Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych

Magdalena Derwojedowa, Witold Kieraś,
Danuta Skowrońska i Robert Wołosz

Instytut Języka Polskiego UW

Konferencja *Przyszłość językoznawstwa – językoznawstwo przyszłości*
Kraków, 11-12 czerwca 2013 r.

Projekt

Projekt finansowany przez NCN w ramach konkursu OPUS 4.

- Tytuł projektu: Automatyczna analiza fleksyjna tekstów polskich z lat 1830-1918 z uwzględnieniem zmian w odmianie i pisowni
- Kierownik: Magdalena Derwojedowa
- Pozostali wykonawcy: Witold Kieraś, Zygmunt Saloni, Danuta Skowrońska, Robert Wołosz

Cele projektu

- Opis systemowych zmian w zakresie odmiany polszczyzny pisanej w latach 1830–1918;
  - zmiana wzorca paradygmatycznego (np. powszechne w XIX w. zakończenia *-em*, *-emi* w formach narzędnika przymiotników),
  - zmiana paradygmatu leksemu (np. KOMETA — w XIX w. leksem rodzaju męskiego, obecnie — żeńskiego),
  - szeroko rozumiane zmiany pisowniane, zarówno ortograficzne (np. zapis z użyciem *x*, *j*, *y*, np. *fantazya*); pisownia łączna, np. *zdawna*, *zlekka*), jak i graficzne, dotyczące oznaczania pewnych zjawisk fonetycznych (samogłoski pochylone).
- Stworzenie słownika fleksyjnego ukazującego ewolucję odmiany.
- Stworzenie milionowego zrównoważonego korpusu polszczyzny lat 1830–1918.

Analiza morfologiczna

Przez analizę morfologiczną słów pojawiających się w tekstach rozumie się przyporządkowanie im odpowiednich form hasłowych leksemów oraz przypisywanie im charakterystyki gramatycznej (ustalanie wartości form wyrazowych). (Wołosz, 2005)

Interpretacją nazywam każde z możliwych odczytań słowa przypisujące mu pewną charakterystykę fleksyjną i leksem, do którego należy. (Rudolf, 2004)

Motywacje

- Automatyczna analiza morfologiczna na użytek przetwarzania języka to problem, którym do tej pory mało się zajmowano.
- Jednocześnie nie warto tworzyć zasobów dla polszczyzny dawnej zupełnie od podstaw, lepiej wykorzystać i dostosować zasoby i narzędzia już istniejące.
- Podstawowym zasobem tego typu dla polszczyzny współczesnej jest *Słownik gramatyczny języka polskiego* i wykorzystujący jego dane analizator morfologiczny Morfeusz SGJP.

Analizatory morfologiczne

- Morfeusz (SIaT i SGJP)
- PoMor
- Morfologik
- Polimorf (Morfeusz Polimorf)

Język dawniejszy w analizie automatycznej

- Podstawą leksykalną SGJP jest SJPDor., którego zakres materiałowy sięga tekstów nawet z połowy XVIII w.
- Dlatego SGJP zawiera wiele słownictwa dawnego i przestarzałego (blisko 16 tys. jednostek oznaczonych kwalifikatorami *daw.* lub *przest.*).
- Te same dane zawierają analizatory morfologiczne Morfeusz SGJP i PoMor.
- Jest to materiał użyteczny z punktu widzenia gromadzenia leksyki na potrzeby automatycznej analizy tekstów dawnych.
- Należy jednak pamiętać, że w SJPDor. i SGJP te jednostki leksykalne zostały zanotowane we współczesnej fleksji i grafii.

Ramy czasowe: 1830–1918

Początek:

- pierwsza instytucjonalna reforma ortograficzna przeprowadzona przez Warszawskie Towarzystwo Przyjaciół Nauk,
- wybuch powstania listopadowego na ziemiach zaboru rosyjskiego.

Koniec:

- odzyskanie przez Polskę niepodległości,
- koniec I wojny światowej rozumiany jako upadek XIX-wiecznego porządku w Europie i początek XX wieku w sensie historycznym, społecznym i kulturowym.

Pan Tadeusz (1834)

Za pomocą Morfeusza SGJP poddaliśmy analizie pierwszych pięć ksiąg „Pana Tadeusza” w pierwodruku oraz w jednym z wydań współczesnych.

W pierwodruku:

- 31059 słów (ciągów od spacji do spacji),
- 1896 ciągów nierozpoznanych (6,1%),
- w tym 1186 ciągów różnokształtnych.

W wydaniu współczesnym:

- 31285 słów,
- 559 ciągów nierozpoznanych (1,8%)
- w tym 434 ciągi różnokształtne.

Pan Tadeusz (1834)

Wojski cicho siedzący s przymrużuném okiem,
Zdawał się pogrążony w dumaniu głębokiém;
Dopiero gdy się Hrabia s Podkomorzym skłócił
I Sędziemu pogroził, Wojski głowę zwrócił.
Zażył dwakroć tabaki i przetarł powieki.
Chociaż Wojski Sędziemu był krewny daleki,
Ale w gościnnym jego domu zamieszkały,
O zdrowie przyjaciela był niezmiernie dbały.
Przypatrywał się zatém s ciekawością walce,
Wyciągnął zlekka na stół rękę dłoń i palce,
Położył nóż na dłoni, trzonkiem do paznokcia
Indexu, a żelazem zwrócony do łokcia,
Potém rękę w tył nieco wychyloną kiwał
Niby bawiąc się, lecz się w Hrabiego wpatrywał.

Pan Tadeusz (1834)

Początek listy frekwencyjnej form nierozpoznanych przez Morfeusza SGJP:

53	jéj	13	nakoniec	8	niemógł	5	roskazy
42	tém	13	któréj	8	lepiéj	5	pokryjomu
35	potém	12	wié	7	stoła	5	niewidział
25	assessor	12	swéj	7	niém	5	gdym
21	téj	11	wtém	7	całéj	5	dziecie
19	horeszków	11	drugiéj	6	zdaleka	5	dłużéj
18	niéj	10	jednéj	6	więcéj	5	assessora
18	daléj	10	dawniéj	6	niebyło	4	waszéj
15	nakształt	9	wielkiéj	6	każdéj	4	szczególniéj
14	zwolna	9	mniéj	5	takiéj	4	skołuba

Pan Tadeusz (1834)

Początek listy frekwencyjnej form nierozpoznanych przez Morfeusza SGJP:

53	jéj	13	**nakoniec**	8	niemógł	5	roskazy
42	tém	13	któréj	8	lepiéj	5	**pokryjomu**
35	potém	12	wié	7	stoła	5	niewidział
25	assessor	12	swéj	7	niém	5	gdym
21	téj	11	wtém	7	całéj	5	dziecie
19	horeszków	11	drugiéj	6	**zdaleka**	5	dłużéj
18	niéj	10	jednéj	6	więcéj	5	assessora
18	daléj	10	dawniéj	6	niebyło	4	waszéj
15	**nakształt**	9	wielkiéj	6	każdéj	4	szczególniéj
14	**zwolna**	9	mniéj	5	takiéj	4	skołuba

Pan Tadeusz (1834)

Początek listy frekwencyjnej form nierozpoznanych przez Morfeusza SGJP:

53	jéj	13	nakoniec	8	**niemógł**	5	roskazy
42	tém	13	któréj	8	lepiéj	5	pokryjomu
35	potém	12	wié	7	stoła	5	**niewidział**
25	assessor	12	swéj	7	niém	5	gdym
21	téj	11	wtém	7	całéj	5	dziecie
19	horeszków	11	drugiéj	6	zdaleka	5	dłużéj
18	niéj	10	jednéj	6	więcéj	5	assessora
18	daléj	10	dawniéj	6	**niebyło**	4	waszéj
15	nakształt	9	wielkiéj	6	każdéj	4	szczególniéj
14	zwolna	9	mniéj	5	takiéj	4	skołuba

Pan Tadeusz (1834)

Początek listy frekwencyjnej form nierozpoznanych przez Morfeusza SGJP:

53	jéj	13	nakoniec	8	niemógł	5	roskazy
42	tém	13	któréj	8	lepiéj	5	pokryjomu
35	potém	12	wié	7	**stoła**	5	niewidział
25	assessor	12	swéj	7	niém	5	gdym
21	téj	11	wtém	7	całéj	5	dziecie
19	horeszków	11	drugiéj	6	zdaleka	5	dłużéj
18	niéj	10	jednéj	6	więcéj	5	assessora
18	daléj	10	dawniéj	6	niebyło	4	waszéj
15	nakształt	9	wielkiéj	6	każdéj	4	szczególniéj
14	zwolna	9	mniéj	5	takiéj	4	skołuba

Pan Tadeusz (1834)

- męszczyzni
- rosprzestrzenił
- rospięta
- roskazał
- ssiadł
- zwycięsca
- zwiąski

Pan Tadeusz (1834)

- męszczyzni
- rosprzestrzenił
- rospięta
- roskazał
- ssiadł
- zwycięsca
- zwiąski
- dobiedz

Pan Tadeusz (1834)

- męszczyzni
- rosprzestrzenił
- rospięta
- roskazał
- ssiadł
- zwycięsca
- zwiąski
- dobiedz

- łapaj
- wziąść
- wrzaśli
- zrobim
- usłyszyć
- biegą
- niesą

Pan Tadeusz (1834)

- męszczyzni
- rosprzestrzenił
- rospięta
- roskazał
- ssiadł
- zwycięsca
- zwiąski
- dobiedz

- łapaj
- wziąść
- wrzaśli
- zrobim
- usłyszyć
- biegą
- niesą

- partyę
- assessor
- wassal
- karabella
- historye
- historja
- bestyi
- xięża

Pan Tadeusz (1834)

- męszczyzni
- rosprzestrzenił
- rospięta
- roskazał
- ssiadł
- zwycięsca
- zwiąski
- dobiedz

- łapaj
- wziąść
- wrzaśli
- zrobim
- usłyszyć
- biegą
- niesą

- partyę
- assessor
- wassal
- karabella
- historye
- historja
- bestyi
- xięża

- otoczon
- pniów
- srebrnemi
- ramiony

Plan prac

- Zgromadzenie korpusu o długości 1 mln słów, na który będzie się składać 1000 próbek po ok. 1000 segmentów, zachowującego zróżnicowanie stylistyczne stosowane w *Słowniku frekwencyjnym polszczyzny współczesnej*.

Plan prac

- Zgromadzenie korpusu o długości 1 mln słów, na który będzie się składać 1000 próbek po ok. 1000 segmentów, zachowującego zróżnicowanie stylistyczne stosowane w *Słowniku frekwencyjnym polszczyzny współczesnej*.
- Przeprowadzenie za pomocą niemodyfikowanego analizatora współczesnego (opartego na SGJP) wstępnej analizy fleksyjnej półmilionowego podkorpusu; jej celem będzie stworzenie listy jednostek nierozpoznanych.

Plan prac

- Zgromadzenie korpusu o długości 1 mln słów, na który będzie się składać 1000 próbek po ok. 1000 segmentów, zachowującego zróżnicowanie stylistyczne stosowane w *Słowniku frekwencyjnym polszczyzny współczesnej*.
- Przeprowadzenie za pomocą niemodyfikowanego analizatora współczesnego (opartego na SGJP) wstępnej analizy fleksyjnej półmilionowego podkorpusu; jej celem będzie stworzenie listy jednostek nierozpoznanych.
- Wzbogacenie listy jednostek o hasła *Słownika warszawskiego* nienotowane w podstawowym leksykonie analizatora.

Plan prac

- Sklasyfikowanie jednostek według przysługujących im paradygmatów.

Plan prac

- Sklasyfikowanie jednostek według przysługujących im paradygmatów.
- Powtórne przeanalizowanie półmilionowego podkorpusu za pomocą zmodyfikowanego analizatora. Kolejne modyfikacje analizatora będą w kolejnych krokach poddawane szczegółowej weryfikacji, aż do osiągnięcia zamierzonego progu poprawności analizy.

Plan prac

- Sklasyfikowanie jednostek według przysługujących im paradygmatów.
- Powtórne przeanalizowanie półmilionowego podkorpusu za pomocą zmodyfikowanego analizatora. Kolejne modyfikacje analizatora będą w kolejnych krokach poddawane szczegółowej weryfikacji, aż do osiągnięcia zamierzonego progu poprawności analizy.
- Przeprowadzenie ostatecznego testu na korpusie testowym o strukturze bliźniaczej do korpusu treningowego.

Oczekiwane efekty

- Opracowanie koncepcji leksykograficznej opisu zmian fleksyjnych i pisownianych w (elektronicznym) słowniku gramatycznym.
- opracowanie wzorców paradygmatycznych nieuwzględnionych w SGJP, ale funkcjonujących w badanym okresie w formacie umożliwiającym automatyczną analizę morfologiczną.
- Poszerzenie słownika analizatora morfologicznego o hasła wyrazów obecnych w korpusie, ale nierozpoznawanych przez analizator, w tym o słownictwo notowane w „Słowniku warszawskim”.
- Opracowanie niewielkiego (1 milion segmentów), oznakowanego fleksyjnie zrównoważonego korpusu tekstów z lat 1830–1918 i udostępnienie go wraz z programem do przeszukiwania według zadanych kryteriów.

Możliwe zastosowania

- Punkt wyjścia dla rozszerzania słownika morfologicznego o okresy wcześniejsze, przede wszystkim o I połowę doby nowopolskiego.
- Obszerny zrównoważony korpus polszczyzny XIX w. znakowany morfologicznie.
- Wykorzystanie w zasobach tworzonych w ramach nowej dziedziny o nazwie *digital humanities* — np. w narzędziach do przeszukiwania zdigitalizowanych tekstów dawnych.

Dziękujemy za uwagę!

Literatura

- Bajerowa, I., Polski język ogólny XIX wieku: stan i ewolucja. T. 1-3, Katowice 1986-2002.
- Klemensiewicz, Z., Historia języka polskiego, Warszawa 2002.
- Rudolf, M., Metody automatycznej analizy korpusu tekstów polskich, Warszawa 2004.
- Saloni, Z., M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, Słownik gramatyczny języka polskiego, Warszawa 2012.
- Woliński, M., Morfeusz — a practical tool for the morphological analysis of Polish [w:] Intelligent Information Processing and Web Mining, Advances in Soft Computing, [red.:] M. A. Kłopotek, S. T. Wierzchon, K. Trojanowski, s. 503–512. Springer-Verlag, Berlin 2006.
- Wołosz, R., Efektywna metoda analizy i syntezy morfologicznej w języku polskim, Warszawa 2005.