POLONICA ???
PL ISSN 0137-9712

ADAM PRZEPIÓRKOWSKI
PIOTR BAŃSKI
ŁUKASZ DĘBOWSKI
ELŻBIETA HAJNICZ
MARCIN WOLIŃSKI

Konstrukcja korpusu IPI PAN

Niniejszy artykuł stanowi krótki opis projektu realizowanego w Instytucie Podstaw Informatyki PAN i zmierzającego do konstrukcji dużego anotowanego korpusu języka polskiego. Projekt ten jest częściowo finansowany w ramach grantu KBN 7 T11C 043 20 realizowanego w IPI PAN od kwietnia 2001 do marca 2004, częściowo zaś — w ramach działalności statutowej IPI PAN.

Poniżej przedstawiamy genezę projektu (p. 1), jego zakres (p. 2) i przewidywane zastosowania (p. 3). Rozwiązania szczegółowe dotyczące opracowania zbioru znaczników morfologicznych (tzw. *tagsetu*), problemu dehomonimizacji oraz systemu znaczników XML-owych omówione zostały w osobnych artykułach (Woliński 2002, Przepiórkowski 2002 i Bański 2002).[1]

1. Geneza projektu

Inżynieria lingwistyczna, zajmująca się przede wszystkim analizą i syntezą wypowiedzi w językach naturalnych, przeżywa w ostatnich latach intensywny rozwój na całym świecie. W ciągu ostatniej dekady dziedzina ta, będąca wcześniej głównie domeną laboratoriów naukowych, wkroczyła w etap różnorodnych zastosowań, od systemów rozpoznawania i generowania mowy, poprzez systemy umożliwiające dialog z programem w języku naturalnym, do systemów tłumaczenia maszynowego.

Do tego rozwoju w dużej mierze przyczyniło się zastosowanie metod statystycznych w przetwarzaniu języka naturalnego i odejście od metod wyłącznie symbolicznych, które — choć bliższe istniejącym teoriom lingwistycznym — okazały się mało efektywne.

Możliwość zastosowania metod statystycznych w przetwarzaniu danego języka zależy od istnienia tzw. korpusu, czyli zbioru tekstów tego języka. Aby taki zbiór mógł posłużyć za podstawę do modelowania statystycznych cech języka, musi on spełniać pewne warunki. Przede wszystkim, musi on być duży (co najmniej kilkadziesiąt milionów słów), musi być reprezentatywny (tj. musi zawierać różne style i rejestry danego języka) oraz musi być anotowany. Ten ostatni wymóg oznacza, że oprócz samego tekstu, korpus musi zawierać

[1] Projekt ten jest obecnie, na przełomie lat 2002 i 2003, w trakcie realizacji. Oznacza to, że wszystkie rozwiązania szczegółowe prezentowane tutaj i w cytowanych powyżej artykułach mogą w ostatecznej wersji korpusu ulec zmianie.

różnorakie informacje lingwistyczne o tekście, przede wszystkim informacje o morfologicznej i składniowej charakterystyce poszczególnych słów (ang. *POS tags*, czyli *part-of-speech tags*). Aby taki korpus mógł być pożyteczny dla społeczności naukowej, musi on być dostępny publicznie.

Język polski należy do tych nielicznych języków europejskich, które do tej pory nie doczekały się takiego korpusu. Prace korpusowe w Polsce są bardzo rozproszone, a istniejące zbiory tekstów języka polskiego są często niewielkie (rzędu kilku milionów słów), albo nie są anotowane, albo nie są publicznie dostępne. Celem naszych prac jest wypełnienie tej luki i, w konsekwencji, przetarcie drogi dla zastosowań metod statystycznych w przetwarzaniu języka polskiego.

2. Zakres i przewidywane wyniki

2.1. Rozmiar i reprezentatywność

Rozmiar współczesnych korpusów mierzy się w dziesiątkach milionów słów, korpusy o rozmiarach rzędu 100 mln. słów określa się jako ‘duże’ (ang. *large*).

Celem niniejszego projektu jest stworzenie korpusu polszczyzny pisanej liczącego około 75–100 mln. słów. W czasach, gdy większość tekstów tworzona jest w postaci elektronicznej, stworzenie korpusu tego rozmiaru nie jest zadaniem bardzo trudnym. Trudne jest jednak zapewnienie, żeby taki korpus był reprezentatywny (ang. *balanced, representative*), czyli żeby w odpowiednim stopniu reprezentował różne style i rejestry języka.

Dlatego z całego korpusu wykrojony zostanie mniejszy korpus, liczący około 15–25 milionów słów, który zaprojektowany zostanie w taki sposób, aby jak najlepiej reprezentować stan współczesnego języka polskiego. (Podkorpus ten będzie więc tzw. reprezentatywnym korpusem synchronicznym.) Nie jest celem niniejszego projektu zdefiniowanie pojęcia *reprezentatywność* w odniesieniu do języka polskiego. Wykorzystane zostaną natomiast prace przeprowadzone w Instytucie Języka Polskiego PAN (Górski, 2002) oraz doświadczenia reprezentatywnego Brytyjskiego Korpusu Narodowego (British National Corpus), Czeskiego Korpusu Narodowego (Český Národní Korpus) i korpusów innych języków.

Pozostała część korpusu będzie reprezentować przede wszystkim polszczyznę współczesną, choć niekoniecznie różne jej style w równym wymiarze, i w ograniczonym zakresie także teksty historyczne. W ten sposób struktura korpusu będzie przypominać strukturę tworzonego obecnie Narodowego Korpusu Amerykańskiego (American National Corpus): korpus jako całość ma charakter tzw. ‘oportunistyczny’ (ang. *opportunistic*), ale istnieje starannie dobrany synchroniczny podkorpus reprezentatywny w stylu Narodowego Korpusu Brytyjskiego. Taka struktura zapewni, iż korpus będzie przydatny zarówno w tych zastosowaniach, w których ważna jest reprezentatywność, jak i w tych, w których większa rolę odgrywa wielkość korpusu.

Pewnym ograniczeniem reprezentatywności w niniejszym projekcie jest to, że zebrany tekst będzie tekstem istniejącym w postaci pisanej — rozszerzenie korpusu o język mówiony wymagałoby rozwiązania szeregu problemów, które wykraczają poza ramy tego projektu. Aby zmniejszyć efekty tego ograniczenia, planowany jest duży udział tekstów ‘przybliżających’ język mówiony, czyli sztuk teatralnych, ścieżek dialogowych oraz transkrypcji programów telewizyjnych i radiowych.

2.2. Anotacja lingwistyczna

Korpus tworzony w ramach niniejszego projektu będzie anotowany[2] informacjami morfosyntaktycznymi, tj. każdemu słowu w korpusie przypisany zostanie znacznik (tzw. tag) odzwierciedlający klasę gramatyczną, do której dane słowo należy (np. 'przyimek', 'rzeczownik'), oraz wartości poszczególnych kategorii gramatycznych takich jak przypadek, rodzaj i liczba.

W ramach projektu opracowany został nowy tagset dla języka polskiego, różniący się od tagsetów stworzonych dla innych języków słowiańskich przede wszystkim metodą wyodrębnienia klas gramatycznych, opartą tutaj na pojęciu *fleksemu* (Bień, 1991), oraz zestawem kategorii gramatycznych i ich wartości, opartym w niniejszym projekcie przede wszystkim na pracach Zygmunta Saloniego i jego współpracowników (m.in. Saloni 1976, Saloni 1977, Gruszczyński i Saloni 1978 oraz Bień i Saloni 1982).

Aktualna wersja tagsetu opisana jest w artykułach Woliński 2002, Przepiórkowski i Woliński 2003a,b, zaś porównanie z tagsetami dla innych języków słowiańskich — w pracy Woliński i Przepiórkowski 2001. Ponadto artykuł Przepiórkowski 2002 przedstawia niektóre reguły wyboru odpowiednich interpretacji spośród wszystkich możliwych interpretacji morfosyntaktycznych dla danego słowa.

2.3. System znaczników XML

Dane o tekstach, ich struktura oraz informacje lingwistyczne reprezentowane będą za pomocą języka XML (Extended Markup Language), obecnego *de facto* standardu reprezentacji korpusów tekstów. Dzięki zastosowaniu XML możliwe będzie wykorzystanie licznych publicznie dostępnych narzędzi obsługujących ten system notacji, na przykład narzędzi stworzonych przez Language Technology Group na Uniwersytecie Edynburskim (m.in. sprawdzających poprawność znakowania, ułatwiających wizualizację korpusu i konwersję na inne formaty i pozwalających przeszukiwać korpus; `http://www.ltg.ed.ac.uk/software/xml/`).

W ramach niniejszego projektu opracowany został schemat znakowania XML wykorzystujący rekomendacje TEI (Text Encoding Initiative) i bazujący na standardzie CES (Ide i in., 1996) stworzonym dla języka SGML przez zespół Expert Advisory Group on Language Engineering Standards (EAGLES) i następnie przeniesionym do języka XML (Ide i in., 2000). Aktualna wersja schematu opisana jest w artykule Bański 2002.

2.4. Narzędzia

Produktem ubocznym niniejszego projektu, choć równie ważnym jak sam korpus języka polskiego, będą informatyczne narzędzia do automatycznej anotacji tekstów polskich oparte na metodach statystycznych, przede wszystkim tzw. 'tager' (ang. *tagger*), tj. dezambiguator współpracujący z analizatorem morfosyntaktycznym (por. Dębowski 2001, 2003).

Jak wiadomo jednak z doświadczeń korpusowych dotyczących innych języków, taki proces automatycznej anotacji nigdy nie jest bezbłędny — najlepsze tagery przypisują słowom właściwe symbole lingwistyczne z prawdopodobieństwem ok. 95–97%, co oznacza,

[2] Lingwistyka korpusowa jest stosunkowo młodą gałęzią nauki, o jeszcze nieustabilizowanej terminologii polskiej. W niniejszym artykule będziemy używać terminu *anotacja* w sensie lingwistycznego znakowania korpusu, zaś termin *znakowanie* zarezerwujemy dla znakowania znacznikami XML (p. 2.3). Będziemy także używać zapożyczonego z języka angielskiego terminu *tagset*, jako wygodniejszego niż *system znaczników lingwistycznych*.

że co 20-te słowo może mieć niewłaściwą anotację. Dlatego z podkorpusu zrównoważonego wyodrębniony zostanie mały (1–2 mln. słów) podkorpus, którego anotacja będzie sprawdzona i odpowiednio skorygowana przez lingwistów.

Planowane jest także stworzenie oprogramowania umożliwiającego udostępnianie i przeszukiwanie korpusu przez Internet, lub też dostosowanie do potrzeb projektu istniejącego oprogramowania tego typu.

3. Przewidywane zastosowania

Trudno jest przecenić zastosowania korpusów języków naturalnych w inżynierii lingwistycznej, a także w innych dziedzinach nauki.

Jednym z najważniejszych zastosowań jest statystyczne modelowanie języka na potrzeby systemów przetwarzania mowy dla celów głosowej komunikacji z komputerem. W szczególności, informacja o częstotliwości występowania zbitek słów (tzw. bi-gramów, tri-gramów itd.) w danym języku wykorzystywana jest w systemach rozpoznawania mowy ciągłej. Informacja ta może być łatwo uzyskana na podstawie dużego korpusu tekstów. Podobnie, narzędzia do automatycznej anotacji tekstów języka polskiego mogą być wykorzystane w systemach syntezy mowy, gdzie warunkiem przetworzenia tekstu pisanego na sygnał akustyczny jest bogata anotacja lingwistyczna, zawierająca informacje m.in. o charakterystyce morfologicznej i składniowej słów.

Inne ważne zastosowanie dotyczy statystycznych metod tłumaczenia maszynowego. Na przykład system tłumaczący z języka angielskiego na polski może przetłumaczyć angielski wyraz *strong* na polskie *silny* albo *mocny*. W wielu kontekstach nie jest istotne, który z polskich ekwiwalentów zostanie wybrany, ale ma to znaczenie przy tłumaczeniu np. *strong wind* i *strong tea* — w pierwszym wypadku tłumaczeniem będzie *silny wiatr*, a w drugim *mocna herbata*, a nie na odwrót. Nie wynika to jednak z żadnych właściwości 'formalnych' języka (tj. z jego składni albo semantyki), a wyłącznie z jego właściwości statystycznych (z częstotliwości występowania poszczególnych zbitek słów, tj. tzw. kolokacji). Taki system tłumaczenia maszynowego powinien więc mieć dostęp do danych statystycznych uzyskanych na podstawie badań korpusowych, albo bezpośrednio do korpusu języka polskiego.

Spośród innych ważnych zastosowań prac korpusowych w inżynierii lingwistycznej wymienić jeszcze trzeba zastosowania w automatycznym pozyskiwaniu informacji z dużych zbiorów tekstu (ang. *information extraction*) i nowe zastosowania w automatycznej indukcji gramatyk (ang. *grammar induction*, *machine-learning of grammars*).

Zastosowania korpusów nie ograniczają się jednak tylko do inżynierii lingwistycznej. Jednym z najważniejszych i najbardziej znanych zastosowań jest leksykografia, gdzie dane korpusowe są podstawą definicji słów i opisów ich własności gramatycznych, a także służą jako źródło przykładów w słownikach. Innym ważnym zastosowaniem jest tworzenie materiałów do nauczania danego języka. Z kolei lingwiści teoretyczni często korzystają z publicznie dostępnych korpusów w celu weryfikacji swoich teorii. Modelowanie statystyczne może też służyć do oceny poprawności tłumaczeń (maszynowych lub wykonanych przez ludzi) z jednego języka na inny. Ponadto dane korpusowe są niezbędne w badaniach psycholingwistycznych, gdzie istotna jest informacja o częstotliwości występowań poszczególnych słów w danym języku. I na koniec — korpusy tekstów są często wykorzystywane w badaniach socjolingwistycznych (np. w Polsce realizowany był projekt „Językowe środki

naukowego wartościowania w krytykach muzycznych. Analiza języka fachowego krytyk muzycznych na korpusie *Neue Zeitschrift für Musik* roczniki 1985–1997").

Poszczególne korpusy często konstruowane są pod kątem tylko jednego z tych zastosowań, najczęściej w celach leksykograficznych albo do nauczania danego języka jako języka obcego. Celem niniejszego projektu jest skonstruowanie korpusu tak, aby zmaksymalizować jego użyteczność w różnych dziedzinach badań i zastosowań.

Bibliografia

Bański, Piotr. (2002) „Anotacja zewnętrzna: wpływ architektury korpusu IPI PAN na efektywność jego tworzenia oraz wykorzystania". Złożone do *Poloników*.

Bień, Janusz S. (1991) *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego.

Bień, Janusz S. i Zygmunt Saloni. (1982) „Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)". *Prace Filologiczne* XXXI: 31–45.

Dębowski, Łukasz. (2001) „Tagowanie i dezambiguacja morfologiczna". Prace IPI PAN 934, Instytut Podstaw Informatyki PAN.

———. (2003) „A reconfigurable stochastic tagger for languages with complex tag structure". *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Górski, Rafał. (2002) „Reprezentatywność i zrównoważenie korpusu". Maszynopis, IJP PAN.

Gruszczyński, Włodzimierz i Zygmunt Saloni. (1978) „Składnia grup liczebnikowych we współczesnym języku polskim". *Studia Gramatyczne* II: 17–42.

Ide, Nancy, Patrice Bonhomme i Laurent Romary. (2000) „XCES: An XML-based standard for linguistic corpora". *Proceedings of the linguistic resources and evaluation conference*. Athens, Greece.

Ide, Nancy, Greg Priest-Dorman i Jean Véronis. (1996) „Corpus encoding standard". Maszynopis, http://www.cs.vassar.edu/CES/.

Przepiórkowski, Adam. (2002) „Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN". Złożone do *Poloników*.

Przepiórkowski, Adam i Marcin Woliński. (2003a) „A flexemic tagset for Polish". *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

———. (2003b) „The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish". *Proceedings of the* 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, EACL 2003*.

Saloni, Zygmunt. (1976) „Kategoria rodzaju we współczesnym języku polskim". *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*. Wrocław: Ossolineum, 41–75.

———. (1977) „Kategorie gramatyczne liczebników we współczesnym języku polskim". *Studia Gramatyczne* I: 145–173.

Woliński, Marcin. (2002) „System znaczników morfosyntaktycznych w korpusie IPI PAN". Złożone do *Poloników*.

Woliński, Marcin i Adam Przepiórkowski. (2001) „Projekt anotacji morfosynktaktycznej korpusu języka polskiego". Prace IPI PAN 938, Instytut Podstaw Informatyki PAN.

SUMMARY

The construction of the IPI PAN Corpus

This article presents a project aiming at the construction of a large morphosyntactically and structurally annotated corpus of Polish. The project, realized at the Institute of Computer Science, Polish Academy of Sciences between April 2001 and March 2004, is partially financed by the State Committee for Scientific Research (KBN).

While particular solutions concerning the linguistic and XML annotation of natural language texts are presented in three separate articles in this volume, this article describes the origin of the project, its scope and some envisaged applications of the results of the project.