Marcin Woliński
(e-mail: wolinski@ipipan.waw.pl)
ORCID: 0000-0002-7498-1484
Witold Kieraś
(e-mail: w.kieras@ipipan.waw.pl)
ORCID: 0000-0002-8062-5881
(Instytut Podstaw Informatyki
Polskiej Akademii Nauk, Warszawa)

DOI: 10.33896/PorJ.2020.8.5

ANALIZA FLEKSYJNA TEKSTÓW HISTORYCZNYCH I ZMIENNOŚĆ FLEKSJI POLSKIEJ Z PERSPEKTYWY DANYCH KORPUSOWYCH

Język polski, jak każdy język naturalny, podlega ciągłym zmianom. Historycy języka do niedawna dysponowali wyłącznie wyrywkowymi danymi, na podstawie których formułowali swoje hipotezy badawcze. Obecnie dostępne są korpusy tekstów dawnych, które można badać metodami automatycznymi na dużo większą skalę. Do prowadzenia takich badań niezbędne są wyspecjalizowane narzędzia informatyczne. Opracowanie narzędzi komputerowego modelowania polskiej fleksji historycznej było przedmiotem projektu „Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja" (projekt Narodowego Centrum Nauki nr 2014/15/B/HS2/03119) prowadzonego w IPI PAN w latach 2015–2019.

Głównym wynikiem projektu jest system komputerowy Chronofleks [http://chronofleks.nlp.ipipan.waw.pl/], który oferuje nowe spojrzenie na korpus diachroniczny. System składa się z dwóch części. Pierwsza z nich może być widziana jako prototyp słownika fleksyjnego prezentującego zmienność paradygmatów fleksyjnych poszczególnych leksemów w czasie. Druga część pozwala badać zmienność w czasie frekwencji wybranych form fleksyjnych konkretnych leksemów lub grup leksemów. Jest to pierwsze narzędzie dające łatwy dostęp do informacji ilościowej – możliwe jest badanie nie tylko tego, jakie formy danych leksemów są potwierdzone, ale też ile jest tych potwierdzeń i jak są one rozłożone w czasie. W dalszej części artykułu przedstawimy możliwości badawcze, jakie się dzięki temu otwierają.

Praca z tekstem historycznym różni się istotnie od pracy z tekstem współczesnym, ponieważ dotyczy ona w istocie języka martwego. Nie ma jego żyjących użytkowników, a badacz, zwłaszcza w wypadku bardziej odległych epok, nie może w pełni polegać na swoim wyczuciu językowym. Rekonstrukcja form niepoświadczonych w materiale może prowadzić do powstawania form fikcyjnych, nigdy niewystępujących w realnych tek-

stach. Dlatego w projekcie przyjęto zasadę uwzględniania w modelowaniu wyłącznie poświadczonych form fleksyjnych. Przedstawiane przez system Chronofleks paradygmaty mają potwierdzenie w postaci twardych danych ekstrahowanych ze znakowanych korpusów. Oczywiście w związku z tym prezentowane paradygmaty są kompletne tylko dla najczęstszych wyrazów.

Aby możliwa była taka analiza, konieczne jest odpowiednie opracowanie korpusu: wszystkie słowa muszą zostać poddane analizie morfologicznej, a więc muszą one zostać przypisane do leksemów (zlematyzowane) i scharakteryzowane co do pełnionej funkcji gramatycznej. W pierwszej części artykułu przedstawiono przyjęte zasady takiego opracowania oraz służące do tego narzędzia.

PRZETWARZANIE FLEKSYJNE TEKSTÓW ŚREDNIO- I NOWOPOLSKICH

Analiza morfologiczna i tagowanie stanowią jeden z podstawowych etapów przetwarzania tekstów zarówno w wypadku budowy korpusu, jak i w innych zadaniach NLP. Dla współczesnych języków o bogatej fleksji zagadnienie to jest zwykle rozwiązywane przez zestawienie obszernego słownika wykorzystywanego przez analizator i budowę modelu ujednoznaczniającego [Kobyliński, Kieraś 2016; Przepiórkowski i in. 2012]. Przy zastosowaniu tej metodologii do tekstów historycznych napotyka się problemy powodujące, że jakość znakowania jest dużo gorsza.

Problemy te obejmują: rozchwianie ortograficzne; dawne formy fleksyjne nienotowane przez współczesne słowniki; problem leksykalny – słownictwo nieobecne we współczesnym języku. Przy przetwarzaniu tekstów dawnych mamy problem języka obcego, dla którego brak odpowiednich zasobów. Rozwiązanie problemu jest analogiczne: próbujemy przekształcić zasoby dostępne dla podobnego języka – w tym wypadku współczesnej polszczyzny. Taką właśnie procedurę zastosowano dla historycznych tekstów polskich.

Brak ustabilizowanej ortografii w tekstach dawnych sprawia, że przy ich przetwarzaniu zwykle jako pierwszy krok stosuje się normalizację zapisu (transkrypcję tekstów). Może ona mieć różny zakres. Niektórzy badacze stosują transkrypcję w pełni uwspółcześniającą tekst (również pod względem fleksyjnym, a nawet leksykalnym) – tak postąpili m.in. twórcy korpusu historycznego języka słoweńskiego [Erjavec 2015]. Dzięki temu automatyczne przetwarzanie może się odbyć wyłącznie za pomocą narzędzi dla języka współczesnego, bez konieczności ich modyfikowania. Odwrotne podejście zastosowano w projekcie f19, w którym założono, że przetwarzaniu podlega tekst niepoddany jakiejkolwiek normalizacji [Bilińska i in. 2016]. W prezentowanej pracy przyjęliśmy podejście pośrednie – normalizacji podlega wyłącznie pisownia, nie ingeruje się zaś w zjawiska fleksyjne oryginalnego tekstu. Pozwala to na wykorzystanie

istniejących rozwiązań technicznych po uprzednim dostosowaniu ich do specyfiki tekstów dawnych, a jednocześnie nie pozbawia możliwości badania zmian fleksyjnych.

Analizę fleksyjną tekstów wykonano za pomocą analizatora Morfeusz [Woliński 2014] z odpowiednio zmienionym słownikiem – przystosowanym do danej epoki. Podstawowym źródłem danych do analizy fleksyjnej współczesnej polszczyzny jest *Słownik gramatyczny języka polskiego* [Saloni i in. 2015]. W wypadku polszczyzny dawniejszej dane SGJP są bardzo pomocne, ale niezbędna jest dodatkowa lista form niewystępujących w tekstach współczesnych. Źródłem takich danych może być inny słownik lub odpowiednia procedura modyfikująca („postarzająca") dane SGJP. Do stworzenia Korbeusza, czyli słownika fleksyjnego Morfeusza dla tekstów z XVII i XVIII wieku, wykorzystano obie metody [Kieraś i in. 2017], choć ogromna większość danych została uzyskana za pomocą drugiej z nich. Ważnym powodem tego jest fakt, że trzon danych SGJP stanowi siatka haseł SJPDor., którego zasób leksykalny sięga ostatniej ćwierci XVIII wieku, dzięki czemu słownik ten zawiera dużo słownictwa dawnego i przestarzałego, które w tekstach z XVII, XVIII czy XIX w. jest jeszcze w ogólnym użyciu.

Modyfikacja danych SGJP polegała przede wszystkim na dostosowaniu tych danych do tagsetu przyjętego w korpusie KorBa. Na tym etapie z form bazowych SGJP tworzone są także automatycznie niektóre dawne regularne formy fleksyjne, nieobecne w paradygmatach prezentowanych w SGJP, np. dla długiej serii rzeczowników żeńskich typu KONSTYTUCJA utworzono dawną formę dopełniacza, celownika i miejscownika zakończoną na *-ej*: *konstytucjej*.

Drugim źródłem danych fleksyjnych dla Korbeusza jest informacja fleksyjna pochodząca z *Elektronicznego słownika języka polskiego XVII i XVIII wieku* (e-SXVII), w którym notowano w poszczególnych hasłach formy zaświadczone w kanonie tekstów słownika. Znaczy to, że paradygmaty fleksyjne w e-SXVII są właściwie zawsze niekompletne, a bardzo często hasła są w postaci zalążkowej i zawierają jedynie formy hasłowe. W efekcie na dane fleksyjne e-SXVII składa się zaledwie ok. 76 tys. form fleksyjnych, choć słownik zawiera prawie 39 tys. haseł.

Dla korpusu tekstów z lat 1830–1918 powstał osobny analizator fleksyjny [Kieraś, Woliński 2018]. Procedura jego tworzenia jest właściwie uproszczoną wersją procedury tworzenia Korbeusza. Dla XIX wieku nie istnieje żaden dodatkowy zasób danych fleksyjnych, zatem analizator XIX-wieczny korzysta wyłącznie ze zmodyfikowanych danych SGJP. Zasób leksykalny SGJP (czyli w praktyce SJPDor.) okazał się do tego celu wystarczający, ponieważ analizator uzyskuje niemal równie dobre pokrycie tekstowe jak współczesny Morfeusz (odsetek form nierozpoznanych przez analizator XIX-wieczny dla tekstów z okresu 1830–1918 jest niemal równie mały jak Morfeusza współczesnego dla tekstów współczesnych).

Do automatycznego ujednoznacznienia fleksyjnego tekstów z XVII, XVIII i XIX w. użyto dwóch tagerów statystycznych: Concraft [Waszczuk

i in. 2018] oraz Toygger [Krasnowska-Kieraś 2017]. Z uwagi na większe zróżnicowanie tekstów, mniejszy stopień ustandaryzowania dawnej polszczyzny, a także mniejszą ilość ręcznie znakowanych wzorcowych danych zadanie automatycznego znakowania tekstów dawnych jest istotnie trudniejsze niż analogiczne zadanie dla tekstów współczesnych. Nie dziwi więc, że jakość takiego znakowania jest zauważalnie gorsza. Warto też jednak zauważyć – nie wdając się w techniczne szczegóły ewaluacji tego zadania – że wykwalifikowani anotatorzy z przygotowaniem filologicznym również słabiej sobie radzą ze znakowaniem fleksyjnym tekstów dawnych niż tekstów współczesnych.

USPÓJNIONY SPOSÓB OPISU (TAGSET)

Jak dotąd, procesowi tak rozumianego przetwarzania – czyli transliteracji, transkrypcji, analizy fleksyjnej i ujednoznaczniania – podlegały teksty z trzech okresów: współczesnej polszczyzny z przełomu XX i XXI w. (NKJP), polszczyzny XIX i początku XX w. (1830–1918) oraz polszczyzny XVII i XVIII w. (do 1772 r.). W dużym uproszczeniu można powiedzieć, że mamy tu do czynienia ze zbiorami tekstów reprezentującymi trzy różne epoki rozwoju polszczyzny: średniopolską, nowopolską i współczesną. Jak łatwo się domyślić, wszystkie trzy projekty różniły się nieco zbiorem stosowanych znaczników, zasadami anotacji i różnie podchodziły do szczegółowych rozstrzygnięć napotykanych w procesie przetwarzania tekstów. Aby móc wykorzystać wszystkie zasoby i narzędzia stworzone dla poszczególnych okresów polszczyzny, konieczne jest zniwelowanie przynajmniej najważniejszych różnic, czyli sprowadzenie anotacji fleksyjnej do wspólnego zbioru znaczników (tagsetu). Istotne jest też to, by proces uspójnienia mógł przebiegać automatycznie, ponieważ ręczna praca anotatorów jest bardzo czaso- i kosztochłonna. Z konieczności zatem uspójnienie prowadzi niekiedy do utraty pewnych szczegółowych informacji uwzględnionych w tagsetach dla konkretnych epok.

Tagsety poszczególnych projektów różnią się przede wszystkim systemem opisu rodzaju gramatycznego, zwłaszcza w obrębie podrodzajów męskich. W NKJP przyjęto klasyczny opis W. Mańczaka, w którym rozróżnia się pięć rodzajów, w tym trzy męskie: osobowy (m1), żywotny (m2) i nieżywotny (m3). Dodatkowo za osobne formy w paradygmacie rzeczownika męskiego osobowego uznano [za Salonim 1988 i SGJP] tzw. formy deprecjatywne, czyli formy mianownika i wołacza liczby mnogiej łączące się z przymiotnikami w rodzaju męskim żywotnym. Na mocy konwencji formom deprecjatywnym w tagsecie NKJP przypisano rodzaj m2. W znakowaniu korpusu XIX-wiecznego zasada jest podobna – system rodzajowy polszczyzny w tym okresie był już ustabilizowany i nie wymagał innego potraktowania rodzajów męskich. Zupełnie inaczej jest jednak w wypadku tekstów z XVII i XVIII w., w którym to okresie

kształtowała się kategoria męskoosobowości i co przejawia się przede wszystkim dużym rozchwianiem form rzeczowników męskich w obrębie paradygmatu. W tagsecie tego projektu przyjęto zatem, że notowany jest po prostu jeden rodzaj męski, a dodatkowo oznacza się jego użycia osobowe, żywotne i nieżywotne w formach diagnostycznych dla tych podrodzajów [Gruszczyński, Bronikowska 2018]. A zatem rzeczowniki męskie w formach biernika liczby pojedynczej mogą mieć oznaczenie manim2 lub mnanim (czyli odpowiednio: rodzaj męski w użyciu żywotnym i nieżywotnym), w formach zaś mianownika, biernika i wołacza liczby mnogiej może się pojawić dodatkowo oznaczenie manim1, czyli rodzaj męski w użyciu osobowym. Nie wyróżnia się oczywiście form deprecjatywnych, które – gdyby istniały w tej funkcji – nie różniłyby się od użyć żywotnych.

We wspólnym tagsecie zasadniczo przyjęto rozwiązanie z korpusu tekstów z XVII i XVIII w. – z jednym uproszczeniem. Za domyślny rodzaj męski uznaje się rodzaj nieżywotny. W diagnostycznych dla żywotności i osobowości formach pojawia się zatem wartość m, jeśli użycia są nieżywotne, lub manim1 i manim2 – w wypadku użyć osobowych i żywotnych. Znika zatem wartość mnanim. W danych korpusów epok późniejszych rodzaje m1, m2 i m3 w znacznikach fleksyjnych zostały skonwertowane na wartość m lub w wypadku biernika liczby pojedynczej oraz mianownika, biernika i wołacza liczby mnogiej – na wartości manim1 lub manim2.

Inną przykładową różnicą pomiędzy tagsetami poszczególnych korpusów jest opis form przymiotnikowych tzw. odmiany niezłożonej. W polszczyźnie współczesnej formy te są obecne już tylko szczątkowo, przede wszystkim w mianowniku męskim (*zdrów*, *godzien*), a także w wyrażeniach przyimkowych z celownikiem (*po angielsku*, *po cichu*) i dopełniaczem (*z wolna*, *od dawna*) o funkcji przysłówkowej. W NKJP formy poprzyimkowe zgromadzono wraz z innymi skostniałymi poprzyimkowymi formami przymiotnikowymi pod wspólną etykietą adjp. Z kolei formy mianownikowe (zakończone na spółgłoskę) rozdzielono między zwykłe formy przymiotnikowe i formy predykatywne o etykiecie adjc. W korpusie XIX-wiecznym etykietę adjp ograniczono już tylko do poprzyimkowych form odmiany niezłożonej wraz z określeniem ich przypadka (dopełniacza lub celownika). Nie wyróżniono też oddzielnej klasy adjc, formy mianownikowe odmiany niezłożonej opisano zaś identycznie jak ich odpowiedniki o odmianie złożonej. Zupełnie inny opis zastosowano w korpusie XVII i XVIII w., ponieważ w tym okresie formy odmiany niezłożonej występowały istotnie częściej niż w późniejszych epokach i obejmowały więcej form paradygmatu przymiotnikowego (w tym przede wszystkim biernikowe formy żeńskie, np. *piękną*). W tym tagsecie odmiana złożona oznaczana była oddzielną etykietą adjb uwzględniającą wszystkie kategorie przymiotnika. Analogicznie postąpiono z imiesłowami przymiotnikowymi, których formy odmiany niezłożonej również w tym okresie sporadycznie się pojawiają. Uspójniając tagsety, postanowiliśmy odtworzyć opis dawnych form odmiany złożonej ze szczątkowych

informacji, które można odzyskać z korpusów późniejszych okresów. I tak, formy predykatywne z NKJP uznano za formy mianownika męskiego liczby pojedynczej. Pozostałe formy przymiotnikowe (mianownikowe i biernikowe) zakończone na spółgłoskę uznano odpowiednio za formy męskie mianownika lub biernika liczby pojedynczej. Formy poprzyimkowe z NKJP i korpusu XIX w. uznano odpowiednio za formy dopełniacza lub celownika liczby pojedynczej – historycznie mogły być to formy rodzaju męskiego i nijakiego, ponieważ jednak zwykle nie występują w uzgodnieniu z żadnym rzeczownikiem, na mocy konwencji przypisuje się im rodzaj nijaki (podobną zasadę stosowano w korpusie XVII i XVIII w.). Wreszcie, formy przymiotnikowe żeńskie w bierniku zakończone na *-ę*, występujące niekiedy jeszcze w tekstach XIX-wiecznych, również uznano za formy odmiany niezłożonej.

Nie da się jednak zniwelować skutków niektórych zasadniczych decyzji, które podejmowane są w poszczególnych projektach związanych z korpusami historycznymi konkretnych okresów. I tak, przykładowo, korpus XVII i XVIII w. w swoim schemacie anotacji uwzględnia liczbę podwójną, która co prawda w średniopolszczyźnie była już zdecydowanie rzadka, ale jej obecność w tekstach była wciąż większa niż w wiekach późniejszych, gdy widać ją jedynie w skostniałych formach kilku leksemów. Dodatkowo zdecydowano, że za formy liczby podwójnej uznaje się również te, które zaczęły być używane w funkcji liczby mnogiej i są w użyciu do dziś, czyli przede wszystkim formy rzeczowników OKO, UCHO i RĘKA. W znakowaniu korpusów tekstów późniejszych tej zasady oczywiście nie stosowano i te same formy opisywane były po prostu jako mnogie. W efekcie zasięg liczby podwójnej w wykorzystanych danych sięga roku 1772 i urywa się nagle. Można by oczywiście ograniczyć się do uznania za formy liczby podwójnej tych trzech leksemów tylko tych, które faktycznie są użyte w takim znaczeniu (np. z liczebnikiem), ale nie da się takiej postaci danych uzyskać całkowicie automatycznie. Dlatego trzeba mieć świadomość, że prezentowane w Chronofleksie wyniki są obarczone wieloma szczegółowymi decyzjami, na które autorzy samej aplikacji nie mieli wpływu. Stosuje się to jednak do wszystkich zaawansowanych narzędzi i zasobów lingwistycznych.

Innym przykładem takich szczegółowych decyzji, które mają wpływ na możliwość śledzenia obecności pewnych form fleksyjnych w czasie mogą być zasady transkrypcji. W wypadku tekstów XIX-wiecznych wyrażenia dawniej pisane niekiedy rozdzielnie, a dziś tylko łącznie, np. *po krótce*, *na prędce*, w których obecne są formy miejscownikowe przymiotników KRÓTKI i PRĘDKI odmiany niezłożonej, uwspółcześniano, zapisując je łącznie i oznaczając jako przysłówki (zresztą miały już one w XIX w. wyraźnie przysłówkową funkcję i charakter). W korpusie XVII i XVIII w. tymczasem pozostawiano je w zapisie rozdzielnym i opisywano oba człony oddzielnie: jako przyimek i formę przymiotnikową odmiany niezłożonej. W tekstach współczesnych te wyrażenia są oczywiście zapisywane tylko łącznie i uznawane za przysłówki, a ich związek z przymiotnikami, od których pochodzą, jest już trudny do uchwycenia automatycznie.

SYSTEM CHRONOFLEKS

System Chronofleks został udostępniony w postaci aplikacji webowej pod adresem http://chronofleks.nlp.ipipan.waw.pl/. Ta instalacja systemu została zasilona danymi omawianych wyżej korpusów historycznych, a więc korpusu barokowego, oznaczonego jako KorBa [https://korba.edu.pl/] i korpusu tekstów XIX-wiecznych [http://korpus19.nlp.ipipan.waw.pl/]. Uwzględniono zarówno teksty znakowane ręcznie, jak i teksty znakowane za pomocą tagera automatycznego. Stanowią one osobne podkorpusy, które można uwzględnić lub pominąć w wyświetlanych wynikach (służy do tego przycisk wyboru korpusów pojawiający się w obu częściach systemu Chronofleks, por. rys. 1. i 4.). Dzięki temu użytkownik systemu może użyć w badaniach albo danych mniejszych, opracowanych dokładniej, albo danych większych, ale wtedy musi się liczyć z wpływem na wyniki mniej dokładnych metod automatycznych. Część współczesną używanych danych stanowią teksty jednomilionowego ręcznie znakowanego podkorpusu NKJP [Przepiórkowski i in. 2012]. W miarę pojawiania się innych korpusów zamierzamy dodawać dane do systemu, wzbogacając tym samym uzyskiwane wyniki.

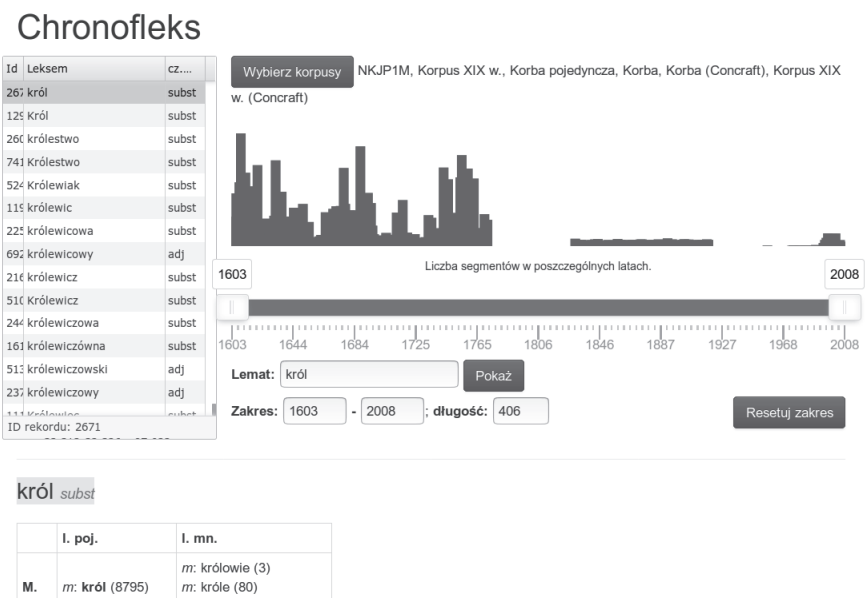
BADANIE ZMIENNOŚCI PARADYGMATÓW FLEKSYJNYCH Z WYKORZYSTANIEM SYSTEMU CHRONOFLEKS

Pierwsza część systemu (dostępna ze strony głównej za pomocą przycisku Paradygmaty) stanowi prototyp przekrojowego słownika fleksyjnego języka polskiego dostępny przez Internet i umożliwiający wizualizację zmian paradygmatów w czasie. Jak wspomniano wcześniej, słownik ten prezentuje wyłącznie formy odmiany poświadczone w zgromadzonych korpusach, podając ich zliczenia, przy czym użytkownik ma możliwość wskazania interesującego go okresu.

Rysunek 1. przedstawia interfejs systemu służący do tego celu. Przewijana lista po lewej stronie rysunku zestawia leksemy poświadczone w przeanalizowanych tekstach. Można z niej wybrać leksem stanowiący przedmiot zainteresowania. Po prawej stronie podano informację o użytych (pod)korpusach. Wykres przedstawia ilość tekstu (liczoną w segmentach) dostępną do analizy w poszczególnych latach (nie jest to wykres dla wybranego leksemu, tylko dla wybranych podkorpusów). Na rys. 1. przedstawiono wykres dla wszystkich dostępnych w systemie korpusów. Widać, że teksty barokowe mają dużo bogatszą reprezentację niż późniejsze. Widać też dwie luki w materiale korpusowym: w zgromadzonych dotychczas korpusach brak tekstów z okresu między rokiem 1773 a 1830 oraz między 1919 a 1956.

Umieszczony poniżej wykresu suwak pozwala wskazać podokres, z którego mają pochodzić wyświetlone niżej formy. Na rysunku 2. przedstawiono trzy migawki z systemu Chronofleks ukazujące formy odmiany

Rysunek 1. Interfejs części systemu Chronofleks prezentującej paradygmaty fleksyjne



leksemu KRÓL z tekstów z XVII, XVIII i XIX wieku. Paradygmat rzeczownika jest prezentowany w postaci tabeli, której kolumny odpowiadają wartości kategorii gramatycznej liczby (jeśli wykryto formy liczby podwójnej, pojawią się one w osobnej kolumnie), a wiersze – kategorii przypadka. W każdej z klatek pokazano różne możliwe wykładniki formy o danej charakterystyce. Przy każdym podano charakterystykę rodzajową (z uwzględnieniem form żywotnych i osobowych rodzaju męskiego) oraz liczbę wystąpień w rozpatrywanym okresie. Pogrubiono wykładnik najczęstszy.

Jeżeli na przykład skupić się na klatce paradygmatu reprezentującej formę narzędnika liczby mnogiej, można zaobserwować, jak zmniejsza się częstość formy *królmi*, która w XVII wieku dominuje nad *królami* (89 wystąpień *królmi* i 38 *królami*), później jednak proporcje się odwracają (w XVIII wieku odnotowano 54 wystąpienia *królami* i 27 *królmi*). Można uznać, że jeszcze w XVIII wieku forma *królmi* jest żywa. Zgromadzone dane nie pozwalają ustalić momentu jej zaniku, bo w tekstach z XIX w. (po roku 1830) jest tylko jedno poświadczenie narzędnika i jest to forma współczesna.

W obliczeniu uwzględniono formy z korpusów znakowanych automatycznie, co nie grozi przekłamaniem dzięki temu, że forma *królmi* nie jest homonimiczna z żadną inną formą fleksyjną.

Rysunek 2. Paradygmat leksemu *król* na podstawie form zanotowanych w tekstach z lat 1601–1700, 1701–1800 oraz 1801–1900

król *subst*

	l. poj.	l. mn.
M.	*m*: **król** (5038)	*m*: królowie (2) *m*: króle (68) *manim1*: **królowie** (403)
D.	*m*: **króla** (3060) *m*: królu (6)	*m*: króli (4) *m*: **królów** (584)
C.	*m*: **królowi** (1300) *m*: królu (12)	*m*: **królom** (180)
B.	*m*: **króla** (1147) *m*: król (87) *manim2*: króla (3)	*m*: królów (6) *m*: **króle** (75) *manim1*: króli (0) *manim1*: królów (24)
N.	*m*: **królem** (853)	*m*: królami (38) *m*: **królmi** (89)
Msc.	*m*: **królu** (225)	*m*: **królach** (24)
W.	*m*: król (0) *m*: **królu** (461)	*m*: królowie (1) *m*: króle (2) *manim1*: **królowie** (10)

król *subst*

	l. poj.	l. mn.
M.	*m*: **król** (3318)	*m*: królowie (1) *m*: króle (10) *manim1*: **królowie** (316)
D.	*m*: **króla** (2464) *m*: królu (6)	*m*: króli (4) *m*: **królów** (766)
C.	*m*: **królowi** (787) *m*: królu (8)	*m*: **królom** (127)
B.	*m*: **króla** (807) *m*: król (43) *manim2*: króla (0)	*m*: królów (9) *m*: króle (11) *manim1*: króli (0) *manim1*: **królów** (83)
N.	*m*: **królem** (603)	*m*: **królami** (54) *m*: królmi (27)
Msc.	*m*: **królu** (123)	*m*: **królach** (55)
W.	*m*: król (1) *m*: **królu** (158)	*m*: **królowie** (1) *m*: króle (0) *manim1*: **królowie** (1)

król *subst*

	l. poj.	l. mn.
M.	*m*: **król** (395)	*m*: królowie (0) *m*: króle (2) *manim1*: **królowie** (10)
D.	*m*: **króla** (213) *m*: królu (0)	*m*: króli (1) *m*: **królów** (28)
C.	*m*: **królowi** (32) *m*: królu (0)	*m*: **królom** (2)
B.	*m*: **króla** (45) *m*: król (0) *manim2*: króla (0)	*m*: królów (0) *m*: króle (2) *manim1*: króli (1) *manim1*: **królów** (5)
N.	*m*: **królem** (52)	*m*: **królami** (1) *m*: królmi (0)
Msc.	*m*: **królu** (11)	*m*: królach (0)
W.	*m*: król (0) *m*: **królu** (14)	*m*: królowie (0) *m*: króle (0) *manim1*: **królowie** (1)

Jeden z problemów, jakie napotykali znakujący korpusy historyczne, ilustruje leksem KOMETA (zob. rys. 3.). W tekstach dawnych leksem ten jest rozchwiany między rodzajem męskim a żeńskim (czasami na formy obu rodzajów natrafia się nawet u tego samego autora). Formy te były, na ile to było możliwe, przypisywane przez znakujących do właściwego rodzaju na podstawie kontekstu, a więc przede wszystkim na podstawie uzgodnień z formami przymiotnikowymi i czasownikowymi. Analogicznych przypisań uczyły się też narzędzia automatyczne. Wyniki obu metod można porównać na rysunku 3.

Rzucającym się w oczy elementem jest obecność kolumny liczby podwójnej w paradygmacie odpowiadającym znakowaniu ręcznemu. Występująca w niej jedna w całym korpusie forma pochodzi z tekstu Mateusza Bembusa *Kometa to jest pogróżka z nieba na postrach, przestrogę i upomnienie ludzkie* z roku 1619:

> *Dwie zasie **Komecie**/ które przyniósł rok Pański 1469. okrutnym krwie rożlaniem osobliwie w Niemczech/ Europie groziłu/jako masz u Kromera.*

Przy znakowaniu automatycznym zjawiska tego nie udało się wychwycić: było ono zbyt rzadkie, żeby odcisnęło się w modelach statystycznych.

Narzędziom automatycznym udało się wyłuskać i poprawnie zinterpretować 8 wystąpień męskoosobowej formy *kometowie* nieobecnej w mniejszych danych znakowanych ręcznie. W wypadku form, które są homonimiczne dla rzeczownika męskiego i żeńskiego (np. *komecie*, *komety*) narzędzia automatyczne przypisywały im rodzaj na podstawie kontekstu. Proporcje form zakwalifikowanych do poszczególnych rodzajów są podobne w danych znakowanych ręcznie i automatycznie, więc można

domniemywać, że również te automatyczne zostały ustalone w większości wypadków poprawnie. Ciekawostkę stanowi forma *komety* interpretowana jako narzędnik. Niestety, chociaż taka jej interpretacja wydaje się możliwa, sprawdzenie cytatów w korpusie ujawnia, że we wszystkich wypadkach narzędzia automatyczne zakwalifikowały ją błędnie. Pokazuje to, że przy formułowaniu wniosków na podstawie znakowania automatycznego należy zachować ostrożność.

Rysunek 3. Paradygmat dwurodzajowego leksemu *kometa* na podstawie danych korpusowych znakowanych ręcznie (po lewej) i automatycznie (po prawej)

kometa *subst*

	l. poj.	l. podw.	l. mn.
M.	*m*: **kometa** (29) *f*: kometa (7)	*f*: **komecie** (1)	*m*: komety (3) *f*: **komety** (6)
D.	*m*: **komety** (15) *f*: komety (7)	-	*m*: komet (4) *f*: **komet** (8)
C.	*m*: **komecie** (1) *f*: **komecie** (1)	-	*m*: **kometom** (1)
B.	*m*: **kometę** (2) *manim2*: kometę (1) *f*: kometę (1)	-	*m*: komety (2) *f*: **komety** (3)
N.	*m*: **kometą** (2)	-	*m*: **kometami** (1)
Msc.	*f*: **komecie** (2)	-	-
W.	-	-	-

kometa *subst*

	l. poj.	l. mn.
M.	*m*: **kometa** (345) *f*: kometa (194)	*m*: komety (50) *manim1*: kometowie (8) *f*: **komety** (117)
D.	*m*: **komety** (325) *f*: komety (62)	*m*: komet (17) *m*: kometów (14) *f*: **komet** (193)
C.	*m*: **komecie** (7) *f*: **komecie** (7)	*m*: **kometom** (5) *f*: kometom (1)
B.	*m*: **kometę** (54) *f*: kometę (19) *f*: kometą (2)	*m*: komety (28) *f*: **komety** (37)
N.	*m*: **kometą** (13) *f*: kometą (10)	*m*: **komety** (18) *m*: kometami (4) *f*: kometami (3)
Msc.	*m*: komecie (9) *f*: **komecie** (27)	*m*: komeciech (1) *f*: **kometach** (37) *f*: komeciech (1)
W.	-	-

BADANIE ZMIAN FREKWENCJI GRUP FORM FLEKSYJNYCH W CZASIE Z WYKORZYSTANIEM SYSTEMU CHRONOFLEKS

Druga część systemu Chronofleks (do której można przejść ze strony głównej http://chronofleks.nlp.ipipan.waw.pl/ za pomocą przycisku Wykresy) przedstawia inne spojrzenie na te same dane. Pozwala ona mianowicie generować wykresy przedstawiające zmienność w czasie frekwencji wybranych grup form fleksyjnych.

Aby przeprowadzić badanie zmian częstości, trzeba zadać trzy elementy odpowiadające ramkom po lewej stronie rysunku 4. Po pierwsze, konieczne jest wskazanie badanego materiału (ramka Ustawienia), a więc wybór używanych podkorpusów i interesującego nas zakresu czasowego. Można także zadać sposób zliczania i grupowania wyników, co omówimy dalej.

Po drugie, należy zadać, jakie formy fleksyjne będą podlegać badaniu (ramka Segmenty). Sposób ich określenia przypomina formułowanie zapytania w wyszukiwarce korpusowej dotyczącego pojedynczego segmentu. W przykładzie pokazanym na rysunku 4. przedmiotem badania są formy o lemacie *król*, którym przypisano w korpusie cechy narzędnika liczby mnogiej. Warunki dookreślające wybierane formy dodaje się za pomocą przycisku opatrzonego symbolem Plus.

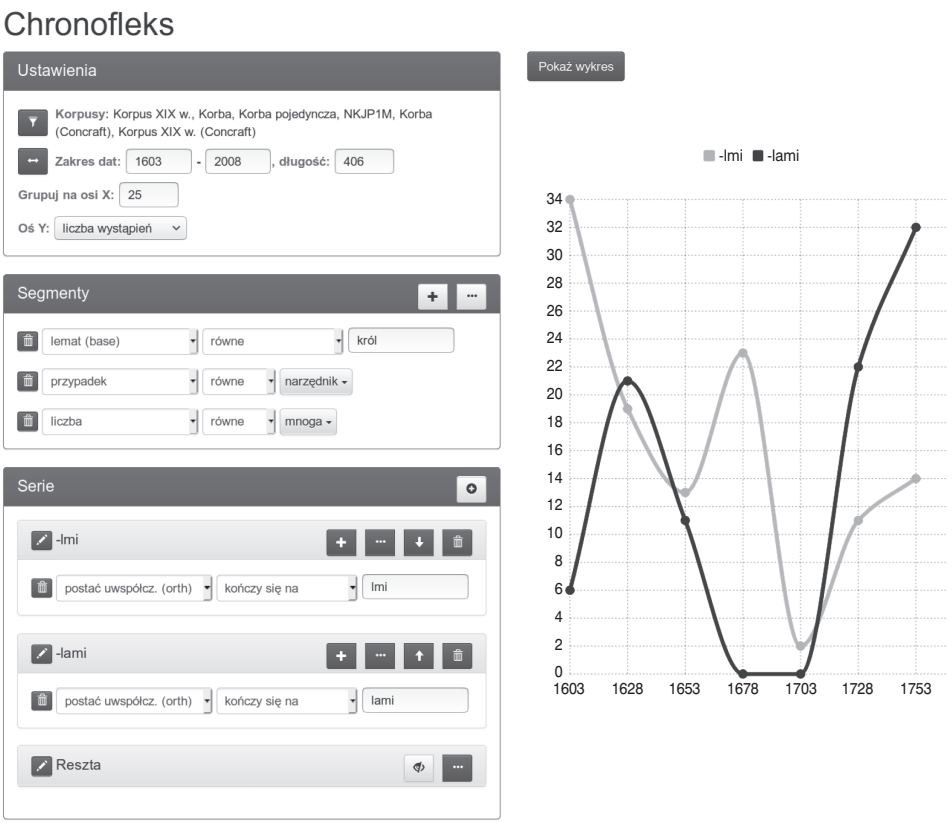
Trzecim elementem konstrukcji badania jest podział wyszukanych form na serie danych (ramka Serie). Każda seria zostanie wyświetlona jako oddzielna linia na wykresie. Określenie serii ma postać warunków podobnych do używanych w części drugiej, ale tym razem wyszukiwanie odbywa się wśród wskazanych uprzednio form (a więc w przykładzie form narzędnika liczby mnogiej leksemu KRÓL). Kolejność określania serii jest istotna, dane wystąpienie formy zostanie zakwalifikowane do pierwszej serii, w której spełnia zadane warunki. Formy niespełniające żadnego z warunków trafiają do predefiniowanej serii Reszta. W szczególności jeżeli nie zada się żadnych warunków w tej części, wszystkie formy znalezione w części drugiej trafiają do serii Reszta i ich frekwencja jest rysowana w postaci pojedynczej linii (por. rys. 5., przedstawiający zmiany frekwencji wszystkich form leksemu KRÓL).

W przykładzie na rysunku 4. zdefiniowano dwie serie danych, rozdzielając formy narzędnikowe na podstawie końcówki. Do serii pierwszej trafiły formy zakończone ciągiem liter *lmi*, a drugiej – *lami*. Serie te obejmują wszystkie wybrane formy, więc domyślna seria Reszta pozostaje pusta – nie pokazano jej na wykresie. Przy prowadzeniu tego rodzaju badań użyteczny bywa przycisk z wielokropkiem znajdujący się przy definicji serii. Pozwala on obejrzeć przykłady form, które trafiły do danej serii. Dzięki temu można sprawdzić, czy zadane warunki działają zgodnie z oczekiwaniami.

Uzyskany wykres przedstawia liczby form o danych zakończeniach w poszczególnych latach. Istotnym parametrem badania jest sposób grupowania danych. W przykładzie przyjęto zbieranie tekstów w grupy po 25 lat. Drobniejsze grupy wprowadzają więcej losowości, związanej z tym, że zgromadzone teksty nie są równo rozłożone w czasie. Grupowanie w dłuższych okresach wygładza wykres, ale zmniejsza jego szczegółowość. W praktyce dobranie najlepszego sposobu grupowania wymaga eksperymentów.

Na osi pionowej wykresu na rysunku 4. zobrazowano zliczenia form. Można więc na przykład odczytać, że w zgromadzonych tekstach z lat 1603–1627 wystąpiły 34 formy *królmi* i 8 form *królami*. Zliczenia te zależą oczywiście od ilości tekstu z danego okresu. Od tego czynnika można się uniezależnić, badając częstości względne form (liczbę wystąpień szukanych form dzieli się przez sumaryczną długość tekstów z danego okresu). W ramce Ustawienia przewidziano także trzeci sposób zliczania. Polega on na podawaniu liczby wystąpień *różnych* form w tekstach. Pozwala to odróżnić zjawiska reprezentowane przez małą liczbę częstych form od zjawisk reprezentowanych przez wiele różnych form.

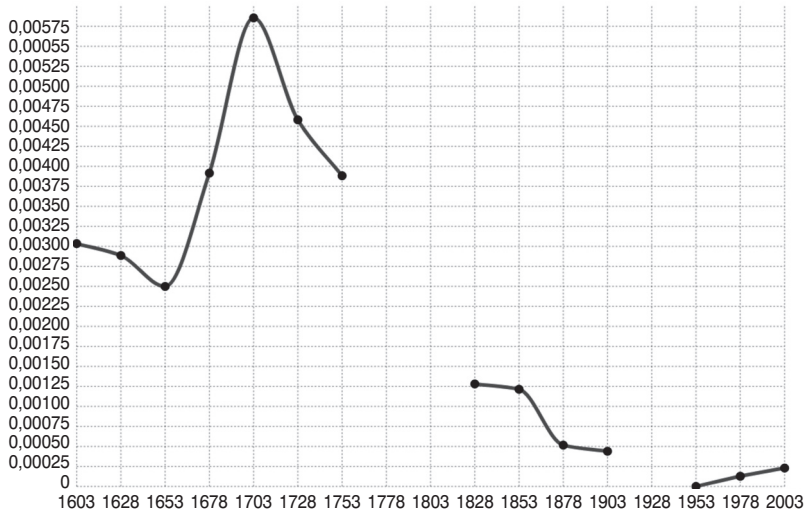
Rysunek 4. Interfejs tworzenia wykresów w systemie Chronofleks



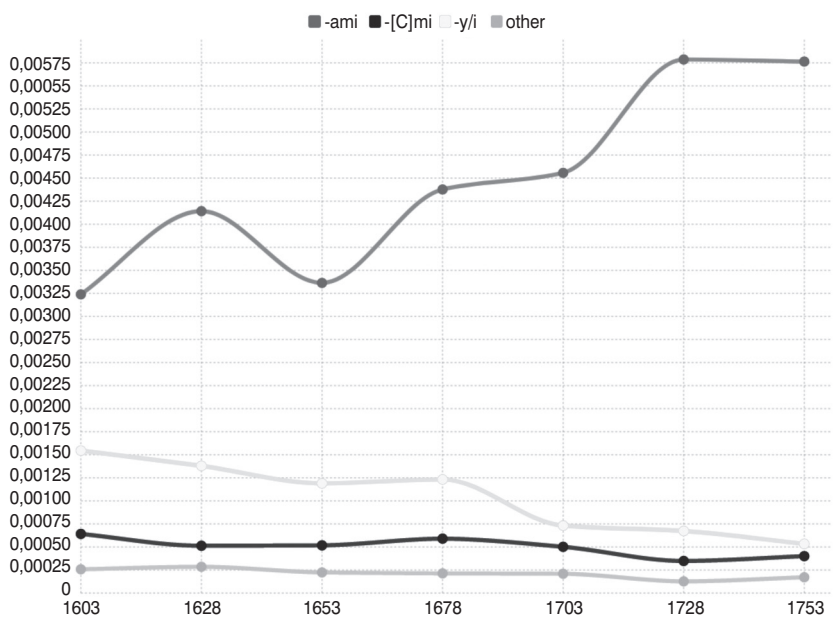
Niezależnie od wyboru omawianych ustawień z wykresu przedstawiającego rozkład w czasie form narzędnikowych leksemu KRÓL nie daje się odczytać użytecznej generalizacji. Form tych jest po prostu za mało. Ciekawszy jest, przedstawiony na rys. 5., wykres częstości względnej dowolnych form leksemu *król* – widoczna jest tendencja zniżkowa frekwencji tego leksemu (co jest zdroworozsądkowe wobec przemian w odzwierciedlanej językowo rzeczywistości).

Prawdziwa siła systemu Chronofleks polega jednak na możliwości zebrania na jednym wykresie form całej grupy leksemów. Na rysunku 6. przedstawiono frekwencje względne form narzędnika liczby mnogiej dla wszystkich rzeczowników męskich i nijakich odnotowanych w korpusie barokowym. Formy te podzielono na cztery serie: o zakończeniu *-ami*, o zakończeniu *-mi* po spółgłosce, o zakończeniu *-y* lub *-i* oraz pozostałe. W badanym okresie widoczna jest bardzo wyraźna tendencja rosnąca końcówki *-ami*, która dominuje współcześnie.

Rysunek 5. Zmienność w czasie frekwencji form leksemu *król*



Rysunek 6. Zmiany frekwencji w tekstach korpusu barokowego form rzeczowników męskonijakich w narzędniku według zakończeń



PODSUMOWANIE

Przedstawiony system umożliwia prowadzenie ciekawych badań fleksji historycznej z perspektywy frekwencyjnej. Jego zaletą jest swobodny dostęp przez Internet, może być on wykorzystywany zarówno w praktyce badawczej, jak i edukacyjnej przez profesjonalistów i amatorów.

Trzeba jednak pamiętać, że wprowadzone do systemu zasoby korpusowe są ograniczone, w związku z czym wyraziste wyniki uzyskuje się jedynie dla badań obejmujących zjawiska odpowiednio częste – dotyczące najczęstszych leksemów lub odpowiednio dużych grup form.

Planowane jest uzupełnianie zasobów korpusowych w miarę ich pojawiania się. Uzupełnienia będą dotyczyły powiększania bazy tekstów z okresów już uwzględnionych, jak i uzupełnienia wspominanych luk w reprezentowanych okresach historycznych [por. Król i in. 2016]. Dostęp do obszerniejszych korpusów znakowanych ręcznie pozwoli też ulepszyć narzędzia znakowania automatycznego, co przełoży się na jakość wyników uzyskiwanych z korpusów znakowanych automatycznie.

Bibliografia

J. Bilińska, M. Derwojedowa, W. Kieraś, M. Kwiecień, 2016, *Mikrokorpus polszczyzny 1830–1918* [w:] Ł. Karpiński, P. Michałowski, *Komunikacja Specjalistyczna* 11, s. 149–161.

W. Gruszczyński, R. Bronikowska, 2018, *Instrukcja korzystania z wyszukiwarki do Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII wieku (do 1772 r.)*, https://www.korba.edu.pl/manual

T. Erjavec, 2015, *The IMP Historical Slovene Language Resources*, „Language Resources and Evaluation" 49(3), s. 753–75; https://doi.org/10.1007/s10579-015-9294-7

W. Kieraś, D. Komosińska, E. Modrzejewski, M. Woliński, 2017, *Morphosyntactic annotation of historical texts. The making of the baroque corpus of Polish* [w:] K. Ekštein, V. Matoušek (red.), *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27–31*, „Lecture Notes in Computer Science" 10415, s. 308–316.

W. Kieraś, M. Woliński, 2018, *Manually annotated corpus of Polish texts published between 1830 and 1918* [w:] N. Calzolari i in. (red.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, s. 3854–3859.

Ł. Kobyliński, W. Kieraś, 2016, *Part of speech tagging for Polish: State of the art and future perspectives* [w:] *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016)*, Konya.

K. Krasnowska-Kieraś, 2017, *Morphosyntactic disambiguation for Polish with bi-LSTM neural networks* [w:] Z. Vetulani, P. Paroubek (red.), *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, s. 367–371.

M. Król, M. Derwojedowa, R.L. Górski, W. Gruszczyński, K.W. Opaliński, P. Potoniec, M. Woliński, W. Kieraś, M. Eder, 2019, *Narodowy Korpus Diachroniczny Polszczyzny. Projekt*, „Język Polski" XCIX(1), s. 92–101.
A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), 2012, *Narodowy Korpus Języka Polskiego*, Warszawa.
Z. Saloni, 1988, *O tzw. formach nieosobowych rzeczowników męskoosobowych we współczesnej polszczyźnie*, „Biuletyn Polskiego Towarzystwa Językoznawczego" XLI, Kraków, s. 155–166.
Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, 2015, *Słownik gramatyczny języka polskiego*, wyd. III on-line, Warszawa; http://sgjp.pl
J. Waszczuk, W. Kieraś, M. Woliński, 2018, *Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields* [w:] P. Sojka, A. Horák, I. Kopeček, K. Pala (red.), *Text, Speech, and Dialogue: 21st International Conference, TSD 2018*, Brno, Czech Republic, s. 188–196.
M. Woliński, 2014, *Morfeusz reloaded* [w:] N. Calzolari i in. (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland, s. 1106–1111.

Inflectional analysis of historical texts and variability of Polish inflection from the perspective of corpus data

Summary

The subject matter of this paper is Chronofleks, a computer system (http://chronofleks.nlp.ipipan.waw.pl/) modelling Polish inflection based on a corpus material. The system visualises changes of inflectional paradigms of individual lexemes over time and enables examination of the variability of the frequency of inflected form groups distinguished based on various criteria.

Feeding Chronofleks with corpus data required development of IT tools to ensure an inflectional processing sequence of texts analogous to the ones used for modern language; they comprise a transcriber, a morphological analyser, and a tagger.

The work was performed on data from three historical periods (1601–1772, 1830–1918, and modern ones) elaborated in independent projects. Therefore, finding a common manner of describing data from the individual periods was a significant element of the work.

Keywords: electronic text corpus – natural language processing – inflection of Polish – history of language

Trans. Monika Czarnecka