POLONICA XXXIV
PL ISSN 0137-9712

MAGDALENA DERWOJEDOWA
WITOLD KIERAŚ
DANUTA SKOWROŃSKA
ROBERT WOŁOSZ

Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych

Wprowadzenie

Współczesna polszczyzna doczekała się narzędzi do przetwarzania dużych zasobów tekstów (analizatorów morfologicznych, tagerów i analizatorów składniowych) oraz zasobów leksykalnych (wordnet, słowniki walencyjne) (Dębowski 2003, Woliński 2004, Wołosz 2005, Woliński 2006, Derwojedowa et al. 2007, Piasecki 2007, Acedański 2010, Maziarz et al. 2012). Badacz polszczyzny dawniejszej jest w trudniejszej sytuacji, bo choć zasobów przybywa (słowniki polszczyzny od XVI wieku, zdygitalizowane słowniki XIX-wieczne oraz korpus polszczyzny średniowiecznej), narzędzi do przetwarzania tekstów ciągle nie ma. Jednak w miarę jak przybywa narzędzi i zasobów do analizy współczesnego języka polskiego, rośnie zainteresowanie automatyczną analizą tekstów dawniejszych, by wymienić tylko prace nad korpusem barokowym[1] będącego częścią programu opisu polszczyzny tego okresu. Sądzimy, że w niedługim czasie można się spodziewać prac zmierzających do stworzenia korpusu polskich tekstów dziewiętnastowiecznych. By korpus ten mógł służyć z powodzeniem zainteresowanym, powinny mu towarzyszyć narzędzia, w tym analizator morfologiczny.

Kilka wstępnych testów przekonało nas, że przynajmniej niektóre istniejące narzędzia i zasoby mogą być użyteczne do analizy tekstów dawniejszych (ściślej: nowopolskich) nawet w obecnej, nastawionej na opis polszczyzny współczesnej, postaci. W artykule przedstawiamy jeden z przeprowadzonych testów oraz zarys koncepcji rozszerzenia istniejących sformalizowanych opisów morfologicznych (przede wszystkim

[1] Projekt „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772)", nr 0036/NPRH2/H11/81/2012, kierownik: Włodzimierz Gruszczyński. Projekt jest prezentowany również w niniejszym tomie.

„Słownika gramatycznego języka polskiego" (dalej: SGJP)), tak by poprawnie analizowały dawne formy odmiany i dawne formy pisowniane[2]. Przez analizę fleksyjną rozumiemy zarówno lematyzację (przyporządkowanie formy hasłowej), jak i przyporządkowanie charakterystyki gramatycznej.

Stosunkowo najmniej zmian i rozszerzeń wymaga opis tekstów względnie bliskich współczesności. Dlatego na pierwszy etap naszych prac wybraliśmy polszczyznę drugiej połowy doby nowopolskiej, a ściślej język lat 1830–1918. Pierwsza z dat wyznaczających wybrane przez nas ramy czasowe to oczywiście rok wybuchu powstania listopadowego na ziemiach zaboru rosyjskiego, ale również rok pierwszej instytucjonalnej reformy ortograficznej w historii polszczyzny przeprowadzonej przez Warszawskie Towarzystwo Przyjaciół Nauk (Bajerowa 1984). I choć reforma nie osiągnęła wszystkich celów, jakie stawiali przed nią jej twórcy, to w dłuższej perspektywie przyczyniła się do stopniowego ujednolicania polskiej ortografii, co w kontekście systematycznego opisu fleksyjnego polszczyzny pisanej jest niezwykle istotne.

Druga z dat — rok 1918 — to cezura wyznaczająca zmiany społeczne, polityczne i kulturowe w całej Europie. W przypadku Polski to również rok odzyskania niepodległości, początek spajania ziem trzech zaborów, jak również spajania polszczyzny i piśmiennictwa. Data ta ma więc wymiar symboliczny, w praktyce wyniki analizy będą mogły z powodzeniem posłużyć do automatycznego przetwarzania tekstów późniejszych — aż do roku 1936, czyli do reformy ortograficznej (Klemensiewicz 2002: 664–665).

Podstawowym celem projektu jest opis systemowych zmian w zakresie odmiany polszczyzny pisanej w latach 1830–1918, a więc w drugiej połowie doby nowopolskiej (Klemensiewicz 2002: 495) w sformalizowanej postaci wzorców paradygmatycznych, w których odnotowane zostaną:

— dawne postaci końcówek fleksyjnych, np. powszechne w XIX w. *-em*, *-emi* w formach narzędnika przymiotników,

— dawne postaci form fleksyjnych, np. narzędnik liczby mnogiej rzeczowników (*słowy*, *usty*, *skrzydły*),

— zmiany wzorców odmiany, np. z męskiego na żeński w KOMETA,

— dawne postaci pisowniane (odmienne oznaczanie głosek, oznaczanie samogłosek pochylonych), także w zakresie pisowni łącznej i rozdzielnej, np. *fantazya*, *xiądz*, *zielonémi*, *zlekka*.

Wzorce te będą odzwierciedlały ewolucję form, tzn. docelowo będzie możliwe prześledzenie przemian poszczególnych jednostek. Równocześnie dane słownika będą stanowić podstawę rozszerzenia analizatora morfologicznego, który w ten sposób zostanie przystosowany do analizy fleksyjnej tekstów dziewiętnastowiecznych.

Ubocznym, choć całkowicie zamierzonym, celem projektu jest stworzenie niewielkiego (1 mln segmentów) korpusu języka polskiego lat 1830–1918. Zasób taki jest niezbędny do identyfikacji ciągów nieznanych współczesnemu analizatorowi oraz do

[2] Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2012/07/B/HS2/00570.

testowania dodawanych wzorców. W toku prac korpus zostanie oznakowany automatycznie, znakowanie to zostanie następnie zweryfikowane ręcznie. Tak opracowany korpus zostanie udostępniony.

Wstępne testy, do których użyliśmy analizatorów Morfeusz[3] (Woliński 2006) i PoMor (Wołosz 2005), przekonały nas, że nie warto tworzyć zamierzanych przez nas zasobów zupełnie od podstaw, lepiej wykorzystać istniejące narzędzia, dostosowując je, gdzie trzeba. Jesteśmy przy tym przekonani, że projektowany słownik czy słowniki powinny ostatecznie przedstawiać nie tylko stany statyczne (synchroniczne), ale też zdawać sprawę z ewolucji form.

W obecnej postaci Morfeusz analizuje zdecydowaną większość słów tekstowych z tekstów polskich wydanych po roku 1830. Nie dziwi to, gdyż podstawę jego słownika stanowi lista wyrażeń hasłowych SJPDor, do której ekscerpowano teksty nawet z połowy XVIII wieku. To, co analizator pozostawia jako nierozpoznane, to — poza oczywistymi w przetwarzaniu tekstów z dowolnej epoki brakami leksykalnymi, szczególnie wśród nazw własnych — dawne postaci form fleksyjnych i postaci pisowniane dziś nieużywane.

Przykładowa próba analizy tekstu dawnego narzędziami współczesnymi

Badając możliwość wykorzystania współczesnych analizatorów, wykonaliśmy kilka testów. Na początek przeanalizowaliśmy za pomocą Morfeusza SGJP[4] dwie wersje dzieła z samego początku badanego okresu, a mianowicie pierwszych pięć ksiąg „Pana Tadeusza": dostępny przez Projekt Gutenberg (http://www.gutenberg.org/) tekst pierwszego wydania i tekst jednego z wydań współczesnych (którego tekst jest zgodny z wersją opublikowaną w IV tomie „Dzieł" Mickiewicza z 1955 roku). Celem analizy było porównanie liczby nierozpoznanych słów tekstowych; oczywiste jest, że spodziewaliśmy się, że liczba ta będzie większa dla tekstu pierwodruku. Trzeba jednak pamiętać, że redaktorzy wersji uwspółcześnionych mogli zachować formy dawne ze względu na rym i metrum. Na przykład we fragmencie: *Chybkim był zostawiony nóżkami drobnemi* / *Od kogoś, co zaledwie dotykał się ziemi* przymiotnik *drobnemi* pozostał w dawnej postaci ze względu na rym, ale już we fragmencie *Oczyma ciekawemi po drożynach gonił* współczesne wydania modyfikują postać przymiotnika na *ciekawymi*. Tak więc można się było spodziewać, że liczba ta będzie również większa niż w wypadku tekstu prozatorskiego z tego samego okresu. Z drugiej strony ze względu na doniosłość tego dzieła dla kultury polskiej można było oczekiwać, że nazwy własne — które siłą rzeczy są w analizatorach pomijane lub reprezentowane przez pojedyncze okazy wyłącznie ze względu na typ odmiany — zostaną rozpoznane lepiej niż w innych typach tekstu.

---

[3] Dostępny w Internecie pod adresem www.sgjp.pl/morfeusz.

[4] Morfeusz SGJP to wersja analizatora oparta na danych SGJP, w odróżnieniu od przestarzałej, nierozwijanej wersji Morfeusz SIaT (opracowanej w oparciu o dane „Schematycznego indeksu a tergo polskich form wyrazowych") oraz wersji Morfeusz Polimorf (ze słownikiem Polimorf).

W wydaniu z 1834 roku znajduje się 31 059 słów tekstowych, w wydaniu współczesnym — 31 285 słów. Niewielka różnica w liczbie segmentów wynika głównie z różnych zasad pisowni łącznej i rozłącznej — w XIX wieku chętniej stosowano pisownię łączną, np. *zlekka*, *zdala*. Analizator nie przypisał żadnej interpretacji fleksyjnej 1896 słowom tekstowym pierwodruku. Odpowiada to 1186 ciągom różnokształtnym i stanowi około 6,1% wszystkich analizowanych słów. W wydaniu współczesnym ciągów nierozpoznanych było jedynie 559 (434 różnokształtne), a więc zaledwie 1,8%. Potwierdza to tezę, że lepiej jest rozszerzać istniejące zasoby i narzędzia niż tworzyć nowe od podstaw.

Kilka grup wyrazów nierozpoznanych zwraca uwagę. Po pierwsze, są to słowa notowane przez słowniki współczesne, ale z *e* pochylonym, takie jak *jéj*, *potém*, *niéj*, *wielkiéj*, *wtém* itd. Część z nich — formy przymiotnikowe — są regularne, tak więc włączenie ich do słownika polega na modyfikacji paradygmatu. Pozostałe można opisać jako osobne jednostki lub włączyć do paradygmatów jako formy wariantywne oznaczone jako dawne.

Druga grupa to ciągi odpowiadające grupom przyimkowo-nominalnym współcześnie zapisywanym rozdzielnie, a w tekstach dziewiętnastowiecznych często zapisywane łącznie, np. *nakształt*, *zwolna*, *nakoniec*, *pokryjomu*. Wymagają one podjęcia decyzji co do tego, do jakiej klasy należą — można je bowiem rozdzielić i opisywać zgodnie z ich współczesną klasyfikacją, zachowując paralelność opisu gramatycznego z opisem współczesnym, można też uznać, że są to po prostu formy leksemów i zakwalifikować je według ich własności fleksyjnych, np. jako przysłówki.

Kolejną grupą, podobną nieco do poprzedniej, są formy czasownikowe pisane łącznie z partykułą *nie*, np. *niemógł*, *niebyło*, *niewidział*, *niema* (lub *niéma*). W tym wypadku trzeba zastosować regułę analizatora analogiczną do reguł umożliwiających analizowanie zaprzeczonych form przymiotników (w tym tradycyjnie rozumianych imiesłowów), wpisywanie wszystkich takich form do słownika jest bowiem przynajmniej nieekonomiczne. Przypadek formy *niema* wskazuje ponadto, że nie wystarczy przeglądać jedynie ciągi niezanalizowane, ale również wyniki pozytywne — jest to bowiem również forma leksemu rzeczownikowego NIEMA i przymiotnikowego NIEMY.

Uwagę zwracają również wyrazy pochodzenia obcego zapisane inaczej niż współcześnie, niekiedy z zachowaniem pisowni języka źródłowego (por. Bajerowa 1984), np. *partyę*, *assessor*, *wassal*, *karabella*, *bestyi*, *xiężą*; zdarza się, że w tym samym tekście są one zapisane różnie, np. *historye*, ale *historja.* Oba te sposoby trzeba uwzględnić jako wariantywne.

Również wyrazy rodzime mają inną niż współcześnie pisownię, często ubezdźwięczniającą, np. *rosprzestrzenił*, *rospięta*, *roskazał*, *ssiadł*, *zwycięsca*, *zwiąski*, zdarzają się też przypadki odwrotne, np. *dobiedz*. W tym wypadku należałoby oznaczyć formy wariantywne dawne.

Innym problemem są obecnie nieużywane postaci form fleksyjnych jednostek współcześnie używanych, np. *ramiony* (N. l.mn. RAMIĘ), *pniów* (D. l.mn. PIEŃ), *stoła* (D. l.p. STÓŁ), lub takich, które obecnie mogą być uznane za regularne, lecz potencjalne, np. *otoczon*. Sądzimy, że racjonalnie byłoby je włączyć do paradygmatów, ozna-

czając jako dawne lub przewidzieć w paradygmatach osobne (potencjalnie niewypełnione lub synkretyczne) klatki dla takich form dawnych. Osobną grupą, choć podobnym problemem, są formy dziś uważane za gwarowe lub substandardowe, np. *łapaj*, *wziąść*, *wrzaśli*, *zrobim*. Jeśli miałyby być one notowane, należałoby je traktować jak formy dawne, ewentualnie dodatkowo oznaczając.

Jak widać, w stosunkowo niedużej próbce tekstu z I połowy XIX wieku znaleźć można wiele typów zjawisk, które muszą zostać uwzględnione w rygorystycznym opisie fleksji II połowy doby nowopolskiej. Oczywiście, im tekst jest późniejszy, tym jego postać fleksyjna i pisowniana jest bliższa współczesnej — np. w połowie XIX wieku zanika pisownia z *e* pochylonym — jednak wiele z opisanych powyżej przykładowych zjawisk w różnym natężeniu utrzymywało się w tekstach różnych autorów do początków XX wieku.

Opis fleksji II poł. XIX wieku

Podjęte przez nas zadanie wymaga systematycznego opisu zjawisk fleksyjnych i wariantów ortograficznych w tekstach II poł. XIX wieku. Stworzenie takiego opisu wymaga zgromadzenia korpusu. Pierwszą zatem fazą prac jest zgromadzenie korpusu o długości 1 miliona słów, na który będzie się składać 1000 tysiącsegmentowych próbek. W korpusie chcemy zachować zróżnicowanie stylistyczne odpowiadające w zasadzie zróżnicowaniu korpusu „Słownika frekwencyjnego polszczyzny współczesnej" (dalej SFPW). Taki dobór stylów jest oczywiście do pewnego stopnia arbitralny, jednak wydaje się, że nieco może konserwatywne rozstrzygnięcia autorów SFPW (reprezentowanie języka mówionego przez dramat) w stosunku do tekstów dawniejszych mają lepsze odniesienie niż na przykład kryterium rozpowszechnienia (czytelnictwa) — inna była bowiem pozycja prasy, inna literatury, a wiele współczesnych typów tekstów po prostu nie istniało (mówione radiowe czy telewizyjne, pisane z forów internetowych, listów elektronicznych itp.).

Połowa korpusu posłuży jako materiał do wzbogacania analizatora, tzn. do wyszukiwania form, którym nie jest on w stanie przypisać żadnej interpretacji fleksyjnej. Po odrzuceniu nazw własnych i elementów obcych (cyfr, cytatów obcojęzycznych itp., por. Saloni 1973) dawne formy fleksyjne i pisowniane zostaną sklasyfikowane. Dodatkowo lista jednostek leksykalnych zostanie wzbogacona o hasła „Słownika warszawskiego" (1927) nieobecne w podstawowym leksykonie analizatora, czyli w zasobach SGJP i wśród jednostek zgromadzonych na podstawie analizy korpusu. Ta część korpusu posłuży też do bieżącej kontroli pracy modyfikowanego analizatora i usuwania błędów aż do osiągnięcia wyznaczonego poziomu poprawności. Ostateczna weryfikacja zostanie przeprowadzona na drugiej połowie korpusu. Takie podejście zapewnia rzetelność analizy i pozwala przyjąć, że zgromadzone dane mogą posłużyć do zadowalającej jakościowo analizy innych tekstów z tego okresu.

Ponieważ podczas pracy nad analizatorem zanalizowany korpus będzie przeglądany, na tym etapie prac możliwe jest też odrzucenie analiz poprawnych fleksyjnie,

ale składniowo w danym kontekście niemożliwych, czyli ich ujednoznacznienie, dane w tej postaci zostaną udostępnione jako korpus.

Oczekiwane efekty i możliwe zastosowania

Najważniejszym celem teoretycznym jest koncepcja leksykograficznego, zdyscyplinowanego opisu zmian fleksyjnych i pisownianych w (elektronicznym) słowniku gramatycznym. Z zagadnieniem tym wiąże się wiele problemów szczegółowych, jak segmentacja jednostek, dołączanie form wariantywnych do paradygmatów i tworzenie nowych klatek w paradygmatach, umieszczanie w paradygmatach form regularnie tworzonych, ale niepoświadczonych tekstowo (a więc rekonstrukcja paradygmatów), notowanie stopnia wariantywności (możliwe są przecież zapisy rzadkie czy osobnicze), dodawanie nowych jednostek i wiązanie ich z istniejącymi, o ile postać graficzna i własności nie pozwalają ich uwzględnić jako form wariantywnych, oraz tworzenie paradygmatów dla jednostek nienależących do żadnego z istniejących w analizatorze.

Dodatkowym efektem prac będzie opracowanie niewielkiego — wielkości około miliona słów — oznakowanego fleksyjnie i ujednoznacznionego, zrównoważonego korpusu tekstów z lat 1830–1918 i udostępnienie go.

Zarówno analizator, jak i korpus mogą stanowić bazę do dalszych rozszerzeń, przede wszystkim rozbudowy analizy fleksyjnej o formy fleksyjne i pisowniane notowane w pierwszej połowie doby nowopolskiej, a następnie o kolejne okresy historyczne. Sam korpus może stać się zalążkiem większego korpusu tekstów dziewiętnastowiecznych, a także posłużyć jako zbiór treningowy dla innych narzędzi do przetwarzania języka naturalnego[5].

Literatura

A c e d a ń s k i S., 2010, A Morphosyntactic Brill Tagger for Inflectional Languages, [w:] Advances in Natural Language Processing. 7th International Conference on NLP, IceTAL 2010.

B a j e r o w a I., 1984, Polski język ogólny XIX wieku: stan i ewolucja. Ortografia, fonologia z fonetyką, morfonologia, t. I, Katowice.

D e r w o j e d o w a M., P i a s e c k i M., S z p a k o w i c z S., Z a w i s ł a w s k a M., 2007, Polish WordNet on a shoestring, [w:] Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology, Tübingen, April 11–13 2007, Universität Tübingen, red. G. Rehm, A. Witt, L. Lemnitzer, s. 169–178.

D ę b o w s k i Ł., 2003, A reconfigurable stochastic tagger for languages with complex tag structure, [w:] The Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL, s. 63–71.

K a r ł o w i c z J., K r y ń s k i A., N i e d ź w i e d z k i W., Słownik języka polskiego, Warszawa 1900–1927.

K l e m e n s i e w i c z Z., 2002, Historia języka polskiego, Warszawa.

[5] Postępy w pracach nad projektem i inne informacje można śledzić na stronie http://www.f19.uw.edu.pl/.

Kurcz I., Lewicki A., Sambor J., Szafran K., Woronczak J., 1992, Słownik frekwencyjny polszczyzny współczesnej, Warszawa.
Maziarz M., Piasecki M., Szpakowicz S., 2012, Approaching plWordNet 2.0, [w:] Proceedings of the 6th Global Wordnet Conference.
Piasecki M., 2007, Polish Tagger TaKIPI: Rule Based Construction and Optimisation, Task Quarterly, t. 11 (1–2), s. 151–167.
Saloni Z., 1973, Alien Elements on Texts of a Natural Language, Biuletyn Polskiego Towarzystwa Językoznawczego XXXI.
Saloni Z., Woliński M., Wołosz R., Gruszczyński W., Skowrońska D., 2012, Słownik gramatyczny języka polskiego, II wyd., Warszawa.
Słownik języka polskiego, red. W. Doroszewski, Warszawa 1958–1969.
Woliński M., 2004, Komputerowa weryfikacja gramatyki Świdzińskiego, Rozprawa doktorska, Instytut Podstaw Informatyki, Polska Akademia Nauk.
— 2006, Morfeusz — a Practical Tool for the Morphological Analysis of Polish, [w:] Intelligent Information Processing and Web Mining, red. M.A. Kłopotek, S.T. Wierzchoń, K. Trojanowski, Advances in Soft Computing, Berlin, Springer-Verlag, s. 503–512.
Wołosz R., 2005, Efektywna metoda analizy i syntezy morfologicznej w języku polskim, Warszawa.

SUMMARY

Contemporary linguistic tools in analysis of historical texts

Keywords: 19th century Polish, morphological analysis, natural language processing, inflection.
Słowa kluczowe: polszczyzna XIX w., analiza morfologiczna, przetwarzanie języka, fleksja.

The paper presents results of preliminary tests of an application of Morfeusz morphological analyzer equipped with the data of "Grammatical Dictionary of Polish" (Woliński 2006, Saloni et al. 2012) to the analysis of texts from the second half of 19th century. We were motivated by the hypothesis that an analyzer with an extensive lexical base can be adapted for an analysis of historical texts, therefore there is no need to create new tools for this task. In the paper we present the base concepts of such analysis with examples from books I–V of "Pan Tadeusz", which served as a test data in the experiment.