Korpus tekstów drugiej połowy doby nowopolskiej (1830–1918)

Joanna Bilińska, Danuta Skowrońska, Witold Kieraś
Magdalena Derwojedowa, Robert Wołosz

Uniwersytet Warszawski
Uniwersytet w Pécsu

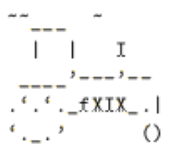
Grammar & Corpora
Warszawa, 25-27 czerwca 2014

Projekt

- *Automatyczna analiza fleksyjna tekstów polskich z lat 1830-1918 z uwzględnieniem zmian w odmianie i pisowni*
- projekt finansowany z środków Narodowego Centrum Nauki — grant nr DEC-2012/07/B/HS2/00570 (kierownik: Magdalena Derwojedowa)

Cele projektu

- opis systemowych zmian w zakresie odmiany polszczyzny pisanej w latach 1830–1918
- stworzenie słownika fleksyjnego ukazującego ewolucję odmiany
- stworzenie niewielkiego korpusu języka polskiego lat 1830–1918



Korpus — ogólnie

- cel 1.: identyfikacja ciągów nierozpoznanych przez analizator morfologiczny (`Morfeusz`, `PoMor`)
- cel 2.: testowanie nowych wzorców
- oznakowany
- udostępniony (przeszukiwanie według kryteriów fleksyjnych i metatekstowych)
- skład: 1 000 losowo wybranych próbek o długości ok. 1 000 segmentów (ok. 1 mln segmentów)

Dostępne zasoby diachroniczne

- „Korpus tekstów staropolskich do roku 1500” (https://www.ijp-pan.krakow.pl/pl/publikacje-elektroniczne/korpus-tekstow-staropolskich)
- „Słownik polszczyzny XVI wieku” (http://korpusy.klf.uw.edu.pl/pl/slownik-polszczyzny-xvi-wieku/)
- „Słownik polszczyzny XVII i 1. poł. XVIII wieku” (http://sxvii.pl/)
- „Elektroniczny korpus tekstów polskich XVII i XVIII w. (do 1772 r.)” (w budowie)

Zróżnicowanie stylistyczne korpusu

Korpus zawiera po 200 próbek wg 5 stylów: „Słownika frekwencyjnego polszczyzny współczesnej” (1993):

1. teksty popularnonaukowe
2. drobne wiadomości prasowe
3. publicystyka (w tym reportaż)
4. proza artystyczna (w tym powieść poetycka i poematy)
5. dramat artystyczny

Pliki

- UTF-8
- na wstępnym etapie pliki tekstowe (.txt)
- minimum dwa pliki: próbka i metryczka (metadane)
  `1830_4.1_sample.txt`,
  `1830_4.1_meta.txt`,
  `1830_4.1_source.txt`
- opcjonalny plik trzeci: cały tekst (różne formaty)
- ok. 1 000 segmentów w próbce
- repozytorium `ownCloud` (`http://owncloud.org/`)

Metadane

- autor
- tytuł
- data wydania
- redaktor
- tytuł książki
- tytuł gazety, czasopisma, serii wydawniczej
- nr czasopisma
- wydawnictwo
- numery stron
- styl
- źródło
- link

Przykładowa metryczka

autor: Słowacki, Juliusz
tytuł: Mnich
data wydania: 1832
redaktor:
tytuł książki: Poezye Juliusza Słowackiego. T. 1.
tytuł gazety, czasopisma,serii wydawniczej:
nr:
wydawnictwo: u Teofila Barrois Syna : u Hektora Bossange i Komp.
numery stron: 171-181
styl: proza
źródło: Polona
link: `http://www.polona.pl/item/1043213/95/`
i `http://pl.wikisource.org/wiki/Mnich_(S%C5%82owacki)`

Przykładowa próbka

Już dawno doznajemy wszędzie i wiele smntku, złorzeczenia i prześladowania i ani na włos lepiéj nam się nie dzieje, jak się działo pierwszym chrześcianom i ich kapłanom za panowania cesarzów pogańskich. Lecz jak owi kapłani cierpiących swych braci i sióstr nie zdołali inaczéj pocieszyć jak słowy wielkiego Apostoła Pawła świętego: „W wszystkiém utrapienie cierpiemy (...)"

Duchowieństwo Rzymsko-Katolickie Wielkiego Xięstwa Poznańskiego, *Odezwa do ludu katolickiego Wielkiego Xięstwa Poznańskiego*, 1848.

Źródła tekstów — przykłady

- Polona (`http://www.polona.pl/`)
- Kujawsko-Pomorska Biblioteka Cyfrowa (`http://kpbc.umk.pl/dlibra`)
- Wielkopolska Biblioteka Cyfrowa (`http://www.wbc.poznan.pl/dlibra`)
- WikiSource (`http://pl.wikisource.org/`)
- WolneLektury (`http://wolnelektury.pl/`)
- Rolnicza Biblioteka Cyfrowa (`http://delta.cbr.edu.pl/dlibra`)
- Katalog HINT (`http://hint.org.pl/`)
- i inne

Metody pozyskiwania tekstów

- kopiowanie zapisu w oryginalnej wersji ortograficznej
- nanoszenie korekty na zapis z nowszego wydania
- OCR za pomocą FineReadera (z korektą)
- przepisywanie

Słaby skan

DZIENNIK POZNAŃSKI.

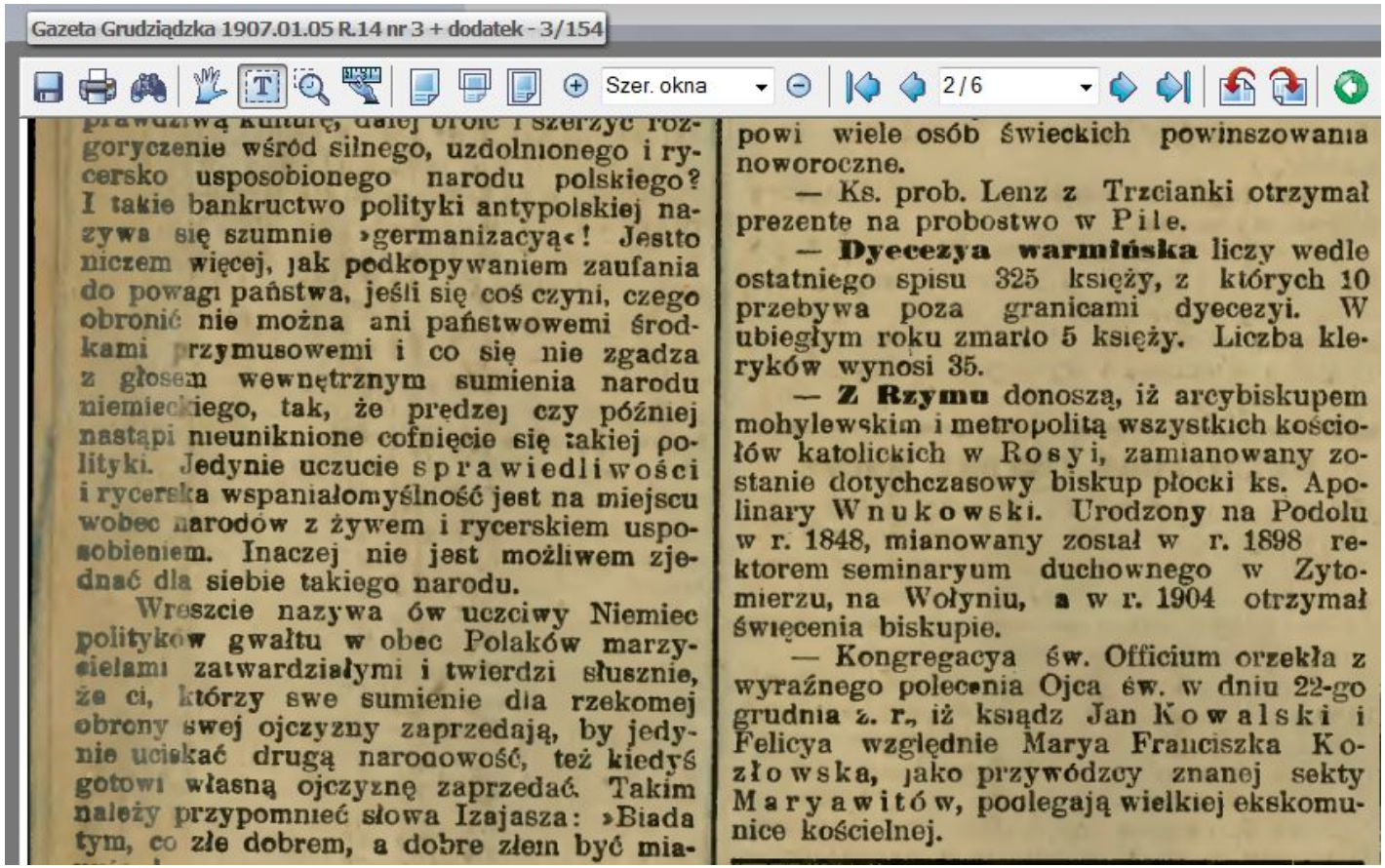
№ 1. Sobota, 1 stycznia. 1859

Z powodu Nowego roku, niedziela, najbliższy numer wyjdzie dopiero d. 4 stycznia.

PROSPEKT*).

DZIENNIK POZNAŃSKI

Dobry skan



prawdziwą kulturę, dalej broić i szerzyć rozgoryczenie wśród silnego, uzdolnionego i rycersko usposobionego narodu polskiego? I takie bankructwo polityki antypolskiej nazywa się szumnie »germanizacyą«! Jestto niczem więcej, jak podkopywaniem zaufania do powagi państwa, jeśli się coś czyni, czego obronić nie można ani państwowemi środkami przymusowemi i co się nie zgadza z głosem wewnętrznym sumienia narodu niemieckiego, tak, że prędzej czy później nastąpi nieuniknione cofnięcie się takiej polityki. Jedynie uczucie sprawiedliwości i rycerska wspaniałomyślność jest na miejscu wobec narodów z żywem i rycerskiem usposobieniem. Inaczej nie jest możliwem zjednać dla siebie takiego narodu.

Wreszcie nazywa ów uczciwy Niemiec polityków gwałtu w obec Polaków marzycielami zatwardziałymi i twierdzi słusznie, że ci, którzy swe sumienie dla rzekomej obrony swej ojczyzny zaprzedają, by jedynie uciskać drugą narodowość, też kiedyś gotowi własną ojczyznę zaprzedać. Takim należy przypomnieć słowa Izajasza: »Biada tym, co złe dobrem, a dobre złem być mia-

powi wiele osób świeckich powinszowania noworoczne.

— Ks. prob. Lenz z Trzcianki otrzymał prezente na probostwo w Pile.

— **Dyecezya warmińska** liczy wedle ostatniego spisu 325 księży, z których 10 przebywa poza granicami dyecezyi. W ubiegłym roku zmarło 5 księży. Liczba kleryków wynosi 35.

— **Z Rzymu** donoszą, iż arcybiskupem mohylewskim i metropolitą wszystkich kościołów katolickich w Rosyi, zamianowany zostanie dotychczasowy biskup płocki ks. Apolinary Wnukowski. Urodzony na Podolu w r. 1848, mianowany został w r. 1898 rektorem seminaryum duchownego w Zytomierzu, na Wołyniu, a w r. 1904 otrzymał święcenia biskupie.

— Kongregacya św. Officium orzekła z wyraźnego polecenia Ojca św. w dniu 22-go grudnia z. r., iż ksiądz Jan Kowalski i Felicya względnie Marya Franciszka Kozłowska, jako przywódzcy znanej sekty Maryawitów, podlegają wielkiej ekskomunice kościelnej.

Przykład OCR-u

– Z Rzymu donoszą, iż arcybiskupem
mohylewc;kim i metropolitą wszystkich kościołów katolickich w Ro s Y i, zamianowany zostanie dotychczasowy biskup płOCki ks
Apo.
linal'Y W n u k o W ski. Urodzony na Podolu
w r
1848, mianowany został w r. 1898 rektorem seminaryum duchownego w Zyto.
mierzu, na W ołyniu, a w r. 1904 otrzymał
śWięcenia biskupie.

Problemy merytoryczne

- zmieszczenie wszystkich lat/dekad w korpusie
- ograniczenia tematyki i leksyki prasowej (prasa religijna)
- nadreprezentacja niektórych autorów
- nadmiar stylizacji (J.I. Kraszewski, J. Słowacki)

Problemy techniczne

- konieczność przepisywania niektórych gazet
- brak zeskanowanych pierwszych wydań
- nanoszenie korekty na podstawie skanu pierwszego wydania
- wybieranie próbki — liczby losowe i intuicja
- konieczność weryfikacji wydania
- niedostateczna długość tekstów
- niska jakość warstwy tekstowej dokumentów w formacie DjVu

Przykładowe zmiany

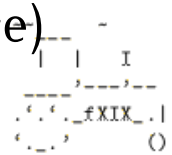
(1) a. Jak poeta... Lecz xiężyc zabłysnął dwurożny
b. Jak poeta... Lecz księżyc zabłysnął dwurożny

(2) a. Kordjan : część pierwsza trilogji. Spisek koronacyjny.
b. Kordian. Część pierwsza trylogii. Spisek koronacyjny

(J. Słowacki, *Kordian* 1834 i 1974)

(3) a. Zkądże, serce, ów dziesiątek?
b. Skądże, serce, ów dziesiątek?

(A. Fredro, *Dożywocie* 1838 i tekst z WikiSource)



Przykładowe zmiany cd.

(4) a. (...) ciągle szła przedemną,
b. (...) ciągle szła przede mną,

(J. Słowacki, *Balladyna*, 1834 i 1974)

(5) a. Matki nieznaleść któraby jéj niezazdrościła
b. Matki nie znaleźć, któraby jej nie zazdrościła

(Z. Krasiński, *Noc letnia*, 1841 i 1904)

Dalsze prace z korpusem

- analiza fleksyjna podkorpusu (0,5 mln) za pomocą współczesnego analizatora morfologicznego (rezultat: nierozpoznane jednostki)
- analiza podkorpusu za pomocą zmodyfikowanego analizatora
- oznakowanie/ujednoznacznienie fleksyjne całego korpusu
- wyszukanie wyników fałszywie pozytywnych

Zapraszamy...

http://www.f19.uw.edu.pl/

Bibliografia I

M. Derwojedowa, W. Kieraś, D. Skowrońska, R. Wołosz, *Zasób leksykalny polszczyzny II poł. XIX wieku a możliwość automatycznej analizy morfologicznej tekstów z tego okresu*, [w:] „Leksyka języków słowiańskich w badaniach synchronicznych i diachronicznych", Toruń 2014. (w druku)

M. Derwojedowa, W. Kieraś, D. Skowrońska, R. Wołosz, *Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych*, Polonica, 2014. (w druku)

I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak, *Słownik frekwencyjny polszczyzny współczesnej*, Z. Saloni (red.), Kraków, Warszawa 1990.

Bibliografia II

Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, *Słownik gramatyczny języka polskiego*, Warszawa, 2012.

M. Woliński *Morfeusz — a Practical Tool for the Morphological Analysis of Polish*, [w:] M. A. Kłopotek, S. T. Wierzchoń i K. Trojanowski (red.) *Intelligent Information Processing and Web Mining*, Springer-Verlag, 2006, 503-512.

M. Woliński, *Morfeusz reloaded*, [w:] N. Calzolari i in. (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014, s. 1106–1111, Reykjavík, Iceland, 2014.

R. Wołosz, *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*, Warszawa, 2005.