morphosyntactic tagger for Polish – [w:] Kłopotek Mieczysław A., Wierzchoń Sławomir T., Trojanowski Krzysztof (2004), 409–413.

Górski Rafał L. (2008): Representativeness of the written component of a large reference corpus of Polish. Primary notes – [w:] Barbara Lewandowska-Tomaszczyk (red.): *Corpus Linguistics, Computer Tools, and Applications – State of the Art*. *PALC 2007*, Frankfurt/M etc: Peter Lang, 119–123.

Janus Daniel, Przepiórkowski Adam (2007a): *Poliqarp: An open source corpus indexer and search engine with syntactic extensions* – [w:] *Proceedings of the ACL 2007 Demo Session*, Prague.

Janus Daniel, Przepiórkowski Adam. (2007b): Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. – [w:] Jacek Waliński, Krzysztof Kredens, Stanisław Goźdź-Roszkowski (red.): *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt am Main: Peter Lang.

Kłopotek Mieczysław A., Wierzchoń Sławomir T., Trojanowski Krzysztof, red. (2004): *Intelligent Information Processing and Web Mining. Advances in Soft Computing.* – Berlin: Springer-Verlag.

Łaziński Marek, Szewczyk Monika (2006): Słowa klucze w semantyce i statystyce. Słowa tygodnia „Rzeczpospolitej". – *Biuletyn Polskiego Towarzystwa Językoznawczego*, LXII: 57–68.

Piasecki Maciej, Godlewski Grzegorz (2006): Reductionistic, tree and rule based tagger for Polish. – [w:] Kłopotek Mieczysław A., Wierzchoń Sławomir T., Trojanowski Krzysztof (2004), 531–540.

Przepiórkowski Adam, Krynicki Zygmunt, Dębowski Łukasz, Woliński Marcin, Janus Daniel, Bański Piotr (2004): A search tool for corpora with positional tagsets and ambiguities – [w:] *LREC* (*LREC* 2004), 1235–1238.

Przepiórkowski Adam (2004): *Korpus IPI PAN. Wersja wstępna / The IPI PAN Corpus: Preliminary version*, IPI PAN. – Warszawa.

Przepiórkowski Adam (2005): The IPI PAN Corpus in numbers – [w:] Vetulani Zygmunt (red.), Proceedings of the 2nd Language & Technology Conference, Poznań, Poland.

Pęzik Piotr, Levin Eric, Uzar Rafał (2004): Developing relational databases for corpus linguistics – [w:] Barbara Lewandowska-Tomaszczyk (red.): *The proceedings of Practical Applications in Language and Computers PALC 2003*, Frankfurt am Main: Peter Lang.

Waliński, Jacek, Pęzik. Piotr (2007): Web access interface to the PELCRA referential corpus of Polish – [w:] Jacek Waliński, Krzysztof Kredens, Stanisław Góźdź-Roszkowski (red.): *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt am Main: Peter Lang, 65–86.