Rozszerzenie opisów morfologicznych w tekstach korpusu słownika frekwencyjnego polszczyzny współczesnej

Maciej Ogrodniczuk

15 maja 2003

Korpus słownika frekwencyjnego polszczyzny współczesnej, mimo że od jego powstania minęło już 35 lat (ciekawą historię korpusu opisuje m. in. Zygmunt Saloni [10]), jest wciąż cennym źródłem inspiracji do badań lingwistycznych. W swojej pracy magisterskiej [9] wykorzystałem dane *Korpusu* jako bazę testową dla tworzonej koncepcji zapisu polskich danych lingwistycznych zgodnego z normą SGML (Standard Generalized Markup Language [2, 5]) i formatem TEI (Text Encoding Initiative [1]).

Jednym z aspektów mojej pracy z korpusem, który przedstawiam poniżej, była próba jego automatycznego wzbogacenia o dodatkowe informacje morfologiczne i postaci hasłowe. Materiałem źródłowym dla tego zadania był korpus frekwencyjny – zbiór 10 000 dobranych metodą losowania próbek ciągłego tekstu o długości ok. 50 słów każda, zgromadzonych w równych transzach odpowiadających pięciu stylom współczesnej polszczyzny pisanej (styl popularnonaukowy, wiadomości prasowych, publicystyczny, prozy artystycznej i dramatu artystycznego). Próbki – oprócz zawartości tekstowej – zawierały również oryginalne, dopisane ręcznie kody morfologiczne oraz oznaczenia dodatkowe (ich powtórzony za autorami *Korpusu* opis zawiera m. in. praca magisterska Marty Nazarczuk [8]). Zgodnie z założeniami projektantów korpusu brak kodu miał oznaczać oczywistość kategoryzacji formy, przetwarzanie automatyczne miało zatem, oprócz rozszerzenia informacji morfologicznej, dodatkowo zweryfikować poprawność tego sądu. W trakcie prac kody oryginalne zostały zastąpione ogólniejszymi oznaczeniami zgodnymi z zestawem zaprojektowanym przez Katarzynę Głowińską i Marcina Wolińskiego [6], wprowadzającym nie uwzględnione przez autorów korpusu kategorie morfologiczne.

Oto fragment jednej z próbek w postaci źródłowej (przed przetworzeniem), nazywanej dalej wersją oryginalną (znak plusa oznacza formę wieloczłonową, ukośnik – nazwę własną):

```
W[66] imieniu dwu[32] tysięcy[122] studentów[122] Uniwersytetu
Warszawskiego[221] zgromadzonych[222] w[66] auli[161] Auditorium[+]
Maximum[&], rektor uczelni[121], profesor Stanisław[/] Turski[/]
powitał wicepremiera[141] Zenona[/][141] Nowaka[/][141].
```

i wynikowej, nazywanej wersją wzbogaconą (tekst skrócony):

```
W[w,66,P-L----------P] imieniu[imię,,SSDN---------P/SSLN---------P]
studentów[student,122,SPGP---------P]
Uniwersytetu[uniwersytet,,SSGI---------P] rektor[rektor,,SSNP---------P]
powitał[powitać,,VS-M-3PONC---P]
wicepremiera[wicepremier,141,SSAP---------P] Nowaka[,141,SSAX---------W].
```

Dla przykładu, opis formy `studentów`, oznaczonej w oryginalnej wersji korpusu kodem `122` odpowiadającym rzeczownikowi (1) w dopełniaczu (2) liczby mnogiej (2), został w wersji wzbogaconej rozszerzony poprzez przypisanie jej postaci hasłowej `student` (pierwsza wartość na oddzielonej przecinkami liście zapisanej w nawiasach kwadratowych po formie analizowanej) oraz rodzaju męskoosobowego; pozostałe wartości kategorii morfologicznych, choć w zmienionym zapisie, zostały przeniesione do wersji wzbogaconej i zapisane w trzeciej wartość listy (`S` to zgodnie z nowym zestawem oznaczeń kod rzeczownika, `P` – liczby mnogiej, `G` – dopełniacza, `P` – rodzaju męskoosobowego; następujący po nich szereg minusów oznacza kategorie nie mające zastosowania do opisywanej części mowy, końcowe `P` to natomiast kod nazwy pospolitej). Druga wartość na liście jest po prostu kopią oryginalnego kodu, zapamiętaną w tekście wynikowym m. in. dla celów porównawczych.

W przypadku niemożności automatycznej identyfikacji formy hasłowej pierwszy element listy jest oznaczany jako pusty; podobnie gdy nie uda się zidentyfikować danej kategorii, oznaczana jest ona znakiem `X` (jak w przypadku formy `Nowaka`).

W samej pracy magisterskiej treść próbek została jeszcze dodatkowo obudowana strukturą SGML-ową zgodnie z mechanizmami oferowanymi przez format TEI, ten jej aspekt nie będzie jednak przedmiotem dalszych omówień.

Pierwsza część zadania rozszerzenia opisów morfologicznych miała na celu przetworzenie form zawartych w tekstach korpusu narzędziem automatycznym w celu ich późniejszego (także automatycznego) porównania z oznaczeniami obecnymi w korpusie. „Nowe" opisy morfologiczne powstały z wykorzystaniem analizatora morfologicznego SAM-99 [11] Krzysztofa Szafrana (który jest prawdopodobnie jedynym analizatorem morfologicznym języka polskiego dostępnym bezpłatnie w Internecie) opartego na schematycznym indeksie a tergo Jana Tokarskiego [12] oraz stworzonych w tym celu i dostępnych w chwili obecnej na licencji GNU programów w języku programowania Perl [13, 14].

Przetworzeniu poddane zostały nie bezpośrednio teksty próbek, lecz lista zawartych w nich form słów, w postaci powstałych na etapie tworzenia słownika pełnych list rangowych form występujących we wszystkich stylach. Oto fragment listy dla słów rozpoczynających się na 'ź':

```
 0.00   0.00
   0    0    0    0    1    1 ździebko
   0    0    0    0    1    1 ździebko
   1
38.76   0.78
   1    0    0    1    0    2 źdźbło 1
   1    0    0    0    0    1 źdźbła 142
   0    0    0    1    0    1 źdźbłami
   2
47.41  33.19
   7    3    6   11   43   70 źle
   7    3    6   11   43   70 źle
   3
```

Listy, oprócz nieistotnych dla bieżącego zadania informacji statystycznych zawierały wykaz słów pogrupowanych według postaci hasłowych, których ekstrakcja umożliwiła wprowadzenie dodatkowego ograniczenia na listę uzyskiwanych w sposób automatyczny analiz.

Powodem sięgnięcia po listy rangowe była chęć uniknięcia nie tylko ekstrakcji form w przypadku przetwarzania całych tekstów (w celu chociażby usunięcia swoistych, nie rozpoznawanych przez analizator morfologiczny oznaczeń dodatkowych), ale przede wszystkim dodatkowego przetwarzania form o dużej frekwencji (które byłyby wówczas, bez wyraźniej potrzeby, analizowane wielokrotnie).

Powstała w wyniku opisanych wyżej zabiegów lista form została poddana analizie morfologicznej analizatorem SAM-99. Jej rezultatem był plik zawierający ciąg analiz o ściśle zdefiniowanej strukturze. Oto ich przykład dla wieloznacznej formy *kurze* (rozpoznanej jako forma mianownika liczby mnogiej rzeczownika *kurz*, miejscownika lub wołacza liczby pojedynczej rzeczownika *kur*, celownika liczby pojedynczej rzeczownika *kura*, a także jako przymiotnik *kurzy* w mianowniku lub bierniku liczby pojedynczej rodzaju nijakiego albo mnogiej rodzaju niemęskoosobowego):

```
kurze
%
{{(lN) < kurz(mII::m3)+ }%
{(LV) < kur(mIV::m2 mIV m3)+ }%
{(D) < kura(żIV::)+ }%
{(5) < kurzy(A::[p])+ } }%
```

Kolejny program w języku Perl umożliwił zamianę kodów zwracanych przez analizator (opartych o oznaczenia indeksu Tokarskiego) na oznaczenia wynikowe i przyporządkowanie im form hasłowych. Oto wynik jego działania dla podanego wyżej bloku (podziału na kilka linii dokonano dla zwiększenia przejrzystości zapisu, w rzeczywistości jest to pojedyncza linia analiz):

```
kurze:kurz/SPNI---------,kurz/SPAI---------,kurz/SPVI---------,
      kur/SSLA---------,kur/SSVA---------,kura/SSDF---------,
      kura/SSLF---------,kurzy/ASNNP--------,kurzy/ASANP--------,
      kurzy/APNOP--------,kurzy/APNRP--------,kurzy/APARP--------
```

Następny etap prac to już sama analiza tekstu wraz z zawartymi w nim pierwotnymi oznaczeniami. Tu także z pomocą przyszły listy rangowe, a dokładnie skojarzone z poszczególnymi blokami haseł ich formy hasłowe. W ten sposób udało się wyeliminować potencjalnie poprawne, lecz w kontekście korpusowych tekstów bezużyteczne interpretacje analizatora (okazało się np. że w tekstach korpusowych forma *kurze* występuje jedynie w postaci rzeczownikowej pochodzącej od słów *kura* i *kurz*, ale już nie od *kur*; ponadto forma ta nigdy nie została użyta w wołaczu).

Po dekompozycji kodów morfologicznych zawartych w tekstach korpusowych oraz ich zamianę na oznaczenia wynikowe kolejny program w języku Perl umożliwił ich porównanie z kodami powstałymi w

wyniku analizy automatycznej, co pozwoliło na przypisanie każdemu wystąpieniu wyrazu w tekstach korpusu nowego kodu morfologicznego. Dzięki wykorzystaniu kodów oryginalnych wyniki analizy morfologicznej zostały ujednoznacznione w maksymalnym stopniu; proces ten nie zakończył się niestety pełnym sukcesem — szczególnie w przypadku przymiotników nie udało się wykluczyć wariantów analiz, które mogłyby zostać wyeliminowane jedynie podczas weryfikacji wyników przez człowieka (warianty te zostały oddzielone w zapisie wynikowym znakiem ukośnika). Inne części mowy uzyskały nowe kody bez większych problemów.

Proces analizy pozwolił na sformułowanie kilku wniosków dotyczących jakości użytego analizatora morfologicznego, specyfiki danych korpusowych oraz możliwości dalszych zastosowań tak wzbogaconych danych.

Analizator morfologiczny SAM-99 doskonale radził sobie z tekstami o tak dużej różnorodności. Jedyną niedogodnością była skąpa klasyfikacja form nieodmiennych; w ich przypadku warto jednak pamiętać o arbitralności rozstrzygnięć. Oryginalne oznaczenia kategorii morfologicznych okazały się dość łatwe w interpretacji, nawet przy uwzględnieniu drobnych zmian klasyfikacyjnych (korpus zalicza np. imiesłowy przymiotnikowe do kategorii przymiotnikowej, SAM natomiast do czasownikowej). Pośrednim wynikiem prac może być wiarygodność użycia korpusu jako zestawu tekstów do testów innych analizatorów morfologicznych.

Przy okazji udało się też pozytywnie zweryfikować wspomnianą opinię o „oczywistości" klasyfikacji morfologicznej w przypadku form, którym nie przypisano w korpusie żadnego oznaczenia kodowego. Formom tym, z uwzględnieniem dostępnej na listach rangowych informacji o formie podstawowej, kody wg nowej taksonomii udało się w większości przypadków dopisać automatycznie.

Literatura cytowana

[1] TEI Guidelines for Electronic Text Encoding and Interchange (TEI P4). Red. C. M. Sperberg-McQueen i Lou Burnard. Oxford — Providence — Charlottesville — Bergen, marzec 2002. The Association for Computers and the Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Lingustic Computing (ALLC). Dostępne pod adresem http://etext.lib.virginia.edu/teip4/.

[2] *ISO 8879 Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. Genewa, 1986. ISO (International Organization for Standardization) z poprawkami SGML TC3:1998 (technical correction).

[3] Bień, Janusz S. *Kodowanie tekstów polskich w systemach komputerowych*. Postscriptum 27-29. S. 4-27. Dostępny też pod adresem ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/ogonki/katow98.ps.

[4] Bień, Janusz S. *Polskojęzyczne reguły składni SGML*. Warszawa, 1998 (tekst nie publikowany).

[5] Goldfarb, Charles F. *The SGML Handbook*. 1990. Oxford University Press.

[6] Głowińska Katarzyna, Woliński, Marcin. *Taksonomia morfologiczna dla Słownika frekwencyjnego.* Tekst opublikowany na płycie CD-ROM *Wzbogacony korpus Słownika frekwencyjnego polszczyzny współczesnej*. Warszawa 2001. Skompresowany obraz płyty dostępny pod adresem http://www.mimuw.edu.pl/polszczyzna/wksf/wksf.iso.bz2

[7] Kurcz, Ida; Lewicki, Andrzej; Sambor, Jadwiga; Woronczak, Jerzy; Szafran, Krzysztof. *Słownik frekwencyjny polszczyzny współczesnej*, red. Zygmunt Saloni. Kraków, 1990. Instytut Języka Polskiego PAN.

[8] Nazarczuk, Marta. *Wstępne przygotowanie korpusu „Słownika frekwencyjnego polszczyzny współczesnej" do dystrybucji na CD-ROM*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa, 1997. Instytut Języka Polskiego Uniwersytetu Warszawskiego.

[9] Ogrodniczuk, Maciej. *Wykorzystanie SGML i TEI do zapisu polskich danych lingwistycznych*. Praca magisterska napisana pod kierunkiem dra hab. Janusza S. Bienia. Warszawa, wrzesień 2000. Wydział Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego.

[10] Saloni, Zygmunt. *Słownik frekwencyjny polszczyzny współczesnej*. W: ComputerWorld. S. 16-17. 4 listopada 1991.

[11] Szafran, Krzysztof. *Analizator morfologiczny SAM-99 — opis użytkowy*. Maj 2000. Instytut Informatyki Uniwersytetu Warszawskiego. Wcześniejsza wersja dokumentu dostępna pod adresem ftp://ftp.mimuw.edu.pl/pub/users/polszczyzna/SAM-95/tr226.ps.

[12] Tokarski, J., *Schematyczny indeks a tergo polskich form wyrazowych.* Opracowanie i redakcja: Zygmunt Saloni. Wydawnictwo Naukowe PWN, Warszawa 1993.

[13] Wall, Larry; Christiansen, Tom; Potter, Stephen; Schwartz, Randal L. *Perl. Programowanie*. ReadMe, Warszawa, 1998.

[14] Wall, Larry; Schwartz, Randal L. *Programming Perl*. 1991. O'Reilly & Associates, Inc.