„LingVaria”
Rok I (2006) nr 1

Marek Świdziński
Uniwersytet Warszawski

LINGWISTYKA KORPUSOWA W POLSCE – ŹRÓDŁA, STAN, PERSPEKTYWY

Wstęp

Niniejszy szkic, adresowany przede wszystkim do młodego pokolenia lingwistów-polonistów, przedstawia wizję językoznawstwa XXI stulecia. Jest to mianowicie lingwistyka korpusowa. Rozwija się ona już od jakiegoś czasu; wynikła w sposób naturalny z dość szczęśliwego splotu różnorodnych okoliczności; zdominowała inne nurty i odmiany językoznawstwa nieodwracalnie. Nie można być dziś językoznawcą i nie otrzeć się o nią choćby jako użytkownik narzędzi. A skoro tak, to warto wejść w ten świat z wyboru, zadań bowiem jest moc i będzie ich coraz więcej. W świecie jest na tę najnowszą lingwistykę, nie na dowolną, mnóstwo pieniędzy – i zajmują się nią całe rzesze; to dowodzi, że coś jest na rzeczy, że czemuś to służy... Niestety, nie w naszej ojczyźnie.

Inżynierię korpusową przyniosła najmłodsza z długiej serii XX-wiecznych rewolucji – rewolucja informatyczna. Ale przed nią zdarzyły się dwie inne rewolucje intelektualne. Marsz przez te rewolucje stał się udziałem pokoleń językoznawców. Drogę tę przeszedłem i ja.

Rewolucja nr 1: strukturalizm

Językoznawstwo jako samodzielna dyscyplina liczy sobie niewiele ponad sto lat. Choć w minionych tysiącleciach znaleźć można wielu ważnych prekursorów, od Paniniego i Arystotelesa poprzez gramatyków starożytnego Rzymu po Arnaulda i Lancelota, twórców *Gramatyki Port-Royal*, to lingwistyka teoretyczna zrodziła się u schył-

ku XIX stulecia. Pierwsza rewolucja jest dziełem Ferdynanda de Saussure'a, językoznawcy szwajcarskiego, profesora uniwersytetów w Genewie i Paryżu. Jemu, jego współpracownikom i wychowankom (którzy zresztą sami spisali i wydali wykłady mistrza), a także grupie wybitnych następców zawdzięczamy sformułowanie koncepcji języka naturalnego jako systemu semiotycznego: to dwuklasowy system znaków służący określonej populacji do komunikacji uniwersalnej. Lingwistyka dostała aparat, który jest dziś oczywistością:

(a) synchronia przede wszystkim,
(b) abstrakcyjny system (*langue*) i konkretny tekst (*parole*),
(c) znak – obiekt o dwóch twarzach: ma kształt i funkcję,
(d) opozycja – różnica kształtu obciążona funkcjonalnie,
(e) paradygmatyka i syntagmatyka,
(f) słownik – zbiór znaków prostych, gramatyka – zbiór instrukcji syntezy i rozbioru wyrażeń,
(g) inwarianty i warianty.

Teoria de Saussure'a dotarła do Polski z górą pół wieku od jej powstania – przekład polski jego wykładów ukazał się w roku 1961 (Saussure 1961). Koryfeuszy ówczesnego językoznawstwa w Polsce nie zainteresowała.

Strukturalizm, który wyrósł z idei de Saussure'a, wyzwolił językoznawstwo – naukę empiryczną – z oków humanistycznej filologii. Filolodzy zajmowali się przez stulecia tekstami (czy kawałkami tekstów) i pochodzeniem; strukturalizm podjął problem budowy i funkcji wyrażeń. Strukturalny opis pewnego obiektu przyrodniczego jako pary <słownik, gramatyka> daje model rodzimego użytkownika języka; elementy tej pary to składniki kompetencji językowej. Doktryna strukturalna opanowała świat lingwistyki pierwszej połowy zeszłego stulecia, a myślenie systemowe, rzecz ciekawa, promieniowało na inne działy humanistyki.

Nie miejsce tu na wykład różnych szkół europejskiego językoznawstwa strukturalnego. Trzeba wszakże wspomnieć o strukturalizmie amerykańskim, czyli dystrybucjonizmie, od niego bowiem do NLP (*Natural Language Processing*) tylko krok. Dystrybucjoniści to pierwsi teoretycy, którzy budowali korpusy – zbiory wyrażeń traktowane jako reprezentujące dany język naturalny. Od nich pochodzi inne niż nasze europejskie rozumienie języka naturalnego. Jest nim zbiór zdań poprawnych i tylko takich. Opis (czy model) danego języka to recepta na wyrażenia tego języka.

Strukturaliści, zapewne jako pierwsi w historii lingwistyki, podjęli się sporządzania wyczerpujących opisów różnych języków naturalnych, opisów *całego* języka. Za przykład niech posłuży monumentalna gramatyka angielska Ottona Jespersena (1909–1949). To, że nowoczesna lingwistyka rozwinęła się najowocniej w kręgu anglosaskim, jest pewnie zasługą Jespersena. Warsztat strukturalizmu to pierwszy składnik kompetencji współczesnego językoznawcy.

Rewolucja nr 2: generatywizm i lingwistyka formalna

Drugą rewolucję potrafimy dokładnie datować. W roku 1957 ukazały się w wydawnictwie Moutona *Struktury składniowe* Noama Chomsky'ego (1957). Ów „przewrót

kopernikański" polegać miał na odrzuceniu tradycji strukturalistycznej. Chomsky uważał, że strukturalizm nie ujawnia tego, że kompetencja językowa jest produktywna; że, innymi słowy, użytkownik języka potrafi interpretować wyrażenia, których nigdy nie słyszał, i nowe wyrażenia produkować. Ale Chomsky'ego krytyka strukturalizmu (niezbyt zresztą sprawiedliwa) dotyka co najwyżej dystrybucjonizmu, i to raczej jego litery. Kamieniem obrazy jest dla Chomsky'ego ograniczoność korpusów, którymi posiłkowali się dystrybucjoniści. Dla nich korpus był źródłem danych empirycznych. Ponieważ podejmowali trud opisywania różnych języków dotąd nie opisywanych, w szczególności języków Indian, którzy wymierali, korpusy z natury rzeczy nie mogły być duże. Wielkich zresztą nie dałoby się objąć oglądem.

Chomsky, krytyk dystrybucjonizmu, korzysta jednak szeroko z tamtej aparatury pojęciowej. Przede wszystkim, idąc śladem dystrybucjonistów, ujmuje język naturalny teoriomnogościowo: to zbiór wszystkich możliwych zdań, zbiór nieskończony. Opis języka, a więc jego gramatyka, jest tego zbioru definicją. Definicję taką nazywamy gramatyką formalną. Słownik jako zbiór pewnych składników prostych należy do gramatyki.

Już od półwiecza Chomsky nosi szatę guru współczesnej lingwistyki. Z gramatyki generatywno-transformacyjnej, której ideę wyłożył w *Stukturach składniowych*, wyrosły kolejne jej mutacje – rozszerzona teoria standardowa (EST), wprowadzona w *Aspektach teorii składni* Chomsky'ego (1966), a także teoria rządu i wiązania (GB; Haegemann 1992) oraz minimalizm; te dwie ostatnie – z nieistotną dla nas tutaj filozoficzną obudową Gramatyki Uniwersalnej. Zrodziły się też w ciągu dziesięcioleci inne teorie, znacznie lepiej dopracowane formalnie, zwłaszcza HPSG (Pollard i Sag 1994). W ramach tych aparatów powstała i powstaje gigantyczna literatura na temat najrozmaitszych języków, od staroislandzkiego po warlpiri. Można powiedzieć bez przesady, że duża część populacji lingwistów na świecie działa w kręgu generatywizmu chomskiańskiego. Nie dotyczy to, niestety, Polski, w której ziemię tę uprawiają niemal wyłącznie angliści. Opisują oni zresztą głównie polszczyznę; obszerny zestaw odesłań do publikacji polskich generatywistów znaleźć można na przykład w tomie studiów poświęconych HPSG (Przepiórkowski i in. 2002). Znamienne, że polski przekład *Aspektów* Chomsky'ego (1982) przeszedł bez echa. Dopiero ostatnio pojawiło się popularne kompendium generatywizmu (Mecner 2004). Generatywizm nie stworzył wszakże wielkich syntez, a więc wyczerpujących opisów poszczególnych języków; choćby tych najważniejszych. Dobra znajomość narzędzi generatywizmu to drugi składnik kompetencji współczesnego językoznawcy.

Rewolucja nr 3: lingwistyka informatyczna

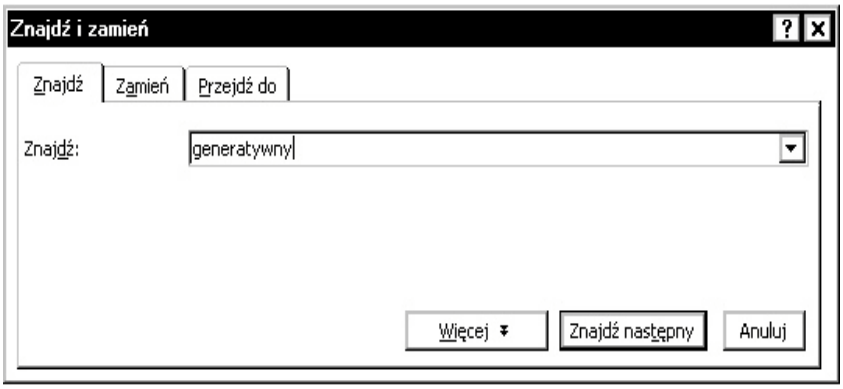
O ile dwie poprzednie rewolucje wynikły, by tak rzec, w toku normalnego rozwoju myślowego pewnej dyscypliny, rewolucja ostatnia przyszła z zewnątrz, i to bardzo niedawno. Komputery, jeszcze w połowie ubiegłego wieku pracujące w Pentagonie, agencjach kosmicznych czy ośrodkach obliczeniowych, trafiły pod strzechy, aby stać się standardowym urządzeniem gospodarstwa domowego. Co więcej, w ciągu paru-

nastu lat wymarł pewien fach: zawód zecera. Skład komputerowy to wyrok śmierci dla drukarstwa Gutenberga.

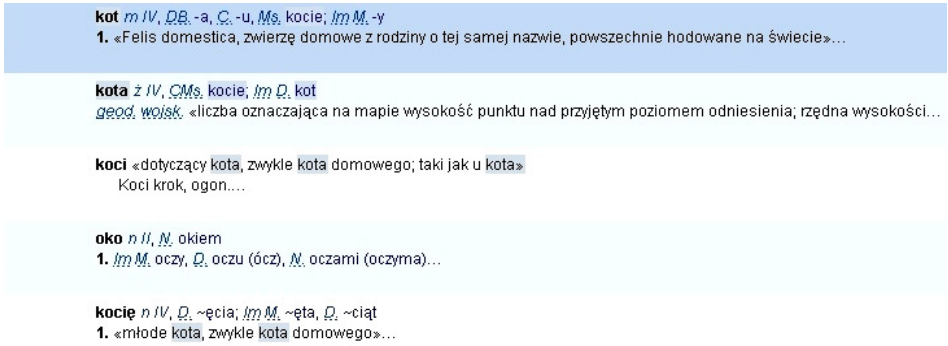
I jeszcze jeden zbieg okoliczności. Oto w latach 70. rozpoczęła się współpraca między grupą informatyków z Wydziału Matematyki Uniwersytetu Warszawskiego i grupą językoznawców Wydziału Polonistyki. Do tej kooperacji obie strony były wtedy dobrze przygotowane, podobnie jak później – do podjęcia zaawansowanych prac w zakresie lingwistyki informatycznej. Niżej będzie mowa o niektórych przedsięwzięciach, które wyrosły z owego zbliżenia dwóch środowisk. Pokażę tu w szczególności prace powstałe w środowisku warszawskim, zwłaszcza w Zakładzie Językoznawstwa Komputerowego, którym kieruję. Skupienie się tutaj na nich nie ma oznaczać, że nigdzie indziej nie działo się i nie dzieje nic, co dla dyscypliny ważne.

Trzy ważne czynniki: eliminacja tradycyjnych technik wydawniczych, błyskawiczny rozwój pamięci masowych i, ostatnio, eksplozja internetu – dały w efekcie dostęp do gigantycznych zbiorów tekstowych. Ręczna ich obsługa jest absolutnie niemożliwa. Nie chodzi przy tym o materiał empiryczny dla naukowców. Teksty są w zasięgu ręki szarego obywatela. Monstrualne zasoby informacji na serwerach internetowych mają przede wszystkim postać zbiorów tekstowych. Narzędzia operowania takimi zbiorami to wyzwanie nie tylko dla informatyka; także (a nawet przede wszystkim) dla lingwisty.

Dostęp do zawartości nośnika umożliwiają specjalne aplikacje – różnego rodzaju wyszukiwarki. Kiedy piszemy tekst w procesorze komercyjnym, pozwala nam on lokalizować interesujące nas napisy: obiekty unilateralne. Wystarczy wpisanie odpowiedniego ciągu znaków, aby dotrzeć do wszystkich tej sekwencji wystąpień. Szukanie jest trywialne, dotyczy bowiem kształtów:



Kiedy jednak korzystamy z encyklopedii czy słownika elektronicznego, nie chodzi nam zwykle o dany napis. Oto wynik poszukiwania jednostki kot w *Komputerowym słowniku języka polskiego* (KSJP):



Szukaliśmy tutaj jakiejś jednostki bilateralnej – zapewne leksemu. KSJP ujawnił nam artykuły hasłowe, w których użyto formy leksemu KOT. Co prawda, jedno z pięciu znalezisk odrzucimy, bo nie chodziło nam raczej o leksem KOTA... Otóż większość wyszukiwań daje rezultaty niechciane; taki jest po prostu język naturalny, który roi się od neutralizacji. Ambitniejsze narzędzia kwerend tekstowych służą poszukiwaniom nie „po kształtach" , jak się mawia w żargonie, tylko po jakichś znacznikach. Wprowadzenie do tekstu znaczników (tagów), które pomogą w ujednoznacznieniach, jest zadaniem lingwistycznym. Świadomość istnienia narzędzi obsługi tekstów to trzeci komponent kompetencji współczesnego językoznawcy.

Wyzwanie homonimii

Zbiór tekstów przygotowany specjalnie do jakiegoś celu nazywamy korpusem. Korpusami posługują się językoznawcy, przede wszystkim leksykografowie. Korpusy lingwistyczne bywają znakowane, czyli wzbogacone przynajmniej o informację gramatyczną; docelowo – pewnie pragmatyczną i semantyczną.

Polszczyzna reprezentuje klasę języków wysoce fleksyjnych. Cechą znamienną takich języków jest homonimiczność słów. Ostrożny szacunek poucza, że w tekście polskim 40 słów na sto to homonimy, czyli słowa będące kształtami więcej niż jednej jednostki systemowej. Fundamentalnym zadaniem lingwistyki informatycznej jest zatem rozwiązywanie homonimii: słabe – przypisanie słowom analizowanego tekstu wszystkich interpretacji; mocne – znalezienie interpretacji właściwej (zob. Świdziński, Derwojedowa i Rudolf 2003).

Warto zaznaczyć, że jedno z pierwszych w świecie przedsięwzięć lingwistyki korpusowej miało miejsce w Polsce i polszczyzny dotyczyło. W latach 1967–1971 powstał w Uniwersytecie Warszawskim półmilionowy zrównoważony korpus znakowany, który posłużył za bazę empiryczną słownika frekwencyjnego języka polskiego. Znakowania dla ujednoznacznienia słów dokonywano ręcznie, ale listy frekwencyjne zostały sporządzone komputerowo. Podstawy gramatyczne projektu były tak solidne, że zachowały aktualność po dziś dzień. Słownik ukazał się najpierw w postaci pięciu tomów (w jedenastu woluminach) pod tytułem *Słownictwo współ-*

czesnego języka polskiego. Listy frekwencyjne (S-LF). Tomy te wyszły potem w postaci zbiorczej pod redakcją Zygmunta Saloniego jako *Słownik frekwencyjny polszczyzny współczesnej* (SFPW). SFPW jest słownikiem form wyrazowych popakowanych w leksemy. Twórcy korpusu przypisywali ręcznie znaczniki słowom, które są homoformami (Awramiuk 1999). Nie jest to zatem znakowanie pełne. Ale początek został uczyniony.

Analizatory i wyszukiwarki

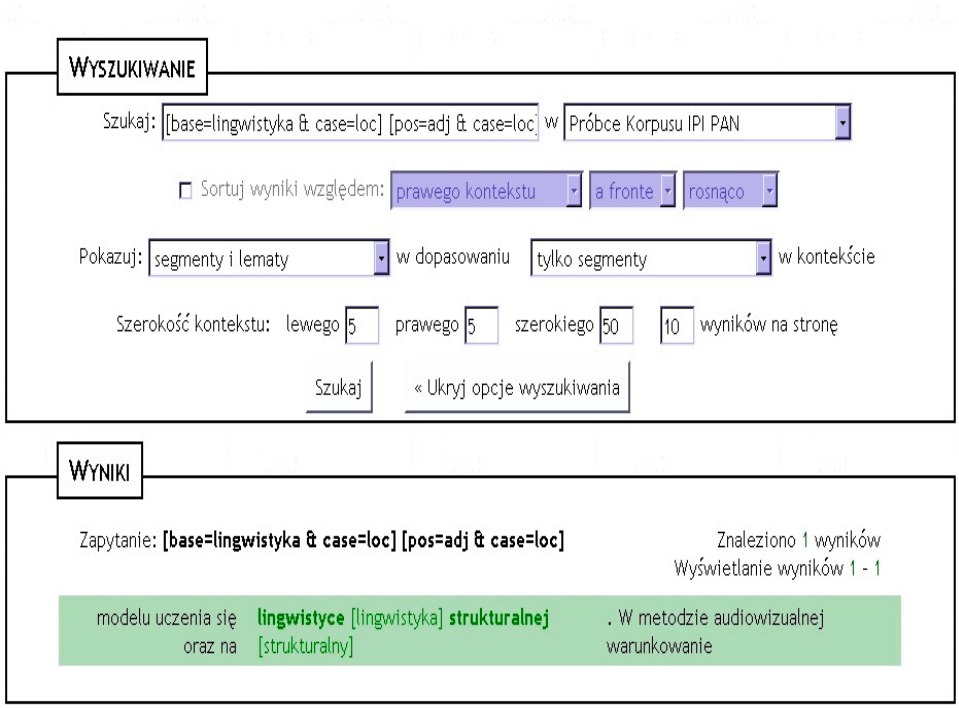
Urządzenie do automatycznego rozwiązywania homonimii to analizator morfologiczny. Musi on opierać się na rygorystycznym opisie gramatycznym danego języka. Dorobek gramatyczny językoznawstwa tradycyjnego, z gramatykami Doroszewskiego, Szobera czy Klemensiewicza na czele, nie spełniał oczywiście warunków pełności i jawności. Polszczyzna doczekała się jednak szczęśliwie zadowalających opisów morfologicznych i składniowych – wymieńmy prace Jana Tokarskiego (SJPD. z tzw. „notacją Tokarskiego", Tokarski 1973 i 1990), Zygmunta Saloniego (1992; 2004), Saloniego i Świdzińskiego (2001), Włodzimierza Gruszczyńskiego (1989), Janusza S. Bienia (1991), a także, z innej szkoły, morfologię z *Gramatyki języka polskiego PAN* (Gramatyka PAN 1984). Morfologię można już było zaimplementować.

Istnieje kilka analizatorów morfologicznych. U schyłku lat 80. powstał analizator Roberta Wołosza, znany dziś pod nazwą Pomor (zob. Wołosz 2005), analizator SAM Krzysztofa Szafrana (1994), Morfeusz Marcina Wolińskiego (2004a), w końcu – AMOR Joanny Rabiegi-Wiśniewskiej i Michała Rudolfa (2003). Analizatory te przypisują słowom zbiory interpretacji gramatycznych.

Analizator dostaje słowo lub listę słów do interpretacji. AMOR na przykład zinterpretuje słowo *jutro* jako należące do leksemu przysłówkowego JUTRO_1 lub rzeczownikowego JUTRO_2, czyli dokona rozpoznania części mowy (*PoS-tagging*) oraz rozpoznania leksemu, do którego forma wyrazowa o takim kształcie należy (*lemmatization*); słowo *szkoły* – jako reprezentujące cztery formy wyrazowe: dopełniaczową w liczbie pojedynczej bądź mianownikową, biernikową albo wołaczową w mnogiej; słowo *czytali* – jako formę wyrazową czasownika CZYTAĆ z pewnym opisem gramatycznym. Program, który zwraca analizowany tekst z odpowiednimi znacznikami poprzypisywanymi wszystkim słowom, nazywany bywa tagerem (*tagger*), a efektem pracy takiego programu jest tekst (czy korpus) znakowany. Na korpusie, znakowanym lub nie, pracują dopiero zaawansowane wyszukiwarki.

Ostatnio zakończyły się dwa projekty naukowo-badawcze, których celem było już to zbudowanie korpusu znakowanego, już to opracowanie narzędzi do obsługi korpusu. Pierwszy z nich realizowany był w Instytucie Podstaw Informatyki PAN pod kierunkiem Adama Przepiórkowskiego. W ramach projektu KBN 7T11C 043 20 powstał w latach 2001–2004 100-milionowy anotowany korpus tekstów polskich (Korpus IPI PAN), który nie ma ambicji bycia korpusem lingwistycznie reprezentatywnym, czyli na przykład zrównoważonym; powstała też wyszukiwarka Poliqarp (zob. Przepiórkowski 2004).

Oto pokaz wyszukiwania:

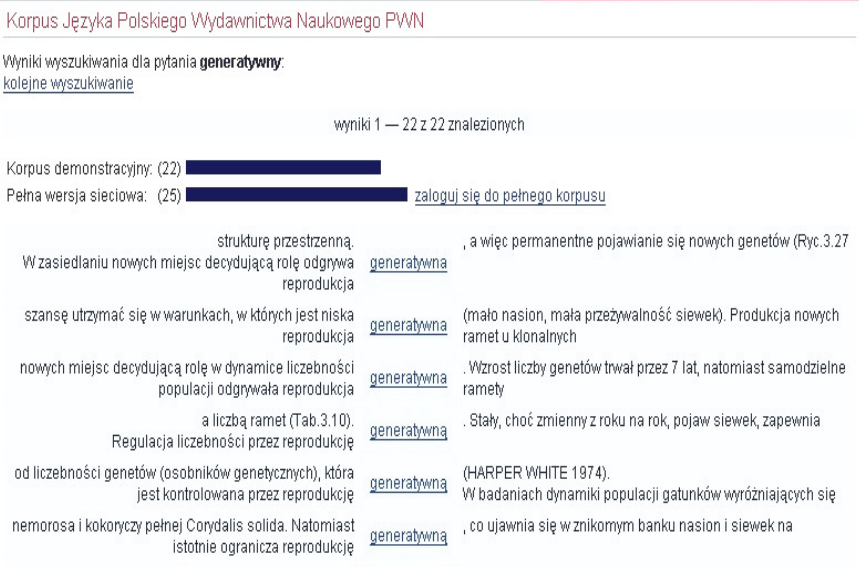


Użytkownik oczekuje przykładów wystąpienia formy miejscownikowej leksemu LINGWISTYKA, po której bezpośrednio następuje miejscownikowa forma wyrazowa przymiotnikowa. Składnia poleceń, jaką dysponuje Poliqarp, jest bardzo rozbudowana, co umożliwia formułowanie wyrafinowanych warunków boole'owskich. Poliqarp nie jest jednak skuteczną maszyną ujednoznaczniania mocnego.

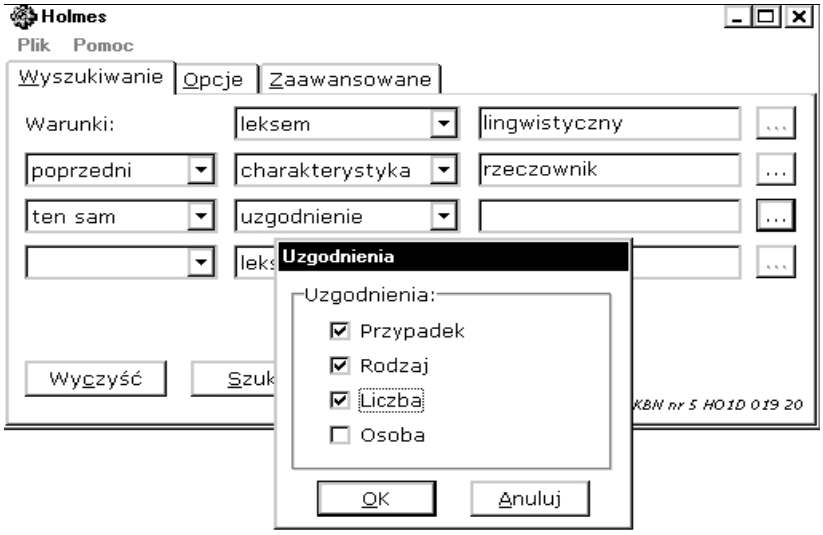
Drugi projekt, kierowany przez Andrzeja Markowskiego, z udziałem m.in. Marka Świdzińskiego i Mirosława Bańki, rozwijał się w tym samym czasie w Instytucie Języka Polskiego Uniwersytetu Warszawskiego – we współpracy z Redakcją Słowników PWN[1]. Redakcja umożliwiła dostęp do obszernych fragmentów własnego korpusu (Korpus PWN). Zrównoważone jego wycinki o długości od 2 do 40 mln słów służyły jako podstawa dla prac programistycznych i testerskich. Korpus PWN z własną wyszukiwarką dostępny jest w internecie oraz na płycie CD.

Oto wynik wyszukiwania leksemu GENERATYWNY w internetowej wersji demo:

[1] Grant KBN 5 HO1D 019 20.



Dla potrzeb leksykografów z Redakcji Słowników PWN stworzono w ramach projektu oprogramowanie służące obsłudze wielkich korpusów tekstów polskich. Jest ono dziełem Michała Rudolfa. Produkt końcowy stanowi aplikacja Sherlock, pracująca w środowiskach Linux, DOS oraz Windows (zob. Rudolf 2004; Świdziński i Rudolf, w druku). Wizualizację wyników umożliwia aplikacja okienna Holmes. Wyniki podawane są natychmiast, między innymi w postaci plików HTML. Oto przykładowe wyszukiwanie leksemów rzeczownikowych, które występują przed formą leksemu LINGWISTYCZNY, z uzgodnieniem przypadka, rodzaju i liczby (leksykograf, być może, szuka dokumentacji ilustrującej łączliwość przymiotnika LINGWISTYCZNY):



I wynik kwerendy:

1. Rejwach głuszy słowa tłustego chłopca w okularach, który usiłuje przedstawić swoją **teorię lingwistycznego** charakteru wampiryzmu. *(Siemion Piotr "Niskie Łąki")*

2. Niewątpliwie **uzdolnienia lingwistyczne** pozwalały jej ponadto porozumieć się po niemiecku, czytać po angielsku. *(various "Przegląd Biblioteczny")*

3. OPRÓCZ **WYKSZTAŁCENIA LINGWISTYCZNEGO** MAJĄ PRZYGOTOWANIE METODYCZNE CO POZWALA IM SPRAWNIE PROWADZIĆ GRUPĘ I JĄ INTEGROWAĆ. *(e0059edu)*

4. Dobre programy OCR korzystają z wbudowanej **bazy lingwistycznej**, analizując nie tylko pojedyncze znaki, ale całe słowa. *(various "reklamy komputerowe")*

5. Jej obrazy przypominają kartki powyrywane z tomiku współczesnej **poezji lingwistycznej**. *(various "Rzeczpospolita")*

6. Czasami jednak trudno przebrnąć przez uczone **wywody lingwistyczne**, w których mowa o derywatach atrybutywnych i palatalizacji. *(various "Rzeczpospolita")*

Leksykograf dostaje listę leksemów rzeczownikowych: TEORIA, UZDOLNIENIE, WYKSZTAŁCENIE, BAZA, POEZJA i WYWÓD.

Holmes, tak jak Poliqarp, dopuszcza zapytania proste i złożone, będące koniunkcją tych pierwszych. Potrafi szukać słów, form wyrazowych zadanych leksemów, form wyrazowych o zadanej charakterystyce gramatycznej, wzorców zadanych przez napis szkieletowy, wyrażeń z uzgodnieniem zadanego typu. Można ustawiać zakres oglądanego kontekstu. Można ograniczać ilościowo i jakościowo zbiór oczekiwanych przykładów. Holmes (a ściślej: Sherlock) jest narzędziem rozwiązującym w miarę skutecznie homonimie i synkretyzmy poprzez obszerny zbiór reguł lingwistycznych wykluczających pewne interpretacje.

Automatyczna analiza składniowa

Można powiedzieć, że problem automatycznej analizy morfologicznej jest dziś dla polszczyzny rozwiązany. Tym, co pozostaje, jest udoskonalanie metod automatycznej dehomonimizacji i desynkretyzacji. Jeśli metody te mają być jakościowe, nie ilościowe, to proces udoskonalania może nie mieć końca, a poszukiwanie dystrybucyjnych wykładników opozycji między najrozmaitszymi jednostkami tekstowymi wymaga najwyższej kompetencji lingwistycznej.

Pozostaje oczywiście osobny problem automatycznej analizy składniowej. Chodzi o narzędzia przypisywania wyrażeniom struktury hierarchicznej. Pamiętajmy, że obok homonimii morfologicznych istnieje homonimia składniowa, czyli zjawisko

identyczności kształtu różnych konstrukcji składniowych. Programy dokonujące analizy syntaktycznej nazywane są parserami.

Dla polszczyzny sporządzono w ciągu ostatniego ćwierćwiecza dwie pełne gramatyki formalne – Stanisława Szpakowicza (1983), z parserem, i Marka Świdzińskiego (1992) (ostatnia to tak zwana GFJP). Dla GFJP analizator składniowy stworzył Marcin Woliński – program Świgra (Woliński 2004b). Od lat trwają prace nad ulepszeniem tej gramatyki. Obecnie testowany jest program Świgra, przede wszystkim po to, aby ograniczyć liczbę dopuszczanych przez GFJP, często jałowych interpretacji. Automatyczna analiza składniowa pozostanie na długo terenem ważnych przedsięwzięć badawczych i technicznych. W dalszej natomiast perspektywie przyjdzie stawić czoło wyzwaniom automatycznej analizy semantycznej.

Zakończenie

XXI wiek jest stuleciem lingwistyki informatycznej. Przetwarzanie tekstów języków naturalnych pozostanie pierwszoplanowym zadaniem dla lingwistów na wiele dekad. Polszczyzna jest dziś dobrze opisana gramatycznie. Niestety, stopień zaawansowania przedsięwzięć wykorzystujących tę wiedzę nie zadowala. Bardzo niepokojące jest zwłaszcza to, że lingwistyką informatyczną zajmują się w Polsce pojedynczy językoznawcy; zupełnie inaczej jest u sąsiadów – Czechów, Węgrów, Niemców czy Rosjan; o świecie anglosaskim już nie mówiąc. Dużo więcej informatyków w Polsce pracuje w tej dziedzinie niż lingwistów. My, w odróżnieniu, powiedzmy, od Czechów, nie mamy powszechnie dostępnego Korpusu Narodowego (por. CNK) – i nie wydaje się, aby coś się zmieniło w najbliższej przyszłości.

Trudno się temu dziwić. W Polsce, inaczej niż w świecie, nie ma właściwie uniwersyteckich studiów lingwistycznych; językoznawstwo wykłada się na wydziałach filologicznych, jak w dobie przedstrukturalnej. Problematyka opisu dystrybucyjnego nie znajduje uznania w polskim środowisku lingwistycznym, dla którego „powierzchniowy" zdaje się znaczyć „powierzchowny". Sam byłem przez dekady namawiany (na szczęście bezskutecznie) do tego, by się zająć rzeczami poważnymi – na przykład semantyką.

Powyższy artykuł, utrzymany w stylistyce popularnonaukowej, pomyślany został jako apel do środowiska polonistycznej młodzieży. To głos językoznawcy, który wkroczył w jesień swego żywota; głos człowieka, który opisane tutaj trzy rewolucje lingwistyczne przeżył w miarę aktywnie i świadomie. Życzyłbym sobie, lingwistyce polskiej – i samej polszczyźnie, abyśmy włączyli się energicznie w to wszystko, co światowa lingwistyka uprawia bujnie i owocnie od dziesięcioleci.

Literatura

AWRAMIUK E., 1999, *Systemowość polskiej hominimii międzyparadygmatycznej*, Białystok.

BIEŃ J. S., 1991, *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji,* „Rozprawy Uniwersytetu Warszawskiego", t. 383, Warszawa.

CHOMSKY N., 1957, *Syntactic Structures*, Mouton.
CHOMSKY N., 1965, *Aspects of the Theory of Syntax,* Cambridge.
CHOMSKY N., 1982, *Zagadnienia teorii składni*, tłum. I. Jakubczak, Wrocław.
CNK: *Český Národní Korpus*, http://ucnk.ff.cuni.cz.
GRAMATYKA PAN, 1984, *Gramatyka współczesnego języka polskiego*, red. nauk. S. Urbańczyk, t. 2: *Morfologia*, red. K. Kallas, R. Laskowski, R. Grzegorczykowa, H. Wróbel, Warszawa.
GRUSZCZYŃSKI W., 1989, *Fleksja rzeczowników pospolitych we współczesnej polszczyźnie pisanej (na materiale „Słownika języka polskiego" pod red. W. Doroszewskiego)*, Wrocław.
JESPERSEN O., 1909–1949, *A Modern English Grammar on Historical Principles*, t. 1-7, Copenhagen.
KORPUS IPI PAN: http://korpus.pl.
KORPUS PWN: http://korpus.pwn.pl.
KSJP, 1997, *Komputerowy słownik języka polskiego PWN*, Warszawa.
MECNER, P., 2004, *Elementy gramatyki umysłu*, Warszawa.
POLLARD, C., SAG I., 1994, *Head-driven Phrase Structure Grammar*, Chicago.
PRZEPIÓRKOWSKI A., 2004, *Korpus IPI PAN – wersja wstępna,* Warszawa.
PRZEPIÓRKOWSKI I IN., 2002, A. Przepiórkowski, A. Kupść, M. Marciniak, A. Mykowiecka, *Formalny opis języka polskiego. Teoria i implementacja*, Warszawa.
RABIEGA-WIŚNIEWSKA J., RUDOLF M., 2003, *AMOR – program automatycznej analizy fleksyjnej tekstu polskiego*, „Biuletyn PTJ", R. LVIII, s. 175–186.
RUDOLF M., 2004, *Metody automatycznej analizy korpusu tekstów polskich*, Warszawa.
S-LF, 1974–1977, I. Kurcz, A. Lewicki, W. Masłowski*, J. Sambor, J. Woronczak, *Słownictwo współczesnego języka polskiego. Listy frekwencyjne*, t. 1–5, Warszawa [*: t. 3].
SALONI Z., 1992, *Rygorystyczny opis polskiej deklinacji przymiotnikowej*, „Filologia Polska. Prace Językoznawcze", nr 16, Gdańsk, s. 215–228.
SALONI Z., 2004, *Czasownik polski. Odmiana – słownik*, Warszawa.
SALONI Z., ŚWIDZIŃSKI M., 2001, *Składnia współczesnego języka polskiego*, Warszawa.
SAUSSURE F. DE, 1961, *Kurs językoznawstwa ogólnego*, Warszawa.
SJP DOR., 1958–1970, *Słownik języka polskiego*, red. W. Doroszewski, Warszawa.
SFPW, 1990,I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, *Słownik frekwencyjny polszczyzny współczesnej*, red. Z. Saloni, Kraków.
SZAFRAN K., 1994, „Automatyczna analiza fleksyjna tekstu polskiego (na podstawie „Schematycznego indeksu *a tergo*" Jana Tokarskiego", niepublikowana rozprawa doktorska, Warszawa.
SZPAKOWICZ S., 1983, *Formalny opis składniowy zdań polskich*, Warszawa.
ŚWIDZIŃSKI M., 1992, *Gramatyka formalna języka polskiego*, „Rozprawy Uniwersytetu Warszawskiego", t. 349, Warszawa.
ŚWIDZIŃSKI M., DERWOJEDOWA, M., RUDOLF M., 2002, *Dehomonimizacja i desynkretyzacja w procesie automatycznego przetwarzania wielkich korpusów tekstów polskich*, „Biuletyn PTJ", R. LVIII, s. 187–199.
ŚWIDZIŃSKI M., RUDOLF M. [w druku], *Narzędzia informatyczne obsługi wielkich korpusów tekstów: wyszukiwarka Holmes*, „Biuletyn PTJ", R. LXI, Warszawa.
TOKARSKI J., 1973, *Fleksja polska*, Warszawa.
TOKARSKI J., 1993, *Schematyczny indeks „a tergo" polskich form wyrazowych*, oprac. i red. Z. Saloni, Warszawa.
WOLIŃSKI M., 2003, http://nlp.ipipan.waw.pl/~wolinski/morfeusz/morfeusz.html.
WOLIŃSKI M., 2004, „Komputerowa weryfikacja gramatyki Świdzińskiego", niepublikowana rozprawa doktorska, Warszawa.
WOŁOSZ R., 2005, *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*, Warszawa.

Corpus linguistics in Poland – the origins, the present, the prospects
Summary

In the article, three sources of corpus engineering are mentioned: (a) theoretical and descriptive achievements of structural linguistics, (b) the formal apparatus of generative theories, and (c) the

development of computational tools. For the last decades, the Polish language has been satisfactorily accounted for both in terms of morphology and syntax. On that basis, two corpus search engines have recently been designed to annotate Polish text corpora (Poliqarp) or to disambiguate them morphologically (Holmes). The prospects of corpus engineering in Poland do not look optimistic, indeed. Unlike in neighbouring countries, not many people work in the area of computational linguistics. The article expresses the author's hope that young Polish linguists may find the job attractive, not only intellectually.