Maciej Ogrodniczuk

Instytut Podstaw Informatyki PAN, Warszawa

maciej.ogrodniczuk@ipipan.waw.pl

Lingwistyka komputerowa dla języka polskiego: dziś i jutro

Słowa kluczowe: lingwistyka komputerowa, lingwistyka informatyczna, przetwarzanie języka naturalnego, technologie językowe, narzędzia i zasoby językowe dla polszczyzny

1. Wprowadzenie

W 2006 r. na łamach LingVariów ukazał się entuzjastyczny tekst Marka Świdzińskiego nt. zdobyczy polskiej lingwistyki korpusowej, w którym autor przedstawiał korzyści, jakie językoznawstwu przyniosła, także w Polsce, rewolucja informatyczna. Sposób, w jaki środowisko językoznawcze te zmiany przyjęło, dość dobrze opisują z kolei publicystyczne artykuły Piotra Żmigrodzkiego (2015a) i Adama Przepiórkowskiego (2015). W niniejszym tekście chciałbym spojrzeć na sytuację lingwistyki informatycznej w Polsce oraz na dalsze kierunki prac nad komputerowym przetwarzaniem polszczyzny z perspektywy 10 lat późniejszej oraz z punktu widzenia przedstawiciela tej dyscypliny, często postrzeganej wyłącznie jako usługowa dla środowiska językoznawczego.

Zacznijmy zatem od kilku słów na temat przedmiotu naszej pracy i trudności, przed jakimi stoimy. Lingwistyka komputerowa (lingwistyka informatyczna, inżynieria lingwistyczna) to dyscyplina z pogranicza obu dziedzin – informatyki i językoznawstwa, której celem jest badanie języka naturalnego z punktu widzenia potrzeb jego przetwarzania i modelowania metodami komputerowymi oraz tworzenie elektronicznych zasobów i narzędzi przetwarzania języka na potrzeby usługowe. Powyższa definicja, łącząca istniejące (w szczególności dwie: Janusza S. Bienia i Bonnie Webber[1]), zwraca uwagę na dwa aspekty problemu: jest to jednocześnie dziedzina teoretyczna i stosowana. O pierwszym zdają się zapominać

[1] (Bień 2000), (Webber 2001), zob. też (Piasecki 2008).

językoznawcy, o drugim – sami informatycy. Nie znam szczegółów tego rozstrzygnięcia, ale o stosunku polskiego środowiska językoznawczego do lingwistyki komputerowej jako dziedziny badań w obszarze nauk humanistycznych najlepiej świadczy niedawne usunięcie jej z panelu HS 2.6 w konkursach Narodowego Centrum Nauki, chociaż w projektach prowadzonych w moim rodzimym Zespole Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN często więcej jest pracy ściśle językoznawczej (wykonywanej przez zaprzyjaźnionych lingwistów) niż informatycznej. Taka jest zresztą specyfika większości polskich zespołów zajmujących się przetwarzaniem języka: naszą domeną jest właśnie lingwistyka informatyczna, a nie informatyka lingwistyczna, z konsekwencjami w postaci rezygnacji z wyścigu w rozwoju technologii anglojęzycznej na rzecz świadomej lingwistycznie technologii dla polszczyzny; stąd też nasza częsta obecność w najważniejszych polskich pismach językoznawczych.

Po stronie informatycznej grzechem głównym jest właśnie trudność do przyznania się do częściowo usługowego wymiaru lingwistyki komputerowej i kłopoty komunikacyjne obu światów. Być może wynikają one z faktu postrzegania informatyków jako wykonawców, a nie partnerów w pracy badawczej. Usługowość rozumiem zatem nie jako gotowość nauczenia lingwistów języka zapytań do wyszukiwarki czy pomoc w stworzeniu własnego korpusu, ale jako rzeczywistą współpracę w badaniach.

Mimo wspomnianych trudności mój punkt widzenia na stan lingwistyki informatycznej w Polsce jest jednak optymistyczny: jesteśmy świadkami bardzo intensywnego rozwoju komputerowych narzędzi i zasobów dla języka polskiego[2], polskie środowisko lingwistyczno-informatyczne bierze udział w licznych grantach polskich i europejskich, uczestniczy w dużych infrastrukturach badawczych CLARIN i DARIAH łączących nauki techniczne i humanistyczne, a sami użytkownicy-humaniści coraz częściej sięgają po metody komputerowe naszego autorstwa. Trwa owocna współpraca między polskimi ośrodkami badawczymi, a komputerowy opis polszczyzny obejmuje coraz głębsze poziomy analizy. W dalszej części tekstu postaram się pokazać, w którym miejscu jesteśmy i zaproponować kilka kierunków badań na najbliższe lata, jeśli nie dla całego środowiska, to przynajmniej dla naszego warszawskiego zespołu.

[2] P. http://clip.ipipan.waw.pl/LRT – strona zawierająca listę dostępnych w sieci narzędzi i zasobów dla polszczyzny.

2. Stan technologii językowej dla polszczyzny okiem lingwisty komputerowego

2.1. Raport językowy

W roku 2011 warszawski Zespół Inżynierii Lingwistycznej Instytutu Podstaw Informatyki PAN brał udział w międzynarodowym projekcie CESAR, którego celem (wraz z innymi inicjatywami podobnego typu, prowadzonymi pod wspólnym szyldem META-NET-u) było zebranie, scalenie oraz zrównoleglenie elektronicznych zasobów i narzędzi językowych dla 30 języków europejskich, a następnie udostępnienie ich we wspólnym repozytorium. Miało to stworzyć nowe możliwości badawcze oraz technologiczne w dziedzinie przetwarzania języków naszego kontynentu. Projekt zakończył się, moim zdaniem, sporym sukcesem: wziął udział w rozwoju wielu ważnych zasobów, jak np. Słownik Gramatyczny Języka Polskiego (Saloni i in. 2007/2015), doprowadził do unormowania kwestii licencyjnych czy odświeżył stronę internetową CLIP (Ogrodniczuk i Przepiórkowski 2013)[3] z rejestrem dostępnych publicznie narzędzi dla polszczyzny oraz polskich zespołów badawczych zajmujących się przetwarzaniem języka i prowadzonych przez nie projektów.

Innym ważnym wynikiem prac był tzw. raport językowy (Miłkowski 2011) upowszechniający wiedzę nt. technologii językowych dla polszczyzny i ich potencjalnych zastosowań. Jego ważnym zadaniem było też przedstawienie obrazu bieżącej sytuacji w dziedzinie przetwarzania języka w Polsce. W oparciu o ocenę ekspercką w skali od 0 (ocena bardzo niska) do 6 (bardzo wysoka) przedstawiał dane nt. dostępności, jakości i elastyczności technologii i zasobów w zgrubnie zarysowanych dziedzinach. Wyniki tej oceny przytaczam w tabeli 1.

[3] Computational Linguistics in Poland, czyli Lingwistyka komputerowa w Polsce – p. http://clip.ipipan.waw.pl (strona w jęz. angielskim).

	Liczba	Dostępność	Jakość	Zakres	Dojrzałość	Trwałość	Elastyczność
Technologie językowe (narzędzia, technologie, aplikacje)							
Rozpoznawanie mowy	1	2	3	4	3	2	4
Synteza mowy	4	3	6	5	4	4	3
Analiza gramatyczna	4	4,5	4,5	4,5	4	4	3
Analiza semantyczna	1	1	3	1	1	2	2
Generowanie tekstu	1	1	1	1	1	1	2
Tłumaczenie maszynowe	3	4	3	3	3	4	3
Zasoby językowe (zasoby, dane, bazy wiedzy)							
Korpusy tekstów	3	2	4	4	5	5	3
Korpusy równoległe	3	1	4	4	5	5	5
Korpusy języka mówionego	1	0	3	3	2	2	2
Zasoby leksykalne	3	3	4	4	4	4	3
Gramatyki	3	2	4	4	3	2	2

Tabela 1. Stan dostępnych technologii językowych dla jęz. polskiego (przedruk ze str. 31 raportu)

Wykaz ten dobrze obrazuje stan prac sprzed 5 lat: już wtedy doskonale radziliśmy sobie z syntezą mowy[4], nieźle z analizą morfoskładniową (tu oznaczoną jako gramatyczna), jako dobrą ocenialiśmy dostępność jednojęzycznych korpusów tekstów[5], jako gorszą – korpusów równoległych; z dostępnością korpusów mówionych było gorzej niż źle. Raport zawiera więcej podobnych wniosków, głównie pesymistycznych: brak standaryzacji zasobów, konieczność intensyfikacji prac nad głęboką analizą tekstu, zasobami ontologicznymi i elektronicznymi słownikami walencyjnymi.

[4] Był to moment triumfu polskiego systemu Ivona we wielu międzynarodowych konkursach syntezy mowy, po którym firma będąca jego autorem została przejęta przez globalną firmę Amazon.
[5] Ze względu na moment zakończenia prac nad Narodowym Korpusem Języka Polskiego (Przepiórkowski i in. 2012).

Ważnych informacji dostarcza także porównanie stanu technologii dla języka polskiego i innych języków europejskich; tabela 2 przedstawia jego wyniki dla zadania analizy tekstu.

Doskonała jakość	Bardzo dobra jakość	Dobra jakość	Średnia jakość	Słaba/zerowa jakość
	angielski	francuski hiszpański niderlandzki niemiecki włoski	baskijski bułgarski czeski duński fiński galisyjski grecki kataloński norweski **polski** portugalski rumuński słowacki słoweński szwedzki węgierski	chorwacki estoński irlandzki islandzki litewski łotewski maltański serbski

Tabela 2. Stan technologii językowych dostępnych dla 30 języków europejskich dla zadania analizy tekstu (przedruk ze str. 33 raportu)

2.2. Sonda środowiskowa

Z perspektywy czasu widać, że zarówno usytuowanie polszczyzny w rodzinie języków europejskich, jak i wnioski nt. stanu rozwoju technologii były właściwe; kierunki najintensywniejszych prac prowadzonych w ciągu ostatnich lat pokrywają się z powyższymi postulatami. Powstaje duży semantyczny słownik walencyjny Walenty (Przepiórkowski i in. 2014), z którego zaczynają korzystać parsery, czyli narzędzia do głębokiej analizy tekstu (Patejuk i Przepiórkowski 2014, Jaworski i Przepiórkowski 2014, Woliński 2015), od wielu lat dynamicznie rozwija się Słowosieć (Piasecki i in. 2014), trwają prace nad poprawą jakości i dostępności korpusów równoległych i mówionych (Pęzik 2015 i 2016).

Wiarygodną ocenę stanu obecnego najlepiej jednak przeprowadzić metodą konsultacji ze środowiskiem, postanowiłem zatem powtórzyć ocenę metodą prostej sondy. Na etapie tworzenia pytań na podstawie tabeli z raportu językowego okazało się, że określenia rodzajów narzędzi i zasobów są zbyt ogólne (stąd zapewne dobra ocena jakości „analizy gramatycznej”,

które dziś większość respondentów zinterpretowałaby jako pytanie o analizę składniową, a nie morfoskładniową), postanowiłem zatem przejąć listę bardziej szczegółowych kategorii z istniejących badań zagranicznych (np. Strik i in. 2002) i dostosować ją do specyfiki polskiego środowiska. Sonda w części dotyczącej obecnego stanu prac składała się z 2 pytań:

1. Jak ocenia Pan/Pani dostępność narzędzi i zasobów dla polszczyzny?
   Proszę o podanie Państwa opinii nt. stopnia możliwości wykorzystania istniejących narzędzi i zasobów dla języka polskiego.
   a. Niedostępne
   b. Dostępne w ograniczonym stopniu (np. za opłatą, w ograniczonym zakresie)
   c. Dostępne w wystarczającym zakresie
   d. Nie wiem
2. Jak ocenia Pan/Pani jakość technologii językowych dla polszczyzny?
   Proszę o podanie Państwa opinii nt. ogólnej jakości dostępnej obecnie technologii (darmowej lub komercyjnej) dla języka polskiego biorąc pod uwagę liczbę i wielkość zasobów, różnorodność i zakres dziedzinowy itp.
   a. Brak lub słaba jakość
   b. Średnia jakość
   c. Dobra jakość
   d. Doskonała jakość
   e. Nie wiem

Natomiast lista ocenianych narzędzi i zasobów liczyła 29 pozycji:

1. Rozpoznawanie mowy
2. Synteza mowy
3. Segmentacja tekstu
4. Analiza morfologiczna
5. Synteza morfologiczna
6. Lematyzacja
7. Tagowanie (ujednoznacznianie morfoskładniowe)
8. Parsowanie zależnościowe
9. Parsowanie powierzchniowe
10. Parsowanie głębokie
11. Rozpoznawanie nazw własnych
12. Ujednoznaczenianie sensu słów
13. Analiza sentymentu/opinii
14. Głęboka analiza semantyczna
15. Analiza referencji
16. Analiza pragmatyczna
17. Analiza dyskursu

18. Korekta pisowni
19. Generowanie tekstu
20. Tłumaczenie maszynowe
21. Streszczanie tekstu
22. Ekstrakcja informacji
23. Korpusy języka pisanego
24. Korpusy języka mówionego
25. Korpusy równoległe
26. Słowniki elektroniczne
27. Gramatyki formalne
28. Tezaurusy
29. Wordnety

Ankieta została rozesłana do 200 osób związanych ze środowiskiem lingwistyczno-informatycznym w Polsce i zebrała 23 odpowiedzi z 14 jednostek (uczelni, instytutów badawczych i firm). Tabela 3 przedstawia wyniki sondy ograniczone do kategorii użytych w raporcie z 2011 r. i przeskalowane w sposób umożliwiający porównanie wyników obecnych z tymi sprzed 5 lat.

	Liczba	Dostępność	Jakość	Zakres	Dojrzałość	Trwałość	Elastyczność
Technologie językowe (narzędzia, technologie, aplikacje)							
Rozpoznawanie mowy	1	3	3	4	3	2	4
Synteza mowy	4	4	5	5	4	4	3
Analiza gramatyczna	4	4,5	5	4,5	4	4	3
Analiza semantyczna	1	1	2	1	1	2	2
Generowanie tekstu	1	3	2	1	1	1	2
Tłumaczenie maszynowe	3	3	2	3	3	4	3
Zasoby językowe (zasoby, dane, bazy wiedzy)							
Korpusy tekstów	3	4	4	4	5	5	3
Korpusy równoległe	3	3	3	4	5	5	5
Korpusy języka mówionego	1	3	3	3	2	2	2
Zasoby leksykalne	3	3	4	4	4	4	3
Gramatyki	3	3	3	4	3	2	2

Tabela 3. Stan dostępnych technologii językowych dla jęz. polskiego na podstawie sondy przeprowadzonej wśród przedstawicieli środowiska lingwistyczno-informatycznego

Oznaczenia zielone odpowiadają poprawie oceny, czerwone – jej pogorszeniu. Wynik negatywny należy raczej interpretować jako zmianę w sposobie postrzegania danej dziedziny

(w porównaniu z technologiami podobnego typu albo dostępnymi dla innych języków) niż faktyczny spadek jakości danego narzędzia czy zasobu. Ze względu na obszerniejszą listę kategorii warto przyjrzeć się także krótko wynikom dla technologii, które nie były uwzględnione na liście z 2011 r.: wśród nich za najbardziej dostępne (o dostępności w wystarczającym zakresie) uznano oprócz wskazanych w tabeli także mechanizmy segmentacji tekstu, tagowania i wordnety. Za niedostępne uznano z kolei narzędzia do głębokiej analizy semantycznej i pragmatycznej oraz analizy dyskursu. Najlepszą jakość przypisano natomiast technologii syntezy mowy, segmentacji tekstu, analizy i syntezy morfologicznej oraz korekty pisowni (doskonała jakość). Spośród technologii dostępnych najsłabiej oceniono jakość technologii do streszczania tekstu, a niewiele lepiej do parsowania głębokiego, ujednoznaczniania sensu słów, analizy sentymentu/opinii, generowania tekstu i, co już zostało pokazane w tabeli, tłumaczenia maszynowego. W mojej opinii oceny te dobrze odpowiadają obecnemu stanowi rozwoju wymienionych technologii.

3. Zadania dla lingwistyki komputerowej w Polsce

Postulaty, które teraz przedstawię, należy odczytać jako w dużej mierze subiektywną prezentację kierunków rozwoju lingwistyki komputerowej w Polsce. Mam nadzieję, że moje propozycje będą atrakcyjne dla obu środowisk naukowych – humanistycznego i informatycznego, pomogą odpowiedzieć na ich potrzeby i pozwolą im zbliżyć się jeszcze bardziej. W pewnym stopniu poruszają one problemy zasygnalizowane w odpowiedziach ankietowych, ale przedstawiam je jednak z perspektywy planów (i aspiracji) warszawskiego zespołu.

3.1. Korpus narodowy

Pierwsza wersja Narodowego Korpusu Języka Polskiego była organizacyjnym i badawczym przełomem, który jeszcze długo będzie wpływał na środowisko humanistyczne i lingwistyczno-informatyczne w Polsce. Prace, prowadzone w ramach projektu rozwojowego MNiSW, doprowadziły do powstania nie tylko samego zasobu korpusowego, ale także wielu narzędzi informatycznych do przetwarzania języka polskiego. Warto jednak zwrócić uwagę, że tryb projektowy doprowadził do powstania zbioru, który od zakończenia prac trwa w postaci zamrożonej – ostatnie teksty pochodzą z roku 2011, zaś automatyczne analizy lingwistyczne zostały wytworzone narzędziami kilka generacji wcześniejszymi niż używane obecnie.

Zapewnienie aktualności i stałego monitoringu NKJP, który stanowi przecież jeden z podstawowych zasobów referencyjnych dla rzeszy użytkowników, wydaje się koniecznością. A potrzeb jest przecież dużo więcej: równie ważne wydaje się wielokierunkowe rozszerzanie bazy materiałowej, opracowanie metod reprezentacji i wizualizacji danych lingwistycznych, rozwój narzędzi analitycznych i eksploracyjnych. Próbę dokładniejszego zdefiniowania potrzeb w tym zakresie podjęliśmy już w ramach grupy roboczej „Korpusowa infrastruktura badawcza dla polszczyzny" powołanej w ramach konsorcjum DARIAH-PL, warto zatem powtórzyć kierunki, w których widzimy rozwój korpusu narodowego:

1. Rozszerzenie bazy materiałowej Narodowego Korpusu Języka Polskiego (o podkorpusy diachroniczne, podkorpusy równoległe, specjalistyczne, w tym gwarowy, dane mówione, podkorpus synchroniczny tworzony na bazie bibliotek cyfrowych i Internetu).
2. Opracowanie korpusowych standardów znakowania dla języka polskiego:
    a. opracowanie formatu znakowania wszystkich tekstów (także historycznych i gwarowych) rozszerzonymi metadanymi i znacznikami morfosyntaktycznymi,
    b. opracowanie formatu znakowania tekstów współczesnych znacznikami składniowymi, semantycznymi i dyskursywnymi.
3. Opracowanie metod reprezentacji danych multimedialnych:
    a. opracowanie metod zrównoleglania tekstów ze źródłami (dane mówione, wideo, skany).
    b. opracowanie metod wyszukiwania w plikach źródłowych.
4. Opracowanie interfejsu użytkownika korpusu uwzględniającego proste formułowanie zapytań, oraz atrakcyjną i czytelną wizualizację wyników:
    a. opracowanie intuicyjnych nakładek graficznych i interakcyjnych,
    b. opracowanie metod wizualizacji zaawansowanych struktur lingwistycznych (drzewa składniowe, koreferencje itp.)
5. Rozwój narzędzi do samodzielnego tworzenia korpusów.
6. Rozwój narzędzi do eksploracji i analizy danych korpusowych.
7. Monitoring źródeł internetowych na potrzeby korpusowe i leksykograficzne.

O istotności prac w każdym z tych kierunków świadczy fakt uruchomienia wielu projektów finansowanych ze środków Narodowego Programu Rozwoju Humanistyki oraz Narodowego

Centrum Nauki, w ramach których powstają obecnie korpusy diachroniczne: XVI w., XVII–XVIII w. i XIX w., korpus polskich tekstów prasowych (aktualnie z lat 1945–1954) czy korpus gwarowy. Równolegle do tych prac powstają również reprezentacje zaawansowanych struktur lingwistycznych w rodzaju relacji referencyjnych czy metafor wraz z narzędziami analitycznymi. Projekty te, nawet jeśli wykorzystują dane korpusu narodowego, są jednak prowadzone niezależnie, bez intencji jego wzbogacenia, co wydaje mi się dużą stratą. A przecież podobnie do Wielkiego Słownika Języka Polskiego (Żmigrodzki 2015b) wielki korpus polszczyzny jest podstawowym zasobem językowym, który powinien istnieć w sposób stały i stanowić rodzaj „instytucji narodowej". Prace nad nim nie mogą ograniczać się do jednego projektu – wymagają nadzoru przez środowisko humanistyczne, doradztwa naukowego na najwyższym poziomie i, co trzeba powiedzieć, stałego finansowania. Szczególnie ten ostatni postulat wydaje się trudny do spełnienia, jak mogłoby się wydawać po lekturze długiej listy potrzeb, jednak najpilniejsza kwestia rozpoczęcia prac nad utrzymaniem korpusu nie wydaje się aż tak mało realistyczna i w moim przekonaniu wystarczyłaby do niej jednoczesna realizacja trzech celów:

1. zapewnienie ciągłości życia NKJP, co stanowiłoby spełnienie części postulatów autorów monografii o jego ustawicznym nadzorowaniu i monitorowaniu oraz uzupełnianiu i ulepszaniu, w fazie wstępnej ograniczonym do dostosowania opisów lingwistycznych do najnowszych narzędzi analitycznych; tego rodzaju działania wymagałyby pracy zaledwie kilku osób,
2. zapewnienie nadzoru nad długofalowymi kierunkami rozwoju NKJP przez radę programową złożoną z wybitnych językoznawców,
3. opracowanie metody konstruowania grantów obecnych jednostek współfinansujących badania podstawowe w sposób nagradzający przejmowanie wyników tych projektów do zasobu korpusu narodowego i zapewniający finansowe wsparcie tego procesu.

Powyższa formuła jest oczywiście luźną propozycją, której celem jest rozpoczęcie dyskusji nad stworzeniem zasad funkcjonowania i rozwoju korpusu narodowego. Na jej bazie chciałbym niniejszym rozpocząć lobbing na rzecz tego ważnego zasobu i poprosić czytelników o poparcie dla tej inicjatywy.

3.2. Nowe tematy – i mozolnie naprzód

Drugim wg mnie istotnym kierunkiem prac jest formalny opis polszczyzny w zakresie od dawna podejmowanym dla języków zachodnich, a u nas wciąż traktowany jako pieśń przyszłości – mam tu przede wszystkim na myśli kwestie związane z analizą dyskursu rozumianą jako generowanie struktur wypowiedzi wykraczających poza poziom zdania. Temat ten wpisuje się w prowadzone obecnie prace nad komputerowym opisem takich zjawisk jak referencja, metafora czy argumentacja w języku.

Oczywiście mój subiektywny punkt widzenia może nie odzwierciedlać istotności tematów dla pozostałej części lingwistyczno-informatycznego środowiska, sięgnę zatem raz jeszcze do ankiety, w której zadałem też trzecie pytanie, prosząc o podanie informacji nt. planów rozwoju danej technologii w zespole respondenta, ze wspomnianą wyżej szczegółową listą 29 technologii i następującymi odpowiedziami:

1. Jakość technologii jest wystarczająca, nie ma potrzeby prowadzenia dalszych prac
2. Obecnie prowadzimy prace w tej dziedzinie
3. Zamierzamy rozwijać narzędzia/ zasoby z tej dziedziny w przyszłości
4. Temat jest ważny, ale nie planujemy prowadzić prac w tej dziedzinie
5. Nie uważamy tego tematu za istotny
6. Trudno powiedzieć

Intencją pytania było sprawdzenie, które problemy uważamy jako środowisko za rozwiązane, które są dla nas ważne (bo prowadzimy już prace w danej dziedzinie, zamierzamy je prowadzić lub uważamy temat za istotny), a które nie są warte badawczej uwagi. Ciekawym wynikiem sondy okazało się stwierdzenie, że praktycznie nie ma dziedzin, które zostałyby uznane za nieważne; za kwestie rozwiązane największa liczba ankietowanych uznała korektę pisowni (5 odpowiedzi z 23 udzielonych), dostępność korpusów języka pisanego (4), narzędzi do segmentacji tekstu, lematyzacji, analizy morfologicznej oraz takich zasobów jak korpusy równoległe i wordnety (3). Za tematy ważne uznano tłumaczenie maszynowe (19), rozpoznawanie mowy (17), syntezę mowy, tagowanie, analizę sentymentu/opinii i ekstrakcję informacji (16), a także streszczanie tekstu, rozwój korpusów języka pisanego, słowników elektronicznych i gramatyk formalnych (15 odpowiedzi). Wskazania te nie zmieniły się znacząco przy ograniczeniu odpowiedzi do technologii „przyszłościowych", czyli odjęciu odpowiedzi uwzględniających aktualnie prowadzone prace. W kontekście poprzedniego

rozdziału komentarza wymaga obecność na obu listach korpusów języka pisanego; ocena ich dobrej dostępności może wynikać z zaufania do NKJP z jednoczesnym wskazaniem na potrzebę rozwoju korpusów dla polszczyzny – w tym również korpusu narodowego.

Oczywiście przy okazji kreślenia planów na przyszłość nie można pominąć wątku stałego poprawiania jakości dostępnych narzędzi i zasobów. Niedościgły w niektórych dziedzinach poziom 90% dokładności jest dla wielu celów niewystarczający, gdyż nawet jedno błędne wskazanie na każde dziesięć to zbyt wiele, zwłaszcza gdy wynik jednego narzędzia (np. ujednoznacznienie części mowy w przetwarzanym tekście) jest źródłem danych dla kolejnego (np. analizatorze składniowym wykorzystującym informację o częściach mowy). Poprawa narzędzi działających na różnych poziomach wiedzy lingwistycznej – analizatorów składniowych i semantycznych, mechanizmów do wykrywania nazw własnych, koreferencji – choć mało spektakularna, jest niezwykle ważna dla możliwości wykorzystania ich wyników w praktyce.

Oprócz dziedziny stricte badawczej nie można też zapomnieć o niemniej ważnej części „usługowej". Szlak jest już przetarty; środowisko informatyczne bierze udział w dwóch ważnych inicjatywach dot. humanistyki cyfrowej – infrastrukturach badawczych CLARIN-PL i DARIAH-PL. CLARIN-PL jest inicjatywą środowiska informatycznego wspierającą projekty badawcze w naukach humanistycznych i społecznych poprzez umożliwienie wykorzystania dostępnych zasobów i narzędzi w zadaniach związanych z przetwarzaniem języka. Inicjatorem DARIAH-PL jest natomiast środowisko humanistyczne działające na rzecz tworzenia, rozwoju i utrzymania infrastruktury humanistyki cyfrowej w Polsce. Co ważne, instytucje reprezentujące główne polskie centra badawcze zajmujące się lingwistyką informatyczną[6] są członkami obu inicjatyw, włączając się w badania humanistyczne w zakresie własnych zainteresowań i dostępnych środków.

Mimo wielu kłopotów właściwych polskiej nauce, polskie środowisko lingwistyki informatycznej wydaje się działać bardzo prężnie, a trzy główne tematy: składnia, semantyka i dyskurs wydają się wyznaczać kierunki prac na najbliższe lata. Muszę jednak po raz kolejny zastrzec, że powyższy tekst przedstawia przede wszystkim moje osobiste poglądy, a w drugiej kolejności (na podstawie odpowiedzi z sondy) punkt widzenia informatyków zajmujących się przetwarzaniem języka, może zatem nie odzwierciedlać stanowiska użytkowników, których

[6] P. http://clip.ipipan.waw.pl/Centers.

priorytety mogą być zgoła inne. W związku z tym zachęcam humanistów do współpracy, dzielenia się swoimi problemami, ale także wynikami własnych prac. To najlepsza metoda na poprawianie jakości narzędzi dla polszczyzny, z których oba środowiska będą mogły korzystać.

Bibliografia

Bień J. S. 2000: *Zestaw testów do weryfikacji i oceny analizatorów języka polskiego*, Tekst wystąpienia podsumowującego projekt KBN, Zakopane 2000.

Jaworski W. i Przepiórkowski A. 2014: *Syntactic approximation of semantic roles*, [w:] A. Przepiórkowski i M. Ogrodniczuk (red.): *Advances in Natural Language Processing,* Lecture Notes in Artificial Intelligence, t. 8686, s. 193–201. Springer International Publishing, Heidelberg.

Ogrodniczuk M., Przepiórkowski A. 2013: *CLIP – portal internetowy łączący projekty, narzędzia, zespoły związane z komputerowym przetwarzaniem języka polskiego*. Język Polski XCIII (3), s. 242.

Miłkowski M. 2012: *Język polski w erze cyfrowej*. Springer, Berlin–Heidelberg. http://doi.org/10.1007/978-3-642-30811-6.

Patejuk A., Przepiórkowski A. 2012: *Towards an LFG parser for Polish: An exercise in parasitic grammar development*. [w:] Materiały konferencji LREC 2012 (8th International Conference on Language Resources and Evaluation), ELRA, Stambuł, s. 3849–3852.

Pęzik P. 2015: *Spokes – a Search and Exploration Service for Conversational Corpus Data,* [w:] Selected Papers from CLARIN 2014, s. 99–109. Linköping University Electronic Press, Linköpings universitet.

Pęzik P. 2016: *Exploring Phraseological Equivalence with Paralela,* [w:] E. Gruszczyńska i A. Leńko-Szymańska (red.): *Polish-Language Parallel Corpora*, s. 67–81. Warszawa, Instytut Lingwistyki Stosowanej UW.

Piasecki M. 2008: *Cele i zadania lingwistyki informatycznej*, [w:] P. Stalmaszczyk (red.): *Metodologie językoznawstwa. Współczesne tendencje i kontrowersje*, Lexis, Kraków.

Piasecki M., Maziarz M., Szpakowicz S., Rudnicka E. 2014: *PlWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources*, [w:] Materiały konferencji GWC 2014 (7th International Global Wordnet Conference), s. 304–312.

Przepiórkowski A., Bańko M., Górski R. L. i Lewandowska-Tomaszczyk B. (red). 2012: *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Przepiórkowski A., Skwarski F., Hajnicz E., Patejuk A., Świdziński M., Woliński M. 2014: *Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego*, Polonica, XXXIII, s. 159–178.

Przepiórkowski A. 2015: *Inżynieria lingwistyczna a obecna sytuacja językoznawstwa polskiego*, LingVaria 10 (2), s. 135–145.

Saloni Z., Woliński M., Wołosz R., Gruszczyński W., Skowrońska D. 2007/2017: *Słownik gramatyczny języka polskiego*, Warszawa, wydanie III.

Strik H., Daelemans W., Binnenpoorte D., Sturm J., de Vriend F., Cucchiarini, C. 2002: *Dutch HLT resources: From BLARK to priority lists*, [w:] Materiały konferencji ICSLP (7th International Conference on Spoken Language Processing), Denver, USA, s. 1549–1552.

Świdziński M. 2006: *Lingwistyka korpusowa w Polsce – źródła, stan, perspektywy,* LingVaria 1, s. 23–32.

Webber B. L. 2001: *Computational perspectives on discourse and dialogue*, [w:] D. Schirin, D. Tannen i H. Hamilton (red.): *The Handbook of Discourse Analysis*, Blackwell Publishers Ltd.

Woliński M. 2015: *Deploying the new valency dictionary Walenty in a DCG parser of Polish*, [w:] M. Dickinson, E. Hinrichs, A. Patejuk i A. Przepiórkowski (red.): Materiały konferencji TLT 14 (14th International Workshop on Treebanks and Linguistic Theories), s. 221–229, Warszawa, Instytut Podstaw Informatyki PAN.

Żmigrodzki P. 2015a: *Wielki słownik języka polskiego PAN a obecna sytuacja językoznawstwa polskiego*, LingVaria 19 (1/2015), s. 109–119.

Żmigrodzki P. 2015b: *Wielki słownik języka polskiego PAN – historia, stan obecny i perspektywy rozwoju po 2018 roku*, Biuletyn Polskiego Towarzystwa Językoznawczego 71, s. 177–187.

Streszczenie w języku polskim

Tekst jest publicystyczną próbą nakreślenia dalszych kierunków prac nad komputerowym przetwarzaniem polszczyzny w obliczu intensywnego rozwoju cyfrowych narzędzi i zasobów dla języka polskiego oraz zacieśniającej się współpracy między polskimi ośrodkami badawczymi zajmującymi się lingwistyką komputerową. Za najważniejszy temat autor uważa wznowienie prac nad korpusem narodowym, który jako zasób podstawowy dla językoznawstwa polskiego wymaga stałego poszerzania bazy materiałowej i opisu lingwistycznego, włączenia podkorpusów diachronicznych, gwarowych i równoległych.
W sferze technologii językowej autor postuluje wzbogacenie formalnego opisu polszczyzny o głęboki poziom składniowy, semantykę i dyskurs oraz zwraca uwagę na konieczność stałego poprawiania jakości dostępnych narzędzi i zasobów metodą współpracy środowiska językoznawczego z informatycznym.

Streszczenie w języku angielskim

Tytuł: Natural language processing for Polish: today and tomorrow

Słowa kluczowe: computational linguistics, natural language processing, language technology, language resources and tools for Polish

The article attempts at framing directions for future work on computational processing of Polish in the face of recent intensive development of electronic tools and resources and close co-operation between Polish research centres involved in computational linguistics. The author regards renewing the work on the National Corpus of Polish as the most important topic, naming it the basic resource for Polish linguistics and listing the most urgent objectives: extension of the sources and linguistic representation as well as inclusion of diachronic, dialectical and parallel corpora. With respect to language technology, the author cals for enrichment of formal description of Polish with syntactic, semantic and discourse-feature representation and constant improvement of quality of tools and resources by means of co-operation between linguists and computer scientists.