ZMIANA W JĘZYKU

Studia kwantytatywno-korpusowe

PRACE
INSTYTUTU JĘZYKA POLSKIEGO PAN
151

POLSKA AKADEMIA NAUK • INSTYTUT JĘZYKA POLSKIEGO

ZMIANA W JĘZYKU

Studia kwantytatywno-korpusowe

Rafał L. Górski
Magdalena Król
Maciej Eder

Kraków 2019

Recenzenci
Magdalena Derwojedowa
Włodzimierz Gruszczyński

Opracowanie redakcyjne
Krzysztof Porosło

ISBN 978-83-64007-53-8

Wydawca
Instytut Języka Polskiego PAN
al. Mickiewicza 31
31-120 Kraków
www.ijp.pan.pl

Spis treści

Wstęp

Niniejsza praca stawia sobie za cel przybliżenie czytelnikowi wybranych metod korpusowych i kwantytatywnych przyjętych w językoznawstwie historycznym i omówienie możliwych zastosowań tych metod do opisu historii języka polskiego. Książka stanowi pokłosie prac prowadzonych w ramach projektu *Przebiegi zmian gramatycznych i leksykalnych w historii języka polskiego – metody korpusowe i kwantytatywne w językoznawstwie diachronicznym*, finansowanego przez Narodowe Centrum Nauki (nr UMO-2013/11/B/HS2/02795).

Polskie językoznawstwo historyczne może poszczycić się wieloma osiągnięciami, a przeszłość języka polskiego jest bardzo dobrze poznana. Nie oznacza to jednak, że nasi poprzednicy nie pozostawili nam niczego, co warto by zbadać. Dowodzi tego choćby niemała liczba prac poświęconych historii polszczyzny, jaka ukazuje się co roku w obiegu naukowym. Z jednej strony współcześni badacze uzupełniają luki w zagadnieniach szczegółowych, z drugiej strony rozwój metodologii językoznawczej wyznacza nowe tematy badawcze. Jak zauważają Krystyna Kleszczowa i Tomasz Mika, w kształtowaniu refleksji historycznojęzykowej widać wpływy strukturalizmu, generatywizmu czy kognitywizmu (Kleszczowa i Mika, 2018). Językoznawca badający przeszłość języka otrzymuje coraz to nowe narzędzia – dotąd dysponował naukową edycją tekstu oraz słownikiem wraz ze stworzoną do niego kartoteką. I choć edycja i słownik nadal zajmują poczesne miejsce w instrumentarium lingwisty, przestały być jedynymi dostępnymi badaczowi narzędziami.

Druga połowa XX wieku to czas polskich słowników historycznych, ale także czas, w którym na świecie powoli rozwija się technika badania tekstu za pomocą komputera. W tym miejscu warto przypomnieć, że filologiczne zastosowania „maszyn cyfrowych" – by użyć tego archaizmu – zaczęły się już w roku 1949, gdy o. Roberto Busa SJ przekonał dyrektorów korporacji IBM do tego, by za pomocą komputera stworzyć konkordancję do tekstów Tomasza z Akwinu (Winter, 1999). Komputer powoli stawał się istotnym narzędziem w badaniach filologicznych. Bodaj pierwszym elektronicznym korpusem języka polskiego był korpus, który stanowił bazę empiryczną dla *Słownika frekwencyjnego polszczyzny współczesnej* (Kurcz, Lewicki, Sambor, Szafran i Woronczak, 1990).

O ile w językoznawstwie polonistycznym wiek XX zwano często „wiekiem słowników", to na pewno wiek XXI – jeśli nie w całości, to przynajmniej jego

dwie pierwsze dekady – będzie zwany „czasem korpusów tekstowych". Niemal równolegle z pierwszymi korpusami współczesnej polszczyzny powstał w Instytucie Języka Polskiego PAN *Korpus tekstów staropolskich*, przygotowany przez zespół kierowany przez Wacława Twardzika (Twardzik i Górski, 2003). Kolejny korpus historyczny to *KorBa* (Gruszczyński, Adamiec i Ogrodniczuk, 2013), powstały pod kierunkiem Włodzimierza Gruszczyńskiego, także w IJP PAN. Pierwszy obejmuje okres od pierwszych ciągłych tekstów polskich po rok 1500, drugi z kolei zawiera źródła z okresu 1600–1772. Lukę pomiędzy nimi wypełnia korpus tworzony w pracowni *Słownika polszczyzny XVI wieku* Instytutu Badań Literackich PAN w ramach projektu kierowanego przez Patrycję Potoniec.

Dzięki opisanym wyżej narzędziom sytuacja lingwisty pracującego nad historią języka polskiego zmieniła się w ostatnich latach diametralnie, ma on bowiem na wyciągnięcie ręki dane, których pozyskanie jeszcze niedawno wymagało wielomiesięcznej czy nawet wieloletniej ekscerpcji z tekstów drukowanych. Mamy nadzieję, że udostępniając nasze teksty, a także stworzone przez nas słowniki morfologiczne, choć częściowo spłaciliśmy nasz dług wdzięczności. Odczuwamy też niewątpliwą satysfakcję, że owocna współpraca z zespołami korpusów historycznych stanowiła impuls dla inicjatywy stopniowego zespolenia ich w jedną całość obejmującą całą historię polszczyzny piśmiennej (zob. Król i in., 2019).

W niniejszej książce chcielibyśmy pokazać zarówno możliwości, jak i ograniczenia, jakie stawiają metody korpusowe i kwantytatywne w badaniu dziejów polszczyzny. Chcemy pokazać, że dzięki korpusowi możemy zarówno śledzić zmiany językowe, o których istnieniu badacze wiedzą od dawna (przy czym skala i dokładność czynionych obserwacji jest nieporównanie większa), ale także, że jesteśmy w stanie śledzić procesy historyczne, których istnienia wcześniej nie podejrzewaliśmy.

Pierwszy rozdział przybliża osiągnięcia metod korpusowych i kwantytatywnych w językoznawstwie historycznym, głównie anglistycznym, jako że – zgodnie z naszą wiedzą – badania nad przeszłością tego właśnie języka z użyciem metod ilościowych są najbardziej zaawansowane. Przyczynia się do tego nie tylko prestiż języka angielskiego, ale także jego długa, udokumentowana bogatym piśmiennictwem historia.

W rozdziale drugim opisujemy sam korpus, który stanowił materiał empiryczny, oraz proces jego tworzenia. Trzeba tu zaznaczyć, że nie jest to za każdym razem dokładnie ten sam korpus. W zależności od założeń konkretnego eksperymentu niekiedy wybiegaliśmy poza połowę XIX wieku, niekiedy też sięgaliśmy do danych staropolskich, innym razem – gdy istotne było wykorzystanie pisowni nieznormalizowanej – tworzyliśmy podkorpus tekstów spełniających ten warunek. Ponadto cały dostępny korpus był rozmaicie dzielony na odcinki czasu następujące kolejno po sobie.

Rozdział trzeci poświęcamy modelowaniu przebiegu zmiany językowej. Od pewnego czasu w literaturze przedmiotu pojawia się teza, że zastępowanie formy

recesywnej formą innowacyjną można opisać za pomocą regresji logistycznej, czyli pewnego matematycznego modelu opisującego fazową zmianę różnych zjawisk. Choć koncepcja tego rodzaju modelowania powstała dobrych kilka dekad temu, to jednak wciąż jest ona poznawczo atrakcyjna dzięki dostępności coraz obfitszych danych empirycznych. Jakkolwiek takie podejście wydaje się głęboko uzasadnione, trzeba jednocześnie pamiętać, że zmiana językowa jest procesem społecznym, którego przebieg może być zakłócany przez różne czynniki, a przede wszystkim procesem, który może zachodzić z różną prędkością w obrębie grupy społecznej lub na określonym obszarze. Wszystko to zaburza modelowy przebieg takiej zmiany, a w konsekwencji matematyczna idealizacja może stać się dość odległa od świadectwa tekstów. W rozdziale trzecim badamy przebieg kilku współbieżnych zmian językowych, przy czym niekiedy nie chodzi tu ściśle o zmianę językową, ale o konkurencję dwu wyrazów o odmiennej etymologii, z których w końcu jeden wypiera drugi. Dla językoznawcy zapewne ciekawsze od samej regresji logistycznej będzie to, co omawiane podejście mówi o dynamice zmian językowych, a więc zarówno o ich współbieżności, jak i też o rozmaitym tempie przebiegu.

Kolejny rozdział jest poświęcony metodom odnajdywania momentu największej dynamiki zmian w dziejach języka. Mimo że język ewoluuje, a zmiany, które w nim zachodzą, w przeważającej większości tworzą *continnum*, językoznawcy często poszukują najbardziej adekwatnej periodyzacji dziejów języka na potrzeby opisu strukturalnego zachodzących w nim przekształceń. Proponujemy pewną metodę odnajdowania takich momentów w długotrwałych procesach rozwojowych, które wskazują punkt największej zmiany. Oczywiście tam, gdzie chodzi o ewolucję systemu, takim naturalnym momentem jest zakończenie istotnej zmiany językowej. Gdy jednak dochodzimy do podziałów o większej ziarnistości, to zazwyczaj sam system jest dość stabilny, natomiast intuicja językoznawcy podpowiada, że teksty powstałe przed jakąś datą i po niej są jednak odmienne. Proponowana przez nas metoda ma za zadanie zobiektywizować tę intuicję, wskazując datę, która dzieli chronologicznie uporządkowany zbiór tekstów na dwa najbardziej odmienne podzbiory.

Ostatni rozdział poświęcony jest zagadnieniu zmian w produktywności pewnej formy gramatycznej. Imiesłów uprzedni – bo o nim mowa – jest formą fleksyjną, której frekwencja ulegała bardzo dużym wahaniom. Zrazu bardzo rzadka, forma ta zyskała niezwykłą popularność w XVII wieku, by później zanotować znaczący spadek liczby wystąpień. Ów spadek osiągnął najniższy punkt na przełomie wieków XVIII i XIX, następnie imiesłowy uprzednie zaczęły (częściowo) odzyskiwać utraconą pozycję. Rodzi się więc pytanie, czym były spowodowane zarówno wzrost, jak i spadek. Nasuwają się tu dwa przypuszczenia: albo mamy do czynienia z czymś, co można by nazwać językową modą, albo też ze zmianami w produktywności. W pierwszym wypadku oznaczałoby to, że użytkownicy języka wciąż posługiwali się podobnym zasobem słów, lecz używali ich z różną

częstotliwością; w drugim – że zasób wyrazów zwiększył się znacząco, przypuszczalnie dlatego, że pewnemu rozluźnieniu uległy ograniczenia (zapewne przede wszystkim semantyczne) nakładane na czasowniki, które przyjmują tę formę. Zagadnienie to badamy za pomocą narzędzi wypracowanych w synchronicznym językoznawstwie korpusowym, porównując produktywność imiesłowów uprzednich w poszczególnych podkorpusach. Przy tym miary produktywności oparte są jedynie na kryteriach ilościowych.

Autorzy niniejszej książki dużą wagę przywiązują do idei powtarzalności eksperymentu i dostępności danych użytych w badaniach. Naszą ambicją jest bowiem nie tylko zaznajomienie czytelnika z metodologią kwantytatywną i korpusową oraz efektami naszych badań nad zmianą w języku – chcielibyśmy również przyczynić się do rozpowszechnienia idei otwartego oprogramowania i udostępniania danych użytych w eksperymentach (w takim stopniu, w jakim jest to możliwe). Z powyższych powodów do wykonywania obliczeń wykorzystywaliśmy wyłącznie ogólnodostępne oprogramowanie statystyczne i korpusowe (głównie środowisko programistyczne R oraz program AntConc), a rezultaty tych obliczeń może czytelnik łatwo zweryfikować na podstawie danych wejściowych zdeponowanych w internetowym repozytorium. Wprawdzie z powodów prawnych nie możemy udostępnić pełnej wersji używanego przez nas korpusu, ale wszystkie pozostałe dane liczbowe pozyskane z tego korpusu, w tym tabele frekwencji i zbiory surowych danych, jak również wszystkie wykresy oraz ilustracje użyte w niniejszej książce, można znaleźć w ogólnodostępnym repozytorium internetowym GitHub pod adresem: https://github.com/computationalstylistics/diachronia.

Niech nam w tym miejscu wolno będzie podziękować kilku osobom, których pomoc w znacznej mierze przyczyniła się do powstania tej książki. Przede wszystkim chcielibyśmy podziękować twórcom wspomnianych powyżej korpusów za łaskawe użyczenie tekstów na potrzeby niniejszego projektu. Pozwoliły one znacząco uzupełnić zbiór tekstów, który stanowił podstawę empiryczną naszych badań. Chcielibyśmy również wyrazić wdzięczność zespołowi pracującemu nad korpusem XIX-wiecznym za udostępnienie analizatora morfologicznego dla tekstów historycznych jeszcze przed jego oficjalną premierą. Wrszcie pragniemy podziękować dr hab. Magdalenie Derwojedowej i prof. Włodzimierzowi Gruszczyńskiemu za wnikliwe i cenne uwagi zawarte w ich recenzjach. Oczywiście odpowiedzialność za wszystkie niedoskonałości niniejszej książki spada wyłącznie na autorów.

Rozdział 1

Metody korpusowe i kwantytatywne w językoznawstwie historycznym

Językoznawstwo historyczne zawsze było „językoznawstwem korpusowym" w tym sensie, że z zasady odwoływało się do świadectwa tekstu, a nie do kompetencji językowej badacza. Oczywiście mówimy tu o badaniach tych epok, które pozostawiły po sobie świadectwa pisane, nie zaś o epoce przedpiśmiennej, do której badania stosuje się jeszcze inne metody – przede wszystkim metodę historyczno-porównawczą i rekonstrukcję wewnętrzną. Mogłoby się więc wydawać, że dość głęboka ewolucja lingwistyki, jaka ma miejsce pod wpływem szerokiego zastosowania korpusów językowych, nie dotyczy tego jej działu, którym jest językoznawstwo historyczne. Jest jednak niewątpliwie inaczej; co więcej, można dostrzec, jak historia zatacza koło. Oto w synchronii podjęto badania nawiązujące do metody filologicznej stosowanej w językoznawstwie historycznym – mamy tu na myśli wczesne językoznawstwo korpusowe z jego traktowaniem tekstu w gruncie rzeczy jak informatora, tj. rodzimego użytkownika języka. Od tego czasu badania korpusowe przeszły gruntowną ewolucję, a udoskonalone metody wróciły do językoznawstwa historycznego.

Przemiany synchronicznego językoznawstwa korpusowego stanowią zatem niewątpliwy impuls dla powstania przełomowych prac w językoznawstwie historycznym. Drugim impulsem jest stale wzrastająca dostępność tekstów w wersji elektronicznej. Warto na marginesie zauważyć, że proces dygitalizacji w dużej mierze nie jest motywowany potrzebami językoznawców, lecz badaczy z różnych dziedzin czy wręcz „konsumentów kultury" (por. takie inicjatywy, jak Oxford Text Archive, Project Gutenberg, Wolne Lektury itp.). Co więcej, teksty dawne nie są chronione prawem autorskim, co na ogół stanowi barierę w badaniach korpusowych. Sama kwestia dostępności tekstów dla poruszanych tu problemów metodologicznych jest obojętna, jednak ta właśnie stale wzrastająca ilość danych tekstowych umożliwia prowadzenie badań wielkoskalowych i szukanie już nie pojedynczych czy nielicznych potwierdzeń, ale tego, co bardzo częste i typowe. Z kolei znaczące powiększenie skali pozwala na szersze zastosowanie bardziej wyrafinowanych metod statystycznych. Jakkolwiek wszystko to jest dobrze znane z synchronicznego językoznawstwa korpusowego, w diachronii możemy mówić o nowej – w stosunku do tego, co nazywano metodą filologiczną – jakości.

Nie znaczy to oczywiście, że w epoce przedkorpusowej nie istniała w językoznawstwie historycznym świadomość roli danych ilościowych. Znakomitym

studium, w którym przebieg zmian językowych jest śledzony poprzez precyzyjny opis ilościowy, jest praca Bajerowej (1964). Również Wierzbicka w rozprawie o szesnastowiecznej prozie (1966) popiera swój wywód danymi ilościowymi. Podobnie zbiór prac pod redakcją Ostaszewskiej (2002).

1.1. Korpusy historyczne i diachroniczne

Współcześnie termin korpus językowy (tekstowy) implikuje cztery cechy. Powinien być: zdygitalizowany, anotowany, duży i reprezentatywny (McEnery i Wilson, 2001). Dobrze, by taki korpus był również zaopatrzony w moduł konkordancyjny. Żadna z tych cech nie jest jednak konstytutywna. Mianem anotacji określa się zarówno metryczkę tekstu (anotacja zewnętrzna, ang. *external annotation*), jak i anotację gramatyczną. Tak więc każdy tekst włączony do korpusu powinien być opatrzony dokładną informacją bibliograficzną i ewentualnie socjolingwistyczną (data urodzenia i pochodzenie autora, a także typ tekstu). Informacje te są istotne ze względu na oczywisty obowiązek lokalizacji cytatów, ale też dlatego, że pozwala to ograniczać przeszukiwanie korpusu do tekstów spełniających pewien warunek, np. pochodzenie autora z pewnego określonego regionu. Innym elementem anotacji zewnętrznej jest opis struktury tekstu. Chodzi o to, by *explicite*, w sposób zrozumiały dla programów przeszukujących korpus, oddzielić np. tytuły od tekstu, wyznaczyć początek i koniec rozdziału czy też partię tekstu z określonej stronicy druku, co w wypadku danych historycznych może być istotne, gdyż przyjęto, że należy podawać lokalizację cytatu wraz ze stroną.

Z kolei anotacja gramatyczna polega na opatrzeniu każdego słowa jego charakterystyką gramatyczną. Zdanie

Pies mamy szczeka.

to sekwencja rzeczownik – rzeczownik – czasownik, zaś zdanie

Mamy psa.

jest sekwencją czasownik – rzeczownik. Oczywiście charakterystykę fleksyjną każdego z tych wyrazów należy rozbudować o przypadek, liczbę, rodzaj itd. Efektywność i komfort pracy z korpusem znacząco wzrasta, jeśli użytkownik może znaleźć wszystkie wystąpienia danego wyrazu bez względu na formę fleksyjną, ale też wystąpienia zawężone precyzyjnie tylko do szukanego wyrazu (a więc gdy użytkownik szuka *mieć*, powinien dostać zdanie drugie z powyższej pary, ale nie pierwsze). Co więcej, nie jest on ograniczony do przeszukiwania wyłącznie wyrazów, może też szukać danej formy fleksyjnej we wszystkich jej wystąpieniach. Możliwe jest nawet przeszukiwanie korpusu pod kątem składni, przy czym trze-

ba się wtedy uciec do pewnego zabiegu adaptacyjnego, mianowicie poszukiwane konstrukcje składniowe sprowadza się do sekwencji form fleksyjnych[1].

Ujednoznacznienia form fleksyjnych w tekście dokonuje się przy pomocy programu zwanego tagerem. W rzeczywistości tager składa się z dwu programów – analizator morfologiczny rozpoznaje wszystkie potencjalnie istniejące interpretacje form homonimicznych w tekście (*mamy* interpretuje zarówno jako rzeczownik, jak i czasownik), następnie dezambiguator dokonuje ujednoznacznienia, tzn. ustala, która z interpretacji jest w danym kontekście prawidłowa, a ściślej mówiąc: bardziej prawdopodobna. Oczywiście tager jest stworzony dla konkretnego języka i nie może sobie skutecznie radzić, gdy spotyka formy nieprzewidziane przez gramatykę, a takimi są wszystkie formy archaiczne. Tym samym możliwości użycia do tekstów dawnych analizatora zbudowanego dla współczesnego języka są ograniczone, przy czym oczywiście im wcześniejsze teksty, tym bardziej skuteczność tagera maleje (Eder, Klapper i Kołodziej, 2015).

Stworzenie skutecznego tagera (rozumianego jako kompletny system wraz z danymi fleksyjnymi) jest przedsięwzięciem bardzo kosztownym. O ile napisanie programu komputerowego, który przypisuje formom interpretację, jest względnie proste, o tyle stworzenie słownika tych form jest już dużo bardziej pracochłonne. Wszystko to jest jednak ułamkiem wysiłku, jakim jest ręczne oznaczenie form w stosunkowo dużym korpusie (co najmniej 500 000–1 000 000 słów), który następnie służy do wytrenowania tagera. O ile w wypadku tagerów dla języka współczesnego można liczyć na wsparcie agend badawczych i firm informatycznych, choćby z racji tego, że jest to narzędzie użyteczne również biznesowo, o tyle historycy języka rzadko mogą liczyć na taką hojność instytucji finansujących badania. Konieczne staje się zatem adaptowanie narzędzi tworzonych dla współcześnie używanych języków do tagowania tekstów dawnych. Istnieje spora literatura na temat tych rozwiązań (Piotrowski, 2012). W największym skrócie można powiedzieć, że zazwyczaj pierwszym krokiem jest normalizacja grafii, często jej „przekład" na normy współczesne, a także dopisanie wyrazów oraz form wymarłych do słownika fleksyjnego wbudowanego w tager. Skuteczność w wyniku takich zabiegów znacząco wzrasta, jednak wciąż jest niższa niż działanie tagera na tekstach współczesnych[2].

Trzecią cechą współczesnych korpusów (po dygitalizacji i anotacji) jest wspomniana powyżej reprezentatywność, z definicji problematyczna w korpusach obejmujących teksty z różnych epok. By wyjaśnić ten problem, trzeba wcześniej wspomnieć o pewnej istotnej różnicy między korpusem historycznym i diachro-

[1] Na przykład w pewnym uproszczeniu można powiedzieć, że polski czas przyszły to sekwencja słowa *być* w czasie przyszłym i czasownika w czasie przeszłym bądź w bezokoliczniku (w dowolnej kolejności).

[2] Dość powiedzieć, że dla siedemnastowiecznego tekstu niderlandzkiego skuteczność tagowania bez żadnych zabiegów to najwyżej 68% (a więc co trzecie słowo jest błędnie rozpoznane!), zaś po zastosowaniu wszystkich, dość skomplikowanych procedur wzrasta do 89% (Sang, 2016).

nicznym. Ten pierwszy to korpus reprezentujący stan języka w danej epoce historycznej. Ten drugi to właściwie seria podkorpusów reprezentujących poszczególne epoki. W korpusie diachronicznym poszczególne podkorpusy powinny mieć podobną budowę, co zwykle jest postulatem niemożliwym do spełnienia ze względu na to, że po pierwsze w wypadku dawniejszych epok zachowały się nieliczne i dość przypadkowe teksty, a po drugie w dziejach pojawiały się nowe typy tekstów, zaś gatunki niegdyś popularne ulegały marginalizacji. I tu dochodzimy do zagadnienia reprezentatywności korpusu. Jest to kwestia dość kluczowa, skoro korpus ma być wiarygodną reprezentacją języka[3]. W wypadku późniejszych epok, które pozostawiły po sobie bogatsze świadectwo, twórcy korpusu zwykle tworzą próbkę tekstów należących do różnych gatunków i rejestrów. Towarzyszy temu założenie, że dzięki odpowiedniej proporcji poezji, prozy, języka konwersacyjnego, gatunków użytkowych, prasy itd. dostaniemy obraz języka wprawdzie niedoskonały, ale pewnie w jakiś sposób zbliżony do stanu faktycznego.

Nietrudno się domyślić, że powyższa metoda próbkowania jest całkowicie nieadekwatna do materiału staropolskiego czy, powiedzmy, staroangielskiego. W wypadku dawniejszych epok tekstów jest zwykle tak mało, że jedynym sensownym rozwiązaniem jest włączenie ich wszystkich do korpusu. Na przykład *Korpus tekstów staropolskich do 1500 r.* (Twardzik i Górski, 2003) obejmuje okres około dwu wieków, ale stan zachowania średniowiecznego piśmiennictwa polskiego nie pozwala stworzyć kilku reprezentatywnych podkorpusów, nie mówiąc już o tym, że piśmiennictwo z XIV wieku jest w korpusie niemal nieobecne, początek zaś wieku XV zachowany w zasadzie szczątkowo. Z reprezentatywnością wiąże się też dodatkowe utrudnienie, a mianowicie problem zróżnicowania gatunkowego tekstów składających się na korpus. Optymalnie poszczególne podkorpusy następujących po sobie epok powinny mieć identyczną budowę i zawierać podobną reprezentację różnych gatunków – ostatecznie zjawiska językowe występują z różnym natężeniem w różnych typach tekstów, więc tylko spojrzenie przekrojowe, a zatem uśredniające wyniki dla wielu różnych gatunków, jest w stanie dać wiarygodne pojęcie o zjawiskach języka jako takiego. Oczywiście taka idealna sytuacja jest niemożliwa, ponieważ najistotniejsze dla wczesnych epok typy tekstów z czasem tracą na znaczeniu (np. teksty prawne czy homiletyka), podczas gdy inne zyskują na popularności albo wręcz pojawiają się *ex nihilo* na przestrzeni dziejów – np. ogromny wpływ prasy na język ogólny to zjawisko dość nowe. Trudno zatem o zachowanie takiej samej reprezentacji gatunkowej w perspektywie kilkudziesięciu, a tym bardziej kilkuset lat.

Świadomość metodologicznych niebezpieczeństw czyhających na badacza korpusów diachronicznych nie zmienia faktu, że owych niebezpieczeństw w zasadzie nie da się uniknąć. Na szczęście w ostatnich latach pojawiło się kilka

[3] Reprezentatywność może być rozumiana bardzo różnie (por. Biber, 1993; Górski i Łaziński, 2012).

dobrej jakości korpusów historycznych, dzięki czemu dostęp do poszczególnych epok w rozwoju polszczyzny jest dalece łatwiejszy niż jeszcze dziesięć lat temu. Zapewne pierwszym elektronicznym korpusem historycznym języka polskiego był *Elektroniczny korpus tekstów staropolskich do roku 1500* (Twardzik i Górski, 2003). W latach 2009–2012 powstał niewielki korpus w ramach dużego międzynarodowego projektu IMPACT (Bień, 2014), obecnie powstaje *Korpus polszczyzny XVI wieku*[4] oraz *Korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku)*, zwany też przez autorów *Korpusem Barokowym* (KorBa), obejmujący XVII i XVIII wiek (Gruszczyński i in., 2013). W ramach projektu dotyczącego automatycznej analizy morfologicznej polszczyzny XIX wieku powstał również mikrokorpus polszczyzny tej epoki (Derwojedowa, Kieraś, Skowrońska i Wołosz, 2014). Jeśli chodzi o wiek XX, to rzecz jasna należy wymienić NKJP, który wprawdzie nie jest korpusem diachronicznym, ale zawiera źródła obejmujące co najmniej kilkadziesiąt lat. Korpusem *stricte* diachronicznym jest natomiast *ChronoPress* (Pawłowski, 2016), obejmujący lata 1945–1955 korpus diachroniczny prasy polskiej. Po uzupełnieniu *Korpusu Barokowego* o teksty powstałe przed rokiem 1800, co stawia sobie za cel kontynuacja projektu *Korpusu tekstów polskich z XVII I XVIII w.*, jedynie okres 1800–1830 nie będzie reprezentowany przez elektroniczny, publicznie dostępny korpus.

Swoistą namiastką (i uzupełnieniem) korpusu są też kartoteki naukowych słowników historycznych. Tam, gdzie kluczem wyszukiwania jest konkretny leksem, takie kartoteki okazują się niezwykle pomocne. Kartoteki *Słownika staropolskiego*, *Słownika polszczyzny XVI w.*, *Słownika języka polskiego XVII i 1. połowy XVIII wieku* zostały zeskanowane i udostępnione w internecie[5]; słowniki te obejmują polszczyznę od najwcześniejszych poświadczeń pisemnych do połowy XVIII wieku. Dla anglistów podobną rolę spełnia *Oxford English Dictionary*.

Można oczywiście zadać zasadne pytanie, czy cztery korpusy historyczne uzupełnione o kartoteki trzech słowników to wystarczająco dużo, by przeprowadzić wiarygodne badania diachroniczne. Porównanie z najpełniej opracowanym językiem, czyli angielskim, nie wypada jednak niekorzystnie dla korpusów rodzimych. Liczbę bowiem korpusów historycznych i diachronicznych języka angielskiego ocenia się na około 30–40 (Kytö, 2011). Są wśród nich korpusy stawiające sobie za cel równomierne pokrycie kilkuset lat, jak i drobne zbiory tekstów jednej epoki. Warto tu zwrócić uwagę na FLOB (Freiburg LOB Corpus of British English) i FROWN (The Freiburg-Brown Corpus of American English), korpusy powtarzające możliwie dokładnie strukturę pierwszych korpusów języka angielskiego z lat sześćdziesiątych, ale składające się z tekstów powstałych 40 lat później. Jest to o tyle interesujące, że taki okres uchodził dotąd za zbyt krótki,

[4] Informacje na temat tego korpusu można znaleźć pod adresem https://spxvi.edu.pl/korpus/.

[5] Por. http://www.rcin.org.pl/publication/23662; http://www.rcin.org.pl/publication/20029 oraz http://rcin.org.pl/dlibra/publication?id=43801&tab=3.

by go analizować z punktu widzenia diachronii. Tymczasem ostatnie badania (np. Mair, 2006) pokazują, że różnorakie zmiany językowe, choć o charakterze ilościowym raczej niż jakościowym, widać nawet na przestrzeni niewielu lat.

Z coraz lepszą – mimo wysuwanych powyżej zastrzeżeń – dostępnością tekstów nie idzie niestety w parze dostępność narzędzi do anotacji morfologicznej tekstów dawnych. Niemniej i w tej mierze widać wyraźny postęp. Dostrzegając tę dysproporcję między dostępnością materiału badawczego i ograniczeniami technologicznymi, Magdalena Derwojedowa postanowiła stworzyć analizator morfologiczny dla polszczyzny XIX wieku, który pozwoliłby badaczom samodzielnie anotować coraz bardziej dostępne w wirtualnych bibliotekach teksty, zamiast zaczynać pracę od korpusu reprezentującego tę epokę (por. Derwojedowa i in., 2014). Nad analizatorem morfologicznym Chronofleks, przystosowanym do polszczyzny XVII wieku, pracuje z kolei Marcin Woliński z zespołem[6].

1.2. Zmiana językowa

W językoznawstwie diachronicznym mamy do czynienia z ustaleniami typu: „Forma/struktura X zastępuje formę/strukturę Y w czasie T". Znany jest punkt wyjścia i punkt dojścia oraz – w mniejszym czy większym przybliżeniu – czas, kiedy dana zmiana zaszła, a przynajmniej chronologia względna zmiany (zjawisko A nastąpiło po zjawisku B, choć moment przejścia jednego zjawiska w drugie jest nieznany). Elektroniczny korpus pozwala jednak na znacznie więcej, bo daje wgląd w przebieg i dynamikę zmiany, dostarczając przykładów tekstowych w takiej liczbie, o jakiej nie mógł marzyć historyk języka pracujący z drukowanymi wydaniami dawnych tekstów (na ogół zresztą językoznawcy nie pracowali bezpośrednio z wydaniami, lecz ekscerptami w postaci tysięcy fiszek). Dalej: mechanizmy napędzające zmianę można zidentyfikować, jeśli odszuka się wczesne konteksty ją poświadczające, a także cechy formalne, funkcjonalne i pozajęzykowe tych kontekstów (typ tekstu, charakterystyka autora itp.). Wszystko to zachęca, by przesunąć punkt ciężkości z ustalania faktu, że zmiana zaszła, na badanie samego procesu zmiany i – jak zauważa Kytö – na odkrywanie wpływu, jaki użycie języka wywarło na jego strukturze (Kytö, 2011).

Teoria gramatykalizacji liczy sobie z górą 100 lat, niemniej w ostatnich 20 latach nabrała wyraźnego impetu. Badania nad gramatykalizacją mają dość spekulatywny charakter, jeśli są oderwane od danych obserwowanych w tekstach i od analiz ilościowych. Korpusowe badania tego zjawiska są jednym z ważniejszych nurtów współczesnego historycznego językoznawstwa korpusowego. Dodajmy, że Mair (2006) zauważa pewne zbieżności pomiędzy teorią gramatykalizacji i językoznawstwem korpusowym. Oba nurty

[6] Jest to projekt *Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja*, finansowany ze środków Narodowego Centrum Nauki, nr 2014/15/B/HS2/03119.

– za najistotniejszy obiekt badań uznają osadzone w dyskursie wypowiedzenia, nie zaś abstrakcyjne systemy reguł,
– podkreślają rolę danych ilościowych oraz statystyki,
– uznają, że nie ma ostrych przejść od jednej do drugiej kategorii gramatycznej; przeciwnie, to przejście jest stopniowe,
– stają się modne po latach wegetowania na skraju głównego nurtu językoznawstwa (Mair, 2006).

Znakomitymi przykładami tego rodzaju podejścia są m.in. zbiór artykułów *Corpus Approaches to Grammaticalization in English* (Lindquist i Mair, 2004), rozprawa Hilperta (2008) czy dotycząca m.in. polszczyzny praca von Waldenfelsa (2012).

Innym przykładem są badania nad gramatykalizacją *going to* jako wykładnika czasu przyszłego (Danchev i Kytö, 1994). Warto dodać, że analogiczny proces gramatykalizacji zaszedł nie tylko w angielskim, ale także we francuskim (*Je vais le faire demain* 'zrobię to jutro': *je vais* znaczy dosłownie 'idę') czy hiszpańskim (*Voy a llevar a mi hermana a casa* 'wezmę moją siostrę do jej domu'; *voy* to dosłownie 'idę'). Intuicyjnie bardzo łatwo odtworzyć drogę tego rodzaju procesu. W ostatnich latach jednak podjęto próby jego prześledzenia, tak by spekulację zastąpić obserwacją. Danchev i Kytö wykazali, że aż do XVII wieku najbardziej typowymi czasownikami towarzyszącymi *going to* są *bring* 'przynosić, *give* 'dać', *meet* 'spotkać', *see* 'widzieć' oraz *visit* 'odwiedzać', a więc czasowniki denotujące czynności konkretne, otwierające pozycje dla ożywionych, kontrolujących czynność agensów (Danchev i Kytö, 1994). Hilpert przebadał kolejne okresy, wykazując, że do tej listy dochodzą czasowniki o szerszym znaczeniu i denotujące akcje niekontrolowane, jak *happen* 'zdarzyć się' (Hilpert, 2008). Można więc prześledzić, jak konstrukcja wciąż zachowywała preferencje dla czasowników denotujących czynności wykonywane w następstwie ruchu jeszcze długo po tym, kiedy wyraz autosemantyczny stał się wykładnikiem czasu. Jest to ślad dawnego znaczenia owego *going*.

Trzeba tu dodać, że Hilpert stosuje metodę zwaną różnicującą analizą koleksemów (*distinctive collexeme analysis*, Stefanowitsch i Gries, 2003). W skrócie polega ona na tym, że bada się występowanie leksemów o dystrybucji komplementarnej w obrębie danej konstrukcji – w omawianej pracy są to czasowniki występujące po *going to*. Przy czym niekoniecznie chodzi o leksemy najczęściej pojawiające się w tym kontekście, ale raczej współwystępujące z omawianą konstrukcją częściej, niż by to miało wynikać z rozkładu losowego. By ująć rzecz bardziej obrazowo: chodzi o leksemy, które „lgną" do danej konstrukcji bądź są przez nią „odpychane". Metoda ta oczywiście wymaga danych ilościowych i jest wypracowana na gruncie korpusowego językoznawstwa synchronicznego.

Prosta zmiana frekwencji jakiegoś zjawiska łatwo daje się zinterpretować jako pewien trend historyczny. Kilka prac jednak wykazuje, że takie dane, choć praw-

dziwe, mogą prowadzić do fałszywych wniosków. Szmrecsanyi dowodzi, że proporcje w użyciu obu konstrukcji posesywnych w języku angielskim (odpowiednio N *of* N i dopełniacz saksoński) ulegają w dziejach dość gwałtownym zmianom – udział pierwszej konstrukcji w ogólnej liczbie konstrukcji posesywnych w XIX wieku wzrasta, by w XX wieku osiągnąć niemal 90% udziału we wszystkich konstrukcjach wyrażających posesywność, a następnie w naszych czasach spaść do 62% (Szmrecsanyi, 2015). Ów wzrastający trend można uznać za oczywisty proces wypierania jednej synonimicznej konstrukcji przez drugą. W rzeczywistości jednak istnieją pewne wyraźne tendencje wyboru jednej z konstrukcji, na przykład w zależności od charakteru posesywności (rzeczywiste posiadanie, relacja część–całość itp.), od tego, czy posesorem jest obiekt ożywiony, czy nie itp. Warto dodać, że tendencje te zostały zaobserwowane w korpusach współczesnej angielszczyzny.

Przytaczane badania dowodzą, że jednym z najważniejszych czynników, które wpływają na wybór konstrukcji, jest kategoria żywotności w ogóle, a szczególnie charakter bytu ożywionego. Jeśli się weźmie pod uwagę wszystkie te czynniki, które wpływają na wybór, to można dojść do konkluzji, że zmiany w proporcjach użycia obu form wynikają przede wszystkim ze zmiany tematyki utworów, w których używa się konstrukcji posesywnych; w pewnym momencie rozwoju języka autorzy zaczęli opisywać przede wszystkim byty nieożywione, które faworyzują wybór formy z przyimkiem. Ponadto zauważyć można, że niegdyś bardziej rygorystycznie obowiązywały podobne tendencje. Te dwa czynniki się niejako sumowały, powodując przejściowy spadek frekwencji genetiwu saksońskiego.

Powodem, dla którego przywołujemy badania Szmercsany'ego, jest konkluzja badacza, by zbytnio nie ufać samym danym liczbowym bez głębszego spojrzenia na to, co się za nimi kryje (Szmrecsanyi, 2015). Stosunkowo sporo miejsca poświęcamy tej jednej rozprawie również z tego powodu, żeby poczynić następujące uwagi: po pierwsze argumentacja Szmrecsanyiego możliwa jest jedynie przy użyciu dość wyrafinowanych narzędzi statystycznych, po drugie zaś, że badania te to w istocie prowadzona na przestrzeni dziejów obserwacja czynników, które wpływają na wybór jednej z dwu konkurujących konstrukcji; czynników znanych nauce dzięki uprzednim synchronicznym badaniom korpusów współczesnego języka.

1.3. Podsumowanie

Dostęp do dużej ilości danych, a w konsekwencji możliwość użycia zaawansowanych metod ilościowych, zmieniają oblicze językoznawstwa historycznego i przesuwają akcenty w badaniach. Należą do nich niewątpliwie przesunięcie akcentu ze śledzenia zmian jakościowych na ilościowe. Oczywiście to przesunięcie akcentu ma dwojaką przyczynę: zarówno sam fakt, że obecnie dane ilościowe po-

zyskuje się łatwiej, jak i to, że zmiany o charakterze jakościowym w odniesieniu do ważniejszych języków są już dobrze znane i trudno o przełomowe odkrycia w tym zakresie.

Dane ilościowe są bardziej wiarygodne przede wszystkim dlatego, że są oparte na obszernej bazie empirycznej, jaką zapewnia odpowiednio duży elektroniczny korpus. Do tego dochodzi fakt, że przeszukiwanie korpusu trwa sekundy (a nie miesiące, jak w przypadku kwerend wykonywanych bezpośrednio na materiałach źródłowych), choć oczywiście często konieczna jest pracochłonna ręczna kontrola wyników.

Dostępność danych wzrasta wraz ze zbliżaniem się do współczesności – tekstów sprzed 100 lat jest znacznie więcej niż sprzed 300, co wynika zarówno z lepszego stanu ich zachowania, jak i przede wszystkim ze znacznie mniejszej produkcji literackiej w dawniejszych czasach. Z punktu widzenia diachronisty jest to dość smutny paradoks. Oto dla czasów bliskich współczesności, znacznie mniej ciekawych pod względem zmian w systemie gramatycznym, dane były znacznie obfitsze niż dla ciekawszych czasów dawniejszych. Zarazem jednak to właśnie obfitość danych skłaniała do bliższego przyjrzenia się owym pozornie nieciekawym, chronologicznie bliższym nam epokom. To właśnie dzięki metodom ilościowym można stwierdzić, że w krótkiej, nawet zaledwie kilkudziesięcioletniej perspektywie, język nie jest bytem całkowicie statycznym. Z kolei solidne dane liczbowe umożliwiają pogłębione analizy statystyczne, które pozwalają nie tylko odnotować wzrost bądź spadek frekwencji, ale także rozstrzygnąć, które ze splecionych czynników miały inny niż pozostałe wpływ na zmianę.

Z powyższych uwag wynika, że do prowadzenia badań diachronicznych ilościowych potrzebna jest kompetencja w zakresie językoznawstwa historycznego połączona z dobrą orientacją w synchronicznej lingwistyce korpusowej, statystyce i często przynajmniej pewne umiejętności programistyczne. Połączenie tych kompetencji nie jest powszechne, co skutkuje tym, że coraz częściej publikacje naukowe sygnują zespoły wieloautorskie. Niewątpliwie jednak ta dość głęboka transformacja mogła zajść również dzięki temu, że poprzednie pokolenia historyków języka bardzo precyzyjnie zinwentaryzowały zmiany językowe. Tym niemniej jesteśmy przekonani, że zarysowane w niniejszej książce propozycje badawcze mogą dać drugą młodość językoznawstwu historycznemu.

Rozdział 2

Korpus

2.1. Wprowadzenie

Podstawę empiryczną naszych badań stanowił zestawiony specjalnie na potrzeby tego projektu korpus, przy czym dzięki niezwykłemu zrządzeniu losu praca sprowadzała się w dużej części do złączenia istniejących zasobów i późniejszego pozyskania (z różnych źródeł) tekstów uzupełniających braki – już to pod względem pokrycia chronologicznego, już to pod względem reprezentacji różnych gatunków. W momencie gdy planowaliśmy i rozpoczynaliśmy niniejszy projekt, sytuacja badacza zainteresowanego korpusami historycznymi różniła się znacząco: nie istniał żaden publicznie dostępny korpus historyczny, który obejmowałby zakres chronologiczny 1550–1850. W ostatnich latach sytuacja zmieniła się diametralnie: powstały zasoby w różnym stopniu wypełniające tę lukę. Jest to w pierwszym rzędzie *Elektroniczny korpus tekstów polskich XVII i XVIII wieku (do 1772 roku)*[1], wzorowy korpus, anotowany morfologicznie, z tekstami oznaczonymi bogatym zestawem metadanych. Oprócz niego pojawił się *Korpus polszczyzny XVI wieku*[2]. Jego koncepcja jest z gruntu odmienna – jest on nie tyle korpusem językowym, ile wirtualną biblioteką, w której teksty dostępne są z anotacją strukturalną, w wersji transliterowanej, z zachowaniem żywej paginy, adjustacji i grafii. Trzecim korpusem, który powstał równolegle z trwaniem niniejszego projektu, był mikrokorpus tekstów z lat 1830–1918 znany pod nazwą *f19*[3]. Ten korpus z kolei był opracowywany niejako na marginesie tworzenia analizatora morfologicznego dla polszczyzny XIX wieku (Derwojedowa i in., 2014). Został

[1] Projekt realizowany w Instytucie Języka Polskiego PAN pod kierunkiem Włodzimierza Gruszczyńskiego, finansowany przez Narodowy Program Rozwoju Humanistyki (nr 0036/NPRH2/H11/812012), dostępny na stronie https://www.korba.edu.pl/.

[2] Projekt realizowany przez Instytut Badań Literackich PAN pod kierunkiem Patrycji Potoniec, również finansowany przez Narodowy Program Rozwoju Humanistyki (nr 0138/FNiTP/H11/80/2011), którego opracowane teksty dostępne są na stronie https://spxvi.edu.pl/korpus/.

[3] Projekt realizowany na Uniwersytecie Warszawskim pod kierunkiem Magdaleny Derwojedowej, finansowany przez Narodowe Centrum Nauki (nr 2012/07/B/HS2/00570), dostępny przez interfejs Poliqarp na stronie https://szukajwslownikach.uw.edu.pl/f19/.

wyposażony w anotację morfosyntaktyczną oraz lematyzację, składa się natomiast z niewielkich próbek tekstów i w konsekwencji jest bardzo mały. Ponadto dysponowaliśmy tekstami *Korpusu tekstów staropolskich do roku 1500*[4], który jednak nie jest ani lematyzowany, ani anotowany morfosyntaktycznie. Nie podejmowaliśmy prób stworzenia narzędzi do anotacji tego korpusu, ponieważ staropolszczyzna jest na tyle odmienna, że problemy związane z jej analizą składniową, fleksyjną czy leksykalną są nieporównanie większe niż to ma miejsce w odniesieniu do późniejszych etapów rozwojowych polszczyzny, ponadto staropolszczyzna zasadniczo pozostawała poza zasięgiem naszego zainteresowania. Na tyle jednak, na ile to możliwe, modyfikowaliśmy teksty staropolskie tak, by móc poddać je analizie w miejscach, gdzie wydało się to uzasadnione dla przedstawienia trendu lub pewnego ciągu zmian zachodzących w języku. Obecnie w Instytucie Języka Polskiego PAN pod kierunkiem Ewy Deptuchowej powstaje nowy korpus polszczyzny do roku 1500 (Deptuchowa i in., 2019; Jasińska i Kołodziej, 2019), który w zamierzeniu ma być opatrzony anotacją morfosyntaktyczną[5]. Liczymy na to, że teksty z nowego korpusu wkrótce wzbogacą bazę materiałową badań diachronicznych.

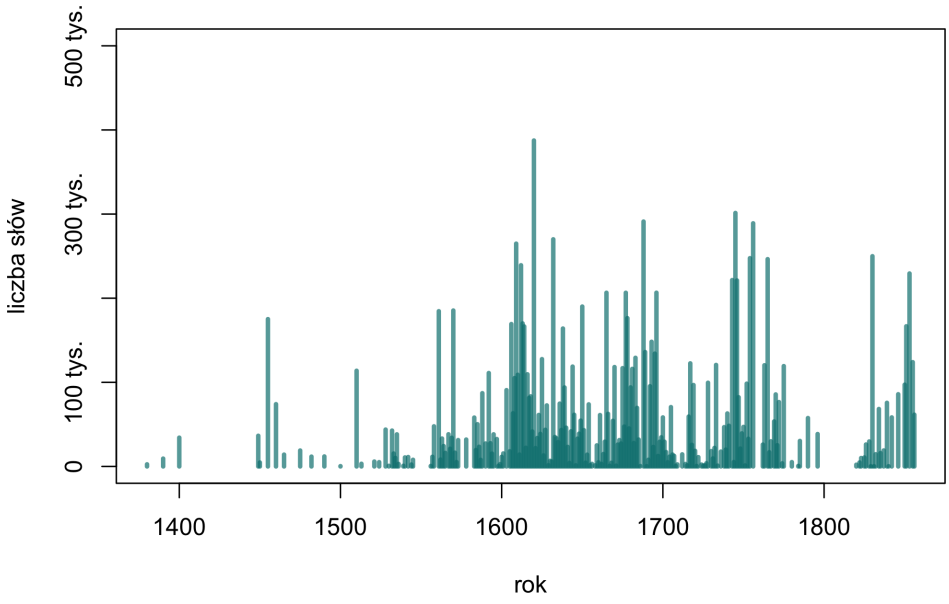
Kierownicy wymienionych projektów z niezwykłą życzliwością, jeszcze przed oficjalnym upublicznieniem korpusów, udostępnili nam wybrane pliki tekstowe (bez anotacji morfosyntaktycznej) do naszych badań.

Ostateczna wersja korpusu składa się z komponentów omówionych powyżej, ale równie ważną jego częścią był *Korpus diachroniczny IJP PAN*, stworzony specjalnie w ramach niniejszego projektu. Choć jest autonomiczny, w zamierzeniu ma na celu uzupełnienie nieciągłości materiałowych powstałych po chronologicznym uporządkowaniu pozostałych podkorpusów w jeden zbiór. Obejmuje on 101 tekstów (40 poetyckich i 61 prozatorskich) o łącznej objętości 1 475 583 wyrazów tekstowych. Najwcześniej datowany utwór pochodzi z 1521 roku (Jana z Koszyczek *Rozmowy króla Salomona z Marchołtem grubym a sprosnym*), najpóźniej – z 1855 roku (Józefa Ignacego Kraszewskiego *Diabeł. Powieść z czasów Stanisława Augusta*).

W skład owego niewielkiego korpusu weszły teksty udostępnione nam dzięki uprzejmości Romana Mazurkiewicza, Marii Piaseckiej, portalu staropolska.pl, opublikowanych w serii wydawniczej Bibliotheca Curiosa, serii Biblioteka Pisarzy Staropolskich, a także z publikacji z przełomu XVIII i XIX wieku umieszczonych w serwisach *Biblioteka literatury polskiej w internecie* (https://literat.ug.edu.pl/) oraz *Wolne Lektury* (https://wolnelektury.pl/). Jeśli można było pozyskać starsze wersje tekstów opublikowane w dwóch ostatnich przytoczonych źródłach,

[4] Projekt realizowany przez Instytut Języka Polskiego PAN pod kierunkiem Wacława Twardzika, finansowany przez Komitet Badań Naukowych, dostępny na stronie https://ijp.pan.pl/publikacje-i-materialy/zasoby/korpus-tekstow-staropolskich/.

[5] Korpus powstaje w ramach projektu *Baza leksykalna średniowiecznej polszczyzny (do 1500 r.) Fleksja*, finansowanego z funduszy Narodowego Programu Rozwoju Humanistyki (nr 1H 17 0201 85).



Rysunek 2.1. Chronologiczny rozkład tekstów w korpusie diachronicznym 1380–1850: liczba wyrazów przypadająca na poszczególne lata.

staraliśmy się do nich dotrzeć i je włączyć do zbioru. Do stworzenia korpusu posłużyły również teksty gromadzone przez projekt Polona (https://www.polona.pl), dygitalizowane w ramach projektu Patrimonium i przeformatowane przez naszych anotatorów w trakcie trwania projektu. Jak zaznaczamy kilkukrotnie na kartach tej książki, różne podejścia wymagały różnych zmian w tekstach: ujednolicenia grafii, uwspółcześnienia form, tak żeby jedna kwerenda objęła wszystkie możliwe formy wyrazów np. piętnasto- i dziewiętnastowiecznych.

Źródła pozyskanych tekstów są bardzo zróżnicowane pod względem celu, jaki przyświecał twórcom kolekcji lub wydawcom, oraz stopnia ingerencji edytorów w oryginalną warstwę językowo-ortograficzną utworów, stąd też konieczne dla naszych badań było takie opracowanie pozyskanych materiałów, by możliwe było ich efektywne przeszukiwanie. Część tekstów opracowana została dzięki ręcznej weryfikacji materiałów poddanych OCR (optycznemu rozpoznawaniu znaków, ang. *optical character recognition*). Umieszczenie surowych źródłowych tekstów na tym etapie wiązałoby się z dużymi problemami w przeszukiwaniu korpusu. Dzięki wprowadzonym przez nas zabiegom ujednolicającym (o czym piszemy nieco niżej) przeszukiwanie jest możliwe przy użyciu języka zapytań z NKJP (Przepiórkowski, Bańko, Górski i Lewandowska-Tomaszczyk, 2012). Wyszukiwanie może odbywać się z uwzględnieniem słowoform lub części mowy. *Korpus diachroniczny IJP PAN* jest opublikowany na stronie http://www.ijp.pan.

pl/publikacje-i-materialy/zasoby. Korpus nie jest duży, ale mamy nadzieję, że wzbogaci już istniejące zbiory i w przyszłości stanie się częścią zintegrowanych zasobów korpusowych polszczyzny.

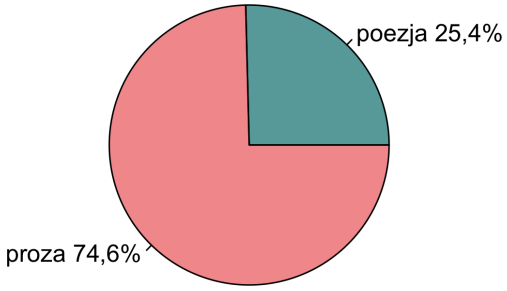
Po złożeniu w jedną całość wszystkich wymienionych powyżej komponentów otrzymaliśmy 12-milionowy korpus diachroniczny polszczyzny pokrywający okres 1380–1850. Objętość danych tekstowych w poszczególnych latach została przedstawiona na Rys. 2.1. Oczywiście obraz rozkładu pozyskanych tekstów na osi czasu daleki jest od modelowego – życzylibyśmy sobie, by dostępne było więcej tekstów z końca XVIII wieku i początku wieku XIX, by luka w pierwszej połowie XVI wieku była mniejsza, by możliwa była prezentacja tekstów w zbliżonym formacie. Wiemy, że niedawno powstały i wciąż powstają inicjatywy, które odpowiedzą na część sformułowanych tu potrzeb, uzupełniając korpusy polszczyzny i zapewniając ciągłość dostępu do tekstów (Król i in., 2019). Celem naszego projektu było modelowanie zmian dających się obserwować na przestrzeni około trzystu lat, dlatego postaraliśmy się uzupełnić omówione w niniejszym rozdziale zasoby w ten sposób, by każda dekada była reprezentowana przez co najmniej kilka tekstów.

2.2. Zawartość korpusu

Zrąb korpusu, na którym prowadziliśmy badania, składał się z 791 tekstów zawierających łącznie 12 325 754 wyrazy rozumiane jako ciąg znaków pomiędzy dwiema spacjami. Mówimy o zrębie, ponieważ dla specyficznych badań rozszerzaliśmy zasób tekstów poza rok 1850.

Tworzony korpus był w założeniu tzw. korpusem oportunistycznym, to znaczy takim, w którym nie zakłada się z góry określonej budowy i nie rozpatruje się jego reprezentatywności. Przeciwnie – włącza się do niego każdy tekst, który jest dostępny. Chodzi o to, by nawet kosztem braku zrównoważenia uzyskać możliwie duży zbiór. Reprezentatywność i zrównoważenie[6] korpusu historycznego jest zagadnieniem delikatnym i trudniejszym do osiągnięcia, niż ma to miejsce w odniesieniu do materiałów współczesnych. Dawne piśmiennictwo było mimo wszystko znacznie mniej zróżnicowane niż współczesne, a literatura piękna zajmowała w nim pozycję istotniejszą niż dziś. Trudno się więc dziwić, że ta ostatnia odmiana tekstów stanowiła zdecydowaną większość materiału, na którym przeprowadzaliśmy badania.

[6] W badaniach korpusowych zwykle pojęcia zrównoważenia i reprezentatywności traktuje się synonimicznie. Warto jednak te dwa pojęcia rozumieć odmiennie (Górski i Łaziński, 2012): o ile reprezentatywność można ogólnie określić jako odzwierciedlanie zróżnicowania języka takim, jakim ono w danej społeczności jest, o tyle przez zrównoważenie należy rozumieć to, że żaden typ tekstów nie dominuje ilościowo całego korpusu. Por. też uwagi Bronikowskiej (2018), a zwłaszcza Adamiec (2015) na ten temat.



Rysunek 2.2. Udział poezji i prozy w korpusie diachronicznym 1380–1850 (12 325 574 wyrazy).

Trzeba również pamiętać, że teksty poetyckie to utwory zawierające neologizmy i modyfikacje wyrazowe, czy – co z naszego punktu widzenia najistotniejsze – nietypowy szyk zdania, których wprowadzanie dyktują niejednokrotnie rytm i rym. Tym niemniej nie odrzucaliśmy w naszych badaniach tekstów wierszowanych. W podejściu ilościowym nazbyt skąpe dane stanowią zwykle większą przeszkodę niż dane gorszej jakości i z tego względu zdecydowaliśmy się na korpus wprawdzie niedoskonały, ale możliwie jak największy. Rozkład gatunkowy w korpusie wykorzystywanym do badań został przedstawiony na Rys. 2.2.

Własności poszczególnych części korpusu, obejmujące kolejne stulecia rozwoju polszczyzny, przekładają się na specyficzne zjawiska wpływające na rezultaty analiz przy użyciu współczesnych narzędzi do przetwarzania języka naturalnego. W dalszej części rozdziału omawiamy procesy przygotowania tekstów na potrzeby omawianego korpusu.

2.2.1. Wieki XIV i XV

Zbiór stanowiący reprezentację dla materiałów z lat 1380–1500 jest jednym z mniejszych podkorpusów użytych w badaniu. Trudno żeby było inaczej, skoro liczba zachowanych tekstów z XIV wieku i pierwszej połowy XV wieku jest szczątkowa; satysfakcjonujące pokrycie pojawia się w zasadzie dopiero w ostatniej ćwierci XV wieku. Zawartość zbioru to zaledwie 17 tekstów, ale aż 393 420 wyrazów tekstowych (na co składa się głównie *Biblia królowej Zofii* i równie obszerne *Rozmyślanie przemyskie*). Obiekty włączone do korpusu są transkrybowane, jednak zachowują oryginalną segmentację i niejednolitą ortografię. Przykładowy tekst zachowany jest w formie:

> Potem z jutra rycerze i panosze poń przyjeli a rzkąc: Błażeju, wynidzi, woła cie książę!
> Teda wystąpiw święty Błażej, przywitał je a rzkąc jim: Dobrzeście przyszli, synowie mili, już widzę, iż mie Bog nie zapomniał, gotowcieśm s wami

jechać, gdziekoli chcecie. Teda tę całą drogę kazania nie przestajał a cuda wielika czynił.
(*Żywot świętego Błażeja*)

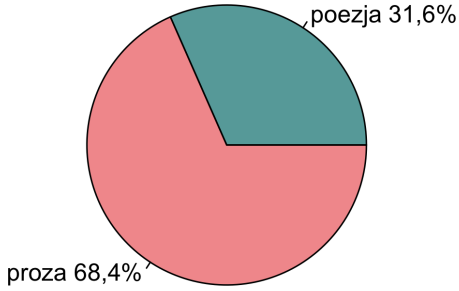
Data początkowa korpusu, tj. rok 1380, jest oczywiście czysto hipotetyczna. Najstarszym zabytkiem reprezentowanym w korpusie są *Kazania świętokrzyskie*, zachowane jedynie fragmentarycznie i trudne do precyzyjnego określenia daty ich powstania. Przyjęliśmy za *Opisem źródeł Słownika staropolskiego* przypuszczalne datowanie zabytku na rok 1380 (Twardzik, 2005: 93). Mimo braku anotacji strukturalnej i morfosyntaktycznej zbiór zabytków języka staropolskiego jest możliwie pełnym i bardzo cennym materiałem do badań nad historyczną polszczyzną; dla nas miał wartość przede wszystkim dlatego, że pozwalał na pełniejsze oświetlenie zmian zachodzących w XVI czy XVII wieku.

2.2.2. Wiek XVI

Wiek XVI jest reprezentowany przez 82 teksty: 49 poetyckich i 33 prozatorskie, które liczą razem 1 514 568 słów. Opracowane teksty, a szczególnie ich aspekt materialny, są możliwie dokładnie odzwierciedlone w formacie XML (a ściślej: w standardzie TEI P5). Część tekstów, w takim zakresie, w jakim pochodzą one z edycji, występuje w czystym formacie tekstowym, co wiąże się z brakiem anotacji strukturalnej, a zarazem stoi w opozycji do bogatej anotacji tekstów pochodzących z materiałów źródłowych do *Słownika polszczyzny XVI wieku*. W przypadku tego drugiego w anotacji zaznacza się np. literę, która stanowi na stronie inicjał, ligaturę, zaznacza się również przeniesienia, podział wierszy na stronie, podział stron itp. Korpus ten natomiast nie jest wyposażony w anotację morfologiczną ani lematyzację. Fragment przykładowych tekstów (*Rozmowa . . . około egzekucyjej* Stanisława Orzechowskiego oraz *Biblia Leopolity*) wchodzących w skład korpusu przytaczamy poniżej:

<milestone unit="matryca" n="K5457d"/>nie máſz wiáry przećiwko Pánu Bogu/ nie
<lb n="18"/>máſz vczćiwośći przećiwko Vrzędom/ nie máſz po=
<lb n="19"/>ſłuſzeńſtwá przećiwko Krolom/ prawdy/ wiáry/
<lb n="20"/>ſpráwiedliwośći miedzy námi nie máſz żadney/ peł=
<lb n="21"/>no wſzędźie niekokoiá y roſterku.

Páweł Apoſtoł piſze do ſiedmi kośćiołow/ bo oſmy liſt co do Zydow/ wiele ich kthorzy go nie kładą w poczet. <milestone unit="matryca" n="T307b"/>Tymotheuſſá y Tytuſá náucża/ Ffilemoná zá ſługą zbiegłym prośi. O ktorym ma zá to że lepiey milczeć/ niżli máło piſáć. <note place="margin" type="gloss"><foreign xml:lang="la" rend="italic">Actus Apoſtolorum</foreign></note>.



Rysunek 2.3. Udział poezji i prozy w XVII wieku (5 903 912 wyrazów).

Podobnie jak w przypadku tekstów piętnastowiecznych, tak i w tekstach z XVI wieku nie znajdziemy konsekwentnej ortografii i fleksji. Materiały z tego podkorpusu zachowują ponadto oryginalne kreskowania i podział wyrazów, które nastręczają pewnych trudności w analizie. Używane przez nas narzędzie normalizujące zapis zostało oryginalnie zaprojektowane do modyfikacji tekstów późniejszych, należało więc uzupełnić je serią wyrażeń regularnych służących zastąpieniu problematycznych dzieleń, kreskowań i pisowni łącznej/rozdzielnej ich współczesnymi odpowiednikami, usunięciu ligatur, specyficznych znaków itp. Konieczne było usunięcie bogatej anotacji, ponieważ jest ona lingwistycznie nierelewantna i utrudnia przeszukiwanie korpusu pod kątem wydobycia informacji istotnych dla analizy zmiany językowej[7]. Procedurę, której poddawaliśmy teksty, opisujemy niżej.

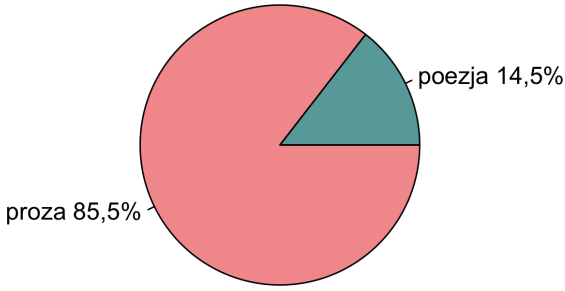
2.2.3. Wieki XVII i XVIII

W podkorpusie XVII wieku pracowaliśmy na zbiorze 5 903 912 wyrazów tekstowych, co przekłada się na 383 teksty: 264 prozatorskie i 119 wierszowanych. Podział ten zaprezentowany jest na Rys. 2.3.

W skład podkorpusu obejmującego wiek XVIII wchodzi 227 tekstów (194 prozatorskie, 33 wierszowane), co daje razem 3 400 817 wyrazów tekstowych. Podział gatunkowy prezentujemy na Rys. 2.4.

Teksty przygotowane w formacie TEI zawierają transkrypcję oraz oznaczenie wyrazów obcego pochodzenia, co znacząco ułatwia analizę. Każdy plik był wyposażony w nagłówek (ang. *header*), tzn. element (w rozumieniu elementu XML), który zawierał metadane tekstu lub tekstów w tym pliku zawartych. Nagłówek można porównać do rozbudowanej strony tytułowej: zawierał autora, tytuł, datę powstania tekstu, wydawcę, wydawnictwo, a także istotne dla właścicieli korpusu

[7] Ponieważ motywacją powstania tego korpusu było stworzenie cytatów dla *Słownika polszczyzny XVI wieku*, jego twórcy starali się możliwie dokładnie odwzorować typografię zabytku literackiego.



Rysunek 2.4. Udział poezji i prozy w XVIII wieku (3 400 817 wyrazów).

informacje o pochodzeniu pliku, prawach autorskich do niego i zakresie, w jakim możemy nim dysponować. Te ostatnie informacje nie są umieszczane w plikach dostępnych publicznie.

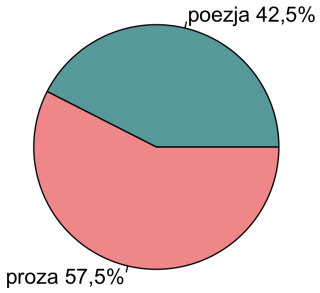
Na potrzeby badania musieliśmy ujednolicić grafię przy pomocy szeregu podmian (wyrażeń regularnych) w oryginalnych tekstach. Przykładowy fragment tekstu włączonego do naszych badań prezentuje się następująco:

> Niech, jako kto chce, cukruje i chwali tego teraźniejszego rokoszu złożenie i to tak nagłe i gwałtowne wnętrzuego pokoju <foreign xml:lang="la" xml:id="txt_2.1-foreign">et status</foreign> Rzpltej poruszenie; trudno to jednak pochwalić baczny może, bo ani przyczyny do tego tak wielkiej nie baczymy jeszcze, dopiero jej szukać się zdadzą i coś ludziom pokazując, deklamacyami raczej i nasadzonemi mowami, a kazaniom podobniejszymi, to coś rozszerzając i farbując słowy, a egzagerując niektóre wszem wiadome rzeczy, ludziom coś wielkiego ukazać usiłują.
> (*Zdanie szlachcica polskiego o rokoszu*)

2.2.4. Wiek XIX

To najbardziej zróżnicowany pod względem typów tekstów podkorpus spośród tych, na których prowadziliśmy badania. Zawiera on 73 teksty: 42 prozatorskie i 31 poetyckich, co daje łącznie 1 112 857 wyrazów tekstowych.

Zróżnicowanie korpusu wynika z kilku czynników. Pierwszym z nich jest wielość źródeł, z których pochodzą teksty. Część z nich umieszczona jest w zasobach *Biblioteki literatury polskiej w internecie*. Teksty pozyskane są z edycji, które ukazały się drukiem i zachowują znormalizowaną ortografię i fleksję. Podobnie jest z tekstami z *Wolnych Lektur*, choć tu w wielu wypadkach archaizmy zastępowane są wyrazami bardziej współczesnymi – oba te serwisy służą udostępnianiu literatury polskiej szerokiemu gronu odbiorców. Modernizujące zmiany tekstowe wynikają oczywiście z tego, że w tych serwisach istotny jest powszechny dostęp do treści, a nie badanie historii rozwoju języka. Zupełnie innym źródłem tekstów



Rysunek 2.5. Udział poezji i prozy w XIX wieku (1 112 857 wyrazów).

jest Polona. Teksty gromadzone przez ten serwis, zarówno w wersji obrazu, jak i w warstwie tekstowej, nie są modyfikowane. Serwis nie stara się dotrzeć do pierwszej edycji tekstów, ale można takowe odnaleźć i opracować na własne potrzeby. Jedyną przeszkodą jest skuteczność optycznego rozpoznawania tekstu – niektóre z zamieszczonych w Polonie tekstów były rozpoznawane przez starsze wersje oprogramowania OCR, co dawało trudne dziś do zaakceptowania rezultaty, np.:

> M e s m e r , k t ó r y p o d a ł n a u k ę i o d k ry c ie swoie
> pod r o z b ió r k ró le w s k ie m u to w a rz y s tw u lekar*
> e k iem u i a k ad em ii u m iejętności w P a r y ż u ; h a n
> ie b n ie z o s ta ł w y ś m ia n y od ty c h z g r o m a d z e ń
> uczonyeh. Nazwano go oszustem i kuglarzem,
> a magnetyzm niebezpiecznym śrzodkicm obłąkania
> umysłów słabych i przeciwnym rozszerzeniu
> oświaty i cyw iliza cji, którey wygórowanie
> daie nam do dziś dnia, okropny przykład nay -
> wyższego zepsucia moralnego, i iest tiieiako źrzódłem,
> z którego sit; wylały klęski na całą Europę.
> (Ignacy E. Lachnicki, *Krótka wiadomość o magnetyzmie zwierzęcym*)

W takim wypadku konieczna była ręczna korekta, przy czym niektóre błędy dawały się wyeliminować za pomocą wyrażeń regularnych; po części były to zmodyfikowane reguły, które służyły do normalizacji podkorpusów z wcześniejszych epok. Poprawione wersje tekstu należało poddać dalszej obróbce. Używając wyrażeń regularnych oraz ręcznie sczytując teksty, anotatorzy doprowadzali je do formy umożliwiającej analizę, konsekwentnie ujednolicając grafię (końcówki *-ey* zmieniając na *-ej*, *-yą* na *-ją*, usuwając pojawiające się w tekstach kreskowania, zmieniając źle rozpoznane znaki, np. *śrzodkicm* > *śrzodkiem* itp.).

2.3. Konstrukcja korpusu

2.3.1. Format tekstów i nazwy plików

Teksty wchodzące w skład korpusu przygotowane zostały w formacie TEI, będącym podzbiorem języka znaczników XML. Każdy plik zawiera jeden tekst oraz informację o metadanych, o których będzie jeszcze mowa poniżej. Ponieważ wiele z opisanych w niniejszej książce eksperymentów opierało się na obserwacji zwykłych form wyrazowych, dysponowaliśmy również uproszczoną wersją korpusu bez anotacji, a zatem i bez konieczności stosowania formatu TEI: w odpowiednich plikach znajdowały się po prostu tekstowe wersje (ang. *plain text*) poszczególnych utworów, nie zawierające żadnego formatowania. W wersji pełnej korpusu anotacja obejmowała oczywiście warstwę morfosyntaktyczną, ale oznakowana została też podstawowa struktura dokumentów i – jeśli to było możliwe – informacja o wtrętach obcojęzycznych.

Sprawą drobną, ale znacząco ułatwiającą dalszą pracę, było przyjęcie konsekwentnej konwencji nadawania nazw plikom. Nazwa zasadniczo zaczynała się od daty powstania tekstu, podkreślnika, nazwiska autora i początku tytułu oraz (tam gdzie zachodziła taka potrzeba) – numeru tomu. Na przykład plik zawierający kazanie Hiacynta Przetockiego z 1650 roku zatytułowane *Miłosierdzie Bogarodzice Maryjej nad Bractwem Szkaplerza Świętego pokazane* otrzymał nazwę `1650_Przetocki_Szkaplerzna.txt` dla tekstu nieanotowanego oraz `1650_Przetocki_Szkaplerzna.xml` dla wersji z anotacją morfosyntaktyczną. Konwencja ta pozwalała na wydobycie daty poświadczenia jakiejś formy bezpośrednio z nazwy pliku, co było wygodne, gdy korpus był przeszukiwany przez program konkordancyjny lub bezpośrednio za pomocą skryptów programistycznych – w naszym wypadku były to odpowiednio program AntConc (http://www.laurenceanthony.net/software.html) oraz środowisko programistyczne R (https://cran.r-project.org/), na ogół w połączeniu z biblioteką `stylo` przeznaczoną do manipulacji danymi tekstowymi (Eder, Rybicki i Kestemont, 2016).

2.3.2. Metadane

Każdy tekst umieszczony w korpusie opatrzony został metryczką z informacjami bibliograficznymi i genologicznymi, co ułatwia identyfikację tekstu oraz jego dokładne cytowanie. Opracowana dla tekstów metryczka zawiera:

– nazwę pliku,
– źródło, z którego został pozyskany,
– informację o prawach autorskich,
– tytuł,
– rok i miejsce wydania,

– datę powstania,
– rok i miejsce pierwszego wydania,
– nazwisko i lata życia autora,
– nazwisko tłumacza,
– przynależność gatunkową.

W dużej większości tekstów powyższe dane nie były trudne do ustalenia. Jednak nie we wszystkich tekstach udało się nam wypełnić metryczkę w całości. Podobnie jak przyjęło się w *Elektronicznym korpusie tekstów polskich z XVII i XVIII wieku* (Bronikowska i Przyborska-Szulc, 2018), utwory wydane na podstawie starodruków jako datę powstania otrzymują datę wydania druku, chyba że wydawca korzystał z późniejszych wersji tekstu. Teksty pochodzące z gazet ukazują się z datą dzienną i inicjałami autora artykułu lub datą i numerem wydania.

Klasyfikacja genologiczna tekstów zaczerpnięta została z NKJP (Przepiórkowski i in., 2012). Staraliśmy się, by mimo olbrzymiej nadreprezentacji tekstów literackich (poezji i prozy) w korpusie znalazły się okazy reprezentujące możliwie wiele innych podtypów piśmiennictwa: wiadomości prasowych, listów, przemów, tekstów użytkowych i poradnikowych, ale także, szczególnie w wypadku wcześniejszych epok – przykłady piśmiennictwa homiletycznego i prawniczego.

2.3.3. Normalizacja ortografii

Jednym z większych problemów, z jakimi spotykają się twórcy korpusów historycznych, jest brak znormalizowanej grafii. Jakkolwiek istnieją pytania, na które można odpowiedzieć tylko wtedy, gdy sięgnąć do oryginalnego zapisu, w większości wypadków badacze potrzebują ujednoliconego kształtu graficznego tekstu. Jednolita i uwspółcześniona grafia jest jednym z podstawowych warunków korzystania z dostępnych analizatorów morfologicznych (tagerów). Dla przypomnienia: zadaniem tego narzędzia jest przypisanie wszystkich możliwych interpretacji każdej z form fleksyjnych w tekście. Narzędzie to porównuje każdy kolejny wyraz napotkany w tekście ze słownikiem fleksyjnym, który w pewnym przybliżeniu można sprowadzić do listy form fleksyjnych przypisanych do danej formy słownikowej (podstawowej), a następnie przypisuje formę słownikową oraz wszystkie potencjalne interpretacje do wyrazu w tekście. I tak np. forma *psa* to potencjalnie dopełniacz i biernik liczby pojedynczej od słowa *pies*. Jest więc rzeczą oczywistą, że każde najdrobniejsze odejście od zapisu przewidzianego przez słownik morfologiczny powoduje, że wyraz pozostaje nierozpoznany. By tak się nie stało, należy albo dodać wariant graficzny do słownika morfologicznego, albo znormalizować grafię przed użyciem analizatora.

Grafia tekstów korpusu została znormalizowana przy pomocy narzędzia opracowanego na potrzeby korpusu barokowego przez Janusza S. Bienia oraz Dorotę

Komosińską z Instytutu Podstaw Informatyki PAN (zob. Bronikowska i Przyborska-Szulc, 2018). Narzędzie to zawiera serię kilkunastu tysięcy komplementarnych podmian grup liter w postaci wyrażeń regularnych, przygotowanych przez zespół opracowujący teksty barokowe, a wyszukujących możliwe zróżnicowania w zapisie wyrazów. Oboczności zamieniane są na jednorodne formy, zatem efektem działania takiego normalizatora będzie sprowadzenie form *posesyą*, *possesyą*, *posessyją* i *posesyją* do zuniformizowanej formy *posesją*.

Ponieważ narzędzie było dostosowane do pisowni siedemnastowiecznej, w naszym projekcie zostało ono zaadaptowane do potrzeb przetwarzania tekstów piętnasto- i szesnastowiecznych[8]. Adaptacja ta była przeprowadzana w wielu kolejnych iteracjach, z zastosowaniem próbnych anotacji za pomocą tagera. Po wprowadzeniu odpowiednich zmian efekty działania narzędzia były sprawdzane na nowo na tym samym materiale porównawczym.

Część tekstów pochodzi ze współczesnych edycji opracowywanych w różnych konwencjach, wersje zaś utworów przygotowane przez Polonę dostępne są zawsze w oryginalnej grafii – już to współczesnej, już to sprzed reformy ortograficznej – jako że ich celem jest jak najwierniejsze oddanie fizycznego kształtu dygitalizowanych zbiorów. Dla zaspokojenia części naszych potrzeb konieczne było ujednolicenie grafii w całym korpusie, wszystkie więc tego typu teksty zostały ręcznie zweryfikowane lub – w sytuacjach skrajnych – przepisane przez anotatorów.

2.3.4. Anotacja morfosyntaktyczna

Cały dostępny korpus – zestawiony z istniejących korpusów historycznych, uzupełniony setką tekstów pozyskanych z przeróżnych źródeł o różnej jakości, w tym tekstami transkrybowanymi specjalnie na potrzeby niniejszego projektu, przynajmniej częściowo zmodernizowanymi pod względem grafii – został również wzbogacony o warstwę anotacji morfosyntaktycznej. Znakowanie wykonano przy pomocy tagera TaKIPI 1.8 (Piasecki, 2007). Jak się nietrudno domyślić, skuteczność tagera była uzależniona od daty powstania anotowanego tekstu: im dawniejszy materiał językowy, tym oczywiście większa liczba błędnie rozpoznanych form. Pierwsze próby anotacji czyniliśmy, używając zupełnie niemodyfikowanej wersji tagera TaKIPI w zastosowaniu do tekstów piętnasto- i szesnastowiecznych. Zaskakująco dobre wyniki, czyli lematyzacja i przypisanie etykiety z kategorią gramatyczną formie wyrazowej w przypadku około 73–80% wyrazów, były zachętą do dalszych prac. Ulepszając badany zbiór, mogliśmy nie tylko zwiększyć skuteczność tagowania (a więc liczbę rozpoznanych form), ale i poprawić jego precyzję (liczbę form rozpoznanych prawidłowo).

[8] Zadanie to wykonali Michał Woźniak i Magdalena Król.

W celu zwiększenia skuteczności tagera teksty z danego podkorpusu najpierw pojedynczo modyfikowaliśmy, usuwając znaczniki XML, przeniesienia oraz znaki specjalne. Później teksty poddawaliśmy modernizacji, czyli stosowaliśmy globalny zbiór reguł opracowanych na potrzeby analizy tekstów siedemnasto- i osiemnastowiecznych. Pierwsze efekty takiej modernizacji nie były satysfakcjonujące – reguły nie uwzględniały zmian zbitek literowych właściwych dla tekstów pochodzących z XVI wieku, np.:

– *fl* > *sł* w kontekście innym niż po *je*,
– *fl* > *śl* po *je*,
– *ffc* > *szcz* w nagłosie i śródgłosie wyrazu,
– *rth* > *rt* w dowolnym miejscu wyrazu,
– *kth* > *kt* w dowolnym miejscu wyrazu,
– *ont* > *ąd* w wygłosie wyrazu,
– *peun* / *peln* / *pełn* > *pełen*,
– *wzw* > *wyzw* / *wezw* w nagłosie wyrazu.

Po wielu próbach i dodaniu nowych reguł do skryptu modernizującego grafię zaczęliśmy uzyskiwać wyniki co najmniej satysfakcjonujące. Na przykład z następującego tekstu wejściowego pochodzącego z Biblii Leopolity:

> Apollonius on/ choćże Mędrzec iáko poſpolity człowiek mowi/ choćże też Filozoph/ iáko Pytágorykowie podáią/ wſſedł do Perſkiey źiemie/ ſchodził gorę co ią Kaukazem zową/ ſchodził Albány/ Tátáry/ Máſſágety/ przebogáte one Indiyſkie Kroleſtwá przeſzedł á na koniec/ przebywſſy bárzo ſſeroką rzekę Ffizon/ záſſedł áż do Bráchmánow/ áby iedno Hiárchi Mędrcá ná ſtolicy złotey ſiedzącego/ y nápijáiącego ſie z ſtudnice Tántálu ſowey/ miedzy nie wielą dyſcypułow/ o Náturze/ o biegániu gwiazd/ o odmienie dniow/ roſpráwiáiącego/ ſluchał.

udało się nam osiągnąć postać:

> Apollonius on choćże Mędrzec jako pospolity człowiek mówi choćże też Filozof jako Pytagórykowie podają wszedł do Perskiej ziemi schodził górę co ją Kaukazem zwą schodził Albany Tatary Maszagety przebogate one Indyjskie królestwa przeszedł a na koniec przebywszy bardzo szeroką rzekę Fizon zaszedł aż do Brachmanów aby jedno Hiarchi Mędrca na stolicy złotej siedzącego i napijającego się z studnice Tantalu z owej między nie wielą dyscypołów o Naturze o bieganiu gwiazd o odmianie dni rozprawiającego słuchał.

Tak przygotowany tekst można było następnie znakować przy użyciu tagera TaKIPI, uzyskując skuteczność tagowania na poziomie 87%. By polepszyć dokładność, a więc szanse poprawnego przypisania etykietki morfosyntaktycznej

(czyli ciągu kodów oznaczających kategorie fleksyjne) do analizowanego wyrazu, spróbowaliśmy rozbudować słownik, na którym bazuje tager. Uzupełnianie słownika analizatora morfologicznego przebiegało następująco: teksty normalizowane poddawaliśmy anotacji morfosyntaktycznej z wyłączonym tzw. guesserem. Guesser to moduł analizatora morfologicznego, który na podstawie pewnych reguł ustala postać lematu („zgaduje", stąd nazwa) oraz możliwe interpretacje wyrazu nieobecnego w słowniku morfologicznym. Jest to funkcja w wielu wypadkach bardzo użyteczna, jakkolwiek należy pamiętać, że co prawda zwiększa ona liczbę rozpoznanych form, ale też wiele form rozpoznaje błędnie. W naszym wypadku istotne jest jednak to, że wszystkie nieobecne w analizatorze formy fleksyjne były oznaczane etykietką *ign*. Pozwalało to stworzyć dla każdego przebiegu programu listę słów, o które słownik analizatora należało uzupełnić. Dotyczy to także form fleksyjnych wyrazów notowanych wprawdzie w analizatorze, ale w formach odmiennych od współczesnych. Poddanie anotacji kolejnych tekstów pozwalało nie tylko uzupełnić analizator, ale także kontrolować, na ile z każdym kolejnym tekstem ubywa słów nieznanych.

Nierozpoznane formy opisywaliśmy, posiłkując się słownikami polszczyzny historycznej i rekonstruując niepoświadczone paradygmaty na podstawie opisanych istniejących form. Przygotowana w ten sposób lista zawierała 3572 formy nieuwzględniane w słowniku TaKIPI.

Teksty dziewiętnastowieczne poddawaliśmy anotacji przy użyciu analizatora morfologicznego *f19*. Dla porównania znormalizowaliśmy je również przy użyciu skryptu normalizującego, by później oznakować je tagerem TaKIPI. Okazało się, że wyniki tych dwóch odmiennych strategii rozpoznawania form w zasadzie nie różnią się, zatem do badań włączyliśmy teksty anotowane wyłącznie przy użyciu skryptu normalizacyjnego i tagera TaKIPI. Normalizacji poddaliśmy również teksty z *Biblioteki literatury polskiej w internecie*.

Rozdział 3

Dynamika zmian językowych

3.1. Wprowadzenie

Dla studenta uczącego się gramatyki historycznej na ogół najistotniejszy jest sam fakt zmiany językowej; może jeszcze stara się on zapamiętać chronologię względną zmiany. Pytanie o jej dynamikę zwykle mu umyka, tym bardziej że o dynamice zmian w epoce przedpiśmiennej możemy powiedzieć bardzo niewiele. Tymczasem pytanie o to, jak szybko zmiana językowa zachodzi, jest pytaniem niezwykle zajmującym, samym w sobie wartym badania. Czy zmiana zachodzi w ciągu jednego pokolenia czy też w dłuższym czasie? Jak szybko konkretna zmiana zachodzi w porównaniu z innymi? Na ile czas trwania danej zmiany nakłada się na czas trwania innych zmian? Oczywiście tego typu pytania były już zadawane przez historyków języka w przeszłości, i to wiele razy. Jednak precyzyjne modelowanie przebiegu zmiany nie było dotychczas możliwe ze względu na trudności w pozyskaniu materiału badawczego. W niniejszym rozdziale będziemy starali się pokazać, że rygorystyczny opis dynamiki zmiany wymaga również zastosowania aparatu matematycznego.

Jedną z takich właśnie prób modelowania dynamiki zmiany jest tzw. prawo Piotrowskiego. Radziecki lingwista o polskich korzeniach Rajmund Piotrowski (1922–2009) zauważył, że niektóre zmiany językowe nie mają przebiegu jednostajnie liniowego, lecz zaczynają się łagodnie i następnie przyspieszają, by znów zwolnić pod koniec dokonywania się zmiany. By wyjaśnić matematyczną ideę stojącą za powyższą obserwacją, warto najpierw rozważyć dwa inne, znacznie prostsze scenariusze. Pierwszy oznacza brak zmiany: gdyby wyobrazić sobie jakąś (dowolną) cechę językową, która w kolejnych stuleciach występuje z idealnie tą samą frekwencją, moglibyśmy tę cechę przedstawić na wykresie obrazującym frekwencję (oś y) w funkcji czasu (oś x) jako linię prostą przechodzącą przez wszystkie punkty czasu dokładnie horyzontalnie. Drugi hipotetyczny scenariusz to sytuacja, w której jakaś inna hipotetyczna cecha językowa nieustannie się wycofuje z systemu. Gdybyśmy zaznaczyli frekwencje tej cechy w kolejnych punktach czasu, wtedy zaobserwowalibyśmy linię prostą nachyloną pod jakimś kątem w stosunku do osi x i w pewnym momencie tę oś przecinającą

(byłby to moment całkowitego zaniku badanej cechy). Tak zwane prawo Piotrowskiego zakłada, że zmiana nie ma charakteru jednostajnie liniowego. Gdyby w takiej sytuacji w równych odstępach czasu zanotować stopień przebiegu zmiany i przedstawić w postaci punktów na wykresie w funkcji czasu (tzn. zmierzyć, jaki w danym momencie jest stosunek procentowy formy innowacyjnej do formy recesywnej na podstawie danych z korpusu), oczom obserwatora ukazałaby się utworzona przez te punkty linia krzywa, przebiegająca po wykresie trajektorią przypominającą literę *s* (por. Rys. 3.1 i następne). Piotrowski zaproponował model matematyczny opisujący tego typu krzywą (Piotrowskaja i Piotrowski, 1974) i wyprowadził odpowiednie równanie:

$$p(t) = \frac{1}{\pi} \arctan \mu(t - t_1) + \frac{1}{2}$$

Jako że w poniższych rozważaniach będziemy się odwoływać do zmodyfikowanej wersji oryginalnego równania Piotrowskiego, nie będziemy w tym miejscu zamęczać czytelników wyjaśnieniem wszystkich jego matematycznych własności. Zamiast tego chcielibyśmy przybliżyć założenia teoretyczne, jakie za nim stoją. Równanie to znalazło bowiem wcześniej zastosowanie w biologii. Żeby przybliżyć wycinek rzeczywistości, który można nim opisać, wyobraźmy sobie kubek mleka, do którego dostały się bakterie *lactobacillus bulgaricus* – zmieniają one świeże mleko w jogurt. Proces zasiedlania kubka będzie zawsze wyglądał tak samo: z początku w kubku znajduje się bardzo niewiele bakterii, może nawet tylko jedna, gdyż do rozpoczęcia procesu namnażania wystarczy pojedynczy organizm. Bakteria dzieli się, więc po chwili kubek zasiedlają już dwie, następnie cztery, osiem itd. Wraz ze wzrostem liczby bakterii wzrasta szybkość, z jaką ich przybywa. Dzieje się tak z przyczyn oczywistych: skoro każda bakteria dzieli się na dwa organizmy potomne, ich liczba w każdym kolejnym pokoleniu podwaja się w stosunku do poprzedniego – wystarczy przypomnieć sobie kolejne potęgi liczby 2, by zdać sobie sprawę, jak gwałtownie proces namnażania będzie przyspieszał. W końcu jednak zaczyna brakować pożywienia i dynamika przyrostu bakterii spada, choć sama liczba tych organizmów wciąż powoli przyrasta. Dochodzimy jednak do momentu, gdy kubkiem mleka nie jest w stanie pożywić się już więcej drobnoustrojów i ich liczba się stabilizuje. *Mutatis mutandis* podobnie dzieje się z zasiedlaniem jakiegoś obszaru przez większe zwierzęta czy ludzi, a także z rozprzestrzenianiem się choroby zakaźnej. Im więcej osób zakażonych, tym większa szansa spotkania chorego, z którym kontakt doprowadzi do zarażenia. Znów jednak, gdy na danym obszarze nie będzie już więcej osób podatnych na chorobę zakaźną, liczba kolejnych zarażonych też nie będzie wzrastać.

W ostatnim przywołanym wypadku podobieństwo do zmiany językowej jest naprawdę spore – zauważmy, że zmianę językową można „odziedziczyć", można się też nią „zakazić". Pierwszy wypadek to taki, gdy innowacja językowa jest używana przez starsze pokolenie i nabywana w procesie akwizycji mowy; drugi

to taki, kiedy użytkownik języka zastępuje w swoim idiolekcie formę recesywną formą innowacyjną pod wpływem ekspozycji na tę ostatnią. Istotną cechą zmiany językowej jest również to, że ma ona naturalny kres, którym jest całkowite wyparcie formy recesywnej[1].

Nieco na marginesie można dodać, że prawo Piotrowskiego opisuje również inny ważny wycinek językoznawstwa historycznego, mianowicie dynamikę zapożyczeń leksykalnych z jednego języka do drugiego. Znów – z początku kontakty pomiędzy dwoma językami są ograniczone i liczba zapożyczeń nie jest duża. Siła oddziaływania języka dawcy trwa jednak przez dziesięciolecia, jeśli nie stulecia. Przyjmując, że raz zapożyczone słowo wchodzi na trwałe do słownika, możemy stwierdzić, że liczba zapożyczeń wciąż się zwiększa. W końcu jednak zmienione warunki polityczne, społeczne lub kulturowe sprawiają, że atrakcyjność języka-dawcy spada i następuje wysycenie słownictwa języka-biorcy, zaś nowe zapożyczenia, jeśli występują, mają inne źródło. Dobrym przykładem takiego zjawiska są zapożyczenia z języka tureckiego w polszczyźnie (Stachowski, 2013).

Powiedzmy tu od razu, że dla językoznawcy zajmującego się diachronią (szczególnie okresu piśmiennego) nie ma niczego zaskakującego w stwierdzeniu, że zmiana językowa nie jest procesem skokowym, ale rozciągniętym w czasie, w którym to forma innowacyjna powoli i stopniowo wypiera formę recesywną i obie przez jakiś czas w języku koegzystują. Językoznawca wie jednak jeszcze jedno: wspólnota językowa nie jest tak homogeniczna jak wspomniane wyżej mleko w kubku. Nie jest nawet tak homogeniczna, jak przyjmujemy, dokonując opisu języka. Zmiana językowa zaczyna się w określonym miejscu i w określonej warstwie społecznej, komunikującej się we właściwy sobie sposób. Z kolei przestrzeń geograficzna i społeczna stanowią barierę, która zmianę spowalnia, stąd próby modelowania tempa zmiany bez uwzględniania tych barier zawsze będą obarczone pewnym błędem. Idealny korpus powinien być bardzo starannie zrównoważony pod względem zarówno socjolingwistycznym, jak i geograficznym, ale oczywiście – jak to opisaliśmy w rozdziale poprzednim – takiego korpusu nie ma, a już na pewno nie należy się go spodziewać dla epok dawniejszych. Należy jednak pamiętać, że w ostatecznym rozrachunku fakt, że dany tekst znalazł się w korpusie bądź nie, jest spowodowany w dużej mierze czynnikami przypadkowymi, tymczasem kilka tekstów reprezentujących obszar bądź socjolekt czy też rejestr, w których zmiana jest nadreprezentowana lub niedoreprezentowana w stosunku do całej wspólnoty językowej, może powodować nieoczekiwany obraz przebiegu zmiany, np. jej cofnięcie w stosunku do lat poprzednich. Nie powinno to jednak w żadnym razie być argumentem, by odrzucić podejście kwantytatywne,

[1] Dodajmy jednak, że refleksy zmiany, która już nastąpiła, mogą się pojawiać w tekstach długo po ustaniu procesu modyfikacji językowej – autorzy mogą np. cytować teksty dawne lub świadomie używać archaizmów.

ale raczej, by przyglądać się dokładniej tym wypadkom, kiedy fakty są wyraźnie różne niż ich przebieg teoretyczny, czyli wyabstrahowany model.

I tu dochodzimy do istoty rzeczy: swoistą wartością dodaną tzw. prawa Piotrowskiego jest fakt, że przebieg procesu diachronicznego możemy wyidealizować, czyli przedstawić w postaci abstrakcyjnego, niezaburzonego przez wyjątki czystego przebiegu zmian językowych. Za pomocą takiego idealnego modelu możemy wygenerować tzw. wartości oczekiwane i zestawić je z wartościami obserwowanymi, czyli danymi pozyskanymi z korpusu. Jeśli rzeczywisty[2] przebieg procesu diachronicznego nie odbiega znacząco od najlepszego teoretycznego modelu – lub też, bardziej precyzyjnie: jeśli dobrany model nie odbiega od danych empirycznych – to wolno zasadnie wnioskować, że zbliżyliśmy się do poznania badanej zmiany językowej.

Oczywiście warto zadać pytanie, czy tego rodzaju modelowanie może cokolwiek wnieść do językoznawstwa diachronicznego ponad to, co już wiemy. W wypadku pojedynczej zmiany zapewne nie, natomiast gdy zachodzi wiele zmian nakładających się w czasie na siebie, pozwala to zaobserwować ich odmienną dynamikę, odmienny punkt startu i całkowitego wyparcia formy recesywnej.

Leopold (2005) stwierdza, że jest niewiele prac, które konfrontują idealizację zaproponowaną przez Piotrowskiego z danymi empirycznymi. Jedną z pierwszych tego rodzaju prac jest rozprawa Besta (1983). Pozycję tę warto omówić szerzej nie tylko dlatego, że wytycza ona pewien szlak w opisywanych badaniach, ale również ze względu na to, że jest przykładem pracy z korpusem niezdygitalizowanym. Autor opisuje mocno rozciągniętą w czasie zmianę w historii języka niemieckiego, mianowicie przejście *ward* w *wurde*. Na opisywany proces składają się w istocie dwie zmiany, przejście *u* > *a* oraz epenteza *-e*. Obie mogły działać niezależnie, dając formy *wurd* i *warde*, choć w końcu razem doprowadziły do jedynie dopuszczalnej obecnie formy *wurde*. Obie też są rozpatrywane odrębnie. Autor starannie dobiera swój korpus, acz przyznaje, że podstawą włączenia bądź niewłączenia poszczególnych tekstów było przede wszystkim to, czy były dostępne w lokalnej bibliotece uniwersyteckiej. Tym niemniej autor zadbał o zrównoważenie czasowe i terytorialne, choć to ostatnie nie zawsze dawało się uzyskać. Badaniu poddano teksty z lat 1430–1939, z tym że w tekstach dwudziestowiecznych zasadniczo można mówić jedynie o archaizacji. Każdy podkorpus obejmował 30 lat i był reprezentowany przez 1000–1200 wystąpień, z wyjątkiem pierwszego (716 wystąpień), co dało w sumie 20 189 poświadczeń. Trzeba od razu zaznaczyć, że są to liczby relatywnie wysokie, przynajmniej w porównaniu do naszego materiału, przedstawionego poniżej. Z drugiej jednak strony formy *warde* i *wurd* (a więc

[2] Zasadniczo lepiej w tym wypadku mówić o przebiegu empirycznym niż rzeczywistym, bo przecież przebieg znamy jedynie z tekstów, które znalazły się w korpusie. Z oczywistych względów dane, które pochodzą z tych tekstów, niekoniecznie precyzyjnie odzwierciedlają rzeczywisty przebieg zmiany, co jest zresztą generalnym problemem badania przeszłości, także przeszłości organizmów żywych.

epenteza bez zmiany samogłoski tematycznej i zmiana samogłoski tematycznej bez epentezy) są reprezentowane w najlepszym wypadku przez trochę ponad 100 wystąpień (wartość maksymalna to 122). Formy te są więc rzadkie i nie pozwalają na wyciąganie pewniejszych wniosków, dlatego nie były poddane osobnemu badaniu, a jedynie powiększały liczbę elementów zbioru form z epentezą i zmianą samogłoski tematycznej. Warto też zauważyć, że badany przez Besta proces jest znacznie rozciągnięty w czasie – materiał wykorzystany w eksperymencie obejmuje 500 lat historii rozwoju języka.

W przypadku przytoczonego studium wartości obserwowane są zaskakująco bliskie wartościom oczekiwanym. Przede wszystkim każdy kolejny (w sensie chronologicznym) podkorpus zawiera wyższy udział formy innowacyjnej niż poprzedzający go. Jedynym wyjątkiem jest czwarty podkorpus, 1506–1535, gdzie udział form epentetycznych jest niższy niż w poprzedzającym podkorpusie. Generalnie przebieg zmiany samogłoski tematycznej jest nieco bliższy wartościom oczekiwanym, niż ma to miejsce w odniesieniu do epentezy, tym niemniej oba procesy stanowią modelowe przebiegi Piotrowskiego.

W badaniach danych diachronicznych wielokrotnie odwoływano się do badań wykonanych przez Ellegårda (1953). Podjął on jedną z pierwszych prób kwantytatywnego opisu zmiany językowej, polegającej na wprowadzeniu we współczesnym języku angielskim tzw. peryfrastycznego *do*, które jest obligatoryjne w zdaniach pytajnych i przeczących, por. następujące przykłady z British National Corpus:

Why do you look on me?

wobec

**Why you look on me?*

Tymczasem jeszcze w tekstach z przełomu XV i XVI wieku spotykamy zdania bez *do*.

Dolores Mortis not touched hym or pynched hym. . .

Należy tu dodać, że peryfrastyczne *do* może się pojawiać w zdaniach twierdzących, jak współcześnie:

You do like Armagnac, I hope?

Najstarszy przykład tej konstrukcji, który zanotował Ellegård, pochodzi z epoki średnioangielskiej, dokładniej z XIII wieku. Od tego momentu udział zdań przeczących i pytajnych zawierających tę konstrukcję wzrastał, tak by stała się ona obligatoryjna w XVIII wieku.

By zbadać kwantytatywnie przebieg tej zmiany, Ellegård zgromadził ponad 15 000 przykładów ze 107 tekstów. Badacz zasadniczo przyjmował, że podkorpus obejmować powinien 25 lat, czynił jednak pewne odstępstwa (w zbiorze danych są

dwa podkorpusy o rozmiarze 10 lat, są też i pięćdziesięcioletnie). Należy jednak dodać, że liczba zdań twierdzących bez *do* jest oparta na ekstrapolacji liczby tych zdań na losowo wybranych 10 stronach z każdego tekstu. W innym wypadku wystąpień byłoby za dużo, by je precyzyjnie policzyć.

Pracę Ellegårda cytowano wielokrotnie, rewidowano też pewne jej ustalenia, np. Rydén zwrócił uwagę na to, że proces wprowadzający obligatoryjne wystąpienie *do* we frazie zaprzeczonej lub pytajnej zakończył się ostatecznie dopiero w XIX wieku, a nie – jak chciał Ellegård – w wieku XVIII (Rydén, 1979). Nas wszakże najbardziej interesuje to, że dane zebrane przez Ellegårda służyły za punkt wyjścia do badań kwantytatywnych, a dla wielu badaczy były także inspiracją do modelowania przebiegu zmiany.

Od czasu publikacji równania Piotrowskiego (1974) prace nad matematycznym opisem zmiany językowej przyspieszyły. Próby takie podejmowali chociażby Kroch (1989), Ogura (1993), Vulanović i Baayen (2007) oraz Vulanović (2007). Równanie zaproponowane pierwotnie przez Piotrowskiego jest jednak obarczone tą zasadniczą wadą, że nie pozwala na dopasowanie teoretycznego modelu do danych empirycznych. Gabriel Altmann, słowacki matematyk o zainteresowaniach lingwistycznych, zaproponował więc równanie, które jest od tej wady wolne (Altmann, 1983):

$$p(t) = \frac{1}{1 + a \times e^{-rt}}$$

W powyższym równaniu parametr a jest arbitralnie dobraną liczbą naturalną (zmiana tego parametru wpływa na przesunięcie wykresu na osi czasu), t oznacza czas, przy czym $t = 0$ dokładnie w miejscu, w którym proporcja formy innowacyjnej do recesywnej wynosi 0,5. Parametr r natomiast należy dobrać eksperymentalnie, tak by osiągnąć największe możliwe dopasowanie modelu do danych empirycznych, e zaś jest podstawą logarytmu naturalnego, w przybliżeniu wynoszącą 2,7182. Oczywistą wadą tego równania jest fakt, że użytym parametrom bardzo trudno dać jakąś sensowną wykładnię lingwistyczną. Jeśli na przykład okaże się, że badana zmiana zakończyła się przy wartości $t = 4$ (por. np. Köhler, 2015: 112), nic nam to nie mówi o rzeczywistym przebiegu tej zmiany.

Czytelnik zaznajomiony z metodologią statystyczną od razu zauważy, że powyższe równanie Altmana przypomina tzw. model regresji logistycznej, doskonale znanej i od dziesięcioleci stosowanej metody modelowania zmian fazowych. I rzeczywiście, w literaturze pojawiają się wzmianki o tym, że model wprowadzony przez Piotrowskiego, a później Altmana, jest w istocie wariantem regresji logistycznej (Köhler, 2015; Vulanović, 2007; Vulanović i Baayen, 2007). Założenie teoretyczne w obu podejściach jest bardzo podobne, przy czym regresja logistyczna ma bardzo intuicyjny przebieg i daje się łatwo objaśnić w kategoriach probabilistycznych. Można mianowicie za jej pomocą obliczyć, jaka jest szansa znalezienia formy innowacyjnej bądź recesywnej w danym roku. Na ogół

szansa natrafienia na formę innowacyjną w początkowych podkorpusach wynosi 0 (lub dąży do 0), w końcowych podkorpusach zbliża się do wartości 1 lub tę wartość osiąga, mniej więcej w połowie pokazując gwałtowny przyrost prawdopodobieństwa – jeśli przebieg zmiany tego prawdopodobieństwa przedstawimy na wykresie, otrzymamy kształt łudząco podobny do krzywej zaproponowanej przez Piotrowskiego. Klasyczny model regresji logistycznej można przedstawić w następujący sposób (James i in., 2013: 132):

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

gdzie oba parametry β_0 i β_1 wyznaczają przecięcie krzywej i jej nachylenie, e jest podstawą logarytmu naturalnego. Przy niewielkim przekształceniu algebraicznym powyższego wzoru możemy dostać jego nieco prostszą wersję, która – pomijając inne oznaczenia zmiennych – jest niemal dokładnym odwzorowaniem przytoczonego powyżej równania Altmana:

$$p(X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}$$

Nie tylko sama krzywa logistyczna, ale i sposoby znajdowania jej obu parametrów są dobrze zbadane i opisane w literaturze przedmiotu (zob. np. Hastie, Friedman i Tibshirani, 2001). Dość powiedzieć, że za pomocą klasycznej metody zwanej estymowaniem największej wiarygodności (ang. *maximum likelihood estimation*) znajduje się takie wartości obu parametrów β_0 i β_1, by minimalizować różnicę między wartościami teoretycznymi i wartościami obserwowanymi (w naszym wypadku: danymi językowymi).

Biorąc pod uwagę powszechność stosowania regresji logistycznej w naukach przyrodniczych i ścisłych, a także łatwość oszacowania parametrów modelu za pomocą wielu narzędzi statystycznych (od arkuszy kalkulacyjnych począwszy, a na programach matematycznych typu Matlab czy Statistica skończywszy), przyjęliśmy, że w naszych badaniach będziemy posługiwać się klasyczną regresją logistyczną z jej estymacją parametrów β_0 i β_1, zostawiając równanie Piotrowskiego na boku i traktując je z należnym szacunkiem jako ważny kamień milowy w historii językoznawstwa kwantytatywnego.

W tym miejscu trzeba poczynić ważne zastrzeżenie. Otóż optymalnie dobrany model spośród wszystkich dostępnych modeli może być nadal modelem złym, niepasującym do danych empirycznych. Dzieje się tak dlatego, że w danych czasem po prostu nie ma żadnych istotnych korelacji, zawsze jednak istnieje jakiś zestaw parametrów β_0 i β_1, który jest lepszy od pozostałych. By zatem uniknąć pułapki złego dopasowania modelu do danych, w statystyce stosuje się dodatkowe procedury oszacowania stopnia pokrycia wartości teoretycznych z wartościami obserwowanymi. W najprostszym rodzaju regresji, tj. w regresji liniowej, od niemal stu lat stosuje się klasyczną i elegancką miarę R^2, która opiera się na obliczaniu

różnicy między poszczególnymi obserwacjami i ich wartościami teoretycznymi (owe różnice są podnoszone do kwadratu i następnie sumowane, a później dzielone przez wariancję wartości obserwowanych). Miara R^2 ma tę zaletę, że bardzo łatwo interpretować jej wskazania: wartość 0 oznacza zupełną nieadekwatność modelu, wartość 1 zaś idealne zlanie się danych empirycznych z wygenerowanymi przez model. Wykazano jednak, że owa miara nie jest wiarygodna w regresji logistycznej, dlatego w naszych poniższych rozważaniach będziemy się posługiwali wariantem tej miary zaproponowanym przez Nagelkerkego, odpornym na nieliniowy kształt krzywej logistycznej (Nagelkerke, 1991):

$$\bar{R}^2 = \frac{R^2}{\max(R^2)}$$

Ilekroć będziemy poniżej powoływać się na miarę dopasowania modelu R^2, będziemy mieli na myśli właśnie jej zmodyfikowaną wersję Nagelkerkego. Mimo że standardowa miara R^2 na ogół zwraca (nieco) wyższe wartości, są to często wartości zafałszowane. To kolejna warta odnotowania różnica między procedurą stosowaną w niniejszej książce i w pracach naszych poprzedników. Do estymowania parametrów modelu logistycznego będziemy posługiwać się środowiskiem programistycznym R i jego wbudowaną funkcją `glm()`, wartość zaś współczynnika determinacji R^2 będziemy obliczać za pomocą funkcji `NagelkerkeR2()` z biblioteki `fmsb`.

3.2. Zmiany

W niniejszym rozdziale przyjrzymy się kilku zmianom, które zaszły w ciągu wieków w polszczyźnie. Spróbujemy znaleźć optymalne parametry modelu logistycznego dla każdej z nich, by następnie oszacować stopień dopasowania najlepszego modelu do danych rzeczywiście pozyskanych z korpusu. Będą to następujące zmiany:

– zmiana grupy *-ir-* > *-er-*,
– zmiana wykładnika stopnia wyższego *na-* > *naj-*,
– formy czasu przeszłego *-bych* > *-bym*, *-bychmy* > *-byśmy*, a także trybu przypuszczającego *-łech* > *-łem*,
– *abo* > *albo*,
– *barzo* > *bardzo*,
– *więtszy* > *większy*,
– *wszytek* > *wszystek*.

Pierwsza z powyższych zmian ma charakter fonetyczny, dwie kolejne to zmiany w zakresie morfologii, ostatnie cztery to zmiany izolowane. Rodzi się w tym miejscu pytanie o dobór tych właśnie, a nie innych zmian. Biorąc pod uwagę

gramatykę, można zasadnie twierdzić, że ważniejszymi od powyższego zestawienia procesami będą chociażby formowanie się współczesnych nam rodzaju gramatycznego, liczby czy przypadka. Naszym założeniem było jednak objęcie badaniem możliwie wielu zjawisk, a to z kolei spowodowało, że ograniczyliśmy się do takich, które można wychwycić z całego wielomilionowego korpusu automatycznie lub posiłkując się dodatkową selekcją ręczną przeprowadzoną na niewielką skalę.

Nasze badania mają podwójny cel. Po pierwsze chcemy stwierdzić, na ile średniopolskie fakty językowe dają się opisać modelem regresji logistycznej. Drugi cel jest z naszego punktu widzenia istotniejszy i sprowadza się do nowego spojrzenia na dynamikę zmian w średniopolszczyźnie. Opisywane powyżej dane chcemy bowiem obejrzeć z lotu ptaka, nie patrząc na każdą ze zmian odrębnie, ale na wszystkie jednocześnie. Oczywiście jeśli mowa o spojrzeniu z lotu ptaka, to nie mamy na myśli oglądu systemu jako całości, ponieważ interesujące nas zmiany nie wpływają na całościową metamorfozę systemu językowego – wskutek ich zachodzenia nie zniknęła ani nie pojawiła się żadna kategoria gramatyczna, żadna też ze zmian nie wymusiła innej[3]. Zmiany, które obserwujemy, są umiejscowione w różnych fragmentach systemu gramatycznego albo wręcz izolowane. Jeżeli więc mówimy o szerszym spojrzeniu, to raczej chodzi nam o dystrybucję form recesywnych i innowacyjnych w tekstach obejmujących duży przedział czasowy, przede wszystkim zaś mamy na myśli skonfrontowanie ze sobą przebiegów wszystkich tych zmian jednocześnie.

3.3. Dane

Dane zebraliśmy z korpusu opisanego w poprzednim rozdziale, z tym że wyłączyliśmy z niego teksty pochodzące z wydań silnie modernizowanych, w których ingerencja w tkankę tekstu dotyczyła nie tylko grafii, lecz także końcówek fleksyjnych i dawnych grup spółgłoskowych. Ponieważ niektóre ze zmian zaczynają się jeszcze w epoce staropolskiej, w opisywanych badaniach sięgaliśmy również do tekstów najdawniejszych.

Procedura przygotowania danych polegała na podziale korpusu na szereg „plasterków", czyli chronologicznie uporządkowanych podkorpusów obejmujących takie same odcinki czasu, np. dwudziestoletnie. Przypomnieć w tym miejscu wypada, że prześledzenie jakiegokolwiek procesu historycznego zawsze będzie wymagało dokonania podziału danych na mniejsze podokresy: czasem takim podokresem będzie jeden rok, czasem całe stulecie. Obserwacje Bajerowej były na przykład oparte na czterech podokresach obejmujących 25 lat każdy, co pozwalało

[3] Przykładem systemowej zmiany z epoki średniopolskiej jest ewolucja kategorii rodzaju gramatycznego.

uchwycić cztery uśrednione punkty na osi czasu (Bajerowa, 1964); podobnie było w wypadku innych prac cytowanych powyżej.

Tu jednak spotykają się dwa sprzeczne wymagania stawiane analizie danych. Z jednej strony chcielibyśmy, by punktów, dla których zbieramy dane, było jak najwięcej. Oznacza to dużą liczbę małych podkorpusów, obejmujących możliwie krótki okres, np. dziesięciolecie. Tak małe podzbiory mają jednak zasadniczą wadę – siłą rzeczy zawierają bardzo niewiele tekstów, a więc ich reprezentatywność jest mocno ograniczona. Mało tego, w wypadku rzadkich form niektórym punktom na osi czasu nie będą odpowiadały żadne dane (tj. brak dla danego okresu jakiegokolwiek poświadczenia). Problem łatwo sobie uzmysłowić przez wykonanie pewnego prostego eksperymentu myślowego: załóżmy, że nasz korpus diachroniczny podzieliliśmy na podkorpusy obejmujące pojedyncze dni – tak jak podzielone są współczesne korpusy obejmujące prasę codzienną. Dla lat 1550–1850 dostaniemy w ten sposób nieco ponad 100 tysięcy jednodniowych podkorpusów. Rzecz jednak w tym, że zaledwie 1% z tych podkorpusów będzie zawierał jakiekolwiek dane – np. między rokiem 1400 i 1449 (przypuszczalne daty powstania odpowiednio *Psałterza floriańskiego* i *Kodeksu Suleda*) dostaniemy 14 700 podkorpusów niewypełnionych żadnymi danymi. Żadne wiarygodne obserwacje statystyczne nie będą w takim wypadku możliwe.

Z drugiej strony kuszące jest zwiększenie badanych podokresów czasowych. Wprawdzie będzie to oznaczało niewiele uzyskanych punktów na osi czasu (gdyż ten sam obserwowany okres dzielimy na mniejszą liczbę podkorpusów), ale za to z dość dobrze uśrednionym wynikiem, bo opartym nie tylko na większej liczbie tekstów, ale też obejmującym większą liczbę lat. Kolejny eksperyment myślowy uzmysłowi jednak niebezpieczeństwo, jakie kryje się za taką strategią. Wyobraźmy sobie, że nasz korpus diachroniczny podzielimy na najmniejszą możliwą liczbę podkorpusów, czyli na jeden (wtedy uzyskany podkorpus będzie tożsamy z całym korpusem). W takiej sytuacji dostaniemy bardzo wiarygodny wynik, oparty na niezliczonych poświadczeniach, tyle że będzie to... zaledwie jeden uśredniony punkt na osi czasu. Jeśli chcemy obserwować zmianę w proporcji form recesywnych i innowacyjnych w kolejnych stuleciach, większa liczba obserwacji jest po prostu warunkiem *sine qua non*.

Sposobem pogodzenia tych sprzecznych wymagań jest podział korpusu na zachodzące na siebie podkorpusy. Wyobraźmy sobie korpus obejmujący lata 1500–1700. Możemy go podzielić na 4 podkorpusy (1500–1550, 1551–1600, 1601–1650, 1651–1700). Można jednak z tego korpusu wykroić więcej podkorpusów, które by na siebie zachodziły, co dawałoby przedziały: 1500–1550, 1525–1575, 1551–1600, 1576–1625, 1601–1650, 1626–1675, 1651–1700. W ten sposób uzyskujemy siedem punktów pomiarowych. Oczywiście oznacza to w konsekwencji, że (poza oboma skrajnymi ćwierćwieczami) dane są pobierane z każdego tekstu dwukrotnie i wchodzą do dwu różnych podkorpusów. Jest to rodzaj procedury zwanej w statystyce wygładzaniem (ang. *smoothing*); jego sposób dzia-

łania najlepiej unaoczni poniższy przykład. Przyjmijmy, że stosujemy podział na cztery podkorpusy, obejmujące materiał z pięćdziesięciolecia. Wyobraźmy sobie teraz trzy teksty napisane kolejno w roku 1502, 1547 i 1551. Między pierwszym i drugim tekstem jest zatem 45 lat odstępu w czasie, między drugim i trzecim – cztery lata. A jednak, zgodnie z naszym podziałem korpusu na półwiecza, to pierwszy i drugi tekst znajdują się w jednym przedziale, a więc stanowią podstawę dla jednego punktu, trzeci tekst wpada do kolejnego podkorpusu – znajduje się za granicą wyznaczającą kolejny przedział. Rozwiązanie z nachodzącymi na siebie podkorpusami niweluje częściowo to swoiste przekłamanie. Pierwszy punkt powstały jako reprezentant przedziału jest bowiem oparty o dane pochodzące z tekstu pierwszego i drugiego, drugi zaś – z tekstu drugiego i trzeciego.

Okres pięćdziesięcioletni – całe półwiecze! – to czas sam w sobie wystarczająco długi, by mogła w nim zajść zmiana językowa (Mair, 2006), zdawaliśmy więc sobie sprawę, że optymalna wielkość poszczególnych podkorpusów powinna być mniejsza. Po przeprowadzeniu przeróżnych dodatkowych testów ewaluacyjnych (opisujemy je poniżej) uzyskaliśmy zadowalające wyniki dla podkorpusów obejmujących okres 20 lat, przy wielkości „zakładki" ustalonej na 10 lat. Wszystkie opisane poniżej modele logistyczne dla poszczególnych zmian językowych opierają się na takim właśnie dwudziestoletnim próbkowaniu.

W naszych rozważaniach przyjęliśmy milcząco, że w każdym podkorpusie zliczamy poszczególne wystąpienia wyrazów. Nie jest to jednak rozwiązanie jedyne: moglibyśmy przecież przyjąć, że podstawową jednostką podlegającą kwantyfikacji nie jest wyraz (ile razy dana forma wystąpiła w podkorpusie?), lecz tekst (ile tekstów z podkorpusu zawierało daną formę?). Poszczególne teksty wchodzące w skład korpusu mają przecież niekiedy diametralnie odmienną liczbę poświadczeń interesujących nas form. Wynika to zarówno z różnej ich długości, jak i z idiosynkratycznych właściwości stylu autora, który nawet w długim tekście może użyć niewiele razy np. słowa *inszy*/*inny*. W tej sytuacji wydawać by się mogło, że zamiast liczby poświadczeń istotniejsza może być liczba tekstów, w których pojawia się forma recesywna bądź innowacyjna. Istnieją jednak powody, by nie iść tą drogą. Po pierwsze, liczba tekstów w danym podkorpusie jest wielokrotnie mniejsza od liczby badanych form, co prowadziłoby do bardzo gruboziarnistych wyników obliczonych proporcji. Po drugie, w rzeczywistości wiele tekstów zawiera poświadczenia tak jednej, jak i drugiej formy, co kazałoby przyjąć istnienie trzech kategorii: teksty z formami wyłącznie recesywnymi, teksty z formami wyłącznie innowacyjnymi oraz takie, w których obie formy koegzystują. I tak np. spośród 498 tekstów, które zawierają słowo *bardzo*, aż 131 poświadcza obie formy. Wśród tych ostatnich bywa tak, że jedna dominuje, druga zaś jest wyraźnie rzadsza, niejednokrotnie jednak ich frekwencja jest rozłożona dość równo – np. u Klonowica są to 3 formy recesywne wobec 6 innowacyjnych. W wypadku innych tekstów rzeczywiście jedna z form dominuje, np. *Dyskurs*

o zawziętych teraźniejszych zaciągach z roku 1606 zawiera 42 formy recesywne i 2 innowacyjne, natomiast późniejsza o ponad wiek *Poczta królewiecka* zawiera 58 wystąpień formy innowacyjnej i 3 recesywnej. Oba teksty poświadczają więc jedną i drugą formę, jednocześnie jednak wykazują wyraźną preferencję dla, odpowiednio, *barzo* lub *bardzo*. To pokazuje, że raczej konkretna forma wyrazowa powinna być jednostką, nie zaś cały tekst.

Na koniec przedstawmy rzeczywistą procedurę gromadzenia i analizy danych. Korpus był przeszukiwany bezpośrednio w środowisku programistycznym R (przy użyciu niskopoziomowych funkcji i niewielkich skryptów tworzonch na potrzeby niniejszego projektu) bądź przy pomocy programu konkordancyjnego AntConc. Nazwa każdego pliku zaczynała się od daty publikacji tekstu, w istocie rzeczy potrzebowaliśmy więc jedynie tych początkowych części nazw plików. I tak np. analizując wystąpienia *-bych* > *-bym*, najpierw tworzyliśmy konkordancję z formami recesywnymi (*-bych*). Od tego momentu jedyną istotną informacją była liczba wystąpień danej daty w konkordancji.

Użycie programu konkordancyjnego było konieczne wtedy, gdy wyniki trzeba było kontrolować ręcznie, jak np. w wypadku zmiany *-ir-* > *-er-*. Przy analizie tego segmentu z konkordancji tworzono listę wyrazów, na której zaznaczano te, które nie były przykładami interesującej nas zmiany, np. liczne wtrącenia obcojęzyczne, głównie łacińskie. Wyrazy te zostały następnie usunięte z konkordancji, która była poddawana dalszej analizie.

Dodajmy, że jakkolwiek w większości wypadków można było pominąć etap przeszukiwania korpusu przez program konkordancyjny i wczytywać dane z tekstów prosto do konsoli środowiska R, nie rezygnowaliśmy z tego, przede wszystkim dlatego, że konkordancja daje pewną kontrolę wyników, umożliwiając przyjrzenie się im z bliska. Ponadto konkordancja pozawala udoskonalić zapytanie, np. poprzez sprawdzanie, czy spodziewane warianty graficzne występują w tekstach i czy rzeczywiście dany wzorzec wyszukiwania zwraca właściwe przykłady.

W skrypcie języka R stworzonym na potrzeby niniejszej analizy danych należało ustalić cztery wartości: datę początkową pierwszego podkorpusu, datę końcową ostatniego podkorpusu, wielkość podkorpusu rozumianą jako liczbę lat, które się na niego składają, oraz wielkość „zakładki", tj. liczbę lat, o którą podkorpusy zachodzą na siebie. I tak np. jeśli pierwsza z tych wartości przyjmuje 1500, druga 1900, trzecia 20, a czwarta 10, to oznacza to, że pierwszy z podkorpusów obejmuje lata 1500–1520, drugi 1510–1530, ostatni zaś 1880–1900. Jak już wspomniano, w niniejszym rozdziale co do zasady będziemy mówili o podkorpusach o wielkości 20 lat, które zachodzą na siebie co 10 lat. Pod koniec odwołamy się do innych rozmiarów podkorpusów, by sprawdzić, w jakim stopniu zmiana założeń wyjściowych wpływa na wyniki. Jest bowiem rzeczą oczywistą, że gdy rozmiar podkorpusu jest duży, zaś poszczególne podkorpusy zachodzą na siebie dość gęsto, to przebieg zmiany będzie znacznie bardziej wygładzony (czyli wyidealizowany), choćby dlatego, że idiolekt pojedynczego autora, który

swoimi preferencjami względem jednej z form odbiega od gustów epoki, ginie w większym zbiorze tekstów.

W kolejnym kroku zliczamy wszystkie wystąpienia szukanej formy, które mieszczą się w następujących po sobie przedziałach czasowych, czyli sumujemy odnalezione wystąpienia w tekstach z lat 1500–1520, potem 1510–1530 itd. W ten sposób otrzymujemy liczbę wystąpień formy innowacyjnej i w każdym podkorpusie; gdy to samo robimy dla formy recesywnej r, obliczamy proporcję tych pierwszych w stosunku do drugich. Przed rozpoczęciem procesu zmiany (na ogół w pierwszych kilku podkorpusach) proporcja i do r będzie wynosić 0 (procentowo: 0% do 100%), w połowie zmiany 0,5 (50% do 50%), by po całkowitym zaniknięciu formy recesywnej osiągnąć wartość 1 (100% do 0%). W podkorpusie notującym 66 razy formę recesywną i 9 razy formę innowacyjną proporcja wyniesie:

$$p(i) = \frac{9}{9 + 66} = 0,12$$

Dla formy innowacyjnej poświadczonej 756 razy i 176 poświadczeń formy recesywnej proporcja wyniesie 0,81, i tak dalej. Warto zauważyć, że obliczone proporcje mają zarazem swoją interpretację probabilistyczną: proporcja 0,12 oznacza bowiem jednocześnie, że prawdopodobieństwo wylosowania formy i w pierwszym korpusie wynosi 0,12, a formy $r = 0,88$. Wzajemna relacja obu form w każdym podkorpusie spełnia oczywiście równanie $p(r) = 1 - p(i)$, z czego siłą rzeczy wynika, że $p(i) = 1 - p(r)$. Podstawowe pytanie, na które szukamy odpowiedzi w tym rozdziale, brzmi: w jaki sposób proporcja i do r rozkłada się w każdym kolejnym podkorpusie?

3.4. Wyniki

3.4.1. *więtszy* > *większy*

Pierwsza z omawianych zmian to przejście *więtszy* > *większy* (wliczając w to wszystkie ich formy fleksyjne w stopniu równym, wyższym i najwyższym). Rys. 3.1 przedstawia proporcję form recesywnych do ich innowacyjnych odpowiedników czy też, mówiąc ściślej, prawdopodobieństwo odnalezienia formy innowacyjnej, $p(i)$, w poszczególnych odcinkach korpusu (por. ciemne punkty na wykresie; linia ciągła zostanie omówiona za chwilę). Pierwsze wystąpienie formy innowacyjnej datujemy na rok 1543 (*Zielnik* Falimirza), ostatnie wystąpienie formy recesywnej to *Maria* Malczewskiego z roku 1825. Dodajmy jednak, że jest to jedno jedyne wystąpienie, nie notujemy poza tym tych form później niż w *Nowych*

Atenach Benedykta Chmielowskiego. Dane liczbowe użyte do wygenerowania wykresu ukazuje Tab. 3.1[4].

Tabela 3.1. Dane liczbowe dla zmiany *więtszy* > *większy* (z uwzględnieniem form fleksyjnych).

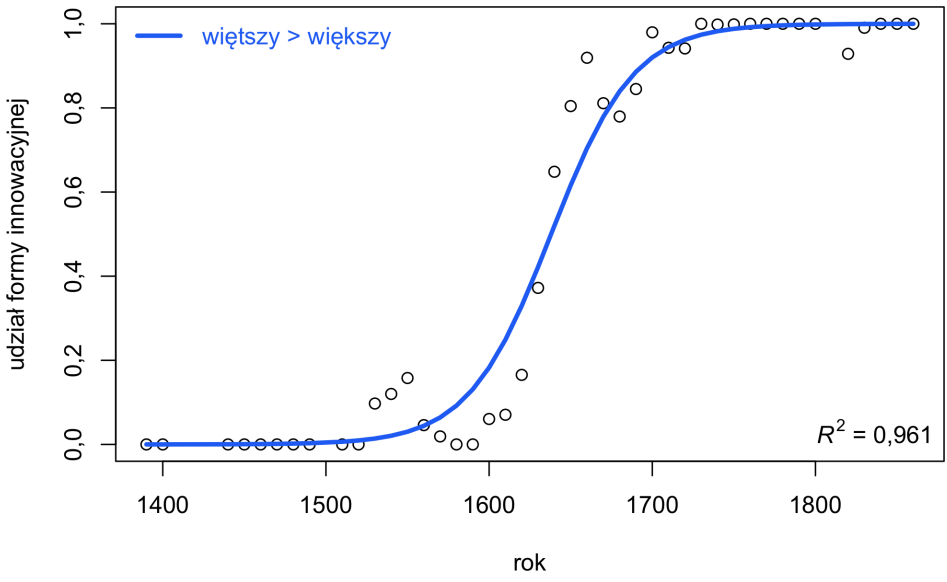
podkorpus	formy recesywne	formy innowacyjne	udział form innowacyjnych
1380–1400	3	0	0
1390–1410	3	0	0
1400–1420	0	0	–
1410–1430	0	0	–
1420–1440	0	0	–
1430–1450	2	0	0
1440–1460	2	0	0
1450–1470	1	0	0
1460–1480	1	0	0
1470–1490	2	0	0
1480–1500	2	0	0
1490–1510	0	0	–
1500–1520	36	0	0
1510–1530	44	0	0
1520–1540	65	7	0,10
1530–1550	66	9	0,12
1540–1560	16	3	0,16
1550–1570	104	5	0,05
1560–1580	206	4	0,02
1570–1590	221	0	0
1580–1600	284	0	0
1590–1610	823	53	0,06
1600–1620	1173	89	0,07
1610–1630	863	171	0,17
1620–1640	442	262	0,37
1630–1650	159	293	0,65
1640–1660	72	296	0,80
1650–1670	37	421	0,92
1660–1680	176	756	0,81

[4] Wprawdzie ze względu na szczupłość miejsca w książce publikujemy wyłącznie dane liczbowe do zmiany *więtszy* > *większy*, wszystkie jednak pozostałe tabele i zbiory surowych danych znajdzie czytelnik w repozytorium internetowym GitHub: https://github.com/computationalstylistics/diachronia/.

podkorpus	formy recesywne	formy innowacyjne	udział form innowacyjnych
1670–1690	431	1524	0,78
1680–1700	289	1574	0,84
1690–1710	14	676	0,98
1700–1720	21	347	0,94
1710–1730	18	289	0,94
1720–1740	0	296	1
1730–1750	2	1090	1
1740–1760	2	1333	1
1750–1770	0	909	1
1760–1780	0	925	1
1770–1790	0	511	1
1780–1800	0	216	1
1790–1810	0	157	1
1800–1820	0	0	–
1810–1830	1	13	0,93
1820–1840	1	105	0,99
1830–1850	0	198	1
1840–1860	0	154	1

Na Rys. 3.1 widać wyraźnie, że wzajemne relacje form *więtszy* oraz *większy* (proporcje w poszczególnych podkorpusach) układają się tak, jakby je nanizano na niewidoczną nitkę czy poprowadzono wzdłuż niewidzialnej krzywej. Nie ma najmniejszej wątpliwości, że mamy tu do czynienia z jakąś prawidłowością. Oczywiście niektóre z punktów znajdują się w pewnym oddaleniu od siebie, jakby nie chciały się do końca poddać owej podskórnie działającej regule przyciągającej punkty do hipotetycznej krzywej. Niewidzialna nitka albo hipotetyczna krzywa – a w istocie ciąg lokalnych uśrednień danych empirycznych – jest właśnie poszukiwanym przez nas modelem. Relacja krzywej (modelu) do punktów (danych) pokazuje, że modelowanie nie jest czysto akademicką zabawą czy kolejną „modą badawczą": jest próbą zrozumienia skomplikowanych procesów zachodzących w przyrodzie przez znalezienie hipotetycznej linii, którą w sposób optymalny można by poprowadzić przez punkty wyznaczające dane empiryczne. Towarzyszy temu założenie, że ukryte dla oka ludzkiego prawidłowości są odpowiedzialne za „wygenerowanie" rzeczywistych zjawisk naturalnych (z całą tych zjawisk niedoskonałością) i że możliwe jest odszyfrowanie pierwotnych prawidłowości przez model – wtórną idealizację danych empirycznych.

Osobnym zagadnieniem jest oszacowanie istotności statystycznej obliczonego modelu, a zatem znalezienie odpowiedzi na pytanie, do jakiego stopnia model potrafi wyjaśnić opisywany przez siebie zbiór danych. Nieco uproszczając: jeśli



Rysunek 3.1. Przebieg zmiany *większy* > *większy* (z uwzględnieniem form fleksyjnych). Punkty ukazują proporcję między formą recesywną i innowacyjną (dane empiryczne), linia ciągła przedstawia model logistyczny.

wysublimowane matematycznie równanie, wyprowadzone przy użyciu zaawansowanej algebry, nie jest w stanie powiedzieć nic więcej o danych niż zwykła średnia arytmetyczna, świadczy to o braku istotności takiego równania. Im więcej wiedzy o danych uzyskujemy przy użyciu testowanego modelu, tym większa jego istotność statystyczna. Podstawową miarą istotności jest tzw. wartość p: im jest niższa, tym lepiej dla testowanego modelu. Przyjmuje się na ogół, że wartość $p < 0{,}05$ pozwala uznać model za wiarygodny[5]. Nie zawsze dane są na tyle klarowne, by wybór optymalnego modelu narzucał się samoistnie – choćby wybór między modelem liniowym, logistycznym i wielomianowym może wymagać namysłu, gdy dane nie układają się wzdłuż czytelnej linii. W takiej sytuacji wartość p oraz współczynnik dopasowania R^2 okazują się bezcennymi narzędziami diagnostycznymi.

W naszym przypadku układ punktów zdaje się przebiegać wzdłuż krzywej, która zaczyna się w lewym dolnym rogu i zmierza do prawego górnego po trajektorii przypominającej literę *s*. Taki układ sugeruje, że optymalnym mode-

[5] Celowo odwołujemy się w tym miejscu do słownictwa potocznego, intuicyjnego, unikając ogólnie przyjętej terminologii statystycznej. Łamy niniejszej książki są zdecydowanie zbyt szczupłe, by omówić tutaj zagadnienie testowania hipotez statystycznych czy choćby wprowadzić podstawowe pojęcia, takie jak „hipoteza zerowa", „stopnie swobody", „przedział ufności" czy podobne. Liczymy na to, że zaznajomieni ze statystyką czytelnicy wybaczą nam powyższe uproszczenia.

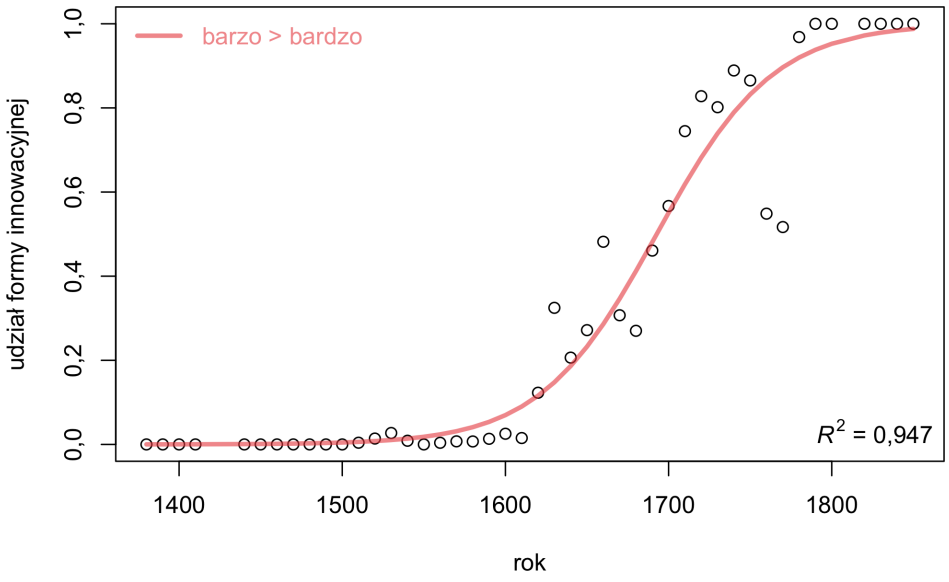
lem może być regresja logistyczna. I rzeczywiście, krzywa logistyczna dobrze opisuje zmianę *więtszy* > *większy*: optymalne parametry modelu ($\beta_0 = -64,569$, $\beta_1 = 0,03942$, przy 42 stopniach swobody) wykazują bardzo wysoką istotność statystyczną, $p < 0,001$. Model został przedstawiony na Rys. 3.1 za pomocą linii ciągłej. Wysoki jest też stopień dopasowania wymodelowanej krzywej logistycznej do obserwowanych danych, o czym świadczy wartość $R^2 = 0,961$. Najsłabszą wartość predykcyjną, jak widać na wykresie, model wykazuje dla danych z początku wieku XVI, późniejszy przebieg krzywej nieuzbrojone oko ludzkie uznałoby za co najmniej akceptowalny.

3.4.2. *barzo* > *bardzo*

Drugą omawianą przez nas zmianą jest *barzo* > *bardzo*. W porównaniu do pozostałych przypadków przedstawionych w tym rozdziale jest to zmiana stosunkowo późna – aż do początku XVII wieku w zasadzie nie można mówić o rozpoczęciu procesu konkurencji form. W korpusie do roku 1600 notujemy zaledwie 25 wystąpień *bardzo* (nie tylko w stopniu równym, a także w formie *bardziej* i *najbardziej*) przy ponad 2213 wystąpieniach formy *barzo* (oraz *barziej* i *najbarziej*). Od tego momentu udział formy innowacyjnej wzrasta, choć w niektórych podkorpusach jest niższy niż w bezpośrednio je poprzedzających. Gwałtowny wzrost udziału formy *bardzo* następuje z końcem XVIII wieku.

Jak widać na Rys. 3.2, mamy tu do czynienia z bardzo czystym procesem Piotrowskiego, choć odstępstwo od modelu w drugiej połowie XVIII wieku mocno rzuca się w oczy (por. dwa odstające punkty na Rys. 3.2). Mimo owego odstępstwa dopasowanie modelu do danych obserwowanych jest bardzo wysokie, $R^2 = 0,947$.

Istnienie dwóch punktów wyraźnie odbiegających od wymodelowanego przebiegu zmiany każe nam się zastanowić nad możliwą przyczyną tego stanu rzeczy. Być może źródła odstępstw trzeba szukać w fakcie, że frekwencje obliczyliśmy zbiorczo dla stopnia równego, wyższego i najwyższego *barzo*, *barziej* i *na(j)barziej* oraz *bardzo*, *bardziej* i *na(j)bardziej*, podczas gdy mogło się zdarzyć, że poszczególne formy zmieniały się w różnym tempie. I rzeczywiście, jeśli potraktujemy stopień równy jako jeden zbiór danych, a stopień wyższy i najwyższy jako drugi, to okaże się, że zmiana następuje nieco prędzej w obrębie tej pierwszej formy (zob. Rys. 3.3). Oba rozdzielne modele mają podobne dopasowanie do danych empirycznych (odpowiednio: 0,943 i 0,928), niewiele gorsze od modelu zbiorczego omówionego powyżej. Udział formy innowacyjnej w obrębie stopnia wyższego i najwyższego (*bardziej* i *na(j)barziej*) jest nieco niższy niż w obrębie stopnia równego (*bardzo*) w tym samym czasie. Również pod koniec XVIII wieku, kiedy udział formy innowacyjnej spada, dla stopnia równego jest on konsekwentnie niższy niż dla pozostałych stopni. Można to więc interpretować tak, że zmiana zaczęła się w obrębie stopnia równego, stopień wyższy zaś i najwyższy wykazywały się większym konserwatyzmem. Być może konserwatyzm



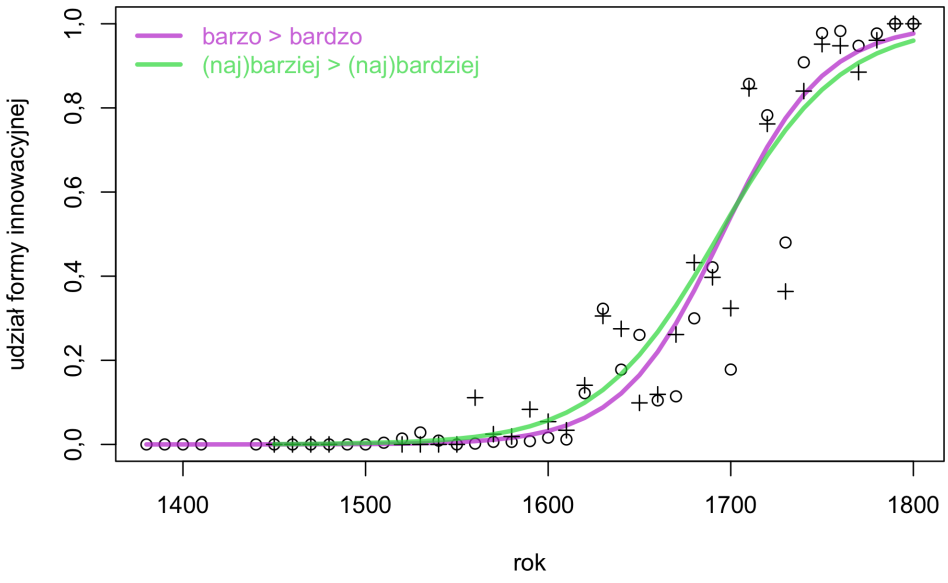
Rysunek 3.2. Przebieg zmiany *barzo* > *bardzo* (także w stopniu wyższym i najwyższym).

ten wynikał stąd, że formy te są znacznie rzadsze (2599 poświadczeń stopnia wyższego i najwyższego wobec 9574 poświadczeń stopnia równego).

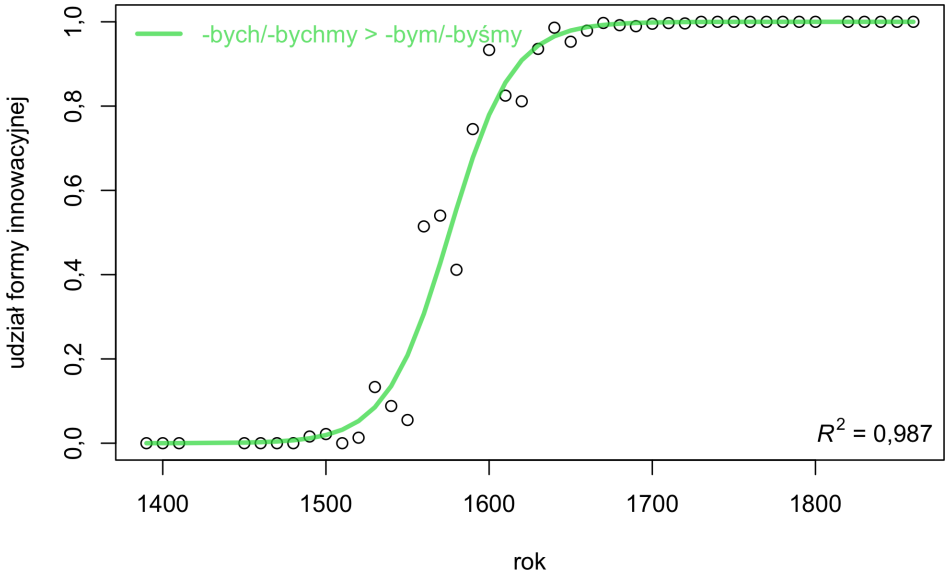
3.4.3. *-bych* > *-bym*, *-bychmy* > *-byśmy*

Na przełomie XV i XVI wieku we fleksji czasownika obok końcówek *-m*, *-smy* i *-swa* zaczęły się pojawiać formy na *-ch*, *-chmy*, *-chwa*, z tym że w tekstach staropolskich forma ta jest rzadka. Nie mamy tu więc do czynienia ze zmianą, ale z konkurencją dwu form, z których jedna jest „nieudaną", odrzuconą innowacją, która jednak w pewnym momencie zdobyła sporą popularność. Konkurencję tych dwu form badali m.in. Taszycki (1946), Kowalska (1978), a także Motyl (2014). Formę z *-ech*/*-chmy* Taszycki łączy z Małopolską, autorzy spoza południa Polski mieli stosować ją rzadko. Zauważono też, że przebieg zmiany w odniesieniu do obu liczb był odmienny – forma ta utrzymywała się dłużej w liczbie mnogiej niż w liczbie pojedynczej.

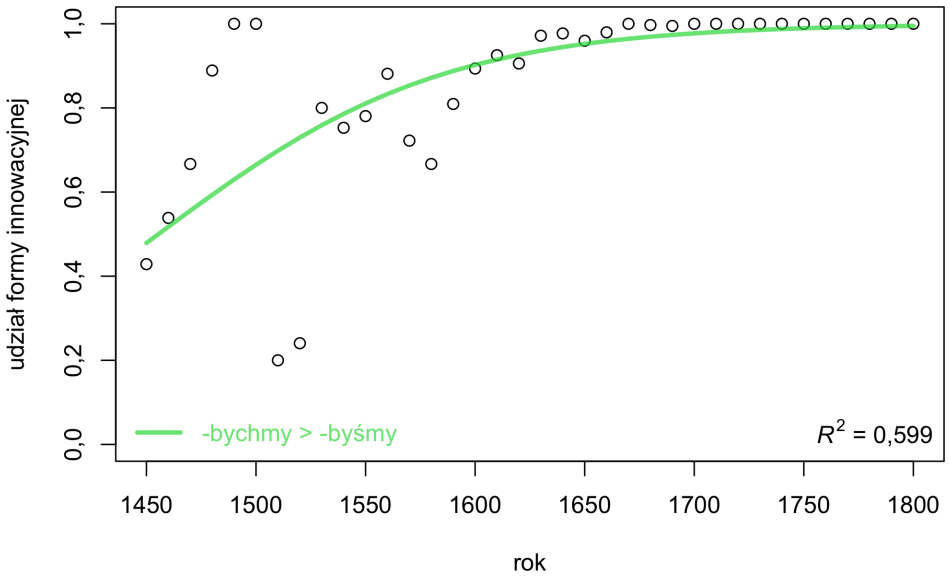
Rys. 3.4 ukazuje przebieg zmiany wykładników obu liczb razem. Pierwsza rzecz, która natychmiast rzuca się w oczy, to bardzo wysokie dopasowanie modelu do danych: słowem, zmiana *-bych* > *-bym*, *-bychmy* > *-byśmy* jest książkowym przykładem regresji logistycznej. Parametr R^2 = 0,987 tylko potwierdza obserwacje poczynione gołym okiem.



Rysunek 3.3. Przebieg zmiany *barzo* > *bardzo* obliczony osobno dla stopnia równego i (naj)wyższego.



Rysunek 3.4. Przebieg zmiany *-bych/-byśmy* > *-bym/-byśmy*.

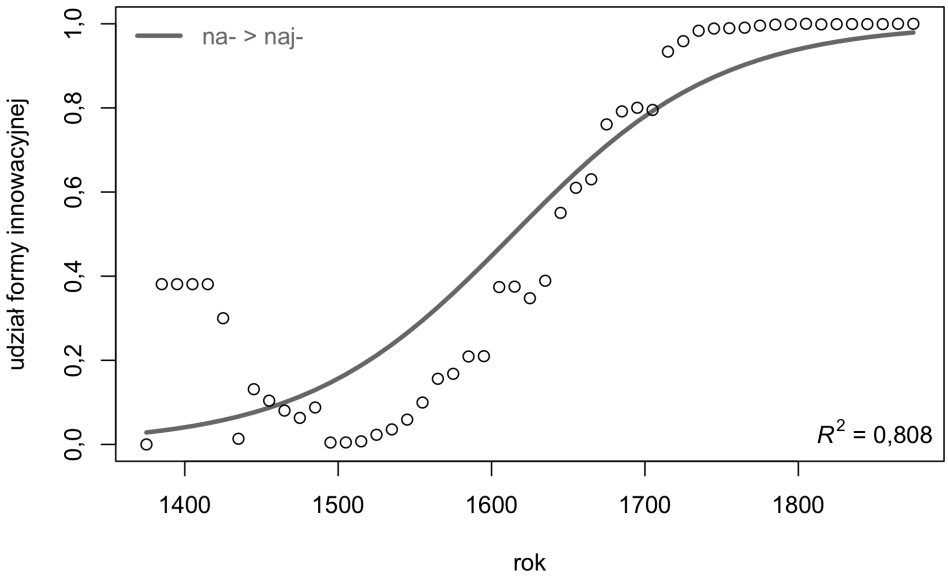


Rysunek 3.5. Przebieg zmiany *bychmy* > *byśmy*.

Nauczeni przykładem omówionej powyżej zmiany *barzo* > *bardzo* i wyczuleni na możliwe różne trajektorie zmian dla różnych form fleksyjnych tego samego wyrazu, końcówki czasownika poddaliśmy dodatkowemu testowi, w którym optymalny model został obliczony dla każdej formy z osobna. Okazuje się, że zmiana *-bych* > *-bym* wydaje się poszukiwaną zmianą modelową – taką, w której dane empiryczne niemal pokrywają się z wartościami oczekiwanymi (wyniki łudząco podobne do tych z Rys. 3.4). Dalece inaczej ma się rzecz z formą *być* w 1. os. liczby mnogiej (Rys. 3.5). O ile końcowa faza zmiany i czas całkowitej dominacji formy innowacyjnej po roku 1650 wyglądają identycznie w obu wypadkach, o tyle początek zmiany *-bychmy* > *-byśmy* jest zupełnie chaotyczny i chyba w ogóle niepodatny na modelowanie. Jedynie dzięki temu, że końcowa faza zmiany przypomina mimo wszystko proces Piotrowskiego, współczynnik dopasowania modelu jest względnie wysoki, $R^2 = 0{,}599$, choć nadal nieakceptowalny i ukazujący znacząco odmienny przebieg zmiany w obrębie liczby pojedynczej i liczby mnogiej.

3.4.4. *na-* > *naj-*

Gdyby nie brać pod uwagę wystąpień *naj-* z korpusu staropolskiego, to pierwsze poświadczenie formy innowacyjnej w korpusie wypadłoby w roku 1535; ostatnie poświadczenie formy recesywnej jest datowane na 1772. Jeśli chodzi o okres staropolski, to trudno wypowiadać się o tej zmianie w sposób wiarygodny ze względu na mocno rozmyte wyniki. Nie możemy przymykać oczu na



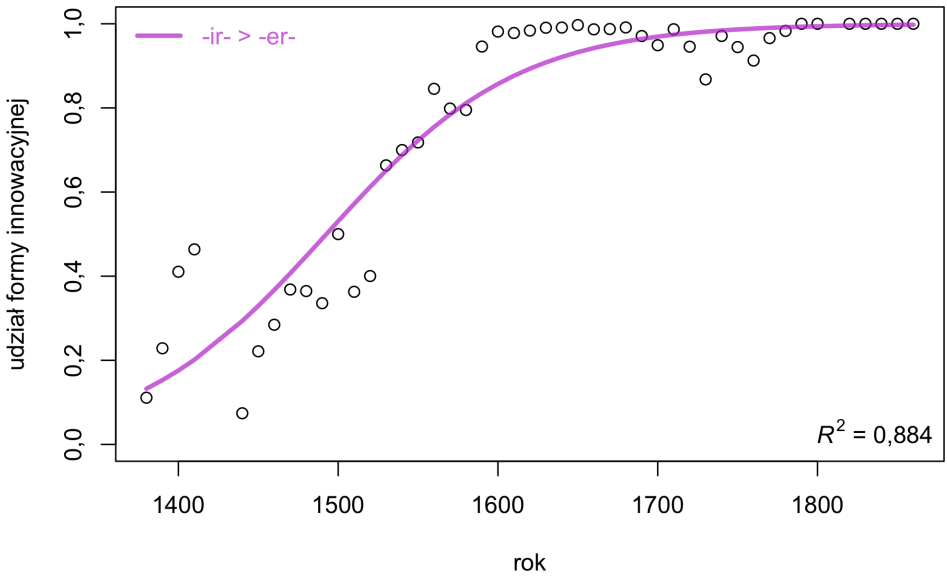
Rysunek 3.6. Przebieg zmiany wykładnika stopnia wyższego *na-* > *naj-*.

ewidentne pojawienie się formy innowacyjnej w dwóch dużych tekstach z epoki: *Psałterzu floriańskim* i nieco późniejszym *Psałterzu puławskim*. Choć oba psałterze są w zasadzie jednym i tym samym tekstem (*Psałterz puławski* opiera się w znacznym stopniu na swoim poprzedniku), a poświadczonych form innowacyjnych jest niewiele (ledwie 20 wystąpień w ciągu całego stulecia), to ogólna szczupłość danych w XV wieku mocno zawyża proporcje form innowacyjnych do recesywnych. Przebieg zmiany w całym korpusie ilustruje Rys. 3.6.

Mimo wspomnianego rozmycia wyników w najdawniejszej polszczyźnie dopasowanie teoretycznego przebiegu regresji logistycznej do danych empirycznych nie jest tak złe, jak można by sądzić, R^2 = 0,808. Gdybyśmy jednak odeszli od naszej procedury badawczej i arbitralnie skrócili badany okres o dwa stulecia (a więc nie brali pod uwagę okresu staropolskiego), dostalibyśmy bardzo wysokie dopasowanie modelu, R^2 = 0,985.

3.4.5. *-ir-* > *-er-*

Konkurencja form *-ir-* > *-er-*, takich jak np. *śmirć* > *śmierć*, *cirpieć* > *cierpieć*, *pirwej* > *pierwej*, przedstawiona została na Rys. 3.7. Jak wynika z wykresu, mamy do czynienia z niedoskonałym przykładem modelu logistycznego, jako że w korpusie nie został zaświadczony punkt początkowy zmiany – formy innowacyjne notuje już pierwszy podkorpus, a później ich udział konsekwentnie wzrasta. Inną nietypową cechą jest bardzo długi okres, w którym udział formy



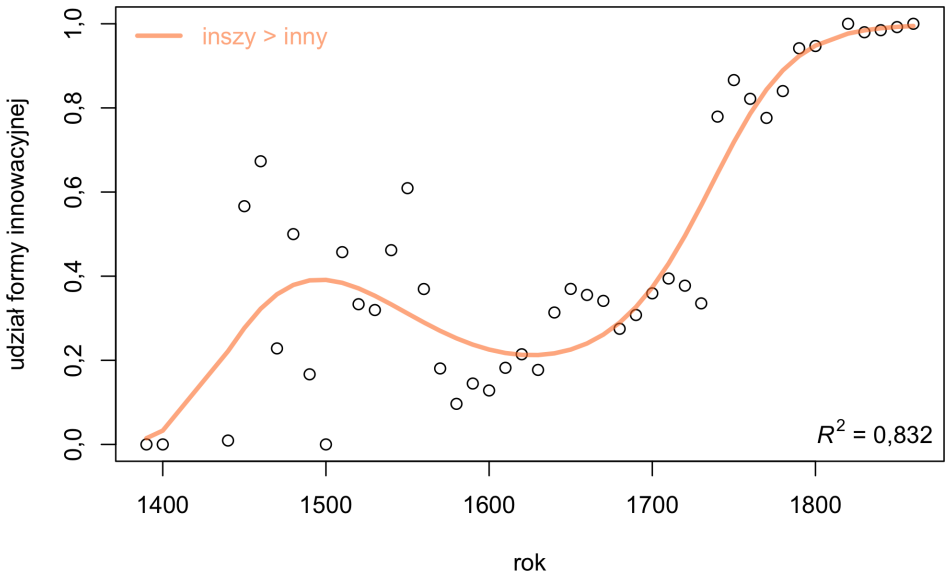
Rysunek 3.7. Przebieg zmiany *-ir- > -er-*.

innowacyjnej jest bliski 100%. Wygląda na to, że zmiana zakończyła się w zasadzie na początku XVII wieku, ale jeszcze przez kolejne dwa stulecia formy archaiczne dawały się zauważyć w tekstach. Być może jest to spowodowane tym, że zastępowanie *-ir-* przez *-er-* przybierało różne tempo w odniesieniu do różnych jednostek leksykalnych, możliwe też, że w XVIII wieku nastąpiło coś w rodzaju czasowego powrotu dawnej formy motywowanego modą. Biorąc pod uwagę fakt, że początek zmiany pozostaje dla nas ukryty w mrokach epoki przedpiśmiennej i przez to niewidoczny dla modelu logistycznego, nie powinno nas dziwić, że dopasowanie modelu, R^2 = 0,884, jest niższe niż w przypadkach omawianych powyżej, choć i tak zastanawiająco dobre.

3.4.6. *inszy > inny*

Za pierwotną formę współczesnego wyrazu *inny* uznaje się formę *iny*. Z niej z kolei wywodzi się zarówno współcześnie używana forma z geminatą *inny*, jak i wyparta przez nią forma *inszy*, która początkowo była formą stopnia wyższego od *iny*. Poza tymi dwiema formami notuje się również formy *inkszy* i *inakszy*, które z kolei mają charakter gwarowy (Sławski, 1952). Zdaniem Karasiowej (1978) współczesna forma wyrazu *inny* jest pochodzenia mazowieckiego.

Nasz korpus nie notuje form *inkszy* oraz *inakszy*. Z kolei *iny* znajduje 153 potwierdzenia, ostatnie w roku 1719, przy czym wydaje się, że forma ta została przez autora użyta dla rymu, por. *Choć bez przyczyny Burzy się iny* (cytat pochodzi



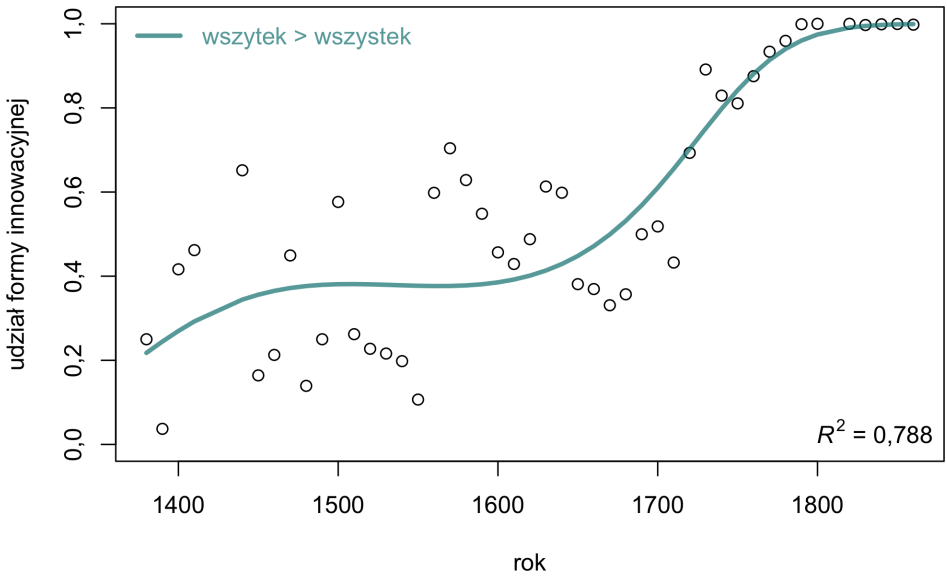
Rysunek 3.8. Przebieg zmiany *inszy* > *inny* (wraz z formami fleksyjnymi).

z utworu *Collectanea, to jest zbierana drużyna poetycka*), Wariant ten jest jednak na tyle rzadki, że nie będziemy go rozpatrywali. Skupimy się na konkurencji form *inszy* > *inny* (wraz z ich wszystkimi formami fleksyjnymi). Wyparcie pierwszej formy przez drugą ilustruje Rys. 3.8.

Nietrudno zauważyć, że mówienie o procesie Piotrowskiego jest w tym wypadku problematyczne. Gdybyśmy w naszych rozważaniach ograniczyli się do procedury takiej jak opisana w powyżej omówionych przykładach, niewiele więcej dałoby się o tej zmianie powiedzieć; stopień dopasowania tak dobranego modelu wynosiłby zaledwie $R^2 = 0{,}55$.

Poczyniona jednak powyżej obserwacja, że tzw. prawo Piotrowskiego jest nieco inaczej sformułowaną odmianą regresji logistycznej (Köhler, 2015; Vulanović, 2007; Vulanović i Baayen, 2007), pozwala nam skorzystać z uogólnionych modeli regresyjnych (James i in., 2013: 265–300), w tym przede wszystkim z modelu wielomianowej regresji logistycznej. Zasada działania takiego modelu w pewnym skrócie polega na tym, że zamiast ograniczać modelowaną krzywą do kształtu litery *s*, dopuszczamy większą jej giętkość. W zależności od stopnia wielomianu linia może się zakrzywiać wielokrotnie; każdy kolejny parametr wielomianu pozwala na dwa dodatkowe zakrzywienia. Na Rys. 3.8 pokazano model wykorzystujący wielomian IV stopnia.

Jak widać, modelowana linia w miarę dokładnie idzie za danymi empirycznymi, co potwierdza stosunkowo wysokie dopasowanie modelu, $R^2 = 0{,}832$. Za-



Rysunek 3.9. Przebieg zmiany *wszytek* > *wszystek* (wraz z formami fleksyjnymi).

uważmy jednak, że po roku 1550 przebieg zmiany układa się według klasycznej krzywej logistycznej: jeśli uwzględnilibyśmy dane tylko od tego roku i użyli modelu klasycznego dla danych 1550–1850, to wartość R^2 wzrośnie do 0,944.

3.4.7. *wszytek* > *wszystek*

Zmiana *wszytek* > *wszystek* (wraz z formami fleksyjnymi, a także leksem *wszytko* > *wszystko*) jest taką zmianą, w której trudno doszukiwać się procesu Piotrowskiego (Rys. 3.9). Wprawdzie trend rozwojowy stabilizuje się w drugiej połowie XVIII wieku i zbliża się istotnie do spodziewanego kształtu litery *s*, w większej swej części jednak wyniki są bardzo mocno rozmyte. Nawet użycie wielomianowego modelu regresji logistycznej na niewiele się zdaje: mimo przyzwoitego stopnia dopasowania modelu, $R^2 = 0{,}788$ przy zastosowaniu wielomianu III stopnia, gołym okiem widać, że dla kilkusetletniego okresu obejmującego wieki XV–XVII rozkład punktów nie przypomina wymodelowanej linii krzywej. Dopasowanie do danych empirycznych można sztucznie zwiększyć przez użycie modelu o wyższym stopniu wielomianu (np. dla wielomianu IV stopnia $R^2 = 0{,}805$), taki model przestaje jednak wykazywać jakąkolwiek istotność statystyczną (wartość p gwałtownie rośnie), co oznacza po prostu, że model nie opisuje danych w dostatecznym stopniu.

Mimo że z matematycznego punktu widzenia problem zmiany *wszytek* > *wszystek* można by uznać za zamknięty, językoznawca diachroniczny z całą

pewnością odczuwa niedosyt. Jak bowiem wyjaśnić dalece nieczytelny charakter tej zmiany? Dlaczego niektóre zmiany w polszczyźnie przypominają modelowy przebieg krzywej Piotrowskiego, a inne nie poddają się modelowaniu? Zamiast odpowiedzieć wprost, można zadać inne pytanie, a mianowicie: dlaczego właściwie mielibyśmy oczekiwać, by proporcje pomiędzy formą recesywną i innowacyjną w kolejnych następujących po sobie podkorpusach układały się zawsze wzdłuż linii logistycznej?

Kryje się za tym milczące założenie, że wspólnota językowa jest homogeniczna. W konsekwencji należałoby się spodziewać, że wzrost liczby użytkowników języka, którzy przyjmują w swoim idiolekcie innowację, jest stały. Dodatkowo im więcej osób używa formy innowacyjnej, tym silniejsza jest ekspozycja na nią pozostałych użytkowników języka. Wszystko to są założenia zdroworozsądkowe. Co prawda nie znajdują one bezpośredniego potwierdzenia w faktach – nie znamy przecież motywacji użytkowników języka – ale stanowią przekonujące, jak się wydaje, wyjaśnienie typowego przebiegu zmiany językowej.

Tymczasem wspólnota językowa rzadko jest homogeniczna, przeważnie wykazuje spore zróżnicowanie geograficzne i społeczne. Co więcej, jedna z form może być preferowana w którymś z typów tekstów, w innych zaś odrzucana[6]. Być może więc problem leży w odmiennym składzie poszczególnych podkorpusów, na jakie podzieliliśmy wszystkie teksty. Wiadomo na przykład, że w kolejnych latach zmienia się reprezentacja autorów pochodzących z danego regionu. Niewykluczone więc, że gdyby udało się stworzyć idealnie zrównoważony korpus (oczywiście hipotetycznie, gdyż takie idealne zrównoważenie nigdy nie jest możliwe, a już na pewno nie w wypadku korpusu historycznego), przebieg zmiany przypominałby pozostałe omówione w tym rozdziale. Nawet jednak gdybyśmy dysponowali idealnie reprezentatywnym korpusem, możliwy byłby taki scenariusz, że w pewnym momencie do głosu dochodzi pokolenie pisarzy z obszaru, na którym preferencje względem jednej z form są odmienne niż w reszcie wspólnoty językowej; scenariusz taki znajdzie swoje odzwierciedlenie w przebiegu trendu. Wreszcie nie można wykluczyć, że na chwilowe cofnięcie się zmiany może mieć wpływ siła oddziaływania autorytetu normatywnego (choć mówienie o normie w epokach przednowopolskich jest oczywistym anachronizmem). Wreszcie w wypadku źródeł pozyskanych z dawnych druków (XVI–XVIII wieku) nie jesteśmy w stanie przekonująco określić, co pochodzi spod pióra autora, a co jest wynikiem pracy zecera: ma to znaczenie szczególnie w wypadku pisarzy pochodzących z prowincji, a drukujących swe dzieła w dużych miastach (Kraków, Wilno, Warszawa), co z pewnością skutkowało uleganiem językowym wpływom metropolii.

[6] Jako przykład tego zjawiska przytoczmy fakt, że dopełniacz liczby mnogiej *ziarn* w NKJP jest wyraźnie nadreprezentowany w tekstach naukowych.

Wszystko to są oczywiście spekulacje, z pewnością jednak nietypowy przebieg krzywej sygnalizuje potrzebę pogłębionych badań nad różnymi możliwymi czynnikami napędzającymi zmianę językową.

3.4.8. Konkurencja *abo* i *albo*

W wypadku form *abo* i *albo* nie można mówić o zmianie językowej, ale o konkurencji dwu form, którą w końcu wygrywa współczesne *albo*. Jak zwraca uwagę Michalska (2013), dystrybucję tych wariantów rozpatrywano w dwu kontekstach. Po pierwsze, dystrybucja obu form jest zdeterminowana czynnikami przestrzennymi: *abo* to forma wielkopolska, czy może szerzej północnopolska, *albo* zaś to forma małopolska (Kuraszkiewicz, 1953). *Abo* z końcem staropolszczyzny upowszechnia się na pozostałe regiony Polski. Z kolei Rospond (1950) uważa *abo* za formę potoczną, zaś *albo* za literacką. Nieobecne w klasycznych tekstach religijnych XIV i XV wieku, *abo* powoli toruje sobie drogę do powszechnego używania w tekstach literackich. Warto tu zacytować przypuszczenie Urbańczyka (1953), że *abo* w XVI wieku rozpowszechniło się w wyniku pomyłki – w drugim wydaniu Rejowej *Postylli* drukarz zastąpił *albo* przez *abo*. Michalska korzystając z danych pochodzących z kartoteki *Słownika polszczyzny XVI wieku* ustaliła, że jakkolwiek *abo* spotkać można u autorów każdego regionu Polski, to jednak tylko w Wielkopolsce i na Mazowszu, a także w mniejszym stopniu w Małopolsce przewaga *albo* nie jest znacząca (Michalska, 2013). Z kolei materiały *Słownika języka polskiego XVII i XVIII wieku* dowodzą, że *abo* zyskuje wyraźną przewagę w Małopolsce i na Mazowszu. Pihan-Kijasowa podsumowuje: „norma *albo* ustaliła się ostatecznie po XVII w. *Słownik Lindego* [. . .] notuje obydwa warianty 93 razy [. . .], przy czym *abo* dokumentuje przykładami głównie z XVI i XVII w., zaś *albo*, choć notowane przez słownikarza już od XVI w., ma dokumentację głównie z XVII i XVIII w. *Słownik języka polskiego* pod red. W. Doroszewskiego [. . .], uwzględniający materiał leksykalny od połowy XVIII w., podaje już tylko postać *albo*" (Pihan-Kijasowa, 1992: 128). Na koniec warto przytoczyć opinię Osiewicza głoszącą, że „zecer mógł, a – w przypadku krótkiego rozmiaru wersu nawet musiał – ingerować w postać graficzną tekstu, i że ingerencja ta mogła polegać zarówno na likwidowaniu redundancji zapisu już utrwalonego, tradycyjnego, jak i – co dla językoznawcy najistotniejsze – na korzystaniu z całego repertuaru takich form obocznych interpretowanych przez nas jako warianty graficzne, fonetyczne czy morfologiczne, które różniły się szerokością zapisu" (Osiewicz, 2012).

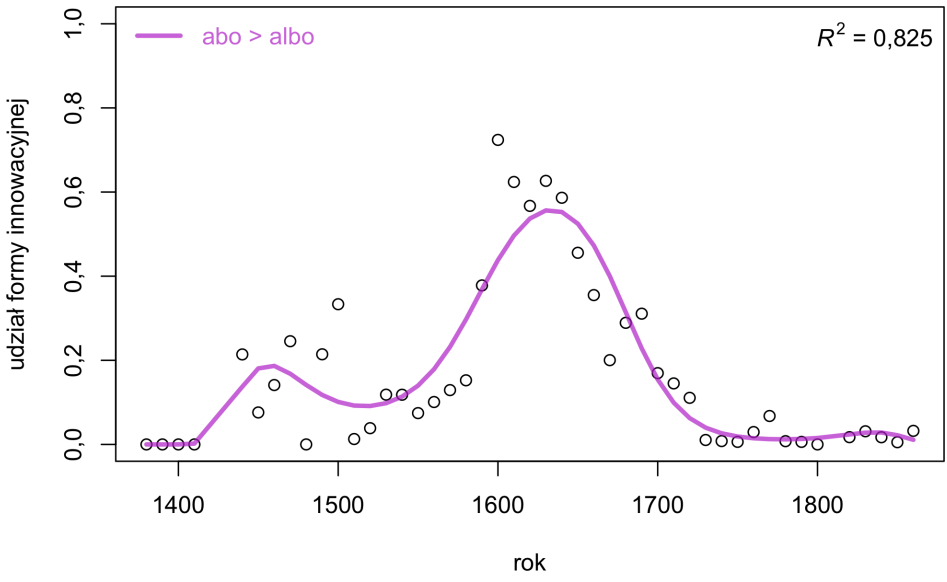
Z przytoczonych powyżej opinii można by wnioskować, że fundamentalne znaczenie dla dystrybucji *abo* i *albo* ma region pochodzenia tekstu. Jeśli tak było w istocie, to sporą przeszkodę stanowiłby dla nas fakt, że nasz korpus w najmniejszym stopniu nie dba o zrównoważenie regionalne. Tym samym wiarygodność naszej argumentacji byłaby obniżona. Jeśli bowiem korpus w poszczególnych

latach zawiera w odmiennych proporcjach teksty powstałe (bądź drukowane) w odmiennych regionach (a tak jest z pewnością), to odnotowywane przez nas zmiany we frekwencji mogłyby zależeć nie tylko od procesu historycznojęzykowego, ale także od budowy korpusu. Tym niemniej Michalska konstatuje, że żadnej z tych dwu form nie można uznać za regionalizm, co więcej, bywa, że obie występują w jednym tekście (Michalska, 2013: 92).

Przebieg zmiany pokazany został na Rys. 3.10. Udział obu form w kolejnych stuleciach okazał się dla nas zupełnie zaskakujący. Z lektury słowników historycznych można odnieść wrażenie, że będziemy mieli do czynienia z typowym przejściem jednej formy w drugą, tymczasem *abo* okazuje się być pewną rozpoczętą innowacją, która po przekroczeniu bariery 60% udziału we wszystkich wystąpieniach *abo* i *albo* uległa odwróceniu. Początkowy wzrost jest bardzo nierównomierny: udział tej formy w wielu punktach jest wyższy niż w poprzednim. Tym niemniej po roku 1500 notujemy stały wzrost, aż do momentu, kiedy w pierwszej połowie XVII wieku frekwencja używania formy osiągnęła około 60% zbioru wszystkich wystąpień. Szacowane prawdopodobieństwo, że w tekście trafimy na *abo*, wynosi w drugiej połowie tego wieku około 50%. Od tego momentu zaczyna się gwałtowny spadek udziału tej formy, choć jeszcze pojawia się ona w XIX wieku jako stylizacja archaiczna w *Pamiątkach Soplicy* Rzewuskiego czy gwarowa w *Diable* Kraszewskiego. Wydaje się, że w istocie rzeczy mamy tu do czynienia z dwoma nieco odrębnymi procesami: konkurencją synonimicznych *abo* oraz *albo* – tutaj stopniowo popularność zyskuje pierwsza z form – oraz procesem zastępowania *abo*, przy czym możemy spekulować, że była ona już odczuwana jako archaizm, stopniowo zastępowany przez formę *albo* odczuwaną jako innowacyjną.

Nieoczywisty przebieg konkurencji *abo* i *albo* najwygodniej będzie wyrazić modelem wielomianowej regresji logistycznej, który stosowaliśmy już powyżej. Dopasowanie modelu (zob. Rys. 3.10) do danych obserwowanych jest względnie wysokie, $R^2 = 0{,}825$, jeśli użyć wielomianu VI stopnia. Jest to wprawdzie model skomplikowany, wymagający użycia wielu parametrów i przez to trudny do obrony przed brzytwą Ockhama, ale mimo to przebiegający precyzyjnie przez środek danych empirycznych.

Z matematycznego punktu widzenia konkurencję *abo* i *albo* można sobie również wyobrazić jako złożenie dwu krzywych logistycznych, z których jedna byłaby lustrzanym odbiciem drugiej: w sytuacji idealnej krzywa typu *s* powinna przejść w krzywą typu *z* (Vulanović, 2007: 113). Na Rys. 3.11 przedstawiono dokładnie te same dane dotyczące *abo* i *albo*, ale tym razem zostały podzielone na dwie części i modelowane niezależnie od siebie przy pomocy klasycznej regresji logistycznej. Jak się okazuje, pierwsza część (1380–1610) poddaje się modelowaniu w dalece mniejszym stopniu ($R^2 = 0{,}47$) niż druga ($R^2 = 0{,}921$). Znacznie gorsze jednak jest to, że teoria mówiąca o przystających do siebie dwóch lustrzanych krzywych logistycznych nie ma pokrycia w faktach: w różnych przetestowanych przez nas



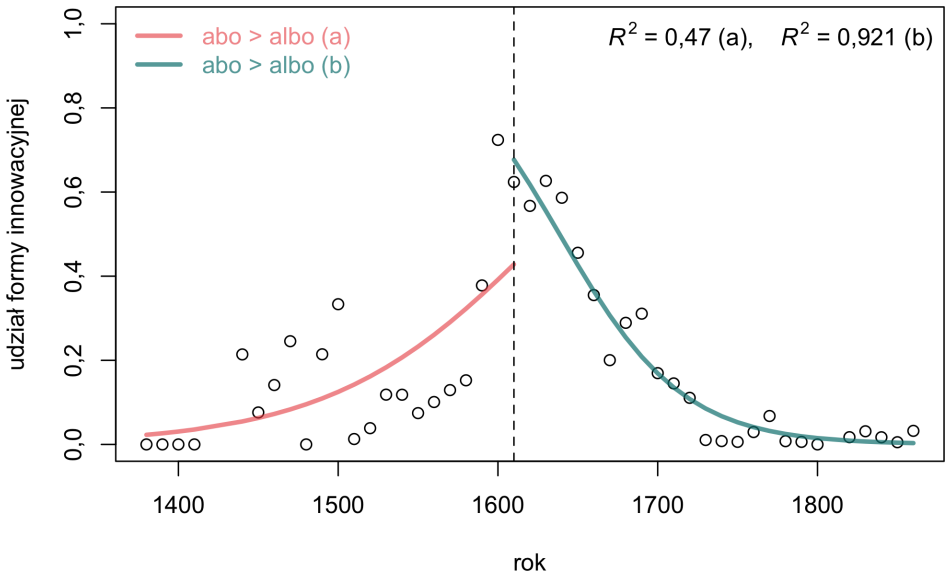
Rysunek 3.10. Konkurencja *abo* i *albo*: model wielomianowej regresji logistycznej VI stopnia.

scenariuszach oba najlepiej dopasowane modele nigdy nie utworzyły pary, bez względu na to, gdzie wyznaczaliśmy podział danych na dwie części.

Porównanie Rys. 3.10 i 3.11 pokazuje jednak jeszcze jedną rzecz – przy czym jest to rzecz wielkiej wagi. Mianowicie przyglądając się obu wykresom, trudno się oprzeć wrażeniu, że mamy tutaj do czynienia z dwoma zupełnie odmiennymi przebiegami zmiany *abo* > *albo*. Tymczasem jest to wyłącznie błąd oka, ponieważ dane (czarne punkty) na obu wykresach są identyczne. Problem ten, zwany z angielska *visual rhetoric bias*, polega na tym, że nałożone na wykres rozmaite linie, kolory czy kształty bardzo mocno wpływają na naszą percepcję rzeczywistych danych. Prowadzi nas to do dwóch kwestii. Po pierwsze, modele omawiane powyżej generują odmienne wyidealizowane trajektorie zmiany – w naszej sytuacji jest to albo przebieg dwugarbny (Rys. 3.10), albo jednogarbny (Rys. 3.11) – a wcale nie jest oczywiste, który z obu modeli jest bliższy prawdy. Po drugie, ocena dopasowania modelu gołym okiem jest podatna na przekłamania – jak to widzieliśmy powyżej – stąd konieczność kierowania się miarami dopasowania modelu do danych, takimi jak R^2 lub któryś z jego wariantów.

3.4.9. Wielkość podkorpusu

Zanim przejdziemy do wniosków scalających przestawione powyżej jednostkowe obserwacje, należy zadać pytanie o stabilność otrzymanych wyników. W ni-



Rysunek 3.11. Konkurencja *abo* i *albo*: dwa niezależne modele logistyczne obliczone osobno dla zakresów 1380–1610 oraz 1610–1850.

niejszym podrozdziale będziemy się zatem starali sprawdzić, czy przy wykorzystaniu tych samych danych wejściowych byłoby możliwe uzyskanie wyników znacząco innych niż te, które otrzymaliśmy.

Co najmniej trzy aspekty całej procedury mają bezpośredni wpływ na ostateczny kształt modelu. Po pierwsze, kształt krzywej logistycznej jest pochodną oszacowanych parametrów β_0 i β_1, dlatego należy zapytać o możliwość przekłamań na etapie ich obliczania. Tu akurat odpowiedź jest prosta: estymacja parametrów β_0 i β_1 polega na rozwiązaniu równania, które minimalizuje różnicę między modelem i danymi empirycznymi (James i in., 2013: 133–134). Istnieje jedna i tylko jedna para liczb spełniająca to równanie, nie ma zatem możliwości, by wynik został zmanipulowany czy „podrasowany" na tym etapie procedury. Dany zbiór wartości wejściowych zawsze da jedno optymalne rozwiązanie, nawet jeśli dla oka ludzkiego przebieg krzywej nie wygląda optymalnie.

Drugie miejsce, w którym można upatrywać potencjalnego źródła zafałszowań, to oczywiście same dane wejściowe. Była o tym mowa w wielu miejscach niniejszej pracy, w szczególności zaś w rozdziale na temat korpusu. Mimo że ten element procedury badawczej ma ogromny wpływ na wynik końcowy, niewiele mogliśmy w tej sprawie zrobić poza zebraniem tak dużej liczby tekstów, jak to było możliwe, i skontrolowaniu jakości transkrypcji w każdym z nich. W szczególności nie byliśmy jednak w stanie zapewnić reprezentatywności korpusu, a także wypełnić luk materiałowych w wypadku epok najdawniejszych. Wybraliśmy jed-

nak do bliższego oglądu takich siedem studiów przypadku (spośród bardzo wielu zmian w polszczyźnie), by miały stosunkowo dużą liczbę poświadczeń, nawet w najsłabiej reprezentowanych rejonach naszego korpusu.

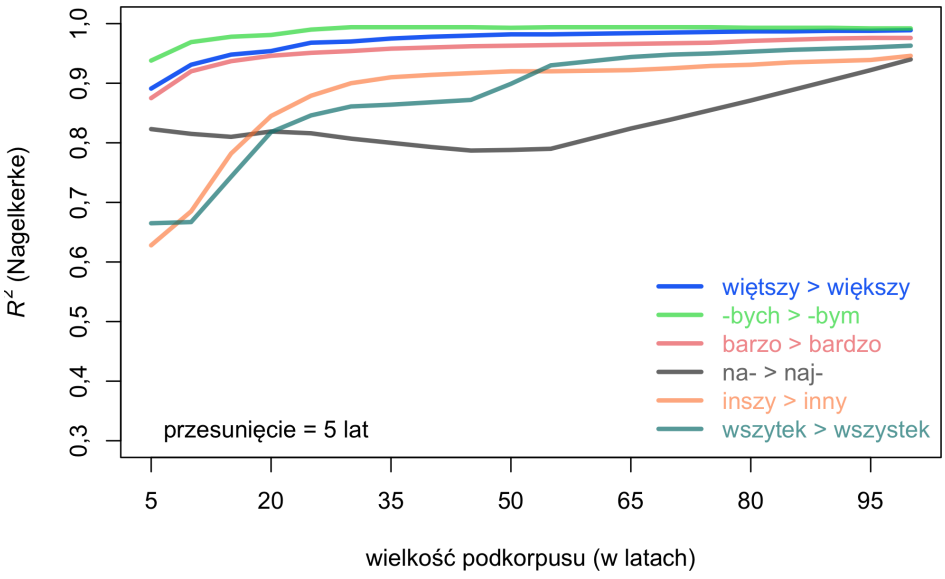
Trzecie wreszcie, z pozoru mało istotne, źródło możliwych przekłamań to nasza procedura dzielenia danych na mniejsze podkorpusy; pisaliśmy o tym szczegółowo w podrozdziale *Dane*. Obserwacja zjawisk diachronicznych zawsze wymaga podziału danych na mniejsze podokresy (takim naturalnym podokresem jest np. jeden rok), więc nie tyle sam fakt segmentacji danych można zakwestionować, ile wielkość pojedynczego podokresu. Należałoby tu podkreślić, że wielkość podkorpusu oraz przesunięcie (czyli gęstość próbkowania przy podziale podokresów „na zakładkę") to jedyne zmienne dobrane przez nas w sposób arbitralny. Staraliśmy się jednak, by ów arbitralny wybór miał empiryczne uzasadnienie, skoro nie sposób wyprowadzić go z założeń teoretycznych.

Przeprowadziliśmy zatem serię eksperymentów dla 20 różnych wielkości podkorpusu: 5, 10, 15, 20, . . . , 100 lat oraz 20 różnych wielkości „zakładki", modelując 6 zmian językowych omówionych powyżej, co dało razem 2400 modeli do porównania. W praktyce byliśmy w stanie obliczyć parametry dla jedynie 1200 modeli, ponieważ rozmiar „zakładki" nie może przekroczyć bieżącej wielkości podkorpusu. Dla każdego z modeli obliczaliśmy następnie stopnień jego dopasowania do danych R^2. Przykładowe wyniki dla podkorpusu pięćdziesięcioletniego z przesunięciem 20 lat oraz dwudziestoletniego z przesunięciem 5 lat przedstawia Tab. 3.2.

Tabela 3.2. Dopasowanie modelu logistycznego do danych empirycznych (wartość R^2) przy różnych sposobach podziału danych na podkorpusy.

zmiana	podkorpus 50 lat	podkorpus 20 lat
więtszy > *większy*	0,984	0,954
-bych > *-bym*	0,995	0,981
barzo > *bardzo*	0,963	0,946
na- > *naj-*	0,837	0,819
inszy > *inny*	0,917	0,845
wszytek > *wszystek*	0,867	0,818
-ir- > *-er-*	0,914	0,880

Trudno będzie omówić wyniki uzyskane we wszystkich 1200 testach, dlatego na Rys. 3.12–3.13 pokazujemy w sposób syntetyczny stopień dopasowania modelu w zależności od wielkości podkorpusu – przy wielkości „zakładki" pięcioletniej na Rys. 3.12 i przy dwudziestopięcioletnim przesunięciu na Rys. 3.13. Jak wspomnieliśmy, podkorpusy obejmujące swym zasięgiem długi okres co do zasady

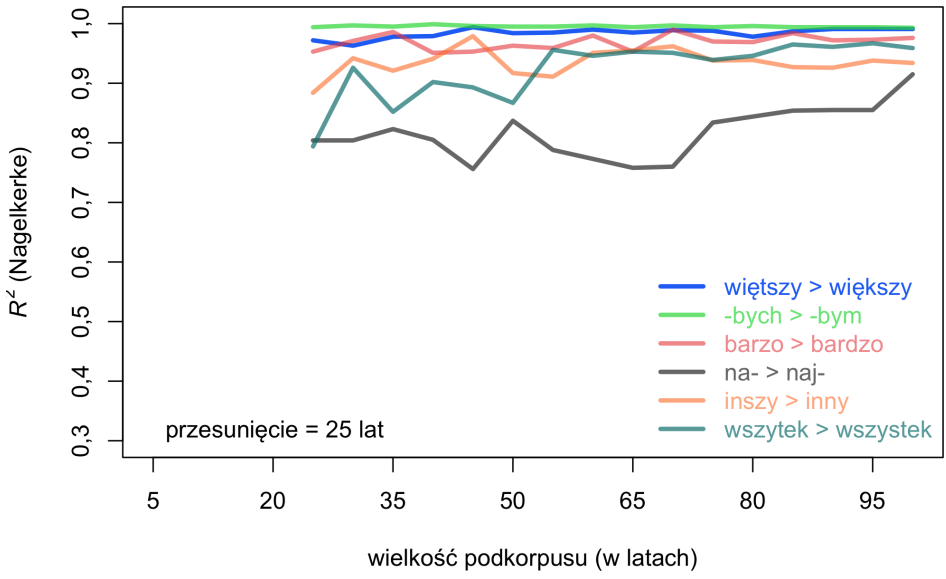


Rysunek 3.12. Wpływ wielkości podkorpusu na dopasowane danych oczekiwanych do danych empirycznych (przesunięcie pięcioletnie).

dają stabilniejsze wyniki, gdyż produkcja poszczególnych autorów w ich obrębie się uśrednia. Z drugiej strony chcielibyśmy otrzymać możliwie precyzyjne dane o dużej ziarnistości. Wynika z tego, że poszukiwana przez nas optymalna wielkość podkorpusu to najmniejszy okres czasu, przy którym współczynnik R^2 zachowuje się stabilnie. Z wykresu na Rys. 3.12 wynika jasno, że taka optymalna wielkość to mniej więcej 20 lat: powyżej tej wartości model nie staje się dokładniejszy (wyjątkiem zmiana *wszytek* > *wszystek*), a jedynie tracilibyśmy ziarnistość obserwowanych wyników (czyli gęstość punktów na Rys. 3.1–3.11).

Nasuwa się kilka dodatkowych wniosków. Pierwszy jest dość oczywisty – tam, gdzie dane empiryczne są bliskie danym oczekiwanym (a więc w naszym wypadku przy $R^2 > 0{,}9$), wielkość podkorpusu nie ma praktycznie znaczenia, co pokazuje stabilna wartość współczynnika R^2. Jeśli jest jakiś powód, by zmniejszyć wielkość podkorpusów, to jest nim ziarnistość wyników. Po drugie, porównanie Rys. 3.12 i 3.13 pokazuje intuicyjną skądinąd prawdę, że gęste próbkowanie (a więc mała „zakładka") daje bardziej wygładzone wyniki niż duże przesunięcia. Wykres na Rys. 3.13 zdaje się jednak sugerować, że już przy 25 latach przesunięcia dopasowanie modeli R^2 zaczyna się wahać znacząco.

Wreszcie, dzięki powyższej serii systematycznych testów zyskaliśmy empiryczne uzasadnienie dla obranej przez nas arbitralnej wielkości podkorpusu – 20 lat. Poniżej tej wartości załamanie współczynnika zgodności danych z warto-



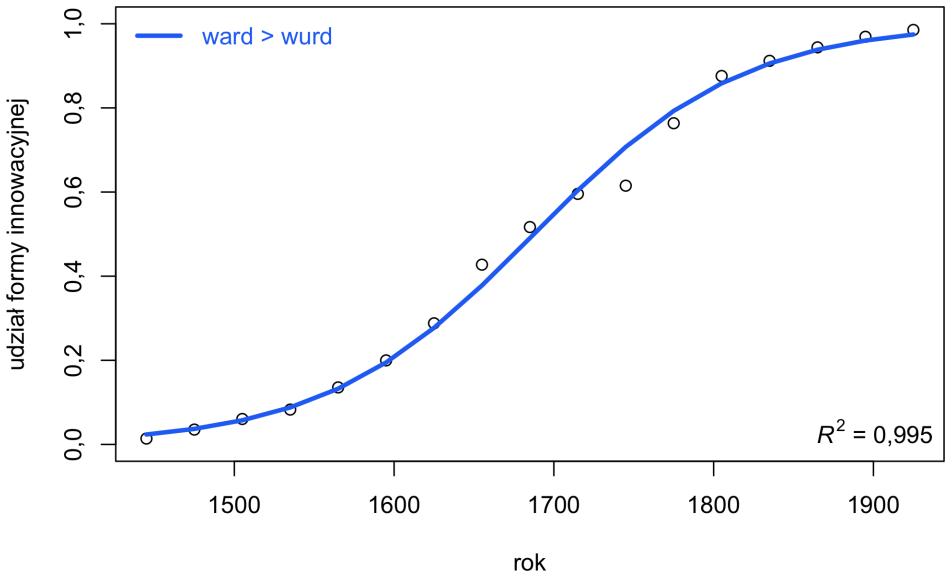
Rysunek 3.13. Wpływ wielkości podkorpusu na dopasowane danych oczekiwanych do danych empirycznych (przesunięcie dwudziestopięcioletnie).

ściami oczekiwanymi jest wtedy niemal zawsze dość wyraźne. O ile pozostałe wnioski być może są specyficzne dla naszego korpusu, o tyle ostatni – zaryzykowalibyśmy to stwierdzenie – ma szersze zastosowanie, gdyż wykazuje, że okres 10 lat jest zbyt krótki, by móc obserwować zmianę w normie językowej.

3.4.10. Dodatek: *ward* > *wurd(e)*

Tytułem uzupełnienia przeprowadziliśmy jeszcze jedną serię testów, w których wymodelowaliśmy za pomocą regresji logistycznej zmianę *ward* > *wurd* i epentezę *-e* w *wurde*, a także regułę składniową nakazującą wstawianie peryfrastycznego *do* przy zdaniach pytajnych i przeczących (Best, 1983; Ellegård, 1953). Ponieważ jednak dane zaczerpnęliśmy nie bezpośrednio z korpusów, ale z publikacji, musieliśmy przyjąć taką wielkość podkorpusu, jaką ustalili ich autorzy. Dane te były punktem wyjścia dla modelowania przebiegu zmiany za pomocą tej samej procedury, którą posługiwaliśmy się do modelowania zmian znanych z historii polszczyzny. Celem owych dodatkowych testów było, po pierwsze, powtórzenie oryginalnych eksperymentów sprzed kilku dekad, a po drugie sprawdzenie, czy obrana przez nas metoda klasycznej regresji logistycznej daje podobne wyniki do modelu zaproponowanego przez Piotrowskiego-Altmana.

Zdecydowanie najlepsze dopasowanie wyników obserwowanych do oczekiwanych wykazuje zmiana *ward* > *wurd* z wartością R^2 na poziomie 0,995.



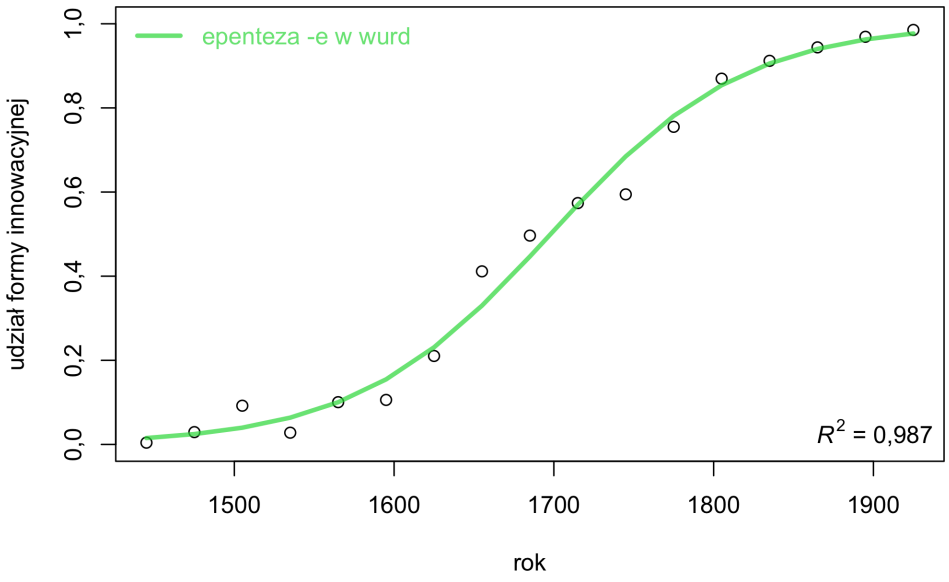
Rysunek 3.14. Przebieg zmiany *ward* > *wurd*.

Przedstawiliśmy ją na Rys. 3.14. W porównaniu do omówionych powyżej zmian w rozwoju polszczyzny uwagę zwraca bardzo łagodny przebieg krzywej dla *ward* > *wurd*, świadczący o tym, że zmiana zachodziła bardzo powoli na przestrzeni wielu stuleci.

Minimalnie niższą wartość R^2 = 0,987 wykazuje epenteza *-e* w omawianej formie (Rys. 3.15), choć łagodny kształt krzywej i powolne tempo zmiany są niemal identyczne jak w przykładzie omówionym bezpośrednio powyżej.

Dane zebrane przez Ellegårda odbiegają od oczekiwanego przebiegu w sposób bardziej znaczący. Zacznijmy od tego, że dostarczone przez niego dane nie podają punktów, w których 100% wystąpień to konstrukcje innowacyjne. Spośród trzech procesów, które zostały zakończone, tj. wprowadzenia *do* do zdań zaprzeczonych, zaprzeczonych pytajnych i pytajnych twierdzących, najbliższy oczekiwanemu jest przebieg zdań pytajnych twierdzących (R^2 = 0,965), dalej zdań pytajnych zaprzeczonych (R^2 = 0,871) i zdań oznajmujących zaprzeczonych (R^2 = 0,825).

Nasuwają się tu następujące wnioski: niemieckie zmiany w obrębie morfologii dają się bardzo precyzyjnie modelować za pomocą regresji logistycznej. Dopasowanie modelu do danych, a więc wartość współczynnika R^2 dla zmiany *ward* > *wurd*, przewyższa dopasowanie wszystkich polskich danych, w wypadku epentezy *-e* jej wartość R^2 jest porównywalna jedynie ze zmianą *-bych* > *-bym*. Z kolei dane angielskie znacznie bardziej oddalają się od wartości oczekiwanych, niż ma to miejsce w odniesieniu do danych polskich. O ile więc w zmianach



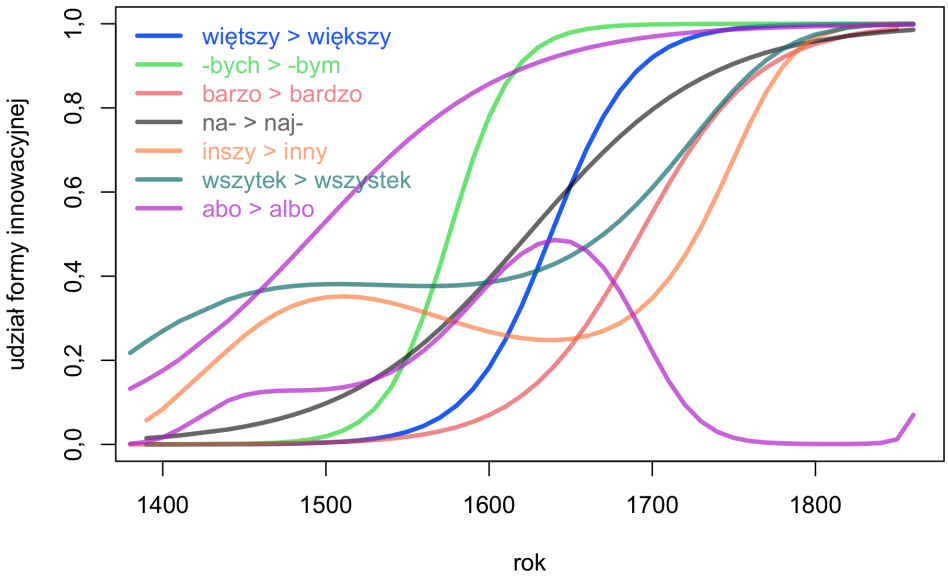
Rysunek 3.15. Przebieg zmiany epentezy *-e* w *wurd*.

niemieckich nie należy się doszukiwać dodatkowych mechanizmów, które by nią sterowały, o tyle w wypadku *do* bardziej nieprzewidywalny przebieg zmiany wymaga rozpoznania czynników odpowiedzialnych za jej przebieg.

3.5. Podsumowanie

Po przedstawieniu kilku zmian w polszczyźnie, z których część jest wzorcową realizacją procesu Piotrowskiego, a część z pewnym trudem poddaje się modelowaniu, należy zadać pytanie o wartość dodaną takich operacji na danych diachronicznych – wychodzimy bowiem z założenia, że interesuje nas coś więcej niż prezentacja ciekawostki naukowej lub przyczynkowego komentarza do opisu procesu zmiany językowej.

Pierwszą i podstawową wartością dodaną, której upatrujemy w modelowaniu zmiany językowej, jest tendencja modelu do znajdowania lokalnych uśrednień danych empirycznych. Mówiąc prościej: podczas gdy obserwator-badacz podświadomie skupia uwagę na poszczególnych faktach językowych, model jest w stanie „dostrzec" prawidłowości na wyższym poziomie uogólnienia. Z tego punktu widzenia jest rzeczą oczywistą, że jeśli w danych empirycznych w niektórych podkorpusach udział formy innowacyjnej jest niższy niż w poprzednim, nie podważa to wartości metody. Wręcz przeciwnie – wyznaczone na podstawie pełnego zbioru danych prawdopodobieństwo pozwala uchwycić proces jako ca-



Rysunek 3.16. Przebieg siedmiu zmian w średniopolszczyźnie.

łość, bez względu na jego lokalne niekonsekwencje. Ujmując rzecz jeszcze inną metaforą: badacz-językoznawca ma skłonność do skupiania się na pojedynczych drzewach, podczas gdy model pozwala dostrzec cały las.

Drugi powód korzystania z modeli wynika bezpośrednio z pierwszego. Jeśli model sam w sobie jest uogólnieniem, to zestawienie obok siebie większej liczby modeli będzie uogólnieniem wyższego stopnia, pewnego rodzaju wstępem do całościowego spojrzenia na zmianę językową. We wszystkich bez mała dotychczasowych ujęciach diachronicznych polszczyzny poszczególne procesy były zawsze rozpatrywane w odosobnieniu. Pewnym wyjątkiem jest niedawne studium na temat składni staropolskiej, w którym autorzy proponują spojrzeć na rozwój języka przez pryzmat ogólnych tendencji rozwojowych, niekoniecznie skupiając uwagę na zjawiskach jednostkowych (Krążyńska, Mika i Słoboda, 2015). Wydaje się, że dzięki nałożeniu na siebie większej liczby modeli, tak jak to czynimy na Rys. 3.16, będziemy w stanie dostrzec niektóre z owych tendencji.

Na wykresie naniesiono prawdopodobieństwa wstąpienia różnych form recesywnych i innowacyjnych w funkcji czasu. Dopiero spojrzenie na symfonię poszczególnych modeli – a może raczej kakofonię? – uzmysławia, jak odmienna może być dynamika procesu diachronicznego. Gdy *na-* > *naj-* jest powolnym, rozciągniętym w czasie procesem, tak z kolei *-bych* > *-bym* jest zmianą gwałtowną, podobnie jak *więtszy* > *większy*. To, co różni te procesy, to fakt, że dzieją się one w innym czasie – gdy prawdopodobieństwo wystąpienia w tekście *-bych*

spada do wartości 0,5, prawdopodobieństwo napotkania formy *więtszy* jest już bliskie 1. Można to ująć nieco prościej: gdy pierwszy z powyższych procesów jest mniej więcej w połowie, drugi się dopiero zaczyna. Jeśli z kolei *-bych* > *-bym* odniesiemy do *barzo* > *bardzo*, to stwierdzimy, że ów drugi proces jest jeszcze późniejszy (zaczyna się wtedy, gdy *-bych* znika z systemu) i dodatkowo nieco wolniejszy, co widać z mniejszego nachylenia modelowanej linii. Wszystko to daje się zaobserwować, jeżeli zrobimy odpowiednią wizualizację danych.

Na koniec łyżka dziegciu. Na podstawie wyników naszych eksperymentów można powątpiewać, czy mamy tu do czynienia z jakimkolwiek lingwistycznym prawem ilościowym, w takim sensie jak mówimy o prawie Zipfa, Menzeratha czy Heapsa. W klasycznym bowiem rozumieniu kwantytatywne prawo językowe, by mogło być takim nazwane, musi zachodzić w każdych warunkach i w każdym języku – słowem, musi być uniwersalne. Tymczasem tzw. prawo Piotrowskiego dotyczy tylko pewnej grupy zmian językowych. Warto też powtórzyć za Stachowskim (w druku), że prawo to (w przeciwieństwie do wielu innych praw lingwistyki kwantytatywnej) nie dotyczy w istocie języka i jego samoorganizacji, ale związanych z językiem procesów społecznych.

Rozdział 4

Klasyfikacja maszynowa w językoznawstwie diachronicznym

4.1. Wprowadzenie

Klasyfikacja maszynowa jest jedną z najgwałtowniej rozwijających się dziedzin współczesnej statystyki; pojęcie to stosuje się do całej grupy metod analizy danych ilościowych. Mowa tutaj zarówno o prostych technikach eksploracji danych, jak i metodach tzw. uczenia maszynowego (ang. *machine learning*), których z kolei podzbiorem są algorytmy sztucznej inteligencji (ang. *artificial intelligence*, w skrócie AI), wykorzystujące zaawansowane sieci neuronowe, zwane też czasami sieciami uczenia głębokiego (ang. *deep learning*). Wszystkie te techniki łączy zdolność do znajdowania podobieństw pomiędzy elementami zbioru danych i grupowanie ich we względnie jednorodne klasy.

Dzięki metodom klasyfikacyjnym jesteśmy na przykład w stanie wykryć preferencje polityczne użytkowników sieci społecznościowych, przeprowadzić wczesną diagnozę różnych schorzeń na podstawie analizy materiału DNA (tzw. ekspresji genów), umożliwić autoryzację usług bankowych przez automatyczną analizę odcisku palca, wykryć nieznane wcześniej substancje niebezpieczne na podstawie ich wzoru chemicznego itp. Przechodząc na grunt bliższy filologii: jeśli elementami zbioru będą teksty z korpusu, możemy użyć metod klasyfikacyjnych do znalezienia różnic pomiędzy – przykładowo – stylem kobiet i mężczyzn albo ustalić autorstwo anonimowego tekstu. Jeśli badanymi elementami będą słowa w korpusie, możemy użyć technik klasyfikacyjnych do znalezienia niewidocznych gołym okiem relacji semantycznych między słowami i użyć tak wytrenowanego modelu w generowaniu języka naturalnego. Możemy wreszcie użyć automatycznej klasyfikacji do wytrenowania tagera morfosyntaktycznego (o tagowaniu pisaliśmy w rozdziale *Korpus*) albo do odnalezienia relacji składniowych między elementami zdania. W niniejszym rozdziale spróbujemy użyć technik klasyfikacyjnych do wykrycia zmian chronologicznych w języku. Zanim jednak przejdziemy do poszczególnych eksperymentów, poprzedzimy wywód krótkim wprowadzeniem do problematyki klasyfikacji.

Jedną z własności metod klasyfikacyjnych jest ich zdolność do porównywania obiektów pod względem wielu ich cech jednocześnie. Zilustrujmy to hipotetycz-

nym przykładem z zakresu filologii, w którym „obiekty" będą oznaczać teksty, a „cechy" – używane w tekstach słowa. Załóżmy, że chcielibyśmy się dowiedzieć, czy felietony i reportaże w gazecie różnią się od siebie słownictwem. Nie chodzi tu nawet o wyraźnie odrębną leksykę, tj. słowa, które występują jedynie w jednym z tych gatunków dziennikarskich, ale również (a może przede wszystkim) o słownictwo wspólnoodmianowe, czyli takie, którym posiłkują się twórcy w obu gatunkach. W tym wypadku teksty może różnicować nie tyle obecność bądź nieobecność pewnych wyrazów, ile ich częstość – w jednym typie ich frekwencja będzie odmienna niż w drugim. By przetestować naszą hipotezę, wybierzemy z korpusu pewną liczbę felietonów i reportaży (ściślej: tekstów oznaczonych jako felietony i reportaże przez autorów korpusu). Każdy z tekstów potraktujemy jako odrębny obiekt, bo na tym etapie jeszcze nie wiemy, czy rzeczywiście oba gatunki różnią się między sobą. Następnie dokonamy wyboru cech (w tym wypadku leksyki), które mają różnicować teksty. Tworzymy dla każdego tekstu listę frekwencyjną i porównujemy częstość każdego z elementów owej listy frekwencyjnej. Najczęstszym słowem w tekstach będzie zapewne *w*. Bierzemy więc liczbę wystąpień tego słowa (ze względów praktycznych częstość wystąpień dzielimy dodatkowo przez długość całego tekstu, żeby uzyskać tzw. frekwencję względną) i kolejno porównujemy każdy tekst z każdym tekstem. To samo robimy z kolejnym słowem (będzie to zapewne *i* albo *się*), z którym czynimy to samo itd. Opisane powyżej kolejne kroki procedury można by – wprawdzie niemałym nakładem czasu – wykonać ręcznie, za pomocą kartki i ołówka. Wielką zaletą metod klasyfikacyjnych jest jednak ich zdolność do operowania na olbrzymich zbiorach danych w bardzo krótkim czasie, a co dla naszego wywodu ważniejsze, są one w stanie uogólnić pojedyncze różnice między cechami (tutaj: słowami), by określić stopień podobieństwa pomiędzy poszczególnymi tekstami. Celowo używamy tu nieco niejasnego wyrażenia „uogólnić", ponieważ w większości metod wielowymiarowych chodzi o coś więcej niż o prostą sumę tych różnic.

Wróćmy jednak do cech różnicujących, czyli dyskryminatorów. Najoczywistszym ich rodzajem jest rzecz jasna słownictwo, a mówiąc precyzyjniej: częstość użycia poszczególnych słów w tekstach. Oczywiście sprawa zaczyna się komplikować, gdy zapytamy, co konkretnie należy rozumieć pod pojęciem „słownictwo". Czy chodzi o wszystkie słowa użyte przez danego autora, czy tylko o jakiś ich podzbiór? Czy powinniśmy rozpatrywać słowoformy w takiej postaci, w jakiej pojawiły się w tekście, czy może raczej wyrazy sprowadzone do formy hasłowej, czyli lematy? Czy w jakiś sposób wyszczególniać jednostki wielowyrazowe? Czy formy homonimiczne (np. spójnik *ale* i partykułę *ale*) rozgraniczać podczas obliczania frekwencji?

Przeróżne studia empiryczne zdają się sugerować przewagę podejścia zgrubnego nad precyzyjnym definiowaniem warstwy leksykalnej w tekstach. Od czasu rozprawy Mostellera i Wallace'a na temat atrybucji autorskiej esejów z cyklu *The Federalist Papers* wiadomo, że bardzo silnym dyskryminatorem są słowa

funkcyjne (synsemantyczne), takie jak spójniki, partykuły, rodzajniki czy zaimki (Mosteller i Wallace, 1964). W połączeniu z wcześniejszymi obserwacjami George'a Zipfa (1949), że słowa funkcyjne są najczęstszymi jednostkami leksykalnymi w każdym języku i w każdym tekście, otworzyło to drogę do niezliczonych prac stylometrycznych badających częstość użycia kilkudziesięciu wyrazów z góry listy frekwencyjnej, często nawet bez jakiejkolwiek ewaluacji tych dyskryminatorów. W konsekwencji zamiast pracowicie rozgraniczać różne leksemy kryjące się za angielską formą *to*, na ogół po prostu zlicza się mechanicznie wszystkie wystąpienia dwuznakowego łańcucha „t" + „o" otoczonego spacjami i tak samo czyni się z pozostałymi „wyrazami". Choć z językoznawczego punktu widzenia można mieć zasadne wątpliwości co do powyższej procedury, kontrolowane eksperymenty atrybucyjne – czyli takie, w których bada się zachowanie metod klasyfikacyjnych na tekstach o znanym autorstwie – pokazują wysoką skuteczność takiego zgrubnego podejścia.

Podobnie rzecz się ma, jeśli chodzi o problem lematyzacji, a więc odfiltrowania fleksji poprzez sprowadzenie wyrazów tekstowych do ich form podstawowych. Nie ma w gruncie rzeczy silnych teoretycznych przesłanek, by przedkładać lematy nad słowoformy (czy odwrotnie), zostają więc podejścia eksperymentalne, pozwalające porównać siłę dyskryminacyjną obu rodzajów dyskryminatorów. Te zaś nie dają jednoznacznych odpowiedzi, okazuje się bowiem, że na przykład dla języka polskiego oba typy znaczników dają mniej więcej takie same wyniki (Eder i Górski, w druku).

Nie tylko pojedyncze słowa, ale i sekwencje kilku następujących po sobie wyrazów tekstowych (ewentualnie lematów) mogą stanowić znacznik stylu. Milczące założenie jest tutaj oczywiście takie, że przecież język nie sprowadza się wyłącznie do sumy frekwencji słów; wręcz przeciwnie, to linearność w języku definiuje zarówno jego własności składniowe, jak i znaczenie poszczególnych wyrazów – choć w tym drugim wypadku chodzi może nie tyle o linearność, ile o bezpośredni kontekst danego słowa. Biorąc pod uwagę sekwencje dwóch lub trzech bezpośrednio sąsiadujących wyrazów[1], dostajemy wgląd nie tylko w samo słownictwo, ale także pośrednio w związki frazeologiczne, które z rozpatrywanego punktu widzenia są powtarzającymi się zbitkami pojedynczych wyrazów.

Innym potencjalnym dyskryminatorem są kategorie gramatyczne – wtedy zamiast analizować frekwencję konkretnych wyrazów, analizuje się liczbę przyimków, rzeczowników, czasowników itp. Można wreszcie badać względną frekwencję zdań o danej długości lub wybranych fraz o danej strukturze, można stosować różne miary bogactwa leksykalnego, koherencji, stopnia nacechowa-

[1] W języku polskim jest niewiele sekwencji czterech (a tym bardziej więcej niż czterech) wyrazów, które się powtarzają na tyle często, by móc stanowić dobry dyskryminator. Większość z nich pojawia się w tekstach jednokrotnie, w związku z tym nie mogą stanowić dobrej charakterystyki ilościowej utworu.

nia emocjonalnego tekstu itd. Zasadniczo dowolna dająca się policzyć cecha językowa lub stylistyczna może służyć jako dyskryminator, choć oczywiście nie każda będzie równie skutecznie różnicować teksty. Współczesne podejścia eksperymentalne sugerują, że tym, co najsilniej odróżnia teksty, jest jednak leksyka; jeśli za dyskryminatory przyjmiemy części mowy, skuteczność atrybucji spada (Baayen, Van Halteren i Tweedie, 1996). Z kolei dla identyfikacji autora wiersza dość skutecznym dyskryminatorem jest rytm (Plecháč, Bobenhausen i Hammerich, 2018).

Podejmowano też próby atrybucji na podstawie cech składniowych. Bezpośrednie odwołanie się do struktury składniowej tekstu jest dość kłopotliwe, ponieważ wiarygodna informacja na ten temat wciąż nie daje się pozyskać automatycznie. Po to, by analizować cechy składniowe tekstu, potrzebny jest tzw. bank drzewek, to jest korpus, w którym w sposób eksplicytny zaznaczono strukturę gramatyczną zdań. Uzyskuje się ją na drodze anotacji dokonywanej – albo przynajmniej weryfikowanej – przez człowieka. W konsekwencji dostępność tekstów z taką informacją jest minimalna. Istnieje bank drzewek dla stadiów historycznych języka angielskiego, mianowicie *Penn Parsed Corpora of Historical English* (Taylor i Kroch, 1994), wydaje się jednak, że poloniści będą musieli długo poczekać na tego rodzaju narzędzie[2].

Pewne przybliżenie struktury składniowej tekstu daje odwołanie się do sekwencji etykietek kategorii gramatycznych (Hirst i Feiguina, 2007). Zaopatrzenie tekstu w taką informację jest proste i dokonywane automatycznie przez tager (zob. rozdział *Metody korpusowe i kwantytatywne w językoznawstwie historycznym*). Na potrzeby takich badań zdanie:

Ala ma kota.

jest reprezentowane przez sekwencję:

rzeczownik czasownik rzeczownik

Informacja ta może być wzbogacona o bardziej precyzyjny opis kategorii fleksyjnych, np.

(rzeczownik + mianownik + liczba pojedyncza) (czasownik + czas teraźniejszy + liczba pojedyncza) (rzeczownik + biernik + liczba pojedyncza)

W konwencji stosowanej w NKJP wygląda to następująco:

[subst:sg:gen:f] [fin:sg:ter:imperf] [subst:sg:acc:m2]

Mimo że taka sekwencja nie daje bezpośredniej informacji składniowej, to zauważmy, że jeśli autor tekstu ma skłonność do tworzenia długich fraz nomi-

[2] Poczekać będą musieli na bank diachroniczny; dla współczesnej polszczyzny bank drzewek istnieje już od pewnego czasu (zob. Wolinski i in., 2011).

nalnych, to zapewne tekst będzie obfitował w sekwencje rzeczowników przeplatanych przymiotnikami; predylekcja do zdań względnych objawi się większą liczbą sekwencji zawierających spójniki, zaś umieszczanie orzeczenia na końcu zdania zdradzą sekwencje biernik + czasownik w formie osobowej + znak granicy zdania. Procedura wydobywania z tekstu każdej możliwej sekwencji dwóch czy trzech następujących po sobie elementów (nazywa się je odpowiednio bigramami i trigramami) niosących informację o częściach mowy nie jest skomplikowana: najpierw dezambiguator przypisuje do wszystkich słów w oryginalnym tekście ich reprezentacje gramatyczne, następnie inny program przesuwa „okienko" o szerokości dwóch (dla bigramów) bądź trzech (dla trigramów) wyrazów przez tekst i zlicza wszystkie znajdujące się w owym ruchomym „okienku" reprezentacje gramatyczne słów.

Powyższa idea pośredniego wglądu w strukturę składniową daje się stosunkowo łatwo zastosować do tektsów synchronicznych; w wypadku danych diachronicznych pojawia się jednak dodatkowa przeszkoda. Milcząco zakładamy bowiem, że o ile wykładniki formy fleksyjnej mogą się zmieniać, o tyle reprezentują one stabilne kategorie. Innymi słowy spodziewamy się, że formy *nawiętszy* i *największy* będą przez tager opatrzone tą samą etykietką gramatyczną. I choć co do zasady powyższa reguła jest prawdziwa, pamiętajmy, że w procesie historycznym dochodzi do powstania bądź zaniku całych kategorii fleksyjnych, w związku z czym pewne etykiety gramatyczne nie dają się użyć do wszystkich epok. I tak nie można przypisać tej samej kategorii formom *dwa słowa* i *dwie słowie*. W pierwszym wypadku mamy do czynienia z liczbą mnogą, w drugim – z podwójną. Tym bardziej dotyczy to sekwencji *para oczu*: forma *oczu* będzie interpretowana jako liczba podwójna do pewnego momentu, potem już jako liczba mnoga. Piszemy „do pewnego momentu", gdyż moment ten musi być w gruncie rzeczy arbitralnie wybrany przez twórcę korpusu. W efekcie sekwencja dwu takich samych lematów będzie (a w każdym razie powinna być) tagowana odmiennie w zależności od daty powstania tekstu. Nie jest więc tak, że tagowanie jest całkowicie nieczułe na zmiany w obrębie morfologii. Tym niemniej z praktycznego punktu widzenia możemy ten problem zignorować, ponieważ liczba tego rodzaju zjawisk jest w tekście bardzo ograniczona.

Mogłoby się wydawać, że struktury składniowe będą lepszym dyskryminatorem niż warstwa leksykalna tekstów, podejścia eksperymentalne wskazują jednak, że bigramy lub trigramy etykietek gramatycznych nie dają tak dokładnych wyników jak najczęstsze wyrazy. Instruktywny dla dalszych rozważań będzie eksperyment klasyfikacyjny przeprowadzony na dwudziestowiecznych tekstach literackich w języku polskim (Eder i Górski, w druku). Klasyfikacji poddano 500 polskich powieści dobranych tak, by jeden autor był reprezentowany przez 3–5 pozycji. Zadaniem było ustalenie autorstwa każdego z tekstów. Warunki tego zadania były bardzo trudne, ponieważ liczba klas przekraczała sto (procedura musiała prawidłowo „odgadnąć" właściwego autora z ponad setki autorów repre-

zentowanych w korpusie). Jednym z celów tego eksperymentu było sprawdzenie, w jakim stopniu za rozróżnienie autorów odpowiada leksyka (już to w postaci wyrazów tekstowych, już to w postaci lematów), a w jakim stopniu sygnał autorski kryje się w składni, której przybliżeniem są opisane wyżej bigramy i trigramy etykietek gramatycznych wyrazów. Wyniki pokazały, że klasyfikacja na podstawie gramatyki jest mniej skuteczna niż na podstawie leksyki, ale spadek nie był dramatycznie duży, wynosił bowiem około 10 punktów procentowych. Wniosek ten jest istotny dla naszej dalszej argumentacji, okazuje się bowiem, że ślad idiolektu autorskiego objawia się nie tylko w leksyce, ale także, choć w mniejszym stopniu, w składni. Fakt, że te dwa typy dyskryminatorów różnią się swoją skutecznością, można wytłumaczyć jedynie tym, że swoboda w zakresie wyboru słownictwa jest większa niż to ma miejsce w odniesieniu do gramatyki.

Tyle o autorstwie. Jest jednak rzeczą oczywistą, że skoro pewne cechy języka da się w sposób zobiektywizowany powiązać z idiolektem, będą istniały też inne cechy, przez które uwidoczni się styl funkcjonalny (Biber, 2012), gatunek (Schöch, 2017; Stamatatos, Fakotakis i Kokkinakis, 2000), ślad tłumacza (Rybicki, 2012), płeć użytkownika języka (Pennebaker, 2011), wiek autora (Hoover, 2007), a nawet schorzenia neurologiczne w rodzaju demencji (Le, Lancashire, Hirst i Jokel, 2011). Nas rzecz jasna będą interesowały znaczniki języka (czy stylu) odpowiedzialne za zróżnicowanie czasowe badanych tekstów (Stamou, 2008). Jeśli – jak dowodzą liczne rozprawy – sygnał autorski ukrywa się w niewielkiej stosunkowo liczbie wyrazów funkcyjnych, być może predyktorem sygnału diachronicznego okaże się jakaś inna warstwa języka, np. składnia? Postaramy się znaleźć odpowiedź na to pytanie w części eksperymentalnej niniejszego rozdziału.

Kilka stron wcześniej napisaliśmy, że automatyczna klasyfikacja tekstów polega na obliczeniu różnic frekwencji cech (znaczników stylu) w poszczególnych tekstach i następnie oszacowaniu „uogólnionej" różnicy między tekstami. Pora przyjrzeć się uważniej temu etapowi procedury. Metody statystyczne, które będziemy tu mieli na myśli, nazywane są wielowymiarowymi właśnie dlatego, że nie rozpatrują pojedynczych cech w izolacji, lecz starają się znaleźć porządek we frekwencjach wielu cech jednocześnie. Ich „wielowymiarowość" polega zaś na tym, że problem uogólnionej różnicy najprościej jest rozwiązać metodami geometrycznymi – za pomocą obliczenia odległości między punktami w przestrzeni kartezjańskiej, reprezentującymi badane cechy. Skoro jednak liczba cech znacznie przekracza trzy, dokonuje się obliczeń w czysto teoretycznej przestrzeni kilkudziesięcio- czy kilkusetwymiarowej (o liczbie wymiarów równej liczbie cech), zachowującej algebraiczne własności przestrzeni kartezjańskiej. Nie będziemy wchodzić w matematyczne szczegóły różnych sposobów obliczania „odległości" w wielowymiarowej przestrzeni, odsyłając do bogatej literatury przedmiotu (zob. np. Baayen, 2009; James i in., 2013; Moisl, 2014).

W niemałym repertuarze metod wielowymiarowych można wyróżnić dwa z gruntu odmienne podejścia. Jedno polega na tym, że algorytm wyszukuje podo-

bieństwa pomiędzy wszystkimi elementami zestawu (np. tekstami albo organizmami żywymi). Podejście to nazywamy klasyfikacją nienadzorowaną. Efektem może być na przykład wykres grupujący teksty, który następnie należy poddać ocenie obserwatora-badacza (zob. np. Rys. 4.1). Czyni się tym samym założenie, że wszystkie zgromadzone dane „same się wypowiedzą" (ang. *the data will speak for themselves*).

W wypadku danych językowych, gdy testowaną hipotezą badawczą jest np. zróżnicowanie gatunkowe reportaży i felietonów, klasyfikacja mogłaby wyglądać tak, że jakiś, najlepiej możliwie duży, zbiór przykładowych reportaży i felietonów poddajemy analizie skupień albo skalowaniu wielowymiarowemu i następnie oceniamy, do jakiego stopnia felietony tworzą jedną grupę, a reportaże drugą. W ten sposób interpretacja wyników grupowania odbywa się niejako *ex post*, bo musimy porównać wyniki klasyfikacji automatycznej z wiedzą aprioryczną na temat obu gatunków. W omawianym przypadku powinniśmy każdy z analizowanych tekstów oznaczyć jako felieton lub reportaż; pozwoli to ustalić liczbę błędów w analizie automatycznej i tym samym ocenić wiarygodność metody.

Zauważmy, że z samej istoty metod klasyfikacyjnych wynika to, że zawsze podzielą one analizowany materiał na jakieś grupy; nie znaczy to wszakże, że grupowanie takie daje się sensownie zinterpretować. Zarazem metody klasyfikacji nienadzorowanej są wrażliwe na algorytm, według którego oblicza się podobieństwo tekstów. Niejednokrotnie dzieje się tak, że zmiana metody skutkuje odmiennym podziałem na grupy i doprawdy trudno powiedzieć, który z nich jest „prawdziwszy". Podobnie rzecz się ma z dodaniem bądź ujęciem kolejnych cech, np. wynik klasyfikacji może być odmienny zależnie od tego, czy bierzemy pod uwagę 500 najczęstszych słów czy 510. Tak więc dobrze rozpoznane cechy różnicujące nie stanowią same w sobie gwarancji, że zastosowana procedura rzeczywiście pogrupuje elementy zbioru w sposób, o który nam chodzi (Eder, 2014). Zilustrujmy to nieco absurdalnym przykładem: chcielibyśmy oddzielić skarpetki od rękawiczek, tymczasem wybrana metoda oddzieli nam białe skarpetki i rękawiczki od kolorowych skarpetek i rękawiczek. Stąd tak ważna jest rola badacza-obserwatora na etapie interpretacji wyników. Największą zaletą podejścia nienadzorowanego jest bezsprzecznie jego intuicyjność i prostota, największą wadą zaś fakt, że oko ludzkie łatwo daje się oszukać.

Inne podejście – zwane klasyfikacją nadzorowaną – zakłada, że istnieje pewna liczba z góry określonych klas. Rola badacza-obserwatora polega tym razem na tym, że projektuje on wzór dla systemu klasyfikacji, na którego podstawie system „uczy się" rozpoznawania przynależności nieznanego elementu do jednej z klas. Oceny jakości klasyfikacji nie dokonuje już jednak badacz (podatny na różnego rodzaju sugestie czy błąd interpretacji), lecz sam „wyuczony" system – ma to zapewnić większy poziom zobiektywizowania wyników. Nie bez powodu tego typu metody nazywa się „uczeniem maszynowym" (ang. *machine-learning methods*).

Spróbujmy przeanalizować procedurę uczenia nadzorowanego krok po kroku. Najpierw należy stworzyć zbiór trenujący, który jest ręcznie wyodrębnionym podzbiorem wszystkich dostępnych danych. Powinien on zawierać przedstawicieli każdej klasy, przy czym idealnie powinni to być typowi przedstawiciele. Czasem przedstawicieli klas można wyróżnić na podstawie obiektywnych kryteriów, czasem przynależność do klas jest nieoczywista lub rozmyta. Przykładem pierwszego scenariusza jest atrybucja autorska (tekst należący do klasy „Sienkiewicz" nie może należeć do żadnej innej klasy), przykładem drugiego – klasyfikacja na literaturę piękną i publicystykę (możemy się spodziewać wielu przypadków nieostrych lub należących do obu klas jednocześnie). Językoznawca diachroniczny znajduje się w dość komfortowej sytuacji, ponieważ klasyfikacja tekstów na podstawie daty ich powstania należy do kategorii ostrych i obiektywnych[3].

Wróćmy do hipotetycznego eksperymentu, którego celem będzie rozgraniczenie tekstów na podstawie jednej z dwóch kategorii: reportażu i felietonu. Pytanie badawcze mogłoby przyjąć postać: czy dowolnie wylosowany tekst z danego zbioru zostanie prawidłowo przypisany przez algorytm do swojej klasy? W tym celu tworzymy podzbiór uczący z pewnej liczby (niech to będzie np. 100) reportaży i felietonów. Teksty muszą być przypisane do danej kategorii na podstawie wiedzy *a priori*. Możemy na przykład poprosić kilku prasoznawców o taką klasyfikację, może ona też być wyciągnięta z nazwy działu gazety, w każdym razie musimy ręcznie przypisać kategorie w zbiorze uczącym – to jest właśnie ów „nadzór" w metodach klasyfikacji nadzorowanej.

Następnie wybieramy cechy, które – jak podejrzewamy – pozwolą odróżnić przedstawicieli poszczególnych klas. I tak, wracając do naszego przykładu z rękawiczkami i skarpetkami – zapewne kolor nie będzie dobrze odróżniał jednych od drugich, natomiast dobrze odróżni je kształt. Co więcej, materiał też w jakiejś mierze pomoże w odróżnieniu, skoro raczej nie ma skórkowych skarpetek i frotowych rękawiczek. Oczywiście takie mogą się trafić i być może klasyfikacja skórkowej skarpetki będzie nieskuteczna. Wspominamy o tym, ponieważ błędy w klasyfikacji mogą wynikać także z tego, że niektóre próbki mogą być, przynajmniej pod pewnymi względami, nietypowe dla swojej klasy.

Co to w praktyce znaczy, że teksty zostały umieszczone w zbiorze uczącym? W rzeczywistości tworzymy tabelę, w której każda kolumna[4] reprezentuje jeden element (w naszym wypadku tekst), a każda komórka w tym wierszu – cechę,

[3] Oczywiście w praktyce bywa i tak, że jakiś tekst powstawał przez wiele dziesiątków lat i przypisanie mu jakiejś konkretnej daty jest kłopotliwe (w takich wypadkach staraliśmy się, o ile to było możliwe, dzielić okres powstawania utworu na pół i tę datę przyjmować roboczo jako oznaczenie klasy). Czasem roku powstania utworu w ogóle nie da się ustalić, jak np. w wypadku *Kazań świętokrzyskich* (wtedy odwoływaliśmy się do szacunkowej datacji zaproponowanej w literaturze przedmiotu).

[4] Dla ścisłości – nie ma tu znaczenia, czy chodzi o wiersz czy kolumnę.

np. znormalizowaną liczbę wystąpień danego słowa, tak jak to przedstawiono w Tab. 4.1 z fikcyjnymi danymi:

Tabela 4.1. Przykładowe dane liczbowe (względne frekwencje najczęstszych wyrazów) wykorzystywane w wielowymiarowych metodach klasyfikacji maszynowej.

leksem	felieton_1	felieton_2	reportaż_1	reportaż_2	...
w	1,1354	1,9424	1,4197	1,0904	...
z	0,9755	1,0330	0,6916	0,6133	...
nie	0,7036	0,9800	0,1820	0,6905	...
a	0,3998	0,7328	1,1649	0,2907	...
się	1,0075	0,8476	0,4732	0,0181	...
...	...	...	...	...	

Mając przygotowany zbiór uczący, możemy przystąpić do trenowania klasyfikatora. Każda metoda nadzorowana robi to nieco inaczej, ale zawsze chodzi o to, by na podstawie tabeli takiej jak powyższa zidentyfikować najskuteczniejsze cechy – takie, które najmocniej różnicują poszczególne klasy. W naszym przypadku będziemy szukali zestawu cech, dzięki któremu najmocniej uwidaczniają się różnice między reportażem i felietonem.

Wreszcie, po tylu krokach przygotowawczych, następuje główny etap procedury, czyli klasyfikacja tekstów pozostawionych do tej pory na boku. Nie wchodząc w techniczne szczegóły, można powiedzieć, że rzecz się sprowadza do podobieństwa: kiedy algorytm dostaje nowy tekst do analizy, sprawdza, do której z kategorii ów tekst jest najbardziej podobny. Zakłada się, że jeśli tekst jest najbardziej podobny do felietonów w zbiorze uczącym, to najpewniej też jest felietonem. Najpewniej, bo zazwyczaj istnieje pewien nieunikniony procent błędów. Mogą one wynikać z niedoskonałości systemu, gdy np. cechy różnicujące nie są dobrane optymalnie. Przyczyna błędów może także leżeć po stronie źle wytrenowanego zbioru uczącego, a nawet po stronie samych danych wejściowych. Poszczególne kategorie mogą przecież stanowić *continuum*, stąd niektóre elementy naszego zbioru będą niemal nierozróżnialne. Wreszcie – u podstaw jakiejkolwiek kategoryzacji, czy to maszynowej, czy też tworzonej przez ekspertów, leży założenie, że elementy przynależne do poszczególnych kategorii są podobne to prototypowego jej członka, tymczasem nie zawsze daje się łatwo określić, który element zbioru jest najbardziej prototypowy. Przykład pierwszy z brzegu: gdybyśmy chcieli włączyć do zbioru uczącego reprezentatywną próbkę twórczości Sienkiewicza, to czy byłoby to *Quo vadis* czy raczej *Krzyżacy*? A może *Listy z Ameryki* albo *Bez dogmatu*? Mówiliśmy nieco wcześniej o tym, że nadzorowane uczenie

maszynowe polega na dobraniu optymalnych parametrów klasyfikacji przez sam system, bez udziału podatnego na błąd interpretacji obserwatora – jak zatem poradzić sobie z obiektywnym doborem reprezentatywnych tekstów do zbioru uczącego?

I tu dochodzimy do kolejnego niezwykle ważnego etapu opisywanej procedury, mianowicie oszacowania skuteczności klasyfikacji przy danych parametrach testowanego systemu. Powszechnie stosuje się tzw. sprawdzian krzyżowy (ang. *cross-validation*). Zamiast ręcznie konstruować zbiór uczący ze starannie dobranych tekstów, by otrzymać idealną reprezentację każdej klasy – co jest raczej mało realne, choćby ze względu na możliwy błąd człowieka w ocenie reprezentatywności – zostawia się tę czynność maszynie. Oczywiście maszyna z wyborem reprezentatywnych próbek radzi sobie jeszcze gorzej niż człowiek, dlatego zakłada się, że cały proces – od skonstruowania zbioru uczącego, przez wytrenowanie klasyfikatora, aż po testowanie modelu na pozostałych tekstach z korpusu – należy powtórzyć wiele razy, przy losowych ustawieniach zbioru uczącego. Za każdym razem trzeba rzecz jasna sprawdzić skuteczność systemu klasyfikacyjnego, by po zakończeniu wszystkich niezależnych iteracji oszacować średnią sprawność klasyfikacyjną modelu.

Wybór tekstów do zbioru trenującego może być w takiej sytuacji zupełnie losowy, bo jeśli powtórzymy całą procedurę wielokrotnie, to w końcu każdy tekst zarówno zostanie poddany klasyfikacji, jak i stanie się częścią zbioru trenującego. W ten sposób zostaje też rozwiązany problem „reprezentatywności": przez wielokrotne sprawdzenie różnych konstelacji tekstów w zbiorze uczącym dostajemy nie tyle „najlepszy" profil dla danej klasy, ile profil uśredniony, a więc z czysto statystycznego punktu widzenia najbardziej „reprezentatywny" (Eder i Rybicki, 2013). Warto dodać, że procedura ta ma szczególne znaczenie tam, gdzie danych jest po prostu mało, bo dzięki niej możemy stworzyć możliwie duży zbiór uczący, a równocześnie poddać klasyfikacji wszystkie teksty, którymi dysponujemy. Możemy bowiem do zbioru uczącego wstawić wszystkie teksty z wyjątkiem jednego – tego, który jest aktualnie klasyfikowany. Oczywiście w takim wypadku klasyfikację musimy powtórzyć tyle razy, ile mamy tekstów w korpusie.

Jak widać z powyższego ogólnego opisu procedury, bezsprzeczną zaletą stosowania metod nadzorowanych jest ich spora dokładność oraz zdolność do samooptymalizacji – najlepsze parametry modelu dobierane są bez udziału obserwatora-badacza, podatnego na błędy interpretacji. Ceną, jaką musimy zapłacić, jest jednak wielokrotnie dłuższy niż w wypadku metod nienadzorowanych czas potrzebny na wykonanie wszystkich obliczeń oraz odpowiednio duża moc obliczeniowa komputera. Niektóre metody (szczególnie te wykorzystujące sieci neuronowe) potrzebują długich miesięcy na wytrenowanie modelu klasyfikacyjnego i ogromnych zasobów obliczeniowych; największy z naszych eksperymentów, opisany poniżej w podrozdziale *Corpus of Historical American English*, nie był

wprawdzie aż tak wymagający, ale i tak nasze obliczenia zajęły około dwóch tygodni przy pełnym wykorzystaniu mocy serwera.

Ostatnia rzecz, która została do omówienia, to ocena skuteczności klasyfikatora. W metodach nadzorowanych przy każdym zestawie cech i przy każdej losowej rekompozycji zbioru uczącego przeprowadza się automatyczne przypisanie pozostałych tekstów do wytrenowanych klas, w metodach nienadzorowanych sprawdzamy po prostu, czy najbliżej umieszczone elementy należą do tej samej klasy. Przypisanie każdego tekstu może być albo prawidłowe, albo błędne, a suma błędnych przypisań staje się ważną informacją na temat skuteczności klasyfikatora. Na ogół ocenie poddaje się dwie wartości: pełność i dokładność[5]. Przypuśćmy, że algorytm zaklasyfikował 100 tekstów jako felietony. Pierwsza z wartości (tj. pełność) powie nam, jak wiele z nich zostało rzeczywiście rozpoznanych jako felietony. Jeśli system prawidłowo rozpoznał 80 z nich, resztę zaś błędnie przypisał do pozostałych kategorii, to pełność określimy na poziomie 0,8 (czyli 80%). Druga z wartości powie, jak wiele z tych tekstów, które zostały zakwalifikowane jako felietony, jest nimi rzeczywiście, a ile jest błędnie zakwalifikowanymi reportażami.

Na skuteczność cech różnicujących można spojrzeć z dwu stron. Kiedy chcemy w automatyczny sposób oddzielić felietony od reportaży, to powinniśmy dobrać taki zestaw cech dyskryminacyjnych, które je najbardziej od siebie odróżniają. Oczywiście stały czytelnik prasy, gdy mu przedstawić przykład reportażu i felietonu, raczej bezbłędnie przypisze jeden i drugi do właściwej kategorii, uczyni to jednak na podstawie kryteriów intuicyjnych. Analiza wielowymiarowa pozwala zobiektywizować tę intuicję.

Często jednak mamy do czynienia z sytuacją odwrotną, mianowicie chcemy się dowiedzieć, jak skuteczna jest klasyfikacja oparta na danym zestawie cech. W hipotetycznej sytuacji odróżnienia felietonu i reportażu możemy zadać pytanie, czy da się odróżnić typ tekstu na podstawie frekwencji – powiedzmy – 100 najczęstszych trigramów kategorii gramatycznych. Jeżeli tak zdefiniowany podzbiór cech okaże się mało skuteczny w klasyfikacji, dostajemy mocną przesłankę, że najczęstsze kategorie gramatyczne nie zawierają informacji o obu gatunkach publicystyki. Powtarzamy potem eksperyment dla innego zestawu cech wejściowych, np. 1000 częstych lematów, szukając przez eliminację takich cech, w których informacja o gatunku jest najmocniej uwidoczniona. Celem takiego przedsięwzięcia nie jest wtedy klasyfikacja nieznanych elementów, ale pytanie o to, czy dane cechy rzeczywiście różnicują klasy.

W dalszych rozważaniach będziemy posługiwać się metodami klasyfikacji właśnie w celu identyfikacji cech – leksykalnych bądź gramatycznych – które są odpowiedzialne za optymalne rozróżnienie tekstów według daty ich powstania.

---

[5] Polskie terminy nie są dobrze ustabilizowane, ich angielskie odpowiedniki to *recall* i *precision*.

4.2. Klasyfikacja nienadzorowana i sygnał chronologiczny

Jeżeli częstość słów, a także sekwencji słów lub sekwencji kategorii gramatycznych, pozwala rozpoznać autora tekstu, to uzasadnione wydaje się pytanie, czy takie same cechy (lub jakiś ich podzbiór) zdradzą też datę powstania tekstu. Rzecz jasna trudno się spodziewać dokładności klasyfikacji na poziomie 100%, ale interesować nas może, czy losowo wybrany tekst z korpusu będzie wykazywał podobieństwo do tekstów sobie współczesnych czy na przykład do tekstów podobnych gatunkowo, bez względu na datę powstania. Słowem, czy z korpusu uda się wyizolować sygnał chronologiczny?

Zapewne od razu nasuwa się odpowiedź, że przecież leksyka z pewnością zdradzi epokę, w której tekst powstał. Przykładowo tekst, który zawiera słowo *pomada*, będziemy intuicyjnie łączyli raczej z XIX niż z XXI wiekiem. Tymczasem rzecz wcale nie jest taka prosta: wśród 80 poświadczeń słowa *pomada* w zrównoważonym podkorpusie NKJP zaledwie 14 pochodzi z tekstów o dacie pierwszego wydania wcześniejszej niż rok 1920. Tak więc próba datowania tekstu na podstawie obecności tego akurat słowa zawodzi. Można się spodziewać, że łatwiej będzie ustalić *terminus post quem*: na przykład NKJP pierwsze wystąpienie słowa *komputer* notuje w roku 1978 w książce Janusza Płońskiego i Macieja Rybińskiego *Góralskie tango*.

Niemniej pierwsza intuicyjna odpowiedź, że to właśnie leksyka w największym stopniu odpowiada za ewolucję języka, jest co do zasady słuszna. Problem polega jednak na tym, że trudno śledzić zmiany pojedynczych słów – chyba że są to starannie dobrane przykłady dające się wymodelować np. regresją logistyczną, tak jak to czyniliśmy w poprzednim rozdziale. Znacznie sensowniej będzie przyjąć hipotezę, że za ewolucją systemu językowego stoi bardzo duża liczba ledwie zauważalnych zmian leksykalnych, które być może da się zaobserwować *en masse*. Znakomitym narzędziem do obserwacji wielu cech jednocześnie są wielowymiarowe metody klasyfikacji – w niniejszym rozdziale skorzystamy z nienadzorowanej techniki zwanej skalowaniem wielowymiarowym (ang. *multidimensional scaling*, w skrócie MDS).

Jedno z pytań, jakie będziemy chcieli zadać, dotyczy wpływu leksyki, a konkretnie tego, czy silniejszy sygnał chronologiczny kryje się w słownictwie rzadkim czy też w bardzo częstych słowach synsemantycznych. Skoro autorów najłatwiej odróżnić przede wszystkim po słownictwie najczęstszym, być może te same wyrazy zawierają w sobie również silną sygnaturę czasu? Jest to przecież – ujmując rzecz diachronicznie – słownictwo szczególnie trwałe, jeżeli nie liczyć zmian w fonetyce i odpowiadających im zmian w grafii. A zatem jeśli metody, które sprawdziły się w kategoryzacji tekstu pod względem autorstwa, okażą się skuteczne w klasyfikacji chronologicznej, zdobędziemy pośredni dowód, że krótkoterminowa zmiana diachroniczna to przede wszystkim niewielkie przesunięcia w randze na liście frekwencyjnej poszczególnych wyrazów, które

same w sobie nie są charakterystyczne dla danej epoki[6]. I odwrotnie – jeśli okaże się, że słowa synsemantyczne niewiele wnoszą do klasyfikacji w porównaniu ze słownictwem rzadkim, dostaniemy ważny argument przemawiający za tezą, że język jest mało podatny na zmiany w swej warstwie gramatycznej, a na jego ewolucję wpływa w dużej mierze stylistyczny gust epoki (a więc słownictwo rzadkie). Podobnie ma się rzecz z frekwencją sekwencji kategorii gramatycznych: na podstawie tych znaczników stylu spróbujemy z kolei przetestować hipotezę, że zmiana językowa dokonuje się również na poziomie składniowym.

4.2.1. Powieści polskie XIX–XX wieku

Pierwszym, pilotażowym, eksperymentem była kategoryzacja przeprowadzona na grupie 76 polskich powieści z czasu pomiędzy połową XIX wieku a przełomem XX i XXI wieku. Ze względu na fakt, że nasz podstawowy korpus diachroniczny 1380–1850 jest wysoce niejednorodny pod względem reprezentowanych gatunków, a także długości poszczególnych tekstów, przygotowaliśmy niewielki korpus *ad hoc* składający się wyłącznie z tekstów literackich, głównie powieści; do oryginalnego korpusu 1380–1850 przyjdzie jeszcze powrócić pod koniec tego rozdziału.

Teksty pochodziły zarówno z bibliotek wirtualnych (przede wszystkim serwisu Wolne Lektury), jak i z zasobów NKJP. Gdy do dyspozycji była większa liczba powieści jednego autora, to staraliśmy się, by nie były to teksty powstałe ani u progu kariery pisarza, ani też u jej schyłku. Chodziło o to, by korpus nie zawierał tekstów, które odzwierciedlają moment akwizycji mowy znacząco różny od momentu powstania tekstu, czyli by tekst nie reprezentował języka bardziej progresywnego bądź konserwatywnego niż teksty jemu współczesne. Dbaliśmy też o to, by zniwelować siłę sygnału autorskiego, który z pewnością był silniejszy niż sygnał chronologiczny, dlatego też każdy autor jest reprezentowany przez nie więcej niż cztery powieści (a przeważnie nie więcej niż trzy). Wreszcie przyjęliśmy w przybliżeniu, że datą powstania tekstu jest data jego opublikowania. Ograniczyliśmy się do powieści, ponieważ zależało nam na tym, by sygnału chronologicznego nie zaburzał sygnał typu tekstu.

Zastosowaliśmy skalowanie wielowymiarowe (MDS), które jest techniką nienadzorowaną, polegającą na kompresji danych wielowymiarowych do mniejszej liczby wymiarów, na ogół dwóch. Wynikiem jest dwuwymiarowy wykres ukazujący badane próbki w ten sposób, by przy utracie jak najmniejszej ilości informacji pokazać wzajemne podobieństwa między próbkami (zob. np. Baayen, 2009). W wielokrotnie powtórzonych analizach testowaliśmy przeróżne parame-

[6] Warto przy tym uświadomić sobie, że taka jest natura listy frekwencyjnej: jeśli ranga jakiegoś wyrazu wzrasta, to innego musi spaść. W tekście bowiem suma frekwencji wszystkich wyrazów jest równa długości tekstu. Jeżeli tekst ma stałą liczbę słów, to wzrost frekwencji jednego słowa może się odbyć jedynie kosztem frekwencji pozostałych słów.

try wejściowe. Jako cech różnicujących używaliśmy zarówno słów tekstowych (słowoform), jak i bi- oraz trigramów etykietek kategorii gramatycznych, a także bi- i trigramów etykietek gramatycznych w wersji zredukowanej do samej tylko części mowy (tj. pozbawione informacji fleksyjnej). W poszczególnych analizach mieliśmy więc do czynienia z jedną z czterech reprezentacji badanych tekstów, np. zdanie otwierające *Kronikę wypadków miłosnych* Konwickiego przyjmowało następującą postać dla słowoform:

Wicio siedział na stromej skarpie koło torów i czekał.

Lematy zaś były reprezentowane przez sekwencję:

[wicio] [siedzieć] [na] [stromy] [skarpa] [koło] [tor] [i] [czekać] [.]

Przy analizie etykietek kategorii gramatycznych powyższe zdanie przyjęło postać[7]:

[adv:pos] [praet:sg:m1:imperf] [prep:loc] [adj:sg:loc:f:pos] [subst:sg:loc:f] [qub] [subst:pl:gen:m3] [conj] [praet:sg:m1:imperf] [interp]

I wreszcie ostatnia reprezentacja, czyli te same etykietki kategorii gramatycznych, ale pozbawione informacji fleksyjnej:

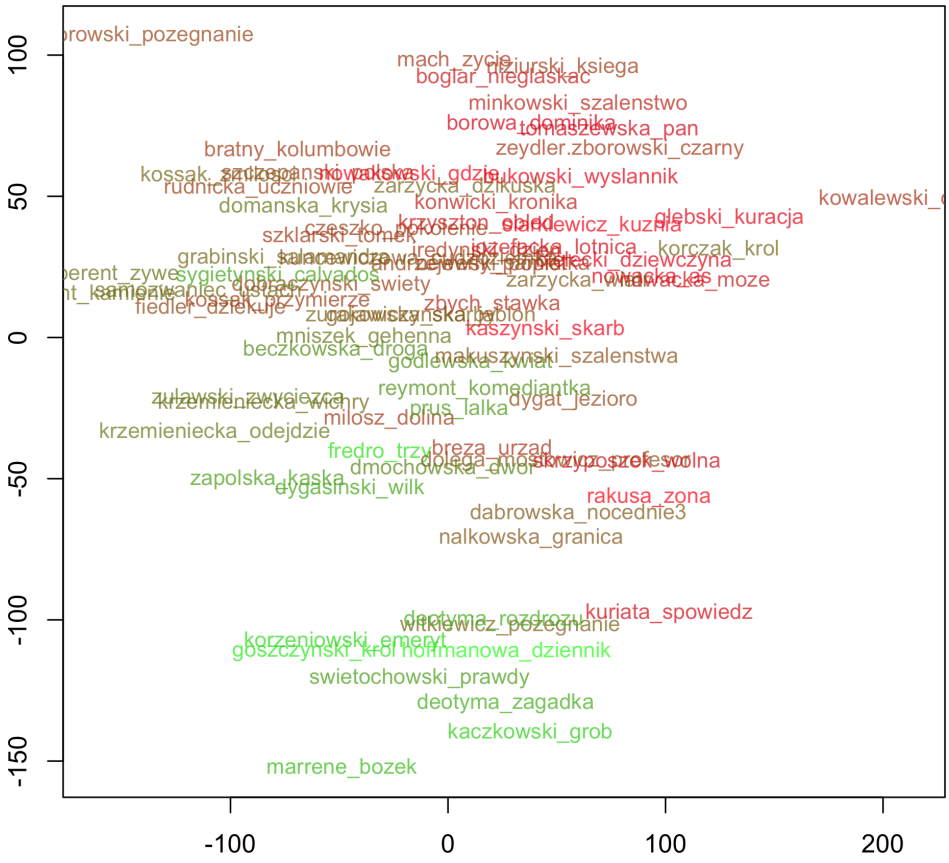
[adv] [praet] [prep] [adj] [subst] [qub] [subst] [conj] [praet] [interp]

Każda z powyższych czterech reprezentacji mogła być następnie podzielona albo na pojedyncze elementy (unigramy), albo na pary sąsiednich elementów (bigramy), albo na sekwencje trzech następujących po sobie elementów (trigramy). Dawało to dwanaście typów cech różnicujących. Pamiętając o tym, że w obrębie poszczególnych typów inne wyniki klasyfikacji dostanie się dla, powiedzmy, 10 najczęstszych cech, a inne dla 5000, przeprowadziliśmy wiele testów dla różnych zbiorów cech.

W każdym kroku badaliśmy, czy skalowanie wielowymiarowe pokaże stopniowe przejście od tekstów napisanych dawniej do tekstów późniejszych, czyli czy na wykresie teksty ułożą się chronologicznie. Przykładowy wykres dla 250 najczęstszych słów – w tej liczbie mieści się około 50 słów synsemantycznych i około 200 częstych słów autosemantycznych – przedstawiony jest na Rys. 4.1.

Ułożenie poszczególnych próbek na wykresie może nie jest całkowicie klarowne, ale mimo to można dostrzec, że teksty powstałe mniej więcej w tym samym czasie wykazują tendencję do pozostawania w swoim sąsiedztwie. Widać to dość dokładnie na skali kolorystycznej – im tekst wcześniejszy, tym zieleńszy,

[7] Nietrudno zauważyć, że leksem *Wicio* został opisany błędnie jako przysłówek w stopniu równym (*adv:pos*). Wynika to z faktu, że kategorie gramatyczne przypisywane są przez dezambiguator na podstawie modelu probabilistycznego, co sprawia kłopot w rozpoznaniu niektórych form. W wersji online korpusu NKJP (http://www.nkjp.pl/poliqarp) wyraz ten jednak został prawidłowo rozpoznany jako [subst:sg:nom:m1].



Rysunek 4.1. Skalowanie wielowymiarowe 76 powieści polskich (250 najczęstszych słów). Obie osie wykresu odpowiadają odległości w przeskalowanej przestrzeni MDS.

upływ czasu sygnalizuje przechodzenie koloru najpierw w brąz, potem w czerwień. Główna linia ewolucji chronologicznej przebiega na wykresie wzdłuż linii wertykalnej, tj. od dołu (wczesne powieści) do góry (późne).

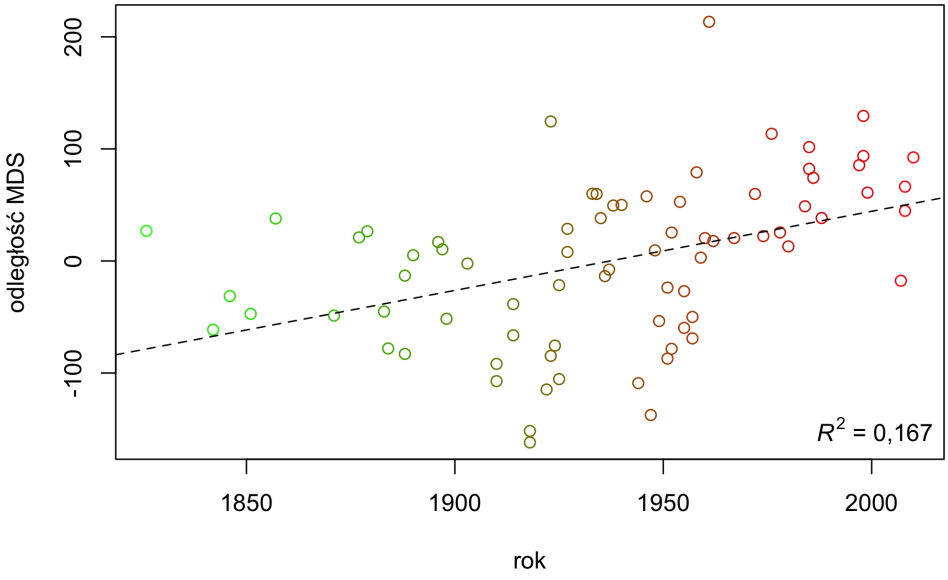
Rys. 4.1 ukazuje zarazem pewne ograniczenie metod klasyfikacji nienadzorowanej (o innych będzie jeszcze mowa). Mianowicie wygenerowany wykres, mimo swych niewątpliwych walorów wizualnych, nie pozwala na sformułowanie żadnych ogólnych wniosków na temat zmiany w języku. Nawet jeśli porównamy kilka wykresów MDS dla różnych parametrów wejściowych, trudno będzie ocenić, który z nich jest „lepszy" albo przynajmniej bardziej wiarygodny.

Spróbujmy zmierzyć się z tym problemem. Jako że proponowane przez nas podejście jest niestandardowe, trzeba będzie poprzedzić je krótkim, jednoakapitowym wprowadzeniem.

Wynik skalowania wielowymiarowego w zasadzie zawsze przedstawia się na dwuwymiarowym wykresie. Wiąże się to nieuchronnie z jakąś stratą informacji, ale taki właśnie jest cel procedury: chodzi o to, by zmaksymalizować przejrzystość wyników przy minimalizacji straty informacyjnej. Nie ma jednak żadnych przeszkód technicznych, by oryginalną przestrzeń *n*-wymiarową zredukować do dowolnej innej przestrzeni. Można na przykład zredukować oryginalną przestrzeń (w naszym bieżącym przypadku 250-wymiarową) do trzech wymiarów – tyle że wtedy wynik będzie trzeba przedstawić na 3-wymiarowym wykresie, ewentualnie w postaci np. płaskorzeźby. Można też dokonać maksymalnej możliwej redukcji: do jednego wymiaru.

Rysunek 4.2 pokazuje, jaki może być cel takiej operacji: dzięki kompresji całej przestrzeni tekstów do zaledwie jednego wymiaru jesteśmy w stanie zobaczyć, czy istnieje jakaś zależność między wynikami MDS (oś *y*) i datą powstania utworu (oś *x*). Im bliżej dwa punkty (teksty) leżą obok siebie na osi *y*, tym bardziej są do siebie podobne. Gdyby sygnał chronologiczny był silny, a więc gdyby teksty powstałe w tym samym czasie były do siebie podobne, to punkty na wykresie reprezentujące poszczególne powieści ułożyłyby się mniej więcej wzdłuż przekątnej. Nawet pomimo straty części informacji w stosunku do 2-wymiarowego MDS widać wyraźnie (Rys. 4.2), że tego typu korelacja uwidacznia się w wypadku MDS dla 250 najczęstszych wyrazów: punkty (teksty) wprawdzie nie układają się dokładnie po przekątnej, ale z pewnością korelują z osią czasu.

Zanim podejmiemy próbę kwantyfikacji współzależności obu zmiennych, nasuwa się pewna obserwacja *ad hoc*: zauważmy, że lewy górny róg wykresu jest pusty, podczas gdy w dolnej części pojawia się więcej obserwacji odstających od modelu. Można to interpretować w ten sposób, że istnieją powieści podobne swoim stylem do wcześniejszych (tj. archaizujące), ale nie da się antycypować stylu. Gdy przesuwamy się w czasie, prawa dolna część wykresu robi się pustsza, być może dlatego, że ubywa powieści o archaizującym stylu. Warto przy tym zwrócić uwagę na lata pięćdziesiąte – tutaj widać szczególnie dużo tekstów bardziej przypominających wiek XIX niż czasy sobie współczesne.



Rysunek 4.2. Skalowanie wielowymiarowe 76 powieści polskich (250 najczęstszych słów) zredukowane do jednego wymiaru.

Czy stopień korelacji obu zmiennych, a zatem wyników MDS i osi czasu, da się w jakiś sposób zmierzyć i porównać z innymi wykresami? Stąd już tylko krok do poczynienia obserwacji, że korelacja między dwiema zmiennymi na ogół bardzo dobrze daje się wyrazić za pomocą modelu regresji liniowej (jest to znacznie prostsza odmiana modeli, o których mówiliśmy w rozdziale *Dynamika zmian językowych*). I rzeczywiście: zależność wartości MDS od daty powstania utworów z naszego korpusu można opisać prostym modelem liniowym, którego zoptymalizowany przebieg zaznaczyliśmy na Rys. 4.2 przerywaną linią. Parametry tego modelu są następujące:

$$y_i = 0,706 \cdot d_i - 1367,684 + \epsilon$$

gdzie d_i oznacza datę publikacji i-tego tekstu, ϵ zaś losowe zaburzenie. Model ma bardzo dużą istotność statystyczną ($p < 0,001$), ale jego dopasowanie do danych empirycznych jest, mówiąc oględnie, niesatysfakcjonujące, przy $R^2 = 0,167$ (używamy tutaj i poniżej miary tzw. *adjusted* R^2, która jest bardziej odporna od klasycznego R^2 na obserwacje odstające).

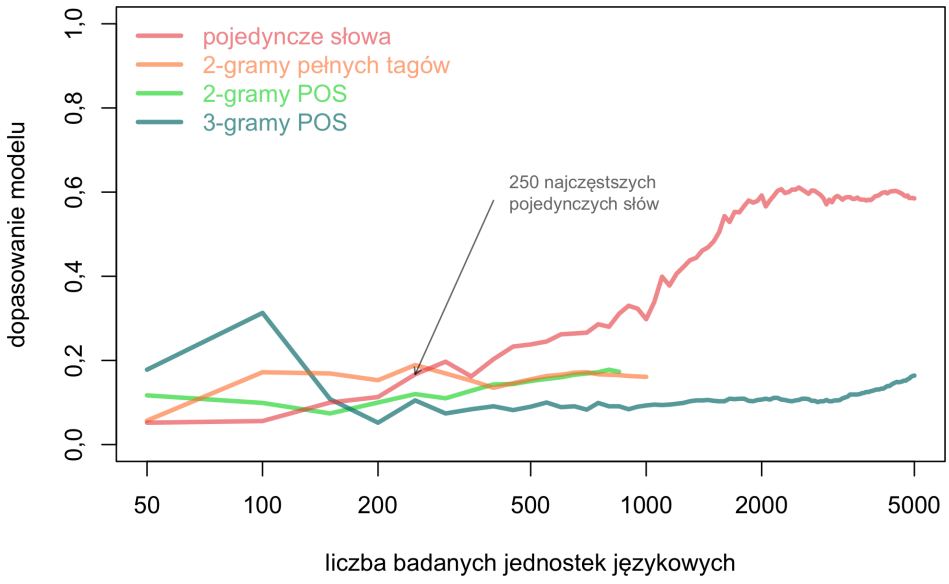
Znając parametry modelu dla 250 najczęstszych wyrazów, możemy teraz powtórzyć całą procedurę dla, powiedzmy, 100 najczęstszych wyrazów i porównać uzyskane wyniki. Żeby zyskać względnie wiarygodną wiedzę o cechach najsilniej zdradzających sygnał chronologiczny, powinniśmy jednak takich porównań

zrobić znacznie więcej. Następnie trzeba by się przyjrzeć bigramom etykietek gramatycznych i dla nich wygenerować kolejną serię wykresów. I tak dalej. Porównanie wszystkich możliwych wykresów na pewno nie jest zadaniem prostym. Wcześniej była mowa o jednej z wad metod nienadzorowanych – właśnie odkryliśmy drugą: klasyfikacja nienadzorowana, mimo swojej oczywistej zalety, jaką jest intuicyjność interpretacji wykresu, staje się kłopotliwa, jeśli eksperyment wymaga porównania wyników dla bardzo wielu pojedynczych analiz.

Skorośmy jednak zaproponowali użycie modelu liniowego do opisu korelacji miary MDS z chronologią, to jesteśmy zaledwie o mały krok od przezwyciężenia wyartykułowanej powyżej wady ręcznego porównywania wykresów. Zauważmy, że stopień dopasowania modelu R^2 może zostać użyty jako syntetyczna miara sygnału chronologicznego – im silniej bowiem dany zestaw cech językowych determinuje zmianę w języku, tym mocniej analizowane teksty będą korelowały z datą wydania utworu. Słowem, im wyższy współczynnik R^2 dla danego zestawu cech, tym większy ich udział w ewolucji języka. Nie oznacza to oczywiście, że samo skalowanie wielowymiarowe jest etapem, który możemy pominąć – przeciwnie, procedurę należy wykonać z całą starannością, tyle że zaproponowana przez nas kompresja MDS połączona z modelowaniem za pomocą regresji liniowej pozwala oszacować siłę sygnału chronologicznego bez porównywania końcowych wykresów, wyłącznie za pomocą miary dopasowania modelu do danych empirycznych.

Przeprowadziliśmy zatem całą serię analiz MDS, by porównać oszacowaną wartość R^2 dla poszczególnych modeli regresji liniowej. W domenie leksyki porównaliśmy zachowanie korpusu dla 50 najczęstszych wyrazów, potem dla 100, 150, 200, 250 itd. aż do 5000. Oznaczało to oszacowanie parametrów dla setki różnych modeli. To samo zrobiliśmy dla bigramów kategorii gramatycznych (pełnych tagów), bigramów samych części mowy, a także trigramów samych części mowy. Dało to w sumie 400 modeli do porównania. Wszystkie obliczone współczynniki dopasowania modeli do danych empirycznych zostały przedstawione na Rys. 4.3 (dla łatwiejszego śledzenia wywodu zaznaczono strzałką model dla 250 najczęstszych wyrazów, któremu poświęciliśmy nieco uwagi powyżej, por. Rys. 4.1–4.2).

Spojrzenie na wykres zbierający wyniki dla kilkuset pojedynczych modeli prowadzi do następujących wniosków. Po pierwsze okazuje się, że najsilniejszym predyktorem ewolucji języka (literackiego) w XX wieku jest przede wszystkim rzadkie słownictwo – dopiero duże podzbiory cech leksykalnych, zawierające 2000–5000 słowoform, dawały mocną korelację MDS z datą powstania powieści. Po drugie: słownictwo częste, a szczególnie najczęstsze, czyli kilkadziesiąt wyrazów synsemantycznych z góry listy frekwencyjnej, w ogóle nie zdradza sygnału chronologicznego (zob. wartości po lewej stronie wykresu). Po trzecie wreszcie: trigramy kategorii gramatycznych zaznaczają swój udział w zmianie diachronicznej w polszczyźnie XX wieku, ale tylko w zakresie kilkudziesięciu



Rysunek 4.3. Stopień dopasowania modeli liniowych dla 50, 100, 200, . . . , 5000 najczęstszych bigramów kategorii gramatycznych (pełnych tagów), bigramów części mowy i trigramów części mowy.

najczęstszych trigramów. Ten ostatni wniosek jest na swój sposób najciekawszy, bo widzimy dokładnie odwrotne zachowanie trigramów kategorii gramatycznych (tj. składni) w stosunku do słowoform (tj. leksyki). Z porównania zdaje się wynikać, że sygnał chronologiczny ulokował się w częstych strukturach składniowych oraz w rzadkim słownictwie. Oczywiście w innych epokach czy w innych typach piśmiennictwa obraz zmian może być nieco inny.

Trzeba również pamiętać, że warunki niniejszego eksperymentu były specyficzne: w korpusie znajdowały się wyłącznie teksty literackie, prawie wyłącznie powieści. Tymczasem styl tekstu, rozumiany jako względna częstość elementów leksykalnych lub gramatycznych, jest uzależniony od bardzo wielu czynników: autorstwa, gatunku, czasu powstania itp. Jeśli jakiś autor pisał zarówno powieści, jak i teksty użytkowe, prawdopodobnie wpłynie to na wybór środków językowych (stylistycznych) w obu typach piśmiennictwa i w konsekwencji ujawni się sygnał gatunkowy, może nawet silniejszy od sygnału autorskiego: powieści będą zapewne podobne do innych powieści, zaś teksty użytkowe do innych tego rodzaju tekstów. Poszukując śladów chronologii w języku, musimy więc pamiętać, że mogą one być zakłócane przez inne sygnały – również takie, których istnienia nie podejrzewaliśmy.

4.2.2. Corpus of Late Modern English Texts

W poprzednim podrozdziale przeprowadziliśmy eksperyment na starannie wyselekcjonowanym podkorpusie zawierającym wyłącznie powieści. Chcieliśmy przetestować skuteczność metod klasyfikacji nienadzorowanej w wykryciu sygnału chronologicznego, w taki jednak sposób, by ograniczyć do minimum liczbę dodatkowych zmiennych, które mogłyby wpłynąć na wynik testów: jedną z takich oczywistych zmiennych wymagających neutralizacji była różnorodność gatunkowa.

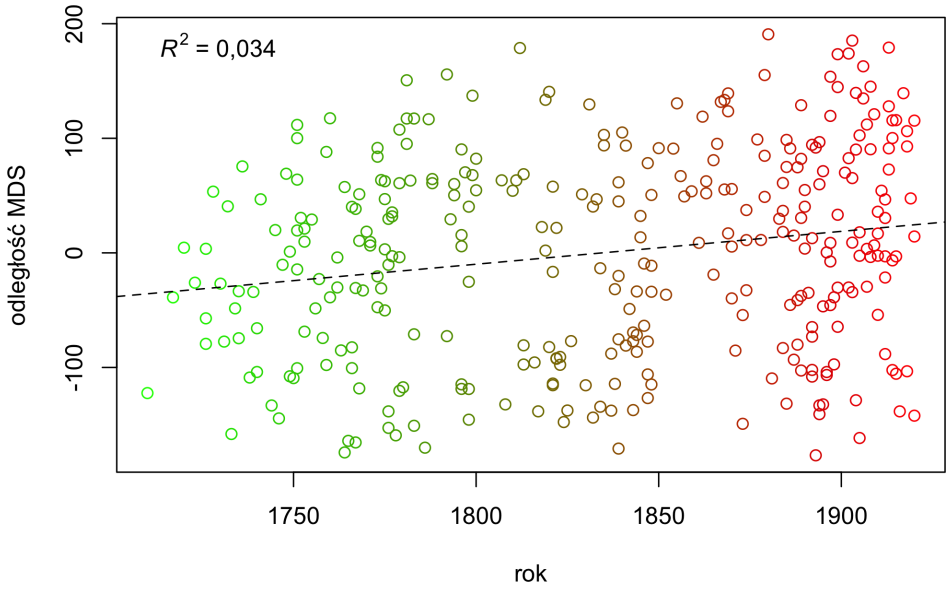
Następnym naturalnym krokiem będzie użycie tej samej co poprzednio metody badawczej, lecz przy rozluźnieniu restrykcji gatunkowej. W tym celu posłużymy się bardzo starannie zrównoważonym korpusem języka późnonowoangielskiego znanym pod nazwą Corpus of Late Modern English Texts albo w skrócie CLMET 3.0 (De Smet, 2005). Korpus ten obejmuje okres 1710–1920; zawiera 333 teksty i 9 818 326 słowoform. Korpus powstawał w sposób możliwie nieskomplikowany dla jego twórców – teksty były pozyskiwane z dostępnych bibliotek wirtualnych, głównie Projektu Gutenberg i Oxford Text Archive. Teksty zostały anotowane morfosyntaktycznie, nie są natomiast lematyzowane, podobnie jak większość korpusów języka angielskiego. Zawarci w korpusie autorzy to wyłącznie Brytyjczycy; chociaż pojedynczy autor może być reprezentowany przez kilka tekstów, to jednak w sumie nie mogą one liczyć więcej niż 200 000 słów.

Najważniejsze różnice między poprzednim eksperymentem i niniejszym to, po pierwsze, znacznie krótsze próbki tekstowe, a po drugie – i w tym wypadku istotniejsze – obecność wielu stylów funkcjonalnych. Celem będzie przetestowanie siły sygnału chronologicznego w różnogatunkowym korpusie. Przykładowe wyniki skalowania wymiarowego dla 250 trigramów kategorii gramatycznych przedstawione zostały na Rys. 4.4.

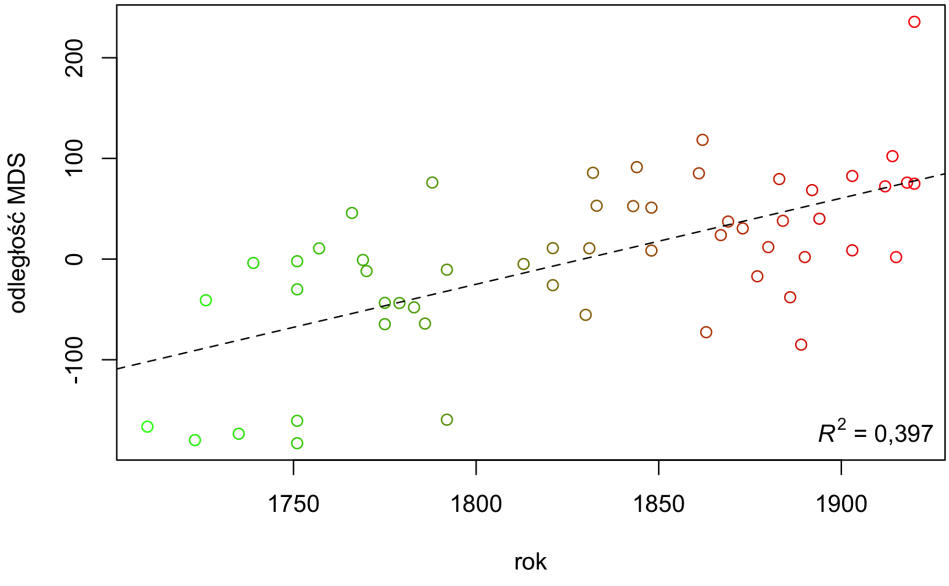
Mimo że model regresji liniowej wykazuje istotność statystyczną, $p < 0{,}01$, od razu widać (nawet gołym okiem) znacząco mniejsze dopasowanie modelu do danych empirycznych. Niski współczynnik dopasowania modelu $R^2 = 0{,}034$ tylko tę intuicję potwierdza. Jeśli jednak wydobędziemy z korpusu tylko jeden styl funkcjonalny, np. traktaty[8], i na takim podzbiorze powtórzymy testy, to sygnał chronologiczny staje się wyraźnie silniejszy. Uwidacznia to Rys. 4.5 oraz współczynnik dopasowania modelu $R^2 = 0{,}397$, wyższy o rząd wielkości w stosunku do poprzedniego przykładu.

Z powyższej analizy MDS na trigramach kategorii gramatycznych – a także kilkudziesięciu innych analiz, wykorzystujących pozostałe znaczniki stylu – jasno wynika, że sygnał gatunku zniekształca obraz diachronii. Wydaje się, że mamy tu do czynienia z sytuacją podobną do sygnału tłumacza tekstu, który prawie

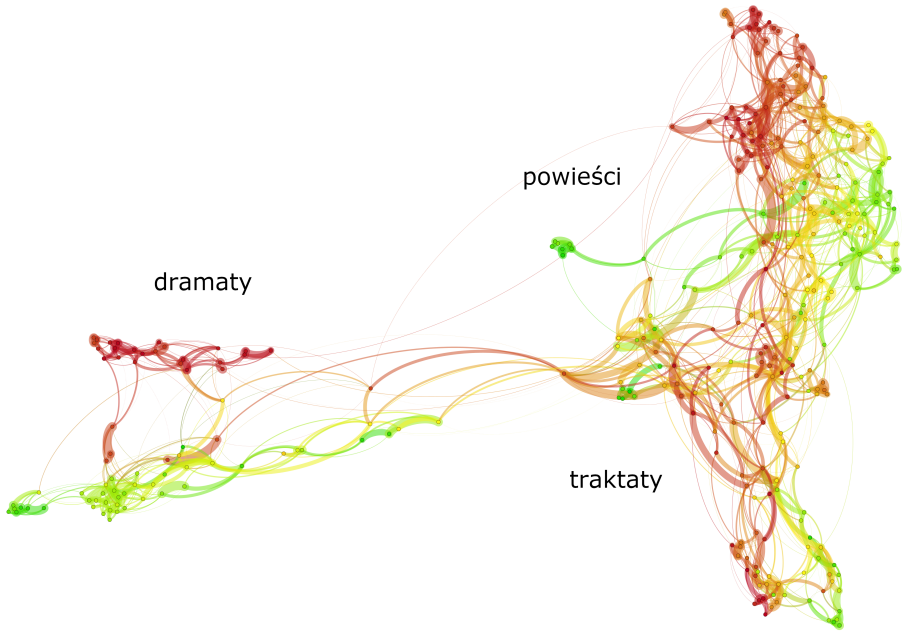
[8] *Traits* w typologii przyjętej przez De Smeta (2005), a więc teksty informacyjne bądź perswazyjne niefikcjonalne.



Rysunek 4.4. Skalowanie wielowymiarowe 333 tekstów angielskich (250 trigramów kategorii gramatycznych) zredukowane do jednego wymiaru.



Rysunek 4.5. Skalowanie wielowymiarowe tekstów przypisanych do kategorii „traktaty” (250 najczęstszych trigramów etykietek gramatycznych) zredukowane do jednego wymiaru.



Rysunek 4.6. Sieć podobieństw 333 tekstów angielskich.

zawsze jest słabszy od sygnału autorskiego (Rybicki, 2012); w naszej sytuacji sygnał gatunkowy dominowałby nad diachronią.

By potwierdzić to przypuszczenie, wszystkie 333 teksty korpusu poddaliśmy analizie za pomocą jeszcze innej metody klasyfikacji nienadzorowanej, tzw. *bootstrap consensus networks* (Eder, 2017). Jest to taki sposób wizualizacji podobieństw tekstowych, w którym informatywne są zarówno odległości pomiędzy węzłami (w naszym wypadku węzeł reprezentuje tekst), jak i grubość linii, która je łączy. Z kolei kolor wskazuje na datę powstania tekstu; *continuum* kolorów od zielonego do czerwonego oznacza czas powstania tekstu (im bardziej zielony, tym wcześniejszy, im bardziej czerwony, tym późniejszy). Jak wynika z automatycznego ułożenia tekstów na sieci podobieństw (Rys. 4.6), teksty grupują się bardzo konsekwentnie według gatunku (dramat, powieści, traktaty), a dopiero później, w ramach poszczególnych gatunków, uwidacznia się grupowanie według czasu powstania tekstu.

Co najmniej dwa wnioski płyną z powyższych obserwacji, jeden praktyczny i jeden ogólny. Po pierwsze zatem, będziemy musieli znaleźć jakiś sposób, by przełamać siłę sygnału gatunkowego, jeśli chcemy w wiarygodny sposób pokazać cechy językowe (lub stylistyczne) odpowiedzialne za zmianę chronologiczną w zrównoważonym korpusie. Temu zagadnieniu będą poświęcone następne pod-

rozdziały. Po drugie, eksperyment niniejszy zdaje się potwierdzać sugestię, którą rzuciliśmy mimochodem kilka stron wcześniej – że ewolucja języka ma często więcej wspólnego ze stylem (a więc z gustem epoki) niż ze zmianami strukturalnymi; mocnym tego dowodem jest właśnie fakt, że chronologia daje się zdominować przez sygnał gatunkowy.

4.3. Klasyfikacja nadzorowana jako narzędzie w periodyzacji

W poprzednim podrozdziale wykorzystaliśmy skalowanie wielowymiarowe do porównania siły dyskryminacyjnej różnych cech językowych, licząc na to, że uda się wyizolować sygnał chronologiczny, tzn. cechy odpowiedzialne za zmiany w języku. Tak zaprojektowany eksperyment nie daje nam jednak odpowiedzi na pytanie o dynamikę tych zmian – o to, czy da się wyróżnić momenty, które stanowią naturalne granice epok (czy ostrożniej: okresów)[9] w dziejach języka.

W podrozdziale niniejszym nieco zmodyfikujemy cel naszych poszukiwań, biorąc pod uwagę powyższe zastrzeżenia. Będzie nas mianowicie interesowało nie tyle wykrycie zmiany w języku, ile znalezienie kryteriów ilościowych, które pozwoliłyby zobiektywizować periodyzację historii języka.

Możemy przyjąć jako pewnik, że ewolucja języka jest procesem ciągłym – to, co obserwujemy jako zmianę systemu, jest sukcesywnym nawarstwieniem się drobnych, ledwo zauważalnych zmian. I o ile takie stwierdzenie jest truizmem, o tyle już mniej trywialne jest pytanie, czy powolne nawarstwianie się zmian doprowadzi w końcu do poważniejszego przełomu, sytuacji, w której teksty powstałe w pewnym momencie na osi czasu okażą się wyraźnie odmienne od tekstów tylko nieznacznie wcześniejszych. Tak zarysowany problem badawczy można też sformułować inaczej: będziemy starali się dowiedzieć, czy tempo tej ewolucji jest stałe czy też składa się z naprzemiennych cykli przyspieszającej i zwalniającej zmiany w języku. Gramatyki historyczne wskazują, że istniały pewne momenty[10] gwałtowniejszych przemian, w których mogło zajść wiele jednoczesnych przeobrażeń w systemie. Czy w krótkim – z punktu widzenia procesu diachronicznego – horyzoncie czasowym będziemy w stanie dostrzec takie momenty przyśpieszenia ewolucji?

Powyżej napisaliśmy, że klasyfikacja – tak nienadzorowana, jak i nadzorowana – sprowadza się jako metoda do wykrywania podobieństw. Jeśli elementy

[9] W dalszym ciągu naszych rozważań będziemy się posługiwali słowem epoka na określenie wszelkich wyróżnionych przez opisywane metody odcinków rozwoju języka. Równocześnie zdajemy sobie sprawę, że epoka sugeruje zmianę radykalną, podczas gdy zmiany w naszych danych mogą być słabo widoczne.

[10] Używamy tu słowa „moment", ale w istocie tą etykietką określamy bardzo różne odcinki czasowe – wiele dziesięcioleci, gdy mowa o głębokiej zmianie systemowej, i dziesięciolecie, gdy mówimy o subtelnych zmianach natury ilościowej.

przynależne do danej klasy są do siebie podobne, to równocześnie będzie się je dało odróżnić od elementów przynależnych do innej klasy. Problem jednak, z jakim mamy do czynienia w naszej sytuacji, polega na tym, że dane diachroniczne nie dzielą się na mocno wyodrębnione klasy, takie jak, dajmy na to, „poezja" vs. „proza". Różnice między klasą „teksty z roku 1800" i „teksty z roku 1801" będą bardzo subtelne lub wręcz prawie niezauważalne. Słowem, trudno oczekiwać, żeby w danych diachronicznych dało się wyodrębnić klasy o ostrych granicach. Jeśli jednak założymy, że interesuje nas nie tyle stuprocentowa rozpoznawalność (tj. dokładność klasyfikacji) tekstów sprzed pewnej daty i po niej, ile raczej rozpoznawalność dla danej cezury czasowej w porównaniu do innych testowanych cezur, dostaniemy do ręki narzędzie dość sprawnie wyznaczające momenty większych przełomów.

Najbardziej naturalną i intuicyjną ścieżką postępowania byłoby w takiej sytuacji przetestowanie dat uznawanych w literaturze przedmiotu za ważne (przełomowe) w dziejach polszczyzny. W naszym podejściu staraliśmy się jednak ograniczyć do minimum liczbę zmiennych zależnych od arbitralnych decyzji badaczy, bez względu na to, jak dużym autorytetem by się cieszyli. Nie oznacza to oczywiście, że hipoteza o przełomowości roku 1543 w historii polszczyzny jest dla nas nieistotna; przeciwnie, zakładamy, że obok uznanych przełomów mogą istnieć inne, równie ważne daty, z różnych powodów przeoczone przez badaczy. Proponowana przez nas metoda nie czyni więc żadnych założeń wstępnych, jeśli chodzi o testowane daty największego przyspieszenia ewolucji języka.

Weźmy dowolny rok – nawet zupełnie losowo – i załóżmy czysto hipotetycznie, że jest to rok największej zmiany w dziejach polszczyzny. Taka robocza hipoteza jest oczywiście z gruntu błędna, ale potrzebujemy jej jako pewnego punktu wyjścia. Mając wybraną datę, np. rok 1603, wybieramy losowo z korpusu jakąś liczbę tekstów napisanych przed tą datą i po tej dacie. Trenujemy następnie klasyfikator nadzorowany (zob. wstęp do niniejszego rozdziału) i pytamy, czy pula losowo wybranych tekstów zostaje prawidłowo rozpoznana przez klasyfikator jako utwory napisane przed lub po roku 1603. Liczbę prawidłowo rozpoznanych tekstów (dokładność klasyfikacji) skrzętnie zapisujemy i formułujemy nową hipotezę roboczą, że główny przełom nastąpił w kolejnym roku, a zatem w naszym wypadku w roku 1604. Znowu wybieramy losowo teksty, tym razem powstałe przed i po 1604, trenujemy klasyfikator, testujemy liczbę prawidłowo rozpoznanych tekstów i zaczynamy procedurę od nowa, tym razem dla roku 1605. I tak dalej. W ten sposób dostajemy skuteczność klasyfikacji dla każdego kolejnego roku i dzięki temu jesteśmy w stanie dostrzec momenty przyspieszającej lub zwalniającej zmiany językowej. Założenie, które tu cały czas czynimy, jest oczywiście takie, że duża skuteczność klasyfikacji świadczy o dużej różnicy stylistycznej: jeśli trafi się jakiś rok o szczególnie wysokim współczynniku dokładności klasyfikacji, będzie to świadczyło o istnieniu silnej cezury stylistycznej w dziejach polszczyzny.

Nakreślony powyżej ogólny zarys metody będzie oczywiście wyglądał nieco inaczej w szczegółach, np. założona hipotetyczna data głównego przełomu nie będzie dobrana losowo, lecz tak, by analizie poddać całe pokrycie chronologiczne korpusu. Przesunięcie hipotezy roboczej także nie musi być jednoroczne (np. 1600, 1601, 1602, ...), lecz zostanie dostosowane do wielkości korpusu i możliwości obliczeniowych komputera (np. 1600, 1610, 1620, ...). Po przeprowadzeniu różnych testów wypracowaliśmy algorym, którego działanie przedstawia następujący opis:

1) tworzymy korpus z tekstami dość równomiernie rozmieszczonymi na osi czasu,
2) dzielimy korpus chronologicznie na dwie części, powiedzmy pierwsze 10 lat i pozostałe lata; data podziału korpusu jest hipotetyczną datą przełomu epok,
3) przeprowadzamy klasyfikację i notujemy jej dokładność,
4) przesuwamy hipotetyczną datę przełomu o pięć lat,
5) powtarzamy pkt 3–4 aż do granicy, za którą daty nie da się przesunąć.

Tak skonstruowany prototyp wymaga jeszcze kilku dalszych wyjaśnień. Pierwsza data w korpusie jest uzależniona od tego, ile tekstów z poprzedzających lat mamy do dyspozycji. Musi ich być tyle, by dało się z nich stworzyć dostatecznie duży zbiór trenujący. *Mutatis mutandis* dotyczy to także daty ostatniej. Również liczba lat, o którą przesuwamy datę graniczną, jest uzależniona od gęstości pokrycia korpusu – można przyjąć okres 5 lat, jeżeli każde kolejne pięciolecie powoduje, że przynajmniej kilka tekstów przesuwa się z epoki późniejszej (nazwijmy ją *post*) do wcześniejszej (*ante*). Klasyfikacja jest przeprowadzana wielokrotnie, tak, by każdy tekst został sklasyfikowany, a zarazem choć raz znalazł się w zbiorze trenującym. Dzięki temu zmniejszamy rolę przypadku w ocenie skuteczności. Podobną rolę odgrywa ponawianie klasyfikacji ze zmienionymi warunkami eksperymentu, a więc zmianą liczby branych pod uwagę słów (lub etykietek gramatycznych bądź ich sekwencji) czy algorytmu klasyfikowania. Każda taka zmiana skutkuje nieco odmiennymi wynikami klasyfikacji. Jeśli jednak wyniki te są do siebie podobne, sugeruje to, że skuteczna klasyfikacja nie jest dziełem przypadku.

Kluczowe jest tu porównywanie skuteczności klasyfikacji. Jeśli jest ona niska, to oznacza, że dwie hipotetycznie wyróżnione epoki nie różnią się między sobą. Ponieważ z zasady przeprowadzamy klasyfikację binarną (wcześniejszy – późniejszy), to gdyby klasyfikować za pomocą rzutu monetą, czyli zupełnie przypadkowo, powinniśmy uzyskać dokładność około 50%, a więc średnio połowa tekstów powinna być zakwalifikowana prawidłowo i do tej wartości powinniśmy odnosić nasze oceny. Oznacza to, że dokładność klasyfikacji niewiele przekraczającą 50% można uznać za dzieło przypadku (zupełny brak sygnału chronologicznego).

Wyższa dokładność oznacza, że cechy różnicujące lepiej zdradzają epokę, do której tekst przynależy. Możemy więc przyjąć, że punktem granicznym epok jest moment, w którym klasyfikacja osiąga najwyższą dokładność. I tu ujawnia

się przewaga zaproponowanej procedury nad zasugerowanym powyżej rozwiązaniem, by sprawdzić, czy rzeczywiście przyjęte w językoznawstwie daty graniczne pozwalają na skuteczną klasyfikację. Przykładowo, jeżeli przyjęlibyśmy, że współczesna polszczyzna zaczyna się od roku 1945, to przeprowadzalibyśmy klasyfikację jedynie dla epoki przed i po roku 1945. Wtedy jednak, po pierwsze, tracimy szansę odkrycia pewnych mniej spodziewanych przełomów, a po drugie – co istotniejsze – nie potrafimy ocenić skuteczności klasyfikacji. Nie chodzi tu bowiem o wartość absolutną, ale o skuteczność wyższą niż ta, którą osiągamy dla innych hipotetycznych dat granicznych.

W gruncie rzeczy jedyna zmienna, jaką jesteśmy w stanie obiektywnie zweryfikować, to skuteczność w klasyfikowaniu tekstu jako przynależnego do epoki *ante* lub *post*: znamy przecież datę powstania tekstu i stwierdzamy jednoznacznie, że albo została zaklasyfikowana zgodnie z rzeczywistą chronologią, albo nie. Nie jesteśmy natomiast w stanie zweryfikować bezpośrednio tego, czy rzeczywiście nasza metoda wskazuje na zmianę językową. Nasze rozumowanie przebiega następująco: korpus nie zawiera dwu klas dających się wyróżnić w oczywisty sposób, lecz stanowi *continuum*. Dopiero w obrębie tego *continuum* doszukujemy się optymalnego pęknięcia na dwie klasy. Poprzednie zdanie zawiera w sobie oczywistą sprzeczność – *continuum* bowiem zasadniczo wyklucza wewnętrzne podziały – zauważmy jednak, że z taką sprzecznością językoznawcy godzą się od dawna, z jednej strony wyróżniając epoki w rozwoju języka, a z drugiej przyznając, że ewolucja językowa nigdy się nie zatrzymuje.

4.3.1. Powieści polskie XIX–XX wieku

Pierwszy z serii eksperymentów został przeprowadzony na korpusie 76 polskich powieści – tym samym, którego użyliśmy w podrozdziale *Klasyfikacja nienadzorowana i sygnał chronologiczny*. Założenie badawcze było również podobne: chodziło przede wszystkim o spojrzenie na zbiór danych jednorodnych pod względem gatunkowym i pod względem długości poszczególnych tekstów.

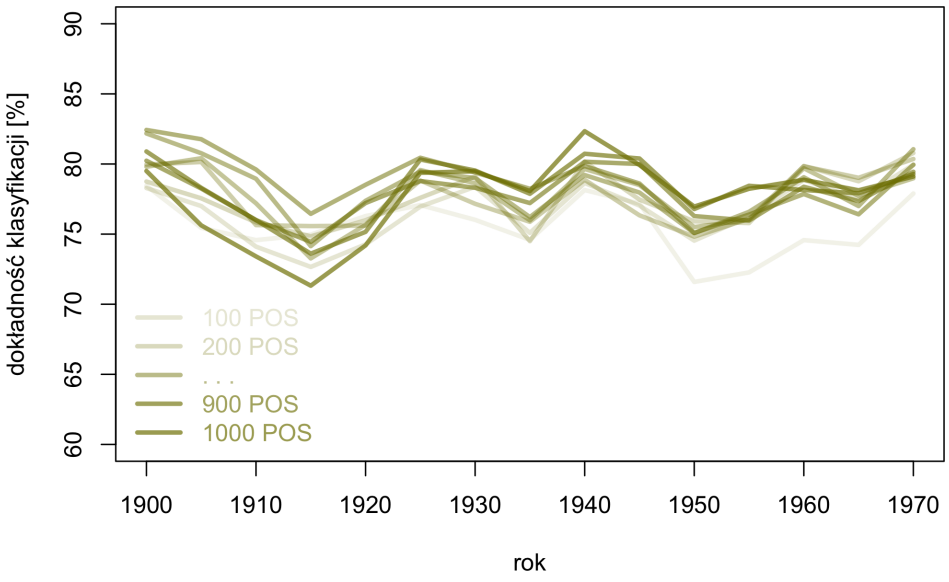
Mimo że korpus pokrywa lata 1828–2010, do testowania hipotetycznych dat przełomu przyjęliśmy – ze względu na ograniczenia metody – okres 1900–1970. Graniczne daty 1900 i 1970 były wyznaczone arbitralnie, jednak w dużej mierze narzucał je sam materiał – zbiór uczący nie mógł być zbyt mały, nie dało się więc przesuwać granic ani na zbyt wczesne, ani zbyt późne lata. Hipotetyczna granica *ante* i *post* była przesuwana co 5 lat. Jako cechy różnicujące wzięliśmy pod uwagę przeróżne kombinacje cech leksykalnych i gramatycznych, połączonych w bi- lub trigramy. Jako metodę klasyfikacji przyjęliśmy algorytm najbliższych skurczonych centroid (ang. *nearest shrunken centroids*), który dobrze sobie radzi z dużą liczbą cech wejściowych (Tibshirani, Hastie, Narasimhan i Chu, 2002). Rezultat zastosowania klasyfikacji trigramów kategorii gramatycznych dla lat 1900, 1905, 1910, . . . , 1970 pokazany został na Rys. 4.7.

Rysunek 4.7. Skuteczność klasyfikacji do grupy *ante* i *post* dla danego roku na podstawie trigramów kategorii gramatycznych.

Przedstawione na wykresie zależności należy interpretować następująco: na osi *y* przedstawiono skuteczność atrybucji liczoną jako procent tekstów prawidłowo przypisanych do grup *ante* oraz *post*; punkty na osi *x* reprezentują lata. Wykres jest zapisem dziesięciu niezależnych eksperymentów – rezultaty każdego z nich ukazuje linia ciągła o innym nasyceniu koloru. W pierwszym eksperymencie wzięliśmy pod uwagę 100 najczęstszych trigramów kategorii gramatycznych, w kolejnym 200 itd. aż do 1000. Każda zatem linia obrazuje skuteczność klasyfikacji przy innej liczbie najczęstszych trigramów.

Jeżeli spróbujemy rozpoznać, czy tekst powstał przed rokiem 1900 czy po nim[11], to skuteczność klasyfikacji dla 100 najczęstszych trigramów wynosi poniżej 70%, dla 200 jest nieco powyżej 70%, zaś dla wartości między 300 a 1000 trigramów mniej więcej 75%. Kiedy zadamy to samo pytanie odnośnie do roku 1905 – skuteczność klasyfikacji nieco wzrośnie, niezależnie od tego, ile najczęstszych trigramów weźmiemy pod uwagę. I tak kolejno, aż do roku 1925, kiedy obserwujemy lokalne maksimum. Skuteczność atrybucji do klas „teksty powstałe przed rokiem 1925" i „teksty powstałe po roku 1925" mieści się w przedziale 80–85%. Dla kolejnego punktu, czyli roku 1930, skuteczność atrybucji będzie minimalnie mniejsza z tendencją ku dalszemu opadaniu, by osiągnąć

[11] W tym miejscu warto przypomnieć, że najstarszy tekst tego korpusu pochodzi z roku 1826, zaś tekstów powstałych przed rokiem 1900 jest 48.



Rysunek 4.8. Skuteczność atrybucji do grupy *ante* i *post* dla danego roku na podstawie najczęstszych leksemów.

wypłaszczenie ciągnące się aż do lat 1955–1960 – momentu drugiego lokalnego maksimum. Po roku 1960 skuteczność znów spada do około 75%. Wprawdzie oba lokalne maksima nie wybijają się znacząco na tle pozostałych wyników, można jednak na ich podstawie ostrożnie wnioskować o dwóch momentach przyspieszenia zmiany: w okolicach roku 1925 i 1960.

Mimo że korpus był nieduży i niezróżnicowany gatunkowo, a metoda (jeszcze) niedopracowana w szczegółach, dało się dostrzec pewne interesujące prawidłowości. Dalecy jesteśmy od prostego łączenia faktów językowych z wydarzeniami historycznymi, ale trudno nie zauważyć, że rok 1925 przypada na okres bezpośrednio po odzyskaniu przez Polskę niepodległości, a lata 1955–1960 nakładają się na okres tzw. odwilży październikowej. W obu wypadkach były to momenty istotne w kształtowaniu się polskiej literatury, w tym oczywiście powieści; wyniki eksperymentu zdają się sugerować, że ważna zmiana nastąpiła też w ich warstwie stylistyczno-językowej.

Wyniki uzyskane dla trigramów kategorii gramatycznych (Rys. 4.7) zyskują nowe spojrzenie, gdy porównać je z przebiegiem zmian leksykalnych opartych na najczęstszych słowach jako predyktorach (Rys. 4.8). Od razu trzeba zaznaczyć, że wyraźnie wyodrębionych maksimów próżno się na tym wykresie doszukiwać, poza lekko zarysowanym wierzchołkiem w okolicach roku 1940; zwraca jednak uwagę fakt, że dwa niezależne testy dla dwóch różnych typów predykto-

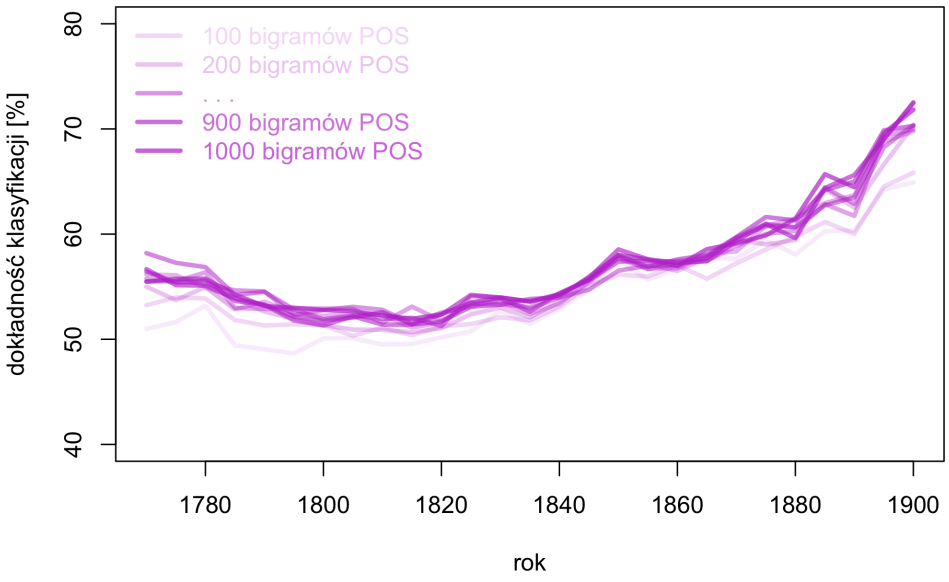
rów, tj. leksyki i gramatyki, dają odmienne wyniki. Szczególnie mocno widać to w pierwszych dwóch dekadach XX wieku: podczas gdy jedna z trajektorii zwalnia (leksyka), druga w tym samym czasie przyspiesza (gramatyka). Rzecz jasna można poddać w wątpliwość zarówno wyniki, jak i samą metodę – spodziewalibyśmy się przecież względnie kolinearnych przebiegów zmian. Zarazem jednak można odwrócić tok rozumowania i postawić następujące pytania (niestety niemożliwe do podjęcia na tym etapie badań): Czy wszystkie poziomy języka muszą ewoluować równolegle? Czy zmiana w obrębie składni nie może zajść w innym czasie niż zmiana w obrębie leksyki?

4.3.2. Corpus of Late Modern English Texts

Drugi eksperyment (zob. Eder i Górski, 2016) jest bliźniaczo podobny do poprzedniego, tyle że interesowało nas poszerzenie bazy materiałowej o różne style funkcjonalne i zróżnicowanie tekstów pod względem ich długości. Dlatego użyliśmy ponownie korpusu tekstów późnonowoangielskich CLMET 3.0: o jego budowie i zawartości była mowa w poprzednim rozdziale. Pokrycie korpusu pozwoliło na przyjęcie lat 1770–1900 za daty graniczne analizy. Jako cech różnicujących ponownie użyliśmy słów oraz etykietek klas gramatycznych, zliczanych w postaci pojedynczych jednostek, ale też łączonych w bigramy i trigramy. Do klasyfikacji, tak jak i poprzednio, użyliśmy metody najbliższych skurczonych centroid. Wyniki dla bigramów kategorii gramatycznych zostały przedstawione na Rys. 4.9.

Mimo pewnych różnic między wynikami dla 100, 300 czy 1000 najczęstszych bigramów ogólna trajektoria poszczególnych linii jest stosunkowo jednorodna – cała wiązka przyjęła kształt łagodnej litery *u*. Jeśli dla roku 1770 skuteczność klasyfikacji oscyluje wokół bardzo niskiej wartości 55%, tak później schodzi jeszcze mocniej w dół, by około roku 1810 osiągnąć poziom zupełnej nieprzydatności algorytmu (przypomnijmy: dokładność klasyfikacji rzędu 50% można uzyskać przez czysto losowe przyporządkowanie tekstów; jeśli algorytm zwraca taką dokładność, oznacza to, że nie jest w stanie wykryć żadnego sygnału). Po roku 1820 widzimy powolny, jednostajny wzrost skuteczności, dochodzący do wartości 70% prawidłowo przyporządkowanych tekstów dla roku 1900 (ściślej: dla tekstów sprzed i po roku 1900). Świadczy to o rzeczywistej różnicy między klasami *ante* i *post* na przełomie wieków XIX i XX, a zatem o dającym się wykryć sygnale chronologicznym. Nie mamy danych wejściowych dla kolejnych dekad XX wieku, nie możemy więc orzec, czy po roku 1900 skuteczność nadal by rosła czy też mamy do czynienia z cezurą, po przekroczeniu której zmiana językowa wytraca impet.

Próbując dokonać interpretacji uzyskanych wyników eksperymentu, zupełnie nieoczekiwanie natrafiamy na fundamentalną przeszkodę metodologiczną, której chyba na obecnym stanie badań nie uda się przezwyciężyć. Gdy na przykład badacz historii języka polskiego proponuje uznać rok 1543 za „umowną" cezurę



Rysunek 4.9. Skuteczność atrybucji do grupy *ante* i *post* dla danego roku na podstawie bigramów kategorii gramatycznych.

między epoką staropolską i średniopolską, ma pewną swobodę w dobieraniu faktów językowych i pozajęzykowych na poparcie swojej tezy, np. rok 1543 można uzasadniać publikacją *De revolutionibus* Kopernika i *Krótkiej rozprawy* Reja, ewentualnie wspomnieć śmierć Klemensa Janickiego i przywołać jakieś ikoniczne manifestacje kultury dojrzałego renesansu. Zupełnie inaczej jest z interpretacją przełomów wyznaczonych przez algorytm uczenia maszynowego: w tym wypadku sytuacja badacza wcale nie jest komfortowa, bo nie jesteśmy w stanie zmodyfikować danej cezury o kilkanaście czy kilkadziesiąt lat, dostosowując ją *ex post* do jakichś faktów językowych i uznając za „umowną" granicę epok.

Paradoks polega więc na tym, że chcąc zobiektywizować badanie zmiany w języku przez użycie metod kwantytatywnych, popadamy w niebezpieczeństwo tzw. efektu potwierdzenia (ang. *confirmation bias*), czyli dostrzegania faktów popierających badaną hipotezę przy jednoczesnym ignorowaniu faktów przeciwnych. W konsekwencji ostatni etap procedury badawczej, jakim jest interpretacja wyników, może całkowicie zniweczyć marzenia o zobiektywizowanym poznaniu zmian językowych. Wspomniany problem pojawił się już w poprzednim eksperymencie, gdy zmianę w polszczyźnie lat 50. XX wieku próbowaliśmy łączyć z tzw. odwilżą październikową, tutaj odzywa się z jeszcze większą siłą. Chcemy więc podkreślić z naciskiem, że mamy świadomość niebezpieczeństw wyjaśniania zmiany językowej *ex post*, dlatego naszym poniższym interpretacjom

(w tym podrozdziale i w następnych) nadajemy charakter eseistyczny i po części spekulatywny.

W interpretacji wyników niniejszego eksperymentu rzuca się w oczy fakt, że trajektoria z Rys. 4.9 wskazuje pośrednio na koniec epoki wiktoriańskiej. Pośrednio, gdyż królowa Wiktoria umiera w roku 1901, jednak ze względu na to, że korpus CLMET 3.0 nie wychodzi poza rok 1920, nie dało się przesuwać daty granicznej dalej. W każdym razie skuteczność klasyfikacji pomiędzy samym tylko rokiem 1860 a 1900 wzrasta o kilkanaście punktów procentowych i być może – choć na podstawie niniejszych badań nie da się tego stwierdzić – wzrosłaby jeszcze bardziej w kolejnych kilkunastu latach. Idąc w tych spekulacjach dalej, można zauważyć, że okres regencji (w szerszym rozumieniu, tj. przed wstąpieniem Wiktorii na tron) to okres, kiedy skuteczność klasyfikacji spada do poziomu przypadku. Chcielibyśmy wszakże podkreślić, że ani wielkość korpusu, ani jego reprezentatywność, ani wreszcie charakter zmiany językowej – która może, ale wcale nie musi być barometrem wydarzeń historycznych – nie upoważniają nas do formułowania daleko idących wniosków.

4.3.3. Corpus of Historical American English

Eksperymenty omówione w dwóch powyższych podrozdziałach pokazały pewien potencjał metody klasyfikacji „kroczącej", ale kilka problemów zostało nierozwiązanych. Przede wszystkim należy zwrócić uwagę, że wyznaczaliśmy hipotetyczną datę przełomu i wybieraliśmy losowo teksty sprzed tej daty i po niej, nie biorąc pod uwagę możliwych dysproporcji okresów *ante* oraz *post*. Przykład z eksperymentu na polskich powieściach: o ile rok 1935 daje dwa podkorpusy mniej więcej równej długości, o tyle dla roku 1970 klasa *ante* obejmuje wielokrotnie dłuższy okres niż klasa *post* (1824–1970 wobec 1970–2010). Oczywiście im bardziej badana data przełomu oddala się od chronologicznego środka korpusu, z tym większą asymetrią obu klas mamy do czynienia.

Drugi problem polega na tym, że teksty o mocno zróżnicowanej długości obciążone są efektem ubocznym w postaci silnego zaburzenia sygnału stylistycznego (zob. Franzini i in., 2018). Gdy teksty do zbioru trenującego są wybierane losowo, istnieje ryzyko, że reprezentacja klas *ante* i *post* dla danego roku będzie dawać wynik fałszywie dodatni. (Eksperyment na polskich powieściach był wolny od tego problemu właśnie dlatego, że był oparty wyłącznie na powieściach, a zatem wyłącznie na tekstach długich.)

Trzeci wreszcie niedostatek prototypu „kroczącej" klasyfikacji, w jakiś sposób związany z drugim, zasadzał się na samej procedurze konstruowania zbioru uczącego z losowo dobranych tekstów. Skoro bowiem kompozycja zbioru uczącego zależy mocno od przypadku, to i końcowa skuteczność klasyfikacji może być zawyżona lub zaniżona – chyba że odwołamy się do sprawdzianu krzyżowego (ang. *cross-validation*), o którym była mowa na początku tego roz-

działu. Zamiast zatem losować teksty tylko jeden raz dla danej daty, trzeba by je losować wielokrotnie i zmierzyć średnie zachowanie klasyfikatora przy różnych kompozycjach klas *ante* i *post*.

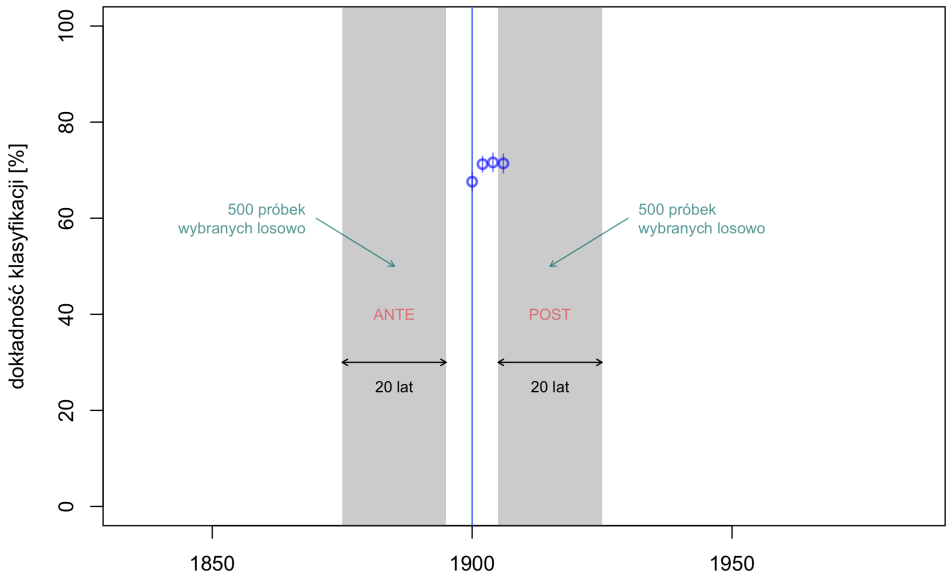
Wyartykułowanie trzech powyższych zastrzeżeń pozwala ulepszyć prototyp metody. Po pierwsze zatem: tym razem nie dzielimy całego korpusu na dwie części, lecz wyodrębniamy dwa podkorpusy obejmujące taki sam okres, np. dwudziestolecie. Dla większej precyzji wyników wprowadzamy dodatkową przerwę czasową – w naszym wypadku dziesięcioletnią – oddzielającą podkorpus *ante* i *post*. I tak, jeśli naszym hipotetycznym rokiem przełomu miałby być 1850, to konfrontujemy ze sobą podkorpus obejmujący lata 1825–1845 oraz 1855–1875. Modyfikacja ta, nieobecna we wcześniejszych eksperymentach, jest umotywowana następująco: zakładamy, że język ewoluuje nieustannie, a więc kolejne fazy rozwoju języka w istocie tworzą *continuum*. Jeśli wprowadzamy dodatkowy odstęp między podkorpusami, to w miejsce *continuum* otrzymujemy dwa bardziej od siebie odmienne okresy dwudziestoletnie.

Po drugie: nie wybieramy losowo pojedynczych tekstów, lecz cały dostępny materiał z bieżącego dwudziestolecia *ante* łączymy w jeden duży metatekst i dopiero z niego losujemy ustaloną liczbę próbek o tej samej długości – w naszym wypadku było to 500 losowych próbek po 1000 słów. To samo robimy dla dwudziestolecia *post*. Dzięki zastosowaniu takiej procedury próbkowania udaje się zniwelować wpływ długich i krótkich tekstów na końcowy wynik, a także wpływ różnych gatunków i stylów funkcjonalnych. W pewnym uproszczeniu można powiedzieć, że losując próbki opisanym powyżej sposobem, dostajemy nie tyle poszczególne teksty, ile język jako taki.

Po trzecie wreszcie: dla każdej testowanej daty hipotetycznego przełomu użyjemy sprawdzianu krzyżowego, czyli powtórzymy całą procedurę wielokrotnie (zdecydowaliśmy się na 100 powtórzeń) i za każdym razem będziemy losować próbki do zbiorów *ante* oraz *post*, by przeprowadzić niezależny test klasyfikacyjny w każdej ze stu iteracji. Oczywisty koszt, jaki musimy ponieść, to stukrotnie dłuższy czas obliczeń w porównaniu do eksperymentów opisanych powyżej, zyskujemy jednak znacznie większą wiarygodność otrzymanych wyników.

Poglądowy schemat nowej wersji procedury został przedstawiony na Rys. 4.10 (dla przykładowego roku 1900, przy czym zaznaczono skuteczność klasyfikacji dla roku 1900 i dodatkowo dla 1902, 1904 oraz 1906). Procedura przedstawiona w postaci algorytmu krok po kroku byłaby zatem następująca:

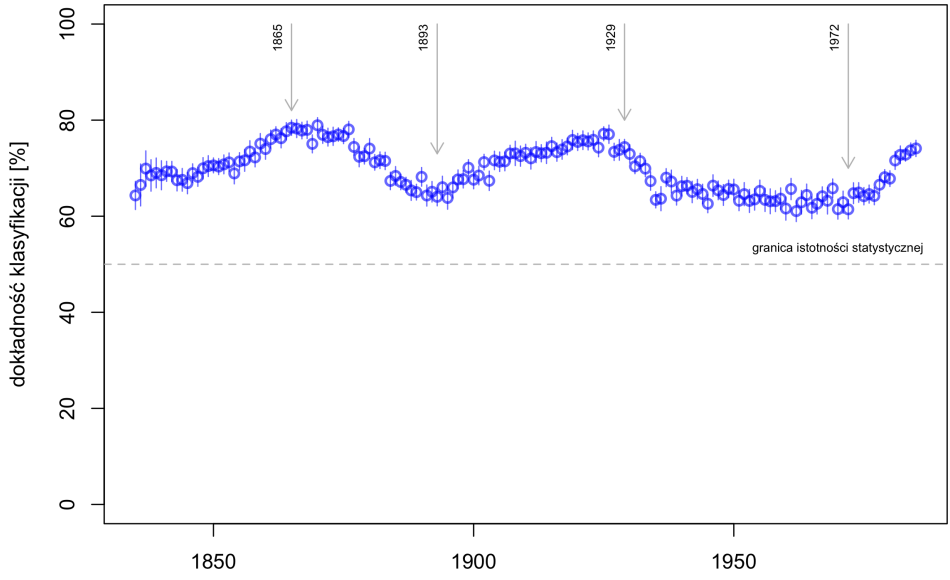
1) tworzymy korpus z tekstami dość równomiernie rozmieszczonymi na osi czasu,
2) z korpusu wydzielamy dwa podkorpusy (*ante* i *post*) obejmujące okres 20 lat przed i po hipotetycznej dacie przełomu epok,
3) w podkorpusie *ante* łączymy wszystkie dostępne teksty w jeden długi ciąg, to samo robimy w podkorpusie *post*,



Rysunek 4.10. Schemat procedury „kroczącej" klasyfikacji. Dla bieżącej daty (tutaj: dla roku 1900) wyłaniane są dwa podkorpusy, na podstawie których dokonywana jest klasyfikacja nadzorowana; następnie w ten sam sposób sprawdzana jest skuteczność dla kolejnych lat (na wykresie zaznaczono wynik dla roku 1900, 1902, 1904 i 1906).

   a) z obu połączonych podkorpusów losujemy po 500 próbek o długości 1000 słów,
   b) wybieramy losowo połowę (tj. 250) próbek z podkorpusu *ante* i połowę próbek z podkorpusu *post* do zbióru uczącącego, pozostałe próbki tworzą zbiór testujący,
   c) przeprowadzamy klasyfikację i notujemy jej dokładność,
4) powtarzamy pkt a–c w 100 iteracjach, notujemy uśrednioną dokładność klasyfikacji ze wszystkich 100 iteracji,
5) przesuwamy hipotetyczną datę przełomu o 5 lat,
6) powtarzamy pkt 2–5 aż do granicy, za którą daty nie da się przesunąć.

Ulepszonej metody „kroczącej" użyjemy w następnym eksperymencie (zob. Eder, 2018), przeprowadzonym na korpusie historycznym angielszczyzny amerykańskiej Corpus of Historical American English (Davies, 2012). Korpus COHA liczy 400 mln słów, a więc więcej niż zrównoważona część NKJP. Składa się nań 115 000 tekstów z lat 1810–2009. Nie dba on w ogóle o zrównoważenie chronologicznie, tj. im bliżej roku 2009, tym więcej tekstów, natomiast w ramach każdej z dekad ma tę samą budowę i składa się z równych proporcji powieści,



Rysunek 4.11. Przyspieszenia zmian leksykalnych w historii angielszczyzny amerykańskiej w latach 1835–1985, na podstawie Corpus of Historical American English. Punkt na wykresie oznacza uśrednioną skuteczność atrybucji, kreseczki przecinające punkt oznaczają jedno standardowe odchylenie w górę i w dół dla danego punktu.

książek niefikcjonalnych, czasopism popularnych i dzienników. Fakt, że korpus nie jest zrównoważony chronologicznie, nie ma większego znaczenia dla opisywanych tu badań – kluczowe jest to, że jego budowa dla każdego dziesięciolecia jest bardzo zbliżona. Ze względu na to, że korpus jest duży, mogliśmy sobie pozwolić na przesuwanie hipotetycznej daty przełomu o jeden rok (podczas gdy w poprzednich eksperymentach stosowaliśmy skok pięcioletni), dzięki czemu byliśmy w stanie znacząco zwiększyć rozdzielczość metody.

Na Rys. 4.11 przedstawione zostały wyniki dla 1000 najczęstszych słów, klasyfikowanych za pomocą tego samego algorytmu co poprzednio, czyli najbliższych skurczonych centroid. Jako cech różnicujących użyliśmy wyrazów tekstowych, nie przeprowadzając dodatkowej lematyzacji. Każdy punkt na wykresie oznacza średnią dokładność klasyfikacji w 100 powtórzeniach sprawdzianu krzyżowego, kreseczki przecinające punkty oznaczają wartość jednego odchylenia standardowego w górę i w dół od wartości średniej (w pewnym uproszczeniu można by powiedzieć, że przedział wyznaczany przez owe kreseczki oznacza szarą strefę niepewności wyników). Warto zauważyć, że nawet jeśli w początkowych dekadach kreseczki są znacząco dłuższe niż pod koniec XX wieku, to i tak klasyfikator

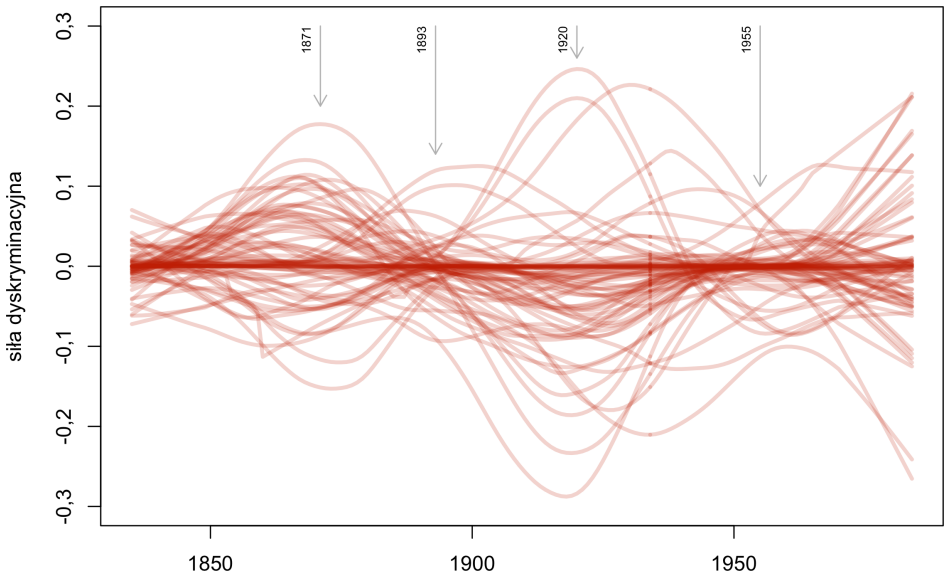
zachowuje się bardzo stabilnie – odchylenia od średniej są w gruncie rzeczy bardzo małe, rzędu 1–2 punktów procentowych.

Pierwsza i oczywista obserwacja jest taka, że skuteczność klasyfikacji nigdy nie spada poniżej 60% (co dowodzi istnienia śladów sygnału chronologicznego w każdym podkorpusie), ale przede wszystkim – że wyniki układają się sinusoidalnie, łagodnie się wznosząc i opadając w nieregularnych interwałach o długości mniej więcej kilkudziesięciu lat. Owa łagodnie sinusoidalna trajektoria stanowi najważniejszy dowód na postawioną powyżej hipotezę roboczą, że ewolucja języka wcale nie musi być liniowa – że równie dobrze można się spodziewać trendu cyklicznego, złożonego z naprzemiennych okresów „burzy i naporu" i okresów względnej stagnacji.

Następna rzecz, która zwraca uwagę, to miejsca kulminacyjne wyznaczające największe przyspieszenie zmian w języku (w tym wypadku: w amerykańskiej odmianie angielskiego). Pierwszy taki językowy – czy raczej stylistyczny – przełom ma miejsce w późnych latach 60. XIX wieku, czyli bezpośrednio po wojnie secesyjnej (1861–1865), drugi w latach 20. XX wieku, a zatem w okresie *prosperity* bezpośrednio przed tzw. wielkim kryzysem w 1929 roku, trzecia zaś kulminacja, jeszcze niecałkowicie uformowana, zaczyna się pod koniec XX wieku i trwa aż do końca pokrycia korpusu COHA. Nie trzeba tu po raz kolejny powtarzać, że doszukiwanie się bezpośrednich korelacji między wydarzeniami historycznymi i przełomami stylistycznymi jest narażone na różnego rodzaju świadome i nieuświadomione uprzedzenia badacza i przez to może prowadzić do zakłamania wyników eksperymentu. Mimo to współbieżność trzech momentów przyspieszenia ewolucji języka i kilku wydarzeń fundamentalnych dla kształtowania się amerykańskiej kultury jest uderzająca.

Jakkolwiek interesujące okazać się mogą dotychczasowe wyniki powyższego eksperymentu – pokazują bowiem zmianę niejako „z lotu ptaka" – językoznawca historyczny może odczuwać pewien niedosyt, a może nawet rozczarowanie. Stosowana przez nas metoda nie mówi mianowicie, co odpowiada za poprawę klasyfikacji, a więc które z dyskryminatorów kryją się za poszczególnymi zmianami językowymi. Doświadczenie w stylometrii uczy, że co prawda sygnał autorski jest rozproszony po wielu cechach tekstu (słowach bądź ich sekwencjach, kategoriach gramatycznych bądź ich sekwencjach itp.), ale na ogół to niewielka ich liczba decyduje o sygnale autorskim (por. np. Koppel i Schler, 2004). Rodzi się więc pytanie, czy tak samo będzie wyglądał sygnał chronologiczny. Czy za zmianę w języku odpowiada niewielka liczba słów czy też, przeciwnie, suma bardzo wielu cech o indywidualnie małej sile dyskryminacyjnej?

By odpowiedzieć na to pytanie, dokonaliśmy ekstrakcji cech o największym wpływie na klasyfikację. Należy tu zastrzec, że nie chodzi w tym wypadku o największe zmiany frekwencji w liczbach bezwzględnych, ale o wagi klasyfikatora. Obserwacja bezpośrednich frekwencji mogłaby wprawdzie pokazać pewne trendy ewolucyjne słów, ale naszym celem było wydobycie cech, których siła dys-

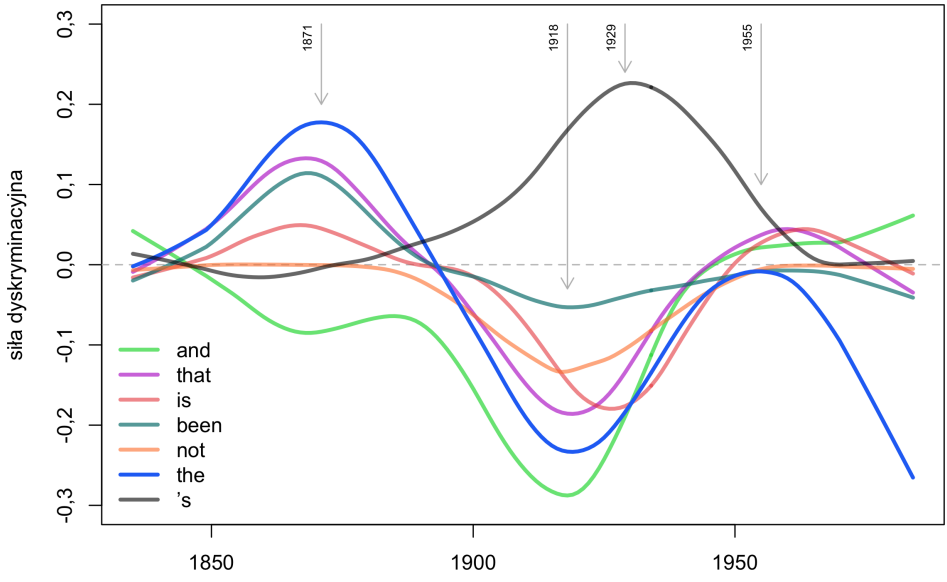


Rysunek 4.12. 76 cech (leksemów), które w największym stopniu odpowiadają za przemiany stylistyczne w korpusie języka amerykańskiego COHA.

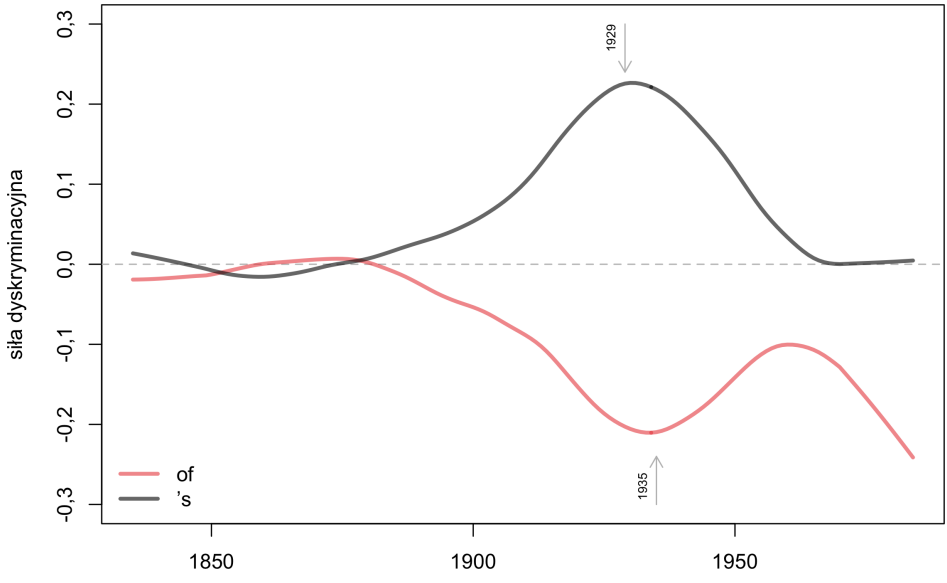
kryminacyjna ujawnia się na tle wszystkich innych predyktorów – dlatego właśnie przyglądamy się wagom cech (ściślej: aposteriorycznym prawdopodobieństwom zwróconym przez klasyfikator). Udział najistotniejszych kilkudziesięciu cech w skuteczności klasyfikacji dla kolejnych dat granicznych został przedstawiony na Rys. 4.12.

Największym zaskoczeniem po wyrysowaniu trajektorii cech dyskryminacyjnych okazał się oczywiście ich sinusoidalny kształt: wprawdzie wyniki skuteczności klasyfikacji (Rys. 4.11) sugerowały, że podobne fluktuacje zauważymy też w zachowaniu samych cech, jednak nie oczekiwaliśmy aż tak czytelnych przebiegów. Druga rzecz, której się nie spodziewaliśmy, to współbieżność wielu trajektorii: okazało się, że słowa mają tendencję do układania się w równoległe wiązki (zjawisko to widać wyraźniej na Rys. 4.13). Trzy wyraźne wierzchołki sinusoidy, które wypadają mniej więcej w latach największej zmiany (por. Rys. 4.12), nie składają się z pojedynczych predyktorów, lecz z całych ich wiązek.

Najmocniejszymi predyktorami okazały się następujące leksemy (w kolejności ich siły dyskryminacyjnej): *the*, *and*, *week*, *that*, *'s*, *last*, *is*, *be*, *of*, *it*, *we*, *i*, *to*, *was*, *mr.*, *our*, *my*, *been*, *not*, *u.s.*, *you*, *new*, *upon*, *there*, *has*. Są to wyrazy o dużej frekwencji, choć nie wszystkie z nich zajmują górę listy frekwencyjnej. Wśród nich warto zwrócić uwagę na kilka słów funkcyjnych, których trajektorie zostały wyrysowane osobno na Rys. 4.13.



Rysunek 4.13. Siedem cech leksykalnych (słów funkcyjnych), które w największym stopniu odpowiadają za przemiany stylistyczne tekstów w korpusie języka amerykańskiego COHA.



Rysunek 4.14. Siła dyskryminacyjna wyrazów *'s* oraz *of* w korpusie języka amerykańskiego COHA.

Mimo że wiele wyników dotychczasowych eksperymentów sugeruje istnienie zmiany raczej stylistycznej niż sięgającej głębszych struktur języka, siła dyskryminacyjna wykazywana przez liczne słowa funkcyjne zdaje się temu przeczyć. Na przykład *has* i *been* zdają się wskazywać na wzrost frekwencji czasów ciągłych, z kolei *'s* może być refleksem najpierw cofania się, a następnie ekspansji tego wykładnika posesywności (Szmrecsanyi, 2015). Widać więc wyraźnie, że przez leksykę przebijają własności gramatyczne tekstów.

Przykład z *'s* jest szczególnie interesujący. Rys. 4.14 pokazuje konkurencję tej formy w porównaniu do alternatywnego sposobu wyrażania posesywności, tj. do formy *of*. Trudno nie zauważyć, że obie formy łączy relacja zgoła toksyczna: przyrost *'s* musi się odbywać kosztem *of* i odwrotnie. Mamy tu oczywiście do czynienia z kolejnym dowodem, że język nie lubi próżni, jeśli jakaś cecha językowa wycofuje się na przestrzeni dziejów, jej funkcję na ogół przejmuje inna cecha. Wspominamy o tym nie bez powodu. Otóż w eksperymencie wyszukiwaliśmy form wyłącznie na podstawie ich kształtu graficznego, co w praktyce oznaczało, że zapytanie *'s* zwracało zarówno formy z dopełniaczem saksońskim, jak i formy ściągnięte (*it's*, *he's* i podobne). I pomimo tej niedoskonałości wyszukiwania otrzymane wyniki okazały się co najmniej satysfakcjonujące. Wielką zaletą wielowymiarowych metod klasyfikacyjnych jest ich zdolność dostrzegania sygnału nawet w zaszumionych danych (Eder, 2013; Franzini i in., 2018), nie jesteśmy więc szczególnie zdziwieni zaobserwowanym przez nas zjawiskiem. Przykład powyższy dostarcza natomiast kolejnych argumentów, że w językoznawstwie korpusowym duża ilość danych pozwala często nadrobić ich niedostatki jakościowe.

Wracając do wyników z Rys. 4.13: uwagę zwraca kolinearność słów takich jak *the*, *and*, *that*, *is*, *been*, jakby stanowiących jedną wiązkę: w tym samym momencie wykazują największą siłę dyskryminacyjną (około roku 1781, po zakończeniu wojny secesyjnej) i również w podobnym czasie zamieniają się w silny predyktor o ujemnej sile (około roku 1918). Nie oznacza to bynajmniej, że wszystkie słowa funkcyjne zachowują się tak samo. Jako przykład słowa zmierzającego przez kolejne stulecia swoją własną trajektorią możemy podać dopełniacz saksoński *'s*. Pokusa – ulegliśmy jej już kilka razy powyżej – by wyjaśniać zmianę językową konkretnymi wydarzeniami historycznymi, pojawia się i tutaj. Czy jednak da się w sensowny sposób wyjaśnić fakt, że dopełniacz saksoński *'s* święcił triumfy za czasów amerykańskiej prohibicji?

4.3.4. Korpus diachroniczny polszczyzny 1380–1850

W kolejnym eksperymencie wracamy do historii języka polskiego. Tym razem dane pochodzą z najpełniejszej wersji korpusu opisanego w rozdziale *Korpus*, pokrywającego lata 1380–1850. Eksperyment ma odpowiedzieć na podobne pytanie, jakie stawialiśmy poprzednim korpusom, mianowicie będziemy chcieli wykryć daty przyspieszenia zmiany w języku, może nawet daty większych

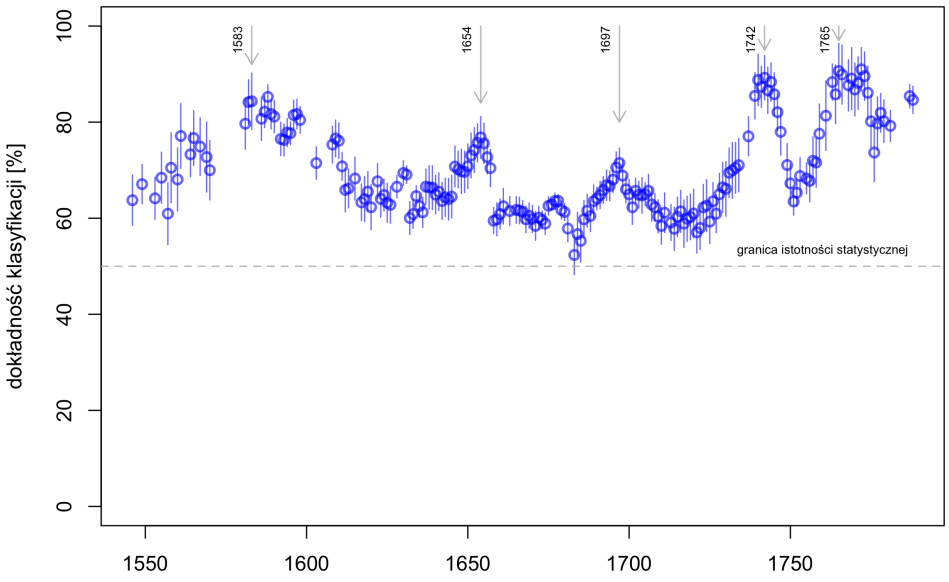
przełomów, bez formułowania żadnych założeń wstępnych. Znamy wprawdzie „umowne" granice epok zaproponowane przez badaczy polszczyzny, ale na potrzeby niniejszego eksperymentu będziemy traktować każdy kolejny rok pokrycia korpusu z taką samą powagą, żeby nie przeoczyć ewentualnych przełomów znalezionych przez algorytm uczenia maszynowego.

Jako cechy odróżniające przyjęliśmy (w różnych analizach) najczęstsze słowa, najczęstsze bi- oraz trigramy kategorii gramatycznych, wreszcie cechy oparte na pełnych tagach NKJP, zawierających znacznie bogatszą informację gramatyczną w porównaniu do samych kategorii gramatycznych. Zastosowaliśmy ulepszoną wersję naszej metody, opisaną w poprzednim podrozdziale, biorącą pod uwagę nie tyle cały korpus, ile dwa okienka czasowe o rozmiarze 20 lat z odstępem 10 lat między nimi (por. Rys. 4.10). Dla każdego testowanego roku użyliśmy sprawdzianu krzyżowego, dzięki czemu trenowanie i testowanie klasyfikatora odbywało się za każdym razem stukrotnie; uśredniona dokładność klasyfikacji dla poszczególnych lat była poszukiwanym przez nas końcowym wynikiem eksperymentu.

Przykładowe wyniki dla bigramów kategorii gramatycznych przedstawione są na Rys. 4.15. Od razu trzeba tu zaznaczyć, że nasz korpus – największy istniejący dziś polski korpus diachroniczny, nie licząc korpusu Chronopress[12] – nadal jest ponad trzydziestokrotnie mniejszy od korpusu COHA, co niestety przekłada się na mniej dokładne wyniki klasyfikacji. Warto na przykład zwrócić uwagę na to, że dla niektórych okresów nie byliśmy w stanie uzyskać żadnych wyników, np. w latach 60. XVI wieku czy w latach 80. XVIII wieku. Podobny kłopot będziemy mieli z interpretacją najmocniejszych predyktorów (zob. Rys. 4.16), jako że niektóre z nich dają dość chaotyczne wyniki. Wierzymy jednak, że mimo niedoskonałości korpusu pewne istotne tendencje dadzą się zauważyć.

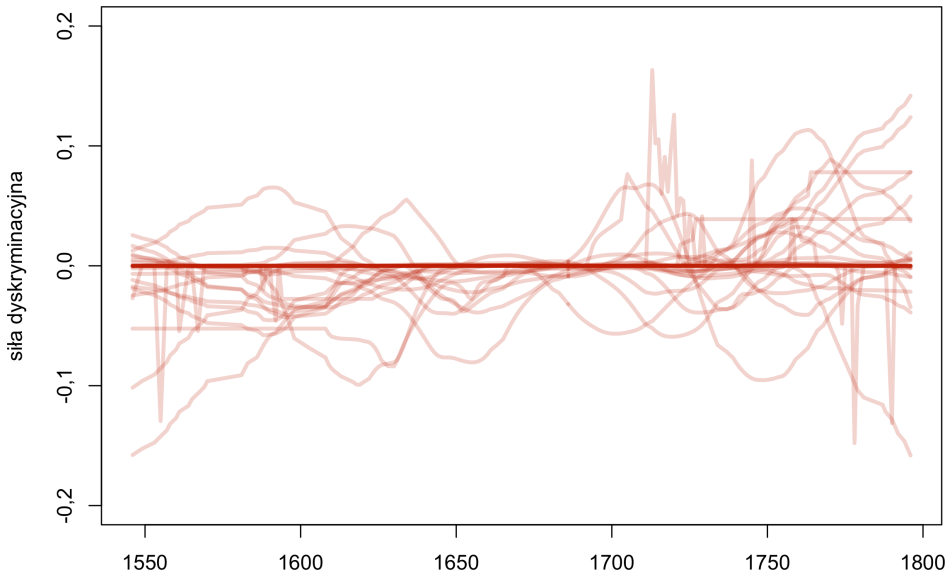
Na wykresie (Rys. 4.15) widać kilka lokalnych maksimów. Pierwsze z nich, wypadające w roku 1583, będzie oczywiście wodą na młyn zwolenników teorii o wpływie wybitnych jednostek na rozwój języka, pokrywa się bowiem niemal idealnie z rokiem śmierci Jana Kochanowskiego (1584). Pamiętając jednak o naszych własnych zastrzeżeniach na temat efektu potwierdzenia, nie będziemy przywiązywać zbyt wielkiej wagi do tej koincydencji, co jednak nie zmienia faktu, że koniec wieku XVI był w Rzeczypospolitej czasem wyczerpania się paradygmatu renesansowego i momentem narodzin nowej estetyki manierystycznej. Zmianę tę widać wyraźnie w podejmowanej przez poetów XVI/XVII wieku tematyce, trudno przypuścić, by tak mocne przewartościowania nie odcisnęły piętna także na języku i stylu.

[12] Korpus Chronopress (http://chronopress.clarin-pl.eu) zawiera 16 milionów segmentów, lecz obejmuje zaledwie jedno dziesięciolecie 1945–1955; nasz korpus ma 12 milionów słów, ale za to pokrywa kilka stuleci (1380–1850).



Rysunek 4.15. Okresy wzmożonych zmian językowych w korpusie polskim 1380–1850, na podstawie bigramów kategorii gramatycznych.

Dwa kolejne lokalne maksima wypadają w roku 1654 i pod sam koniec wieku XVII. Nie mogą się one jednak w żaden sposób równać z następnym, znacznie silniejszym wzrostem dokładności klasyfikacji w połowie wieku XVIII. Trudno jednoznacznie orzec, czy mamy tu do czynienia z jednym rozległym szczytem w latach 1740–1770, czy też z dwoma niezależnymi lokalnymi maksimami w latach 1742 i 1765. Oba te punkty są oddzielone krótkim i gwałtownym spadkiem skuteczności atrybucji, ale doprawdy trudno jednoznacznie stwierdzić, czy jest to artefakt spowodowany szczupłością danych czy też rzeczywiste zatrzymanie zmiany w języku. W każdym razie zwolennicy tradycyjnej periodyzacji znowu poczują się usatysfakcjonowani – rok 1742 (ściślej: rok 1741, gdy ukazało się *O poprawie wad wymowy* Konarskiego) uznaje się przecież powszechnie za początek oświecenia w Polsce, a rok 1765 rozpoczyna tzw. czasy stanisławowskie, wyznaczane koronacją Stanisława Augusta Poniatowskiego, publikacją pierwszego tomu *Monitora* i założeniem Szkoły Rycerskiej. Trochę mniejszą satysfakcję poczują zwolennicy teorii, zgodnie z którą za granicę epoki średniopolskiej i nowopolskiej przyjmuje się koniec wieku XVIII (Klemensiewicz, 1965). Nasz eksperyment pokazuje przecież, że największe pęknięcie ma miejsce w latach 60. XVIII wieku, a więc jakieś trzy dekady wcześniej. Nie zapominajmy jednak, że zaproponowany przełom wieku XVIII i XIX jest kolejną datą „umowną", zaproponowaną w czasach, gdy o badaniach korpusowych na większą skalę nikt

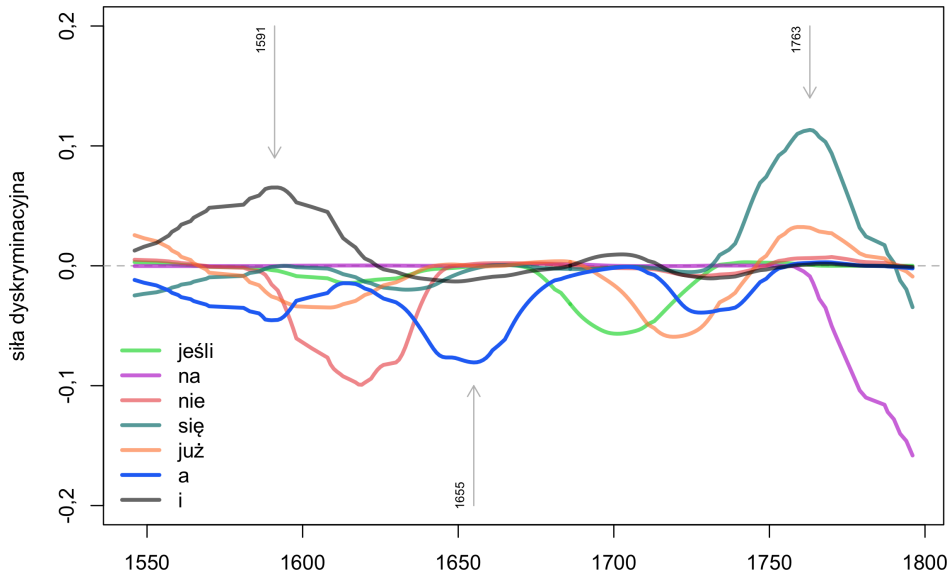


Rysunek 4.16. 34 cechy (leksemy), które w największym stopniu odpowiadają za przemiany stylistyczne w korpusie polskim 1380–1850.

nawet nie marzył. W naszym eksperymencie otrzymywaliśmy zbliżone i mimo szczupłości materiału dość stabilne wyniki dla najczęstszych słów, lematów, a także bi- i trigramów kategorii gramatycznych, co zdaje się potwierdzać istnienie mocnego przełomu w czasach wczesnego oświecenia w Polsce.

Przejdźmy następnie do omówienia najsilniejszych predyktorów. Na Rys. 4.16 przedstawiono siłę dyskryminacyjną 34 najsilniejszych predyktorów leksykalnych w funkcji czasu. Wykres ten ukazuje z całą bezwzględnością, na czym polega różnica między korpusem 400-milionowym (zob. Rys. 4.12) i 12-milionowym (Rys. 4.16): w przeciwieństwie do eksperymentu na korpusie COHA tutaj wyniki są znacznie bardziej rozmyte, a predyktory zachowują się bardziej chaotycznie, choć pewne prawidłowości można, mimo wszystko, prześledzić. Przede wszystkim tym razem nie jesteśmy w stanie dostrzec wiązek cech układających się w czytelne sinusoidy; zamiast tego widzimy całą serię niezależnych trajektorii.

Na Rys. 4.17 przedstawiamy kilka wybranych predyktorów ze zbioru słów funkcyjnych; jak widać, wybierają one na ogół żywot samodzielny, zamiast grupować się w wiązki. Z ciekawszych zjawisk warto zwrócić uwagę na konkurencję spójników *i* oraz *a* w latach 1550–1620 (z kulminacją w roku 1591), która przypomina trudną koegzystencję form *'s* oraz *of* w korpusie COHA. Zaimek *a* idzie potem własną drogą, osiągając sporą negatywną siłę dyskryminacyjną w roku 1655 i potem, choć w mniejszym zakresie, w roku 1725. Krótkotrwała



Rysunek 4.17. Siedem wybranych cech leksykalnych, które w dużym stopniu odpowiadają za przemiany stylistyczne tekstów w korpusie polskim 1380–1850.

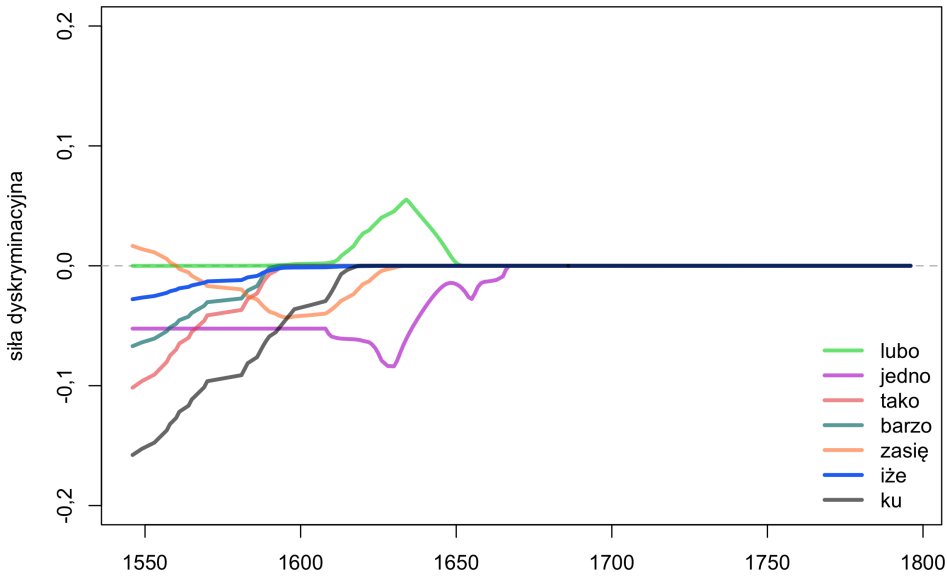
kariera zaimka zwrotnego *się* jako predyktora w drugiej połowie XVIII wieku (z kulminacją w roku 1763) z pewnością jest warta dalszych badań, odwołujących się do bogatszego materiału źródłowego. Wydaje się bowiem, że zwiększający się udział tego zaimka jest czymś więcej niż modą stylistyczną.

Tytułem uzupełnienia pokazujemy trajektorie kilku leksemów wyszłych z użycia w dobie średniopolskiej (Rys. 4.18). Oczywiście na wykresie nie znajdziemy niczego zaskakującego, byliśmy jednak ciekawi, czy archaizmy, o których istnieniu wiedzieliśmy, zachowają się tak, jak przewidywaliśmy. Był to dla nas kolejny sprawdzian działania metody.

4.3.5. Podsumowanie

Przedstawione powyżej cztery eksperymenty skłaniają do sformułowania garści uwag podsumowujących, obracających się, mimo wielu różnic dzielących same eksperymenty, wokół łączącej je problematyki. Zaczniemy od uwag dotyczących metody „kroczącej" klasyfikacji, by następnie dojść do pytań o charakterze bardziej teoretycznym, *stricte* językoznawczych.

Jak pokazały wyniki wszystkich czterech eksperymentów, skuteczność klasyfikacji na ogół mieści się w granicach 50–80%, zwykle zaś nie przekracza znacząco 70%. Oznacza to, że nawet dla optymalnego podziału co czwarty tekst jest klasyfikowany błędnie. Z punktu widzenia metod uczenia maszynowego jest



Rysunek 4.18. Siedem wybranych leksemów archaicznych i ich wpływ na zmianę językową w korpusie polskim 1380–1850.

to wynik zdecydowanie słaby. Te – same w sobie – niezadowalające wyniki mają jednak swoje wyjaśnienie. Przede wszystkim są one efektem nakładania się kilku sygnałów, wśród których sygnał chronologiczny wcale nie musi być (i najczęściej nie jest) najsilniejszy, czego dowodziliśmy w rozdziale *Klasyfikacja nienadzorowana i sygnał chronologiczny* na przykładzie korpusu CLMET 3.0. Po drugie: mimo wszystko mamy do czynienia nie tyle z dwiema wyraźnie odrębnymi klasami, ile z *continuum* (jakkolwiek dodatkowa procedura wprowadzania odstępu pomiędzy podkorpusem *ante* i podkorpusem *post* miała na celu zmniejszenie tego efektu). Nie sama jednak skuteczność klasyfikacji jest dla nas ważna, ale jej zmienność, która – przypomnijmy – wynosi nawet kilkanaście punktów procentowych, co jest skokiem znaczącym.

Podobnie możemy czuć pewien niedosyt, jeśli chodzi o przydatność uzyskanych przez nas wyników w formułowaniu wniosków na temat ewolucji języka. Nasza metoda „kroczącej" klasyfikacji bez większego kłopotu dostrzega sam fakt zmiany, ale już słabiej pozwala uchwycić to, co się zmieniło. Z punku widzenia językoznawcy jest to niewątpliwie nieco rozczarowujące. Tym niemniej zauważmy, że nie czynimy tutaj istotnego wyłomu w dotychczasowej praktyce objaśniania zmiany językowej – badacze często operowali podobnymi argumentami. Na przykład Klemensiewicz pisze: „W obrębie doby średniopolskiej można wydzielić mniejsze okresy. Okres pierwszy, obejmujący cztery dziesięciolecia

XVI w., ma charakter przejściowy. Następny, sięgający lat trzydziestych XVII w., to okres wspaniałego rozkwitu renesansowego i powolnego wyczerpywania się jego żywotności. Okres trzeci trwa do końca XVII w, a znamionuje go obniżanie się poziomu języka na tle ogólnego cofania się życia polskiego pod znakiem sarmatyzmu" (Klemensiewicz, 1965: 10).

Znacznie więcej zastrzeżeń można sformułować pod adresem naszych prób interpretacji uzyskanych wyników. Eksperyment pierwszy jako datę przełomu wskazuje rok 1925 – czy jednak mamy prawo łączyć ów fakt językowy z zakończeniem I wojny światowej i odzyskaniem przez Polskę niepodległości? Jest to w gruncie rzeczy zgodne z tym, co wielu językoznawców uważa za początek współczesnej polszczyzny. W historii zarówno powszechnej, jak i historii kultury przyjmuje się, że I wojna światowa dokonała całkowitego przeobrażenia społeczeństwa. Wiadomo też, że po odrodzeniu Polski język zaczął być używany na terenie jednego państwa, a jego nośnikiem stała się także szkoła. Rok 1925 może być interpretowany jako opóźniony efekt tych zewnętrznych zmian funkcjonowania języka, trudno bowiem oczekiwać, że natychmiast po odzyskaniu niepodległości autorzy napiszą i wydadzą znaczącą liczbę powieści.

Powyższa argumentacja musi się wydawać przekonująca – ale czy jest prawdziwa? Tylko przecież od wyobraźni badacza zależy, jak wiele argumentów uda się znaleźć na powiązanie lokalnego maksimum przypadającego na rok 1925 z wydarzeniami historycznymi. Istota problemu, o którym tu piszemy, staje się jasna, gdy zadamy inne pytanie: dlaczego nie notujemy żadnego lokalnego maksimum bezpośrednio po zakończeniu II wojny światowej, która przecież przeobraziła polskie społeczeństwo w znacznie większym stopniu? W wypadku korpusu COHA mamy dwa lokalne maksima: jedno przypada na okres po wojnie secesyjnej, drugie na czas krótko przed rokiem 1929, a więc dekadę tzw. „szalonych lat dwudziestych" (ang. *Roaring Twenties*). Znowu staramy się owym maksimom nadać jakąś interpretację *ex post*, ale te same wątpliwości pozostają: o ile wojna secesyjna była dla USA ogromnym wstrząsem i właściwie powinni byśmy się dziwić, gdyby nie odcisnęła swego piętna w języku, o tyle fakt, że procedura „nie dostrzega" wojny wietnamskiej i zmian dokonanych wtedy w kulturze amerykańskiej, jest już zastanawiający. Skłania nas to bez wątpienia do ostrożności w prostym kojarzeniu faktów językowych i pozajęzykowych.

Na koniec musimy zadać kluczowe z punktu widzenia językoznawcy pytanie, a mianowicie: gdzie leży obserwowana zmiana: po stronie *langue* czy po stronie *parole*, a może jeszcze gdzie indziej? Czy mamy tu do czynienia ze zmianą językową czy zmianą jedynie pewnej stylistycznej mody, swoistego „kostiumu literackiego"?

Niewątpliwie zmiana systemowa może znaleźć odzwierciedlenie w bigramach lub trigramach etykietek gramatycznych. Tak musi się stać, gdy np. forma syntetyczna ustępuje formie analitycznej. Porównajmy dwa fragmenty zdań wzięte z NKJP:

procesem bardziej złożonym

procesem prostszym niż

Reprezentacjami ich własności gramatycznych są następujące sekwencje:

[subst:sg:inst:m3] [adv:com] [adj:sg:inst:m3:pos]

[subst:sg:inst:m3] [adj:sg:inst:m3:com] [conj]

Podobnie musi się stać z rozprzestrzenianiem się form czasów ciągłych w języku angielskim. Przykładowe zdanie z BNC:

I have been walking

będzie reprezentowane przez ciąg etykietek[13]:

PNP VHB VBN VVG

a więc trzy bigramy [PNP VHB], [VHB VBN] oraz [VBN VVG]. Gdyby to zdanie miało zawierać zamiast czasu ciągłego czas przeszły prosty (*past simple tense*), byłoby reprezentowane przez sekwencję:

PNP VVD

Jest rzeczą oczywistą, że rozprzestrzenianie się czasów ciągłych w pewnym odcinku czasu odciśnie swoje piętno w postaci wzrostu frekwencji bigramów [VHB VBN] i [VBN VVG] oraz spadku frekwencji [PNP VVD], tym niemniej wątpliwe, by właśnie tego rodzaju zmiany stanowiły główny sygnał chronologiczny. W języku polskim takich zmian od XIX wieku nie zanotowano, a mimo to proponowana przez nas metoda okazuje się skuteczna. Pomijamy tu fakt, że rozprzestrzenianie się formy gramatycznej jest zmianą ilościową, która jest możliwa jedynie dzięki temu, że wcześniej zaszła zmiana jakościowa, polegająca na stworzeniu nowej formy[14].

Ponadto – o czym warto tu przypomnieć – gdy cechą różnicującą są etykietki gramatyczne, metoda staje się całkowicie ślepa na zmianę morfologiczną. Tager (program przypisujący wyrazom w tekście etykietki gramatyczne) przypisuje tę samą kategorię gramatyczną wyrazowi *temi* oraz *tymi*.

Tak więc to, co wychwytuje nasza metoda, to nie zmiana systemu. Czy jest to w takim razie zmiana w obrębie *parole*? Również i tę hipotezę łatwo sfalsyfikować. *Parole* bowiem to konkretne instancje użycia *langue*, a te ze swej natury są jednostkowe. Biorąc rzecz po Saussure'owsku, jedyne co w nich wspólne, to

[13] Tagset CLAWS jest dostępny pod adresem: http://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm#tagset.

[14] Takie podejście jest pewnym uproszczeniem. Zmiana jakościowa, jaką jest powstanie czasów progresywnych, jest efektem procesu gramatykalizacji, a więc wzrost frekwencji pewnych sekwencji etykietek gramatycznych powinien poprzedzać zmianę jakościową.

fakt, że stanowią emanację tego samego *langue*. Z jakiego więc powodu teksty powstałe w jednym czasie miałyby być podobne bardziej do siebie niż do tekstów powstałych kilkadziesiąt lat wcześniej, skoro wszystkie są tworzone na podstawie tego samego *langue*?

Przekonujące wydaje się stwierdzenie, że mamy tu do czynienia z ewolucją tego, co pomiędzy *langue* i *parole*, pomiędzy systemem i tworzonym według tego systemu tekstem. Można to nazwać za Hjelmslevem normą (Hjelmslev, 1942). Użytkownicy języka czynią z możliwości, jakie daje system, rozmaity użytek, niektórych konstrukcji używając znacznie rzadziej niż innych. Kompetencja językowa obejmuje zatem nie tylko znajomość systemu gramatycznego i słownika, ale także wiedzę, w jaki sposób z tego systemu korzystać. Ta wiedza również podlega ewolucji i to zapewne szybszej niż ewolucja systemu.

Metoda, którą tu opisujemy, może być czuła na wczesne stadia gramatykalizacji. Użyjmy tu spotykanego w wielu językach wykładnika czasu przyszłego – czasownika *iść*, np. w języku angielskim czy francuskim. Dopóki jest on wyrazem autosemantycznym, będzie etykietkowany jako zwykły czasownik, tym niemniej w procesie gramatykalizacji frekwencja struktur *iść* + bezokolicznik będzie wciąż wzrastała. Oczywiście, gdy proces ten osiągnie dojrzałość, to znaczy *iść* stanie się wykładnikiem czasu, zmieni się jego etykietka gramatyczna[15].

Wreszcie na pewno metoda nasza jest czuła na zmianę tego, co wyżej nazwaliśmy modą literacką, przy czym ta moda, czy też różne mody, obowiązuje nie tylko w tekstach literackich, ale także tekstach użytkowych. Jest to więc także ewolucja wzorca gatunkowego tekstów.

4.4. Przyimki w historii języka polskiego

W niniejszym podrozdziale użyjemy jednej z nienadzorowanych metod klasyfikacyjnych, by spojrzeć na historię polszczyzny poprzez pryzmat przyimków[16], a konkretnie przez pryzmat zmian w ich frekwencji. Przyimki – podobnie jak inne słowa funkcyjne – stoją na pograniczu gramatyki i słownika i choćby z tego powodu stanowią wdzięczny temat badań. Co ważne dla naszych rozważań, stanowią zamkniętą klasę, bardzo trwałą, która powiększa się właściwie jedynie poprzez przyimki złożone.

Zacznijmy od dość zdroworozsądkowego (i zarazem nieco naiwnego) założenia, że liczba funkcji danego przyimka jest ograniczona. Jeżeli założenie jest

[15] Uważny Czytelnik może nam tu zarzucić, że mylimy pewne porządki, mianowicie fakty językowe i ich opis. Zakładamy jednak, że obowiązkiem twórcy korpusu jest adekwatny opis języka, zresztą jednym z dużych wyzwań twórców korpusów diachronicznych jest wyznaczenie momentu, od którego temu samemu ciągowi liter należy zacząć przypisywać inną kategorię gramatyczną.

[16] Warto w tym miejscu przywołać pracę Krążyńskiej o staropolskich konstrukcjach z przyimkami (Krążyńska, 2001) i kolejne części tego studium.

prawdziwe, to wzrost względnej frekwencji danego przyimka byłby równoznaczny z faktem, że przejął on jedną z funkcji innego przyimka lub też przypadka. Pierwszą z tych sytuacji dobrze obrazuje los przyimka *ku*, który w dużej mierze został zastąpiony bliskoznacznym *do* (Krążyńska, 1993). Choć *ku* wciąż istnieje we współczesnej polszczyźnie, jego frekwencja gwałtownie spadła (widzieliśmy to zresztą w porzednim podrozdziale, zob. Rys. 4.18). Drugi proces ilustruje tendencja bardzo świeżej daty, mianowicie zastępowanie dopełniacza frazą przyimkową z *na*, tam gdzie dopełniacz koduje pewną abstrakcyjną relację między desygnatami dwu rzeczowników, a nie posesywność. Por. dwa przykłady z NKJP (w obu wypadkach sprawozdania z obrad Sejmu RP):

W roku 2001 średnia cena wieprzowiny wynosiła ponad 4,30

cena na wieprzowinę w miesiącu lutym to 1 zł 80 gr

W NKJP jest niewiele poświadczeń tej drugiej konstrukcji, jednak codzienne obserwacje każą sądzić, że szybko się ona rozprzestrzenia. Dość powiedzieć, że NKJP nie notuje konstrukcji *propozycja na*, jednak jest ona poświadczona w internecie, np.:

Pensjonat ORKA to doskonała propozycja na wypoczynek nad Bałtykiem!

Jeśli tendencja się utrzyma, to w konsekwencji należy się spodziewać wzrostu frekwencji przyimka *na*.

Wreszcie może się zdarzyć, że wzrost frekwencji danego przyimka idzie w ślad za wzrostem frekwencji czasowników, w których ramie walencyjnej znajduje się ów przyimek. Wszystkie wymienione powyżej czynniki – z osobna bądź wspólnie – mogą spowodować zmiany relatywnej frekwencji przyimków.

Nie zamierzamy się w tym rozdziale zajmować zmianami w poszczególnych przyimkach. Przedmiotem naszej analizy czynimy zmiany względnej częstości polskich przyimków *en masse*, w okresie pomiędzy wiekiem XV a połową wieku XIX, przy użyciu korpusu diachronicznego 1380–1850 opisanego w rozdziale *Korpus*. W ten sposób szukamy po raz kolejny odpowiedzi na pytanie, gdzie stoją słupy milowe w rozwoju języka. Chcemy sprawdzić, czy relatywna frekwencja przyimków (z wyłączeniem przyimków złożonych) w historii języka polskiego była stabilna czy też podlegała pewnym zmianom, a jeśli tak, to czy momenty głębszej zmiany pokrywają się z tym, co w historii języka polskiego wyróżniamy jako granice pomiędzy epokami. Tym razem jednak cechami różnicującymi nie czynimy 100 najczęstszych wyrazów czy 2500 trigramów kategorii gramatycznych, lecz frekwencje 18 wybranych przyimków: *bez*, *dla*, *do*, *ku*, *między*, *na*, *nad*, *o*, *od*, *po*, *pod*, *przed*, *przez*, *przy*, *u*, *w*, *z*, *za*. Nie powinniśmy oczywiście spodziewać się żadnych gwałtownych zmian w obrębie tej klasy gramatycznej – przyimki są przecież dość trwałym elementem języka. Jeśli jakieś tendencje się ujawnią, to z pewnością ich charakter będzie bardzo subtelny. Dlatego znowu

sięgniemy po repertuar metod wielowymiarowych, które sprawdzają się znakomicie w wykrywaniu małych różnic między predyktorami.

4.4.1. Metoda

By uzyskać odpowiedź na pytanie o wpływ przyimków na zmianę w polszczyźnie, posłużymy się nienadzorowaną metodą eksploracji danych zwaną hierarchiczną analizą skupień (ang. *hierarchical cluster analysis*). Polega ona, w największym skrócie, na obliczaniu podobieństw między tekstami – tak jak to robiły wszystkie inne metody przywoływane w tej książce – i następnie na rekurencyjnym znajdowaniu najpierw najbardziej do siebie podobnych pojedynczych tekstów, następnie najbardziej podobnych par tekstów, później zaś coraz większych ich grup, aż do całkowitego połączenia wszystkich dostępnych elementów (zob. np. Baayen, 2009: 118–160). Końcowym efektem analizy skupień jest wykres o kształcie przypominającym drzewo, zwany dendrogramem. W naszym wypadku podstawowym obiektem analizy będą oczywiście nie tyle pojedyncze teksty, ile chronologicznie uporządkowane podkorpusy (por. Rys. 4.19).

W niniejszej analizie posłużymy się wariantem analizy skupień zaproponowanym przez Griesa i Hilperta (2008), zaprojektowanym właśnie do badania danych diachronicznych. Metoda ta opiera się na klasycznej analizie skupień, wprowadza jednak pewną istotną modyfikację. Wynika ona z tego, że oryginalna metoda bierze pod uwagę jedynie podobieństwo pomiędzy tekstami i jest ślepa na inne zmienne, gdy tymczasem dla językoznawcy diachronisty chronologia jest zmienną nie do pominięcia. Jeżeli przeprowadzimy analizę podobieństwa (a do tego przecież sprowadza się analiza skupień) szeregu chronologicznie uporządkowanych podkorpusów, to może się okazać, że najbardziej podobne do siebie są podkorpusy oddalone o setki lat. Co prawda proces diachroniczny jako taki jest zawsze jednokierunkowy i nieodwracalny, ale w danych językowych ujawniają się też inne sygnały, często silniejsze od chronologicznego (np. wpływ stylu funkcjonalnego). Stąd wynika pomysł, by nie porównywać ze sobą wszystkich podkorpusów, a jedynie te sąsiadujące ze sobą chronologicznie. W ten sposób, wolno wierzyć, będziemy w stanie rozpoznać momenty najgwałtowniejszych zmian w ewolucji badanego zjawiska językowego. Gries i Hilpert (2008) proponują zastosować następujący algorytm, by wymusić na analizie skupień łączenie tylko bliskich chronologicznie podkorpusów:

1) oblicz dystans pomiędzy kolejnymi sąsiadującymi ze sobą podkorpusami,
2) anajdź najmniejszy z tych dystansów,
3) połącz dwa sąsiadujące korpusy o najmniejszym dystansie w jeden,
4) ponownie odszukaj najmniejsze różnice pomiędzy sąsiadującymi podkorpusami (przy czym już połączone podkorpusy traktuj jako jeden),
5) powtarzaj, aż połączysz wszystkie podkorpusy.

Algorytm różni się od klasycznej analizy skupień tym, że pozwala na szukanie podobieństw wyłącznie między sąsiadującymi elementami (por. pkt 1 oraz 4). Poza tą jedną cechą dziedziczy po oryginalnej metodzie wszystkie inne jej zalety i niedostatki, m.in. silną tendencję do hierarchizowania podobieństw – podkorpusy są łączone w coraz większe, a zarazem coraz bardziej zróżnicowane grupy, a na szczycie hierarchii mamy zawsze do czynienia z podziałem binarnym. Z zasady więc wykres wykaże pewien podział, nawet w dość homogenicznym materiale. Chcielibyśmy to mocno podkreślić, ponieważ z samego faktu takiego binarnego podziału nie wynika, czy mamy do czynienia z dużą czy z małą różnicą. Jedyne, co można powiedzieć, to że elementy należące do dwu różnych skupień są bardziej od siebie odmienne niż te, które należą do tego samego skupienia.

4.4.2. Wyniki

Napisaliśmy powyżej, że przyimki są dość trwałym elementem języka. Zanim zatem zaczniemy właściwą analizę, spróbujmy przyjrzeć się pokrótce surowym frekwencjom badanych przez nas przyimków, by wstępnie ocenić, do jakiego stopnia w ogóle można mówić o zmienności w tej kategorii gramatycznej. Na razie nie będzie chodziło o ewolucję przyimków w czasie, lecz o zwykłe oszacowanie stopnia zróżnicowania w ich frekwencjach bez względu na chronologię.

Podstawową i szeroko stosowaną miarą dyspersji zmiennej jest odchylenie standardowe, ale jego zastosowanie do danych językowych może być problematyczne ze względu na duże różnice średnich frekwencji słów. W takiej sytuacji bardziej miarodajne jest stosowanie współczynnika zmienności, czyli odchylenia standardowego podzielonego przez średnią arytmetyczną (jest to rodzaj normalizacji, dzięki któremu możemy bezpośrednio porównywać zmienne liczbowe między sobą). Tab. 4.2 przedstawia współczynnik zmienności frekwencji poszczególnych przyimków po normalizacji na 1000 słów w tekście, gdy przyjmiemy jako wielkość podkorpusu 50 lat. Przyimki zostały uszeregowane od najbardziej do najmniej zmiennego między poszczególnymi podkorpusami.

Największą zmienność wykazuje *bez*[17], podczas gdy *za* i *w* są najbardziej stabilne. Obserwowany przez nas procent zmienności jest wprawdzie niewielki, ale mimo to sugeruje, że frekwencja badanych przyimków nie jest zupełnie stała na przestrzeni wieków XV–XIX. Oczywiście tak prosta miara nie powie, czy zaobserwowane fluktuacje układają się w jednostajny trend chronologiczny, wynik daje jednak nadzieję, że badane przyimki sprawdzą się jako predyktory w analizie skupień.

Z powodów, o których pisaliśmy wielokrotnie, nie będziemy analizować poszczególnych tekstów z osobna, lecz podzielimy korpus na podokresy obejmujące równą liczbę lat i przyporządkujemy teksty do wydzielonych w ten sposób pod-

[17] Trzeba w tym miejscu przypomnieć o synonimii *bez* i *przez* w staropolszczyźnie.

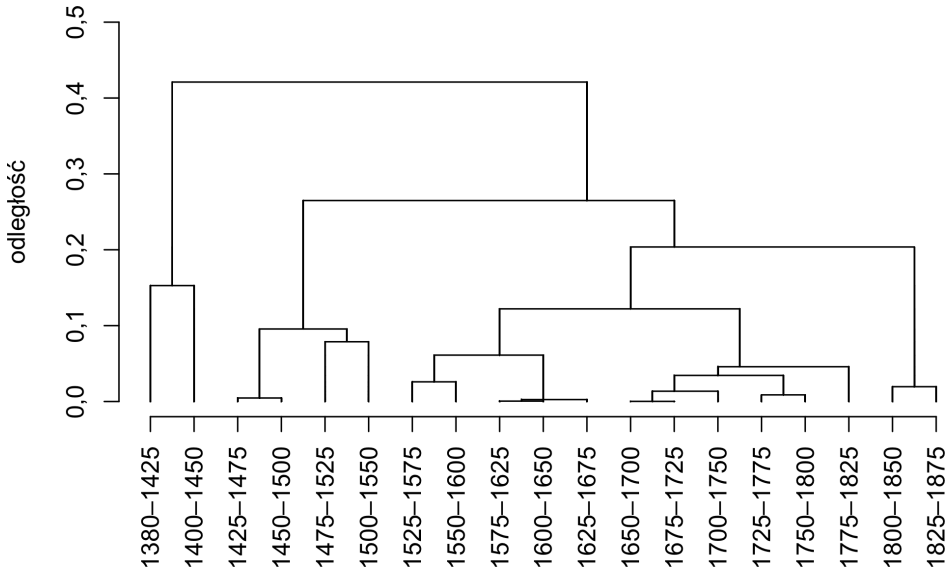
korpusów. W pierwszej próbie dokonamy podziału korpusu na okresy 50-letnie zachodzące na siebie 25-letnią „zakładką", by otrzymać zakresy 1380–1425[18], 1400–1450, 1425–1475, 1450–1500, . . . , 1825–1875. Końcowy dendrogram dla 18 cech dyskryminacyjnych (frekwencji przyimków) badanych w 50-letnich podkorpusach pokazany został na Rys. 4.19.

Tabela 4.2. Zmienność frekwencji 18 przyimków w podkorpusach obejmujacych 50 lat.

przyimek	liczba wystąpień	odchylenie standardowe	współczynnik zmienności
bez	15169	17,46	2,54
przed	17037	18,18	2,39
dla	30076	35,69	2,21
przez	31732	36,43	2,08
od	19620	18,38	2,05
między	13637	15,13	1,99
ku	21822	17,58	1,92
o	66921	37,21	1,74
na	203321	67,32	1,71
przy	19758	4,79	1,55
u	20958	7,13	1,33
nad	20220	6,79	1,31
do	127087	16,39	1,25
z	253736	26,95	1,15
po	46610	4,67	1,13
pod	21166	2,13	1,07
za	51975	4,21	0,94
w	288511	25,53	0,88

Interpretację wykresu zacznijmy od kilku słów wyjaśnienia. To, co w statystyce nazywamy odległością (a więc miara odmienności dwóch lub więcej obiektów, którą można zdefiniować jako odwrotność miary podobieństwa), jest symbolizowane na wykresie przez to, jak wysoko leży na nim horyzontalna kreska – im wyżej jest ona położona, tym różnica między obiektami (podkorpusami) jest większa. Lekturę wykresu na ogół rozpoczyna się od rozpoznania największych „gałęzi" drzewa, które z kolei zawierają zgrupowane „liście", a więc podkorpusy. Na Rys. 4.19 daje się zaobserwować pięć głównych skupień w postaci największych gałęzi

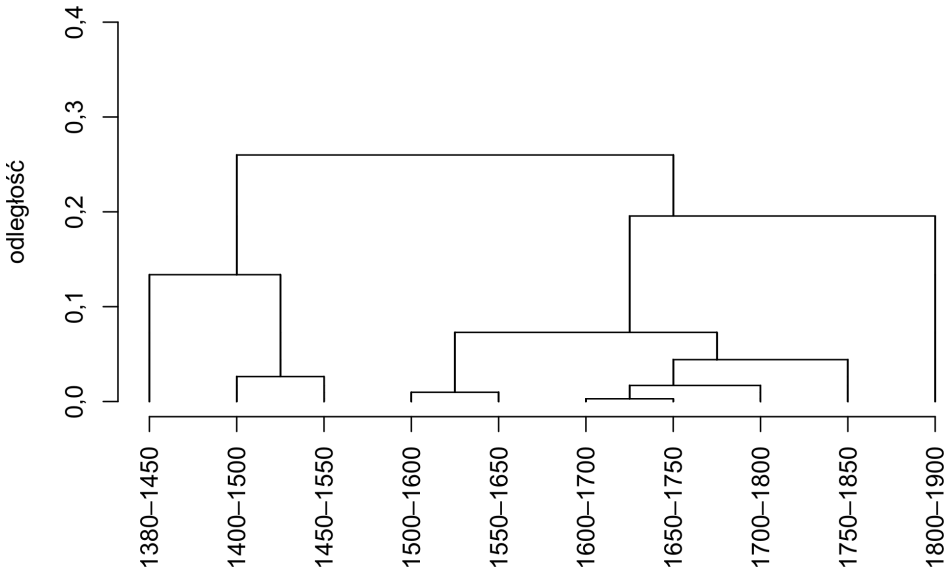
[18] W pierwszym podkorpusie okres 50 lat traktujemy dość umownie ze względu na bardzo małą liczbę dostępnych tekstów.



Rysunek 4.19. Analiza skupień dla 18 przyimków, przy podziale materiału na 50-letnie podkorpusy.

dendrogramu, przy czym największe pęknięcie następuje w połowie XV wieku, a drugie najmocniejsze w połowie XVI wieku. Rzecz jasna liczyliśmy na to, że główny podział nastąpi w okolicach roku 1543, który powszechnie uważa się za granicę między epoką staropolską i średniopolską. Tymczasem sama epoka staropolska, jak się okazuje, jest zróżnicowana wewnętrznie (mowa wyłącznie o przyimkach!) w większym stopniu niż wszystkie inne podokresy razem wzięte. Wynik ten nie zaskoczył nas zresztą aż tak bardzo, podobne zjawisko zaobserwowano bowiem już wcześniej, gdy okazało się, że tager wytrenowany na współczesnej polszczyźnie radzi sobie dość nieźle z tekstami XVI-wiecznymi i późniejszymi, ale wykazuje znaczący spadek skuteczności właśnie w XV wieku (Eder i in., 2015).

Przyglądnijmy się jeszcze następnemu fragmentowi wykresu, a mianowicie podkorpusom obejmującym lata 1525–1825 (Rys. 4.19). Okres ten jest stosunkowo homogeniczny, niemniej znów dzieli się on na podokresy 1525–1675 i 1650–1825. Niewątpliwie różnica pomiędzy tymi epokami jest mała, przynajmniej w porównaniu z głównym pęknięciem omówionym powyżej. Większy podział następuje mniej więcej pod koniec wieku XVII, ale znacznie mocniej zarysowuje się podział na XVII–XVIII wiek z jednej strony i XIX wiek z drugiej. Znów nasuwa się skojarzenie z proponowaną przez badaczy chronologizacją epoki średniopolskiej i nowopolskiej, których granicę ma wyznaczać właśnie przełom XVIII i XIX wieku. Jeśli ograniczyć się do wyłącznie frekwencji przyim-



Rysunek 4.20. Analiza skupień dla 18 przyimków, przy podziale materiału na 100-letnie podkorpusy.

ków, podejście algorytmiczne zdaje się potwierdzać intuicje formułowane przez historyków języka.

Zanim jednak pójdziemy w naszych dywagacjach zbyt daleko, należy poczynić istotne zastrzeżenie. Podejście, które tu opisaliśmy, jest obciążone pewną wadą, mianowicie trudno znaleźć sposób jego weryfikacji. Przypomnijmy, że jeśli chcemy ocenić, na ile wiarygodne są wyniki klasyfikacji nienadzorowanej, należy sprawdzić, w jakim stopniu teksty (czy podkorpusy) o znanej przynależności do klasy rzeczywiście zostały przez algorytm właściwie przyporządkowane. W naszym wypadku nie dysponujemy tego rodzaju wiedzą: nie wiemy przecież, gdzie powinny wypaść pozostałe największe przełomy w korpusie. Do pewnego stopnia można jednak skontrolować wiarygodność wyników, powtarzając eksperyment dla podkorpusów różnej wielkości i o różnych datach granicznych. Jest faktem ogólnie znanym, że analiza skupień bywa niestabilna, jeśli w danych wejściowych nie ma silnego sygnału. To niestabilne zachowanie metody możemy wykorzystać do naszych celów: zaobserwowane znaczące różnice między dendrogramami będą dla nas pośrednim dowodem na to, że data przełomu w okolicach roku 1450 była jedynie artefaktem, wynikiem fałszywie pozytywnym.

Przykładowy wykres dla podkorpusów 100-letnich został pokazany na Rys. 4.20. Widzimy, że tym razem największy podział przypada na połowę wieku XVI, a więc w momencie uznanym przez badaczy za granicę epok staropolskiej i średniopolskiej, podczas gdy drugie główne pęknięcie następuje na przeło-

mie XVIII i XIX wieku, na początku epoki nowopolskiej. Można by rzec, że ten konkretny wykres jest modelowym przykładem potwierdzającym przyjętą chronologizację dziejów polszczyzny, tyle tylko, że inne testowane przez nas wykresy niekoniecznie dawały takie same wyniki. Jeśli chcemy, by nasz eksperyment miał jakąkolwiek wartość poznawczą, trzeba do badanego problemu podejść w sposób bardziej systematyczny.

Przede wszystkim należy zauważyć, że największy wpływ na fluktuacje wyników miała dobrana przez nas arbitralnie wielkość podkorpusów i ewentualnie wielkość „zakładki", czyli zakres chronologiczny, o jaki przesuwaliśmy poszczególne podkorpusy. Zacznijmy od trywialnej w gruncie rzeczy obserwacji, że nie istnieją żadne przesłanki teoretyczne, które kazałyby ustalać granice korpusów w taki sposób, jak to uczyniliśmy powyżej; przecież początek XVII wieku można równie dobrze opisać podkorpusem 1600–1621 jak i, dajmy na to, 1597–1629 czy 1612–1642[19]. Podobnie nie istnieje powód, dla którego podkorpusy miałyby obejmować 50 lat, oczywiście poza jednym, takim mianowicie, by każdy podkorpus obejmował dostatecznie dużą liczbę przykładów. Szczupłość materiału językowego wymusiła na nas rozważenie jeszcze jednego parametru. Ponieważ pierwszy podkorpus jest dość niewielki i obarczony luką we wczesnym XV wieku, zmienialiśmy jego rozmiary dość dowolnie, dzięki czemu początki kolejnych korpusów wypadały w innych latach. Również dzięki temu zabiegowi zmienialiśmy warunki eksperymentu. Systematyczne porównanie kilku wybranych wyników dla 25-, 50- i 100-letniej wielkości korpusu i różnej wielkości „zakładki" przedstawiamy w Tab. 4.3.

Tabela 4.3. Moment podziału na dwa największe skupienia dla 18 przyimków przy różnej wielkości porkorpusu.

rozmiar podkorpusu	wielkość „zakładki"	moment największej zmiany
100	50	1425
100	0	1425
50	25	1440
50	0	1450
25	10	1450
25	0	1450

[19] Każda z tych dat jest związana z rzeczywistym wydarzeniem historycznym, najczęściej z rokiem urodzenia lub śmierci konkretnych postaci historycznych. Nie będziemy odbierać Czytelnikowi przyjemności ich odszyfrowania.

Dane z Tab. 4.3 pokazują, że najbardziej odmienny jest jednak wiek XIV–XV i schyłek wieku XV. Pamiętajmy wszakże, że staropolszczyzna jest bardzo nierówno pokryta przez teksty i że pierwszy podkorpus jest zdominowany przez *Psałterz floriański* (stanowi on około 2/3 tego podkorpusu), a dwa pozostałe teksty to *Kazania świętokrzyskie* i *Kazania gnieźnieńskie*. Te trzy utwory są wyraźnie odrębne od wszystkich późniejszych tekstów, również piętnastowiecznych. Z drugiej strony nawet w tych przypadkach druga połowa XV wieku i pierwsza połowa wieku XVI z reguły są bardzo odmienne od późniejszych podkoropusów; stanowią drugie najmocniejsze pęknięcie w chronologii polszczyzny. Ta natomiast data pokrywa się z powszechnie przyjętym w polskim językoznawstwie momentem przejścia staropolszczyzny w średniopolszczyznę. Jest to wynik warty odnotowania, ponieważ – przypomnijmy – analiza została przeprowadzona na podstawie frekwencji wyłącznie 18 przyimków, co w sposób oczywisty oznacza, że naszym oczom ukazała się zmiana w warstwie gramatycznej języka.

4.4.3. Podsumowanie

Jakkolwiek interesujące by były otrzymane przez nas wyniki, niestety pokazują one również, że metoda hierarchicznej analizy skupień jest podatna na zafałszowanie wyników zwane „zrywaniem czereśni”[20], tzn. wyborem takiej daty, która pasuje do przyjętych wcześniej założeń. Praktycznie każdy rok pomiędzy 1450 a 1550 może być określony jako „punkt zwrotny”, „moment przejścia języka staropolskiego w średniopolski”, wystarczy tak długo zmieniać parametry eksperymentu, aż się uzyska założony wynik. Zafałszowanie tego typu wcale zresztą nie musi oznaczać oszustwa naukowego: badacz najczęściej nieświadomie skłania się ku dostrzeganiu przede wszystkim tych wyników, które potwierdzają sformułowane na początku eksperymentu oczekiwania.

Inny problem to wielkość podkorpusów. Jak wcześniej napisaliśmy, podkorpusy obejmujące niewielką liczbę lat są co do zasady bardziej homogeniczne niż duże podkorpusy, nie stanowią zatem Prokrustowego łoża dla analizy skupień. Z drugiej strony, jak się okazuje, niewielkie podkorpusy skutkują mniej stabilnymi wynikami, ponieważ tekstów staropolskich jest relatywnie niewiele. Oryginalna praca Griesa i Hilperta (2008) przewidywała podkorpusy o rozmiarze 70 lat, podczas gdy Rosemeyer (2015) zdecydował się na 50 lat. Historyk języka rzadko jest w tak luksusowej sytuacji jak autorzy studium diachronicznego wykorzystującego korpus *Time Magazine* o takiej obfitości tekstu, że każdy rok mogli uczynić odrębnym podkorpusem (Kestemont, Karsdorp i Düring, 2014).

Oczywiście wybór odpowiedniego rozmiaru podkorpusu również obarczony jest ryzykiem „zrywania czereśni”, staraliśmy się jednak w wyborze kierować

[20] Ang. *cherry picking*, wyrażenie to wzięło się stąd, że osoba zrywająca owoce zazwyczaj wybiera te ładniejsze, z kolei ktoś, kto ogląda tylko owoce zerwane, a nie te, które rosną na drzewie, może dojść do fałszywego przekonania, że wszystkie one są dojrzałe i czerwone.

zasadą kompromisu pomiędzy ziarnistością podziału a zrównoważeniem poszczególnych podkorpusów, a przede wszystkim przetestować stabilność analizy skupień dla różnych wartości tego parametru. Dodatkowym problemem okazał się pierwszy podkorpus, zawierający najstarsze teksty. I nawet nie tyle jego wielkość miała wpływ na wyniki, ile fakt, że pokrycie tekstami początku dziejów polszczyzny jest bardzo skąpe. W dodatku jedyne dwa teksty z XIV wieku to dwa zbiory kazań, które Pisarkowa (1984) charakteryzuje jako przykłady bardzo prymitywnej składni, dokumentującej przejście między językiem, który zna tylko rejestr mówiony, i tworzącą się polszczyzną pisaną. Ma to oczywisty wpływ na dystrybucję przyimków, a zarazem na wynik całego naszego eksperymentu.

Nie można wreszcie zapominać, że dysponujemy dosyć przypadkowym korpusem – ta przypadkowość jest absolutnie nieunikniona w odniesieniu do tekstów sprzed połowy XVI wieku, których zasób jest niewielki. Być może nasze wyniki w jakimś stopniu odzwierciedlają nierównowagę poszczególnych podkorpusów, a także odmienną ich budowę – wszak niektóre podkorpusy mają znaczący komponent w postaci poezji, gdzie przyimki odgrywają istotną rolę w metryce. Nie jesteśmy w stanie tych czynników wykluczyć.

A jednak mimo tych zastrzeżeń wyniki są w gruncie rzeczy zgodne z dotychczasową periodyzacją stosowaną w językoznawstwie polonistycznym, co wydaje nam się warte odnotowania. Przede wszystkim chcielibyśmy podkreślić fakt, że choć nasze wyniki mogą się wydać przewidywalne z punktu widzenia historyka języka, są one oparte na zupełnie innych kryteriach, a mianowicie jedynie na obserwacji przemian frekwencji przyimków.

Rozdział 5

Studium przypadku: zmiany frekwencji imiesłowu uprzedniego

5.1. Wprowadzenie

W niniejszym rozdziale zajmiemy się zmianą we frekwencji jednej wybranej (dodajmy: dość marginalnej) formy fleksyjnej, a mianowicie imiesłowu uprzedniego. Forma ta, w XV wieku niezbyt częsta i używana głównie z czasownikami ruchu, stawała się coraz bardziej popularna – szczyt popularności osiągając w wieku XVII – by następnie odnotować stopniowy spadek. W kontekście tego faktu rodzi się szereg pytań, w poniższych rozważaniach będziemy szukali odpowiedzi na jedno z nich: czy wzrost frekwencji imiesłowów uprzednich jest spowodowany poluzowaniem warunków, jakie muszą spełniać czasowniki, by użytkownik języka był skłonny użyć ich w postaci imiesłowu uprzedniego, czy po prostu pewną modą, by chętniej sięgać po tę formę fleksyjną przy ograniczonej liczbie leksemów, od których je tworzą? Powyższy problem badawczy można też sformułować nieco prościej: czy dla użytkowników języka danej epoki istnieje ograniczona pula imiesłowów, znacząco mniejsza niż ogólna pula czasowników, czy też od (niemal) dowolnego czasownika tworzone były imiesłowy?

Na podstawie danych pochodzących z NKJP można zasadnie dowodzić, że imiesłów uprzedni jest dość silnie ograniczony leksykalnie we współczesnej polszczyźnie (Górski i Król, 2018). Generalnie rzecz biorąc, czasowniki, które preferują tę formę fleksyjną, to czasowniki czynnościowe, teliczne, rozciągnięte w czasie. Trzeba tu bardzo wyraźnie podkreślić, że nie są to ograniczenia „sztywne", typu regułowego, ale raczej silne tendencje. Oznacza to, że NKJP notuje wprawdzie imiesłowy uprzednie czasowników nietelicznych momentalnych, ale ich wystąpienia są sporadyczne, nawet wśród najczęstszych czasowników. Innymi słowy, kompetencja rodzimego użytkownika języka nie zawiera filtru, który zabraniałby wygenerować zdania z imiesłowem uprzednim czasownika nieteliczego, choć użytkownik będzie dużo bardziej skłonny użyć w tej formie czasownik teliczny niż nieteliczny. Oczywiście tego rodzaju ograniczenia muszą wpływać na ogólną frekwencję imiesłowu uprzedniego, skoro duża grupa czasowników (nieteliczne) tej formy unika.

Rodzi więc się pytanie, czy wzrost frekwencji imiesłowu uprzedniego wiązał się z poluzowaniem ograniczeń semantycznych czy też wynikał po prostu z faktu, że z podobnego (pod względem ilościowym) zestawu czasowników użytkownicy języka czynili częstszy użytek. Jakkolwiek kuszące byłoby powtórzenie na materiale historycznym wyżej cytowanych badań Górskiego i Król (2018), to jest to niemożliwe, ponieważ korpus historyczny jest znacząco mniejszy i poświadczenia dla większości leksemów są nieliczne. W omawianej pracy wnioski są wyciągane na podstawie kilkuset typów, zaś frekwencja typów najczęstszych sięga tysiąca. Trzeba więc poszukać sposobu, który pośrednio da nam odpowiedź na pytanie o to, czy ograniczenia semantyczne się zmieniały. Tą pośrednią odpowiedzią jest produktywność.

Produktywność jest rozumiana bardzo rozmaicie[1]. Pojęcie to najczęściej jest łączone ze słowotwórstwem, choć oczywiście ma zastosowanie w całej gramatyce, a w szczególności w morfologii. Dla naszej argumentacji nie ma znaczenia, czy imiesłów uznamy za produkt derywacji czy fleksji[2], i nie zamierzamy przesądzać tej kwestii, jakkolwiek w dalszych rozważaniach odwołamy się do metod stosowanych w badaniu słowotwórstwa. Nie ma jednak przeszkód, by metody te przenieść na fleksję, a nawet składnię (Zeldes, 2012). Tak jak w sytuacji, gdy użytkownik języka, chcąc nazwać nową rzecz czy czynność, może albo posłużyć się znanymi sobie wyrazami, albo utworzyć neologizm, tak w badanym przez nas wypadku użytkownik może albo wyrazić predykację za pomocą konstrukcji z czasownikiem w formie osobowej, albo utworzyć od tego czasownika imiesłów.

W badaniu produktywności punktem wyjścia może być system. Takie podejście widzi produktywność jedynie w kategoriach jakościowych. Wtedy za formę bardziej produktywną uznamy tę, na którą system nakłada mniej ograniczeń natury semantycznej czy fonologicznej. Jeśli taki brak ograniczeń powoduje, że forma produktywna jest częstsza (przede wszystkim pod względem typów), uznaje się to za epifenomen, ale nie istotę produktywności.

Częściej zakłada się, że miarą produktywności jest liczba typów utworzonych z danym formantem. Z dwu formantów bardziej produktywny jest wtedy ten, który służy do stworzenia większej liczby wyrazów. W takim podejściu produktywność staje się cechą ilościową. (Takie rozumienie produktywności jest najbardziej ugruntowane w polskiej tradycji językoznawczej.)

Z kolei dla Baayena (1993) produktywność to pewna dyspozycja mentalna rodzimego użytkownika języka. Jeżeli dla utworzenia nowego słowa użytkownik ma do wyboru dwa różne formanty, to za bardziej produktywny uznamy ten, po który sięgnie on w pierwszej kolejności. Najpewniejszym sposobem oceny tak rozumianej produktywności jest test psycholingwistyczny. Przed uczestnikami stawia się na przykład zadanie stworzenia szeregu nowych derywatów, jednak

[1] Przegląd stanowisk w zakresie produktywności przytacza Bauer (2001).

[2] Dyskusję na ten temat referuje Rabiega-Wiśniewska (2008).

w taki sposób, by mieli oni wybór formantu, np. dla języka polskiego byłyby to formanty *-owy* oraz *-ny*. Innym rodzajem testu jest badanie przejrzystości słowotwórczej. Wtedy celem jest sprawdzenie, czy nowe słowo utworzone z danym formantem jest od razu zrozumiałe dla użytkownika języka czy też użytkownik potrzebuje dodatkowego czasu (na ogół ułamków sekundy), żeby zanalizować nowe słowo. Formanty produktywne będą słowotwórczo przejrzyste, formanty nieproduktywne – nie.

Zakładamy, że istnieje pewien związek pomiędzy produktywnością w rozumieniu Dresslera (1997), a więc produktywnością rozumianą jako brak ograniczeń, a produktywnością w rozumieniu Baayena, czyli skłonnością do sięgania po pewien formant, gdy trzeba utworzyć nowe słowo. Możliwa jest sytuacja, gdy użytkownik języka ma do dyspozycji dwa synonimiczne (przynajmniej w takim zakresie, w jakim jest to konieczne dla utworzenia neologizmu) formanty, ale jeden z nich jest archaiczny – obecny w wielu słowach, ale raczej nieużywany do tworzenia nowych słów. Najczęściej jednak niższa produktywność jest skutkiem ograniczeń natury semantycznej, fonotaktycznej albo jednego i drugiego. Jeśli tak, to niska produktywność rozumiana w kategoriach ilościowych świadczy o niższej produktywności w kategoriach systemowych. Idąc dalej w tym rozumowaniu, możemy zakładać, że wzrost produktywności (w kategoriach ilościowych) imiesłowów uprzednich oznacza wzrost produktywności w kategoriach systemowych.

W badaniach historycznych eksperymenty psycholingwistyczne są z oczywistych powodów niemożliwe do przeprowadzenia. Baayen twierdzi jednak, że dane korpusowe mogą być źródłem o wiarygodności podobnej do wiarygodności danych psycholingwistycznych (Baayen, 1993). Jego rozumowanie przebiega następująco: jeżeli dany formant jest produktywny, oznacza to, że jest często używany do tworzenia nowych słów. Jednocześnie wyniki eksperymentów psycholingwistycznych sugerują, że wyrazy częste są przechowywane w umyśle jako pewna całostka, natomiast rzadkie są tworzone na bieżąco, tzn. użytkownik języka łączy w trakcie wypowiedzi podstawę słowotwórczą z formantem. Można więc powiedzieć, że słowa rzadkie z punktu widzenia produkcji mowy są niejako wciąż na nowo tworzonymi neologizmami. Duży udział słownictwa rzadkiego wśród wyrazów tworzonych z danym formantem powoduje tym samym, że istnienie tego formantu jest nieustannie wzmacniane w świadomości użytkownika języka przez ponawiane operacje słowotwórcze. Według Baayena formanty, które są obecne w znacznej liczbie słów rzadkich, to także te, które eksperymenty psycholingwistyczne wskazują jako silnie produktywne.

Idąc tym tropem, możemy przyjąć, że dobrym wskaźnikiem produktywności będzie obecność słownictwa rzadkiego wśród słów derywowanych z danym formantem. Na udział tego słownictwa wskazuje z kolei miara stosunku typów do okazów (ang. *type–token ratio*, w skrócie TTR). Najczęściej jest ona stosowana do określania bogactwa leksykalnego danego tekstu, autora czy korpusu. Uzyskujemy ją, dzieląc liczbę wyrazów w tekście (korpusie) przez liczbę typów. Miara

ta przyjmuje wartość pomiędzy 0 a 1, z tym, że obie skrajne wartości (czyli 0 i 1) są niemożliwe w wypadku realnego tekstu, gdyż pierwsza oznaczałaby nieskończenie długi tekst składający się z jednego i tego samego słowa powtórzonego wielokrotnie, a druga – tekst, w którym żadne słowo się nie powtarza. W istocie rzeczy TTR to znormalizowana uśredniona frekwencja słów[3], zauważmy jednak, że tak jak liczymy TTR dla jakiegoś tekstu, tj. bierzemy pod uwagę wszystkie występujące w nim słowa, możemy tak samo obliczyć tę miarę dla pewnej klasy słów, np. dla wszystkich rzeczowników w tekście, czy – jak w interesującym nas przypadku – dla słów z danym formantem. Wysokie TTR, a więc duża liczba typów w stosunku do liczby okazów, sygnalizuje znaczny udział słownictwa rzadkiego wśród wyrazów derywowanych z tym formantem.

Baayen idzie dalej w swojej argumentacji. Jego zdaniem najważniejszym wskaźnikiem produktywności są *hapax legomena* (wyrazy z jednym poświadczeniem w korpusie). *Hapax legomenon* jest przecież z definicji neologizmem – wyraz, który pojawia się raz, siłą rzeczy musi być nowo utworzony.

Naturalnie za tym podejściem kryje się założenie, że korpus to język, a to z kolei oznacza co najmniej dwa problemy metodologiczne. Po pierwsze: słowo, które jest odnotowane w korpusie tylko raz, w rzeczywistości pozakorpusowej może być wcale częste. W końcu to, czy dany tekst trafi do korpusu czy nie, jest w dużej mierze dziełem przypadku – z pewnością wiele słów poświadczonych jeden raz w korpusie straciłoby miano *hapax legomenon*, gdyby inny zestaw tekstów wszedł do korpusu. Tym niemniej słowa naprawdę rzadkie, lecz z przejrzystą budową, są zapewne czymś, co możemy roboczo nazwać „wielokrotnymi neologizmami", tj. wyrazami, które wielu użytkowników języka utworzyło niezależnie od siebie. W wypadku wyrazów naprawdę rzadkich niewielka jest szansa, że przypadkowy członek społeczności językowej wcześniej takie słowo usłyszy, zapamięta, a następnie przywoła z pamięci, kiedy nadarzy się odpowiedni kontekst. Nawet jeśli to ostatnie stwierdzenie nie ma potwierdzenia empirycznego, wydaje się przekonujące, bo przecież większość wypowiedzi dociera do dość ograniczonego kręgu odbiorców. Na szeroki krąg odbiorców i realny wpływ na użycie może liczyć tylko niewielka liczba tekstów.

Po drugie: wiele rzeczywistych neologizmów (tj. wyrazów, które naprawdę po raz pierwszy zostały użyte w jakimś tekście z korpusu) będzie wykazywać stosunkowo wysoką frekwencję. Dzieje się tak dlatego, że pewna liczba nowych wyrazów jest podchwytywana i używana przez innych członków społeczności językowej.

Mimo powyższych dwóch wątpliwości Baayen konkluduje, że słownictwo rzadkie, a w szczególności *hapax legomena*, w praktyce sprawdza się jako dobry wskaźnik produktywności. Jednym z argumentów, by produktywność oceniać na

[3] W wypadku języków o bogatej morfologii oczywiste staje się pytanie, co jest *typem* – słowoforma czy lemat. W zależności od przyjętej definicji wyniki mogą się znacznie różnić.

podstawie hapaksów (ang. *hapax-conditioned productivity*), jest to, że wysoki współczynnik TTR może towarzyszyć formacjom nieproduktywnym i przede wszystkim nieprzejrzystym słowotwórczo, które są rzadkie z tego powodu, że są archaizmami. Zarazem jednak warto pamiętać i o tym, że kilka bardzo częstych wyrazów może znacząco obniżyć TTR bardzo produktywnego formantu.

Biorąc pod uwagę powyższe ograniczenie współczynnika TRR, Baayen proponuje użyć innej prostej miary, dzięki której możliwe jest oszacowanie udziału hapaksów w korpusie (Baayen, 1993). Miarą tą jest prawdopodobieństwo, że kolejny napotkany w tekście wyraz to właśnie *hapax legomenon*. Co prawda nie widzimy przez to rzeczywistego przyrostu nowych wcześniej w korpusie nienotowanych typów, natomiast otrzymujemy pewną konkretną liczbę, którą można porównać z inną. Prawdopodobieństwo to liczymy, dzieląc liczbę hapaksów przez liczbę wszystkich wystąpień wyrazów z danym formantem:

$$P = \frac{n_1}{N}$$

gdzie n_1 to liczba hapaksów, zaś N całkowita liczba wyrazów z danym formantem.

W zarysowanym powyżej podejściu badawczym umknąć może jedna istotna sprawa, a mianowicie fakt, że powyższych obserwacji chcemy dokonywać w ujęciu diachronicznym. Nie porównujemy więc produktywności dwu (lub więcej) formantów między sobą, ale produktywność tej samej formy pomiędzy epokami. Będziemy zatem chcieli wyliczyć TTR oraz miarę P imiesłowów uprzednich dla każdego z sekwencji chronologicznie uporządkowanych podkorpusów.

W związku z powyższym nieco kłopotliwa staje się sprawa rozumienia, czym tak naprawdę jest *hapax legomenon* w korpusie diachronicznym. Skoro każdy z chronologicznie uporządkowanych podkorpusów uznajemy za osobną całostkę, to zasadniczo nie powinno interesować nas, czy dane słowo było odnotowane w którymś z podkorpusów reprezentujących wcześniejsze bądź późniejsze lata. Takie postępowanie może budzić pewne wątpliwości, bo przecież jeśli wyraz jest już wcześniej odnotowany, to z definicji nie może być wyrazem nowo utworzonym. W swojej argumentacji Baayen dodatkowo zwraca uwagę na to, że korpus nie daje świadectwa negatywnego – to, że coś nie jest w nim odnotowane, nie dowodzi, że nie istnieje w języku[4]. Tym niemniej nasze podejście można uzasadnić następująco: tym, co nas naprawdę interesuje, nie jest słownik[5], ale kompetencja użytkownika języka. Jeśli więc dany wyraz został odnotowany choćby kilkukrotnie w poprzednich stuleciach, to jest rzeczą mało prawdopodobną, żeby użytkownik języka wcześniej przeczytał ten wyraz we wcześniejszych tekstach. Raczej można domniemywać, że „dotworzył" go na bieżąco. Przy czym o ile w po-

[4] Naturalnie im większy korpus, tym bardziej przekonujący jest brak jakiegoś faktu. Mimo wszystko musimy pamiętać, że korpus, na którym są oparte niniejsze badania, jest korpusem bardzo niewielkim.

[5] Rozumiany jako część *langue*, a nie publikacja.

jedynczym wypadku jest to domniemanie dość spekulatywne, to jest ono wysoce prawdopodobne w wypadku hapaksów jako całej klasy słownictwa rzadkiego.

Gdybyśmy zatem chcieli porównać produktywność dwu formantów w ujęciu diachronicznym, powiedzmy werbalnych prefiksów *wy-* i *roz-*, staralibyśmy się odnaleźć wszystkie czasowniki z tymi prefiksami w poszczególnych podkorpusach i wyrysować krzywą, która ukaże relację typów do okazów w funkcji czasu, oraz drugą krzywą dla miary *P* Baayena. Zapewne w wypadku jednego z tych prefiksów odpowiednie krzywe będą się wznosiły szybciej niż to ma miejsce w wypadku drugiego z nich. Zapewne też krzywe nie będą całkowicie gładkie, gdyż pojawią się wcześniej nienotowane słowa charakterystyczne dla jakiegoś tekstu, autora czy stylu funkcjonalnego.

5.2. Metoda

Na pytanie o produktywność imiesłowów uprzednich odpowiada zarówno TTR, jak i miara *P* Baayena i do nich odwołamy się w badaniu produktywności imiesłowów. Obie te miary są jednak obciążone pewną wadą, mianowicie są czułe na wielkość populacji. Jeżeli weźmiemy dowolny tekst i obliczymy TTR dla całości i dla połowy tekstu, to w tym drugim wypadku będzie on wyższy niż w pierwszym, choć przecież mamy do czynienia z tym samym tekstem. Zgodnie z tzw. prawem Heapsa (Baayen, 2012) dynamika przyrostu nowych wyrazów w tekście (albo w korpusie) spada, co zresztą oczywiste, jeśli weźmiemy pod uwagę, że słownictwo będące w użyciu jest bądź co bądź ograniczone. Jeśli tak, to siłą rzeczy to samo zjawisko będzie zachodziło w wypadku formantów obecnych w dużej liczbie okazów. Szansa, że trafimy tu na *hapax legomenon*, jest mniejsza, niż to ma miejsce, gdy okazów jest mniej. Zresztą prawdopodobnie gdybyśmy znacząco powiększyli korpus, liczba hapaksów także by spadła – ten fakt będzie istotny dla naszych dalszych rozważań.

By przezwyciężyć powyższe ograniczenie miary TTR, zaproponowano kilka rozwiązań. Jednym z nich jest losowanie stałej liczby wyrazów i obliczanie na ich podstawie stosunku typów do okazów w próbie tej samej wielkości (Tweedie i Baayen, 1998). Inny sposób to obliczanie TTR w przesuwanym okienku o stałej wielkości, a następnie uśrednianie wyników (Kubát i Milička, 2013). Badania wykazują wszakże, że nawet gdy losujemy stałą liczbę słów z tekstów o znacząco różnym rozmiarze, nie da się całkowicie wyeliminować efektu różnicy wielkości tekstu (Tweedie i Baayen, 1998).

Z kolei w odniesieniu do miary *P* zaproponowano, by w celu wyeliminowania efektu wielkości populacji brać pod uwagę tylko tyle okazów z danym formantem, ile się ich znajduje w populacji najmniej licznej (Gaeta i Ricca, 2006). Jeśli na przykład jeden badany formant tworzyłby 300 typów, drugi – 500, trzeci – 3000, bralibyśmy wszystkie wystąpienia pierwszego z nich, zaś z drugiego i trzeciego

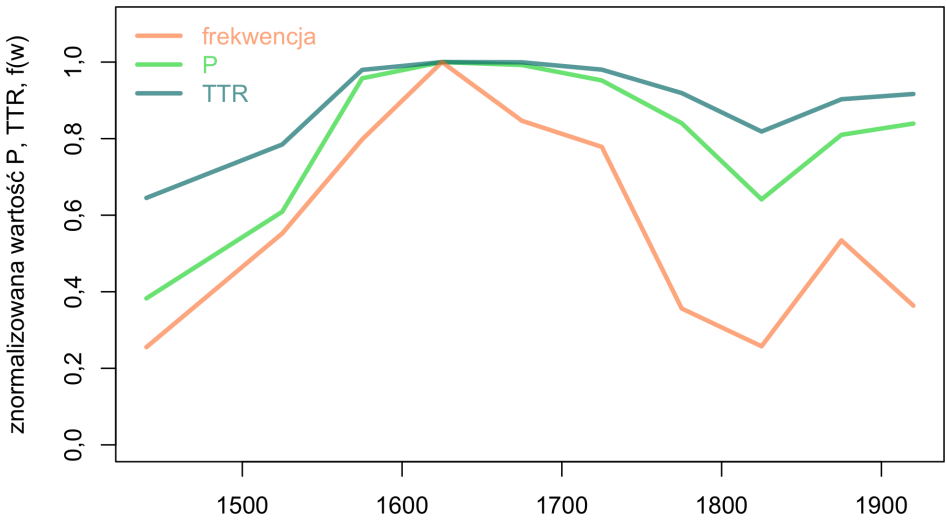
tylko po 300 typów. Dalsze postępowanie jest takie jak w oryginalnym podejściu, czyli liczbę hapaksów, jakie znalazły się wśród tych pierwszych 300 wystąpień, podzielilibyśmy przez 300 (tj. przez wielkość populacji). Zauważmy, że jest rzeczą bardzo prawdopodobną, iż wyrazy, które wystąpiły tylko raz wśród pierwszej trzechsetki, pojawią się wśród pozostałych 200, czy tym bardziej 1700 wystąpień – przypomnijmy, że im bardziej powiększamy tekst, tym wolniej przybywa w nim nowych słów. Nasuwa się tu oczywisty wniosek, że jednostkowość wystąpienia nie jest tu traktowana – by tak rzec – całkowicie serio, tj. nie chodzi o to, by słowo naprawdę tylko raz wystąpiło w całym korpusie (a tym bardziej w całym języku, jeśli tylko dałoby się to stwierdzić), ale by wystąpiło tylko raz w badanej próbce.

Rozwiązanie, które proponują Gaeta i Ricca (2006), obarczone jest jednak inną wadą. W korpusach referencyjnych teksty są zazwyczaj ułożone według pewnego klucza związanego z ich typem, np. na początku idą powieści (dodatkowo uszeregowane autorami), dalej teksty prasowe, naukowe itd. Wiadomo też, że poszczególne formanty są charakterystyczne dla pewnych stylów funkcjonalnych. W praktyce oznacza to, że w wypadku pierwszego z hipotetycznych trzech formantów dane zostaną zebrane z całego korpusu, natomiast dla dwu pozostałych – tylko np. z prozy artystycznej. Problem nie zachodzi wyłącznie wtedy, gdy mamy do czynienia z homogenicznym korpusem, np. zawierającym wydania jednego czasopisma.

By uniknąć takiego przekłamania, zaproponowano, by ową stałą liczbę wyrazów losować z całej populacji (Górski, 2016). Zilustrujmy to znowu tym samym przykładem: dla formantów reprezentowanych przez 300, 500 i 2000 typów za każdym razem losujemy po 300 wyrazów, następnie sprawdzamy wśród nich liczbę słów unikalnych i traktujemy je jako *hapax legomena*. Losowanie możemy powtórzyć wielokrotnie i uśrednić wynik z wielu iteracji. Naturalnie wtedy za każdym razem dostaniemy nieco inną listę tak rozumianych „hapaksów", gdyż w kolejnych losowaniach mogą nie zostać wylosowane wszystkie rzeczywiste hapaksy, a jednocześnie inne słowa zostaną wylosowane zaledwie raz, mimo że w korpusie ich reprezentacja jest bogatsza. Co więcej, jeśli stosujemy losowanie ze zwrotem[6], niektóre rzeczywiste *hapax legomena* mogą być wylosowane więcej niż raz. Niemniej przesunięcia tego typu nie powinny zaburzyć obrazu całościowego przy wielokrotnie powtórzonym losowaniu: ostatecznie jeżeli wyraz jest częsty, to szansa, że zostanie wylosowany tylko jednokrotnie, jest niewielka, a jeśli w populacji jest dużo słów rzadkich, to i szansa na wylosowanie wielu z nich jako hapaksy jest spora.

Biorąc pod uwagę wszystkie powyższe założenia, podzieliliśmy nasz korpus na serię liczących po 50 lat podkorpusów (1380–1500, 1501–1550, 1551–1600,

[6] Losowanie, w którym ten sam element może być wielokrotnie losowany. Ten obrazowy termin nawiązuje do fizycznego losowania przedmiotów – przedmiot po wylosowaniu wraca do koszyka.



Rysunek 5.1. Przebieg zmian frekwencji i produktywności imiesłowów uprzednich. Wartości zostały przeskalowane, każdy punkt oznacza proporcję względem wartości maksymalnej.

1601–1650, 1651–1700, 1701–1750, 1751–1800, 1801–1850, 1851–1900, 1901––1939, pierwszy i ostatni liczą odpowiednio 120 i 38 lat). Dane zostały pozyskane zapytaniem o wszystkie ciągi liter kończące się na *-łszy* lub *-wszy*, które następnie poddano ręcznej selekcji. Później odpowiednim skryptem dokonywaliśmy losowania spośród wszystkich imiesłowów danego podkorpusu 178 z nich i sprawdzaliśmy, ile wśród nich jest hapaksów oraz obliczaliśmy wartość TTR. Liczba 178 to liczba okazów w najmniejszym z podkorpusów, a więc minimalna liczba, jaką można wziąć pod uwagę. Losowanie było następnie powtarzane 100 razy dla każdego podkorpusu, a wynik z wszystkich iteracji uśredniany.

5.3. Wyniki

Rysunek 5.1 przedstawia zmiany frekwencji, a także obu miar produktywności, tj. TTR i miary *P* Baayena wśród 178 okazów. Po to, by wszystkie trzy liczby można było przedstawić na jednym wykresie, należało je przeskalować – każdy punkt oznacza proporcję w stosunku do wartości maksymalnej. Liczba imiesłowów nie jest początkowo wysoka, jednak rośnie, by osiągnąć szczyt w okresie reprezentowanym przez podkorpus obejmujący lata 1600–1650 i następnie spadać. Pewne odbicie widać w drugiej połowie XIX wieku.

Przebieg zmian produktywności (TTR oraz miara *P*) w pewnych granicach odzwierciedla przebieg zmian frekwencji – wszystkie trzy krzywe osiągają swój

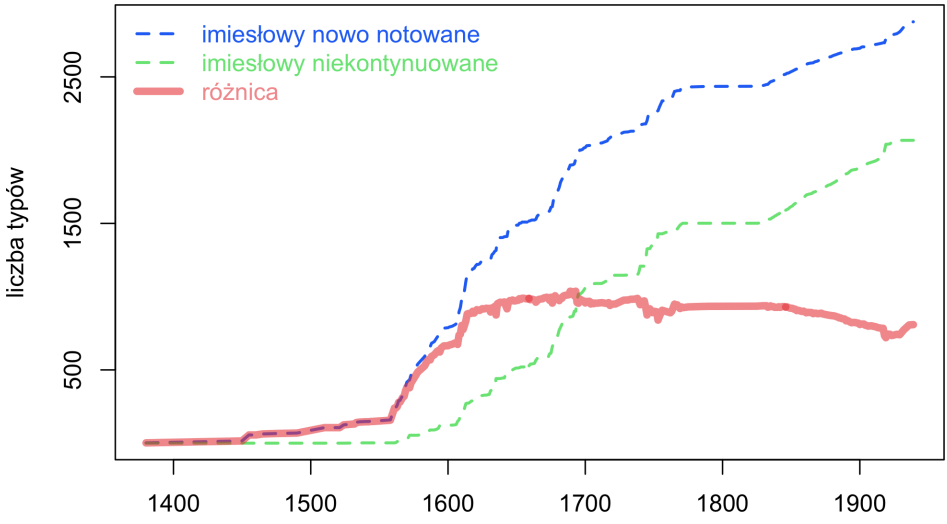
punkt szczytowy w tym samym momencie, a więc w latach 1600–1650 (rok na wykresie oznacza środek danego podkorpusu). Z drugiej strony zmiany frekwencji są nieco bardziej gwałtowne niż zmiany obu miar produktywności. Warto wreszcie zauważyć, że w ostatnim podkorpusie spada frekwencja, ale wzrasta produktywność. Tab. 5.1 i 5.2 przedstawiają omawiane dane (wartości bez skalowania).

Tabela 5.1. Frekwencja imiesłowów uprzednich w poszczególnych podkorpusach.

podokorpus	wielkość podkorpusu	liczba imiesłowów	liczba imiesłowów na 10 000 wystąpień
1380–1500	390940	178	4,5531
1501–1550	328830	324	9,8531
1551–1600	1530644	2178	14,2293
1601–1650	2991911	5336	17,8347
1651–1700	2893467	4369	15,0995
1701–1750	1570196	2180	13,8836
1751–1800	1384867	880	6,3544
1801–1850	840295	386	4,5936
1851–1900	741048	706	9,5270
1901–1939	1278016	829	6,4866

Trudno nie zauważyć, że uzyskane przez nas wyniki są nieco rozczarowujące. Przede wszystkim mimo drobnych odchyleń trajektorie trzech badanych wartości są jednak mocno ze sobą skorelowane. Najbardziej niepokojący jest z pewnością gwałtowny (i czasowy) zanik imiesłowów w pierwszej połowie XIX wieku – mamy uzasadnione powody, by sądzić, że po części odpowiada za to niedostateczne pokrycie korpusu.

Kolejny test będzie miał zatem na celu oszacowanie zmiany zasobów słownictwa występującego w imiesłowie uprzednim (Rys. 5.2). Interesują nas dwie liczby: liczba typów, które do danej daty pojawiły się w imiesłowie uprzednim choć raz (na wykresie oznaczona niebieską przerywaną linią), oraz liczba typów, które pojawiają się w danej dacie ostatni raz (linia zielona). Grubsza, czerwona linia przedstawia różnicę pomiędzy dwiema pozostałymi wartościami. Oczywiście wszystkie *hapax legomena* (w obrębie całego korpusu, a nie któregoś z podkorpusów) są odnotowywane na obu liniach naraz: data jego pierwszego pojawienia się jest równocześnie datą jego ostatniego wystąpienia. Zauważmy, że ten konkretny test stoi niejako w sprzeczności z tym, co założyliśmy wyżej, mianowicie tym



Rysunek 5.2. Liczba wyrazów zanotowanych do danego roku (linia niebieska) i po raz ostatni w danym roku (linia zielona) w latach 1380–1939.

razem korpus traktujemy jako jedną całość. Tu akurat nie przyjmujemy założenia, że wyraz użyty powiedzmy dwakroć w odstępie 200 lat trudno traktować jako trwały element słownictwa. W tym wypadku chodziło jednak o ocenę wielkości słownictwa znajdującego się w obiegu w danym czasie.

Tabela 5.2. Przebieg miary produktywności *P* Baayena w poszczególnych podkorpusach.

podkorpus	losowane *P*	losowana liczba typów
1380–1500	66	113
1501–1550	103	137
1551–1600	164	171
1601–1650	172	175
1651–1700	171	175
1701–1750	164	171
1751–1800	146	161
1801–1850	113	143
1851–1900	138	158
1901–1939	144	160

Wyniki testu trzeba uznać za interesujące. Poza obserwacją oczywistą, tj. dużym tąpnięciem na przełomie epoki staropolskiej i średniopolskiej (około 1550), uwagę zwraca względne ustabilizowanie się repertuaru imiesłowów uprzednich w 300-letnim okresie obejmującym wieki XVI–XIX: korpus notuje mniej więcej tyle samo nowych imiesłowów, ile form niekontynuowanych, co pokazuje stosunkowo stabilna różnica między formami nowymi i archaicznymi. Przy bliższym spojrzeniu owa względna stabilizacja okazuje się jednak powolnym, ale przecież nieuchronnym spadkiem: kolejnym dowodem na powolne wycofywanie się imiesłowu uprzedniego w polszczyźnie.

5.4. Podsumowanie

Na początku tego rozdziału przedstawiliśmy dwa możliwe scenariusze tłumaczące zmiany frekwencji. Scenariusz pierwszy zakłada, że jest to spowodowane wzrostem produktywności, który z kolei powoduje wzrost liczby typów słów, a w dalszej konsekwencji ich okazów. Scenariusz drugi dopatruje się wzrostu liczby słów nie tyle w produktywności, ile w swoistej modzie: liczba typów jest raczej niezmienna, gdyż tworzenie imiesłowów uprzednich podlega rozmaitym ograniczeniom, jednak w pewnych okresach poszczególne typy są używane (średnio) częściej, co w skutkuje wzrostem frekwencji całej formy.

Dane wskazują, że istnieje pozytywna korelacja pomiędzy produktywnością a frekwencją, choć nie jest ona bardzo silna. W szczególności znaczny spadek frekwencji w XVIII wieku nie pociąga za sobą równie dużego spadku produktywności. Wyniki więc sugerują odpowiedź pośrednią, że mianowicie za wahania frekwencji odpowiadają oba czynniki równocześnie, tj. zarówno ograniczenia semantyczne, jak i pewna moda językowa.

Można to ująć inaczej – wzrost frekwencji formy jedynie częściowo był spowodowany zwiększeniem się liczby czasowników, które w tej formie zostały użyte. Dużemu spadkowi frekwencji na przełomie XVII i XVIII wieku towarzyszy stosunkowo mniejszy spadek produktywności. Zarazem jednak w ostatnim podkorpusie, obejmującym okres 1900–1939, spadkowi frekwencji towarzyszy wzrost produktywności. Jeśli nasze rozumowanie jest prawidłowe, to tu mamy do czynienia z pewnym poluzowaniem ograniczeń dotyczących użycia imiesłowów, bowiem zmniejszeniu się liczby okazów towarzyszy zwiększenie się liczby typów.

Niemniej należy podkreślić, że wzrost frekwencji w XVII wieku był poprzedzony wyraźnym wzrostem produktywności. Nie miejsce tu, by rozstrzygnąć, na ile wynikało to z całościowego rozwoju języka, a więc na ile wzrost liczby słów używanych w imiesłowie uprzednim był tylko wynikiem ogólnego wzrostu słownictwa, a na ile z mechanizmów związanych bezpośrednio z tą formą fleksyjną.

Zakończenie

Celem, jaki postawili przed sobą autorzy tej książki, było ukazanie możliwości – ale też ograniczeń – metod ilościowych i korpusowych w diachronicznych badaniach polszczyzny. Nie planowaliśmy odkrywania nieznanych dotąd zmian kategorialnych i w tym sensie książka zapewne wzbudzi pewien niedosyt czytelników. Ale też przeszłość polszczyzny jest intensywnie badana od półtorej setki lat i wiemy na jej temat naprawdę dużo. Prawdopodobnie znamy już wszystkie prawa głosowe, jakie ukształtowały język polski, dobrze też są opisane dawne paradygmaty fleksyjne. Poszczególne zmiany leksykalne mają bogatą dokumentację w literaturze przedmiotu, rozpoznane jest geograficzne zróżnicowanie procesów diachronicznych, znamy wreszcie dość dobrze osobniczy styl ważniejszych autorów. W pewnym stopniu zbadane zostały również niektóre czynniki ilościowe wpływające na zmianę w polszczyźnie.

Wartość dodana naszej książki – chcielibyśmy wierzyć – polega natomiast na tym, że na znane wcześniej zjawiska językowe próbowaliśmy spojrzeć w perspektywie wielokrotnie szerszej niż mogli to robić nasi poprzednicy. Było to możliwe przede wszystkim dlatego, że mogliśmy się opierać na stosunkowo dużych korpusach tekstowych; dzięki korpusowi nie byliśmy skazani na ręczną ekscerpcję danych z kartotek słowników historycznych i ze starodruków. Drugie *novum* naszej książki wynika z zastosowanej przez nas metody, a właściwie kilku różnych metod odwołujących się do rozbudowanego aparatu statystycznego. Wierzymy, że dzięki temu byliśmy w stanie, po pierwsze, zbliżyć się do ideału zobiektywizowanej analizy opartej na danych empirycznych, replikacji eksperymentu i zasadzie falsyfikowalności, a po drugie – spojrzeć na dane językowe w perspektywie uogólniającej, scalającej, dostrzegającej w gąszczu zjawisk jednostkowych pewne ogólniejsze trendy rozwojowe polszczyzny.

Takie właśnie podejście pozwoliło nam dokonać porównania dynamiki kilku zmian językowych, jakie zaszły w okresie średniopolskim. Zwróćmy uwagę, że modelowanie zmian za pomocą regresji liniowej czy logistycznej to właśnie porzucanie szczegółów na rzecz ogólniejszego spojrzenia. Jedna czy druga wartość obserwowana mogła być całkiem odległa od wartości oczekiwanej, taki czy inny autor mógł wyprzedzać swoją epokę bądź przeciwnie – pozostać konserwatystą. Wszystkie te jednostkowe fakty mają oczywiście ogromne znaczenie w historii polszczyzny, dla nas jednak ważniejsza była obserwacja, że owe nieliczne

odstające punkty mówią w istocie znacznie mniej o historii modelowanego procesu niż właśnie sam model. Model bowiem, przypomnijmy, pozwala na uchwycenie ogólnego trendu. Pojedynczy trend zresztą również nie daje takiej wiedzy o zmianach w polszczyźnie jak porównanie przebiegu kilku procesów częściowo na siebie zachodzących.

Innym naszym celem była próba ustalania periodyzacji w oparciu o kryteria nie jakościowe, ale ilościowe. Aby odszukać na osi czasu moment, który dzieli teksty na dwie najbardziej odmienne grupy, zaprzęgnęliśmy metody klasyfikacji nadzorowanej. W takim podejściu znów godzimy się na to, że tracimy z oczu szczegóły, ponieważ metody uczenia maszynowego są skuteczne niezależenie od tego, czy badacz wie, co się zmienia. (Najbardziej wyrafinowane algorytmy sztucznej inteligencji idą w tym względzie jeszcze dalej: badacz nie jest w stanie dociec, w jaki sposób sieć neuronowa znalazła rozwiązanie badanego problemu.) Odnotowujemy tu jedynie sam fakt zmiany. Z jednej strony jest to dla lingwisty rozczarowujące, bo przecież chciałby wiedzieć, czym różnią się teksty z dwu różnych okresów, z drugiej jednak strony wypracowana przez nas metoda pozwala zobiektywizować intuicję, że istnieją daty, które bardziej niż inne oddzielają teksty wcześniejsze od późniejszych. Daty, w których wyraźnie zmienia się sposób tworzenia wypowiedzi, i to zarówno pod względem leksykalnym, jak i składniowym.

Warto przy tym pamiętać, że zmiana językowa rozgrywa się nie tylko w systemie, ale także w obrębie tego, co się lokuje pomiędzy *langue* a *parole*. Chodzi tu o subtelne przetasowania na liście rangowej wyrazów czy zmiany preferencji szyku lub kolokacji bądź stopnia rozbudowy fraz. Są to zmiany przede wszystkim o charakterze ilościowym. Co więcej, każda z nich osobno odgrywa niewielką (wręcz marginalną) rolę, dopiero ich kumulacja sprawia, że teksty powstałe w różnym czasie są od siebie odmienne.

Jak się okazało, wyniki eksperymentu przeprowadzonego na tysiącu częstych wyrazów różniły się nieco od analizy skupień – prostej, ale przez to bardziej intuicyjnej – opartej na kilkunastu zaledwie przyimkach polskich. Wyniki tego bowiem testu pokazały, że największe zmiany nastąpiły jeszcze w wieku XV, a dopiero następne duże pęknięcie miało miejsce w połowie wieku XVI, w momencie zwyczajowo uznawanym za przejście z okresu staropolskiego do średniopolskiego. Uzyskane różnice między oboma podejściami niekoniecznie jednak kwestionują wartość metodologii kwantytatywnej. Raczej wolelibyśmy zadać pytanie (do podjęcia w następnych studiach): czy momenty węzłowe w ewolucji języka muszą przypadać w tym samym czasie w odniesieniu do gramatyki i leksyki?

Rzecz jasna w obrębie samej warstwy leksykalnej możemy mieć do czynienia nie tyle ze zmianą językową, ile raczej ze zmianą gustu literackiego. Istnieją przekonujące argumenty na rzecz stwierdzenia, że często obserwujemy właśnie to drugie. Przede wszystkim eksperyment przeprowadzony na korpusie CLMET

3.0 pokazał, że sygnał typu tekstu jest silniejszy niż sygnał chronologiczny, teksty bowiem grupują się najpierw według typu, a dopiero w obrębie typu chronologicznie. Zauważmy jednak, że wyznaczniki stylu to pewna konwencja, którą mimo wszystko można przełamać, dużo trudniej natomiast uciec od powszechnie obowiązującej gramatyki. Jeśli więc sygnał chronologiczny jest słabszy od sygnału tego, co uznaliśmy za konwencję, to raczej nie możemy go uznać za odbicie zmian gramatycznych. Jest to więc przede wszystkim zmiana „kostiumu literackiego". Przekonująca odpowiedź na pytanie wymaga oczywiście dalszych, pogłębionych badań.

Językoznawca korpusowy prędzej czy później rozbija się o problem niedostatku danych i – paradoksalnie – dzieje się tak również wtedy, gdy pracuje na liczących setki milionów słów korpusach języka współczesnego. W wypadku referowanych tu badań nie mogliśmy w pełni odpowiedzieć na nurtujące nas pytanie dotyczące dziejów imiesłowu uprzedniego: czy w różnych epokach istniały leksemy, które były wyraźnie preferowane w tej formie fleksyjnej i czy preferencje te ewoluowały? Niestety w większości jednostkowe poświadczenia nie pozwalały na wyciągnięcie wiążących wniosków. Równocześnie jednak sięgnęliśmy po metodę badania produktywności, która pozwoliła przynajmniej pośrednio dotknąć tego pytania.

Na koniec garść uwag na temat samego korpusu. Jak wiadomo, wszelkie wnioski z badań są warte tyle, ile dane empiryczne, na których zostały oparte. Zapewne gdybyśmy dysponowali znacząco większym i bardziej reprezentatywnym korpusem, niektóre szczegółowe ustalenia byłyby odmienne. Rodzące się w tym miejscu pytanie o to, czy w takim razie wartości odstające od przebiegu oczekiwanego są jedynie wynikiem niedoskonałości korpusu, musimy pozostawić bez odpowiedzi. Jest jednak rzeczą prawdopodobną, że większy, zrównoważony korpus dałby gładsze (tj. pozbawione wyraźnie odstających punktów) wyniki. I choć marzenie o większym korpusie nie wydaje się szczególnie wygórowane w kontekście najnowszych projektów pozyskania i scalenia dużych ilości danych diachronicznych (Król i in., 2019), o tyle tezy o zbudowaniu reprezentatywnego i zrównoważonego korpusu diachronicznego trzeba chyba między bajki włożyć. Po pierwsze, wciąż przecież nie ma zgody co do tego, jak powinien wyglądać idealny korpus, po drugie zaś, stan zachowania tekstów z epok dawnych sprawia, że w odniesieniu do nich korpusy są z zasady „niereprezentatywne". Cytując obiegowe powiedzenie: zadanie losowania tekstów do korpusów historycznych wzięła na siebie historia.

Mimo wszystko dzięki zastosowaniu korpusów elektronicznych podstawa empiryczna naszych badań była znacząco większa niż to miało miejsce dotychczas. Zadawaliśmy sobie pytanie, czy nie warto było poczekać na to, aż będziemy dysponowali większymi, publicznie dostępnymi korpusami historycznymi. Jesteśmy jednak przekonani, że w językoznawstwie korpusowym musi zachodzić sprzężenie zwrotne – równolegle z tworzeniem korpusów należy prowadzić na

nich badania, te ostatnie wytyczają bowiem z jednej strony szlak, jakim powinni podążać twórcy tych zbiorów danych, z drugiej zaś stawiają pytania, na które da się odpowiedzieć tylko wtedy, gdy się dysponuje lepszym korpusem. Wolno sądzić, że odbywa się tutaj rozpisany na setki lat dialog między twórcami edycji, słowników i korpusów – słowem: uczonymi zajmującymi się podstawą materiałową – a badaczami korzystającymi z tych zasobów w próbach opisu języka. Jedni korzystają z ustaleń tych drugich, dzięki czemu z każdym pokoleniem powiększa się zarówno nasza wiedza o języku (wzbogacona nowymi odkryciami materiałowymi), jak i jakość ogłaszanych edycji czy korpusów (ulepszanych dzięki odkryciom historyków języka). W naszej książce podjęliśmy próbę spojrzenia scalającego, obejmującego kilka stuleci rozwoju polszczyzny, wierząc, że da ona również impuls do stworzenia dużego, obejmującego całość dziejów języka polskiego, anotowanego korpusu diachronicznego.

Mimo tak daleko zakrojonej wizji autorzy uznają, że spełnili swoje zadanie, jeśli ta książka ukaże Czytelnikowi potencjał, jaki tkwi w metodach korpusowych i ilościowych w badaniu przeszłości języka.

Bibliografia

Adamiec, D. (2015). Kryteria doboru tekstów do „Elektronicznego korpusu tekstów polskich z XVII i XVIII w. (do 1772 r.)". *Prace Filologiczne*, *67*, 11–20.

Altmann, G. (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. W: K.-H. Best i J. Kohlhase (red.), *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte* (ss. 54–90). Göttingen: Edition Herodot.

Baayen, H. (1993). On frequency, transparency and productivity. W: G. Booij i J. van Marle (red.), *Yearbook of morphology 1992* (ss. 181–208). Dordrecht: Springer.

Baayen, H. (2009). *Analysing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, H. (2012). *Word frequency distributions*. Springer.

Baayen, H., Van Halteren, H. i Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121–132.

Bajerowa, I. (1964). *Kształtowanie się systemu polskiego języka literackiego w XVIII wieku*. Wrocław: Ossolineum.

Bauer, L. (2001). *Morphological productivity*. Cambridge–New York: Cambridge University Press.

Best, K.-H. (1983). Zum morphologischen Wandel einiger deutscher Verben. W: K.-H. Best i J. Kohlhase (red.), *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte* (ss. 107–118). Göttingen: Edition Herodot.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9–37. https://doi.org/10.1515/cllt-2012-0002

Bień, J. S. (2014). The IMPACT project Polish ground-truth texts as a DjVu corpus. *Cognitive Studies | Études Cognitives*, *14*, 75–84.

Bronikowska, R. i Przyborska-Szulc, A. (2018). Elektroniczny korpus tekstów polskich z XVII i XVIII wieku (do 1772 roku). W: M. Pastuch i M. Siuciak (red.), *Historia języka w XXI wieku. Stan i perspektywy* (ss. 129–135). Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Danchev, A. i Kytö, M. (1994). The construction "be going to" + infinitive in Early Modern English. W: D. Kastovsky (red.), *Bamberger Beitrage zur Englischen Sprachwissenschaft* (ss. 59–77). Berlin – New York: Mouton de Gruyter.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, *7*(2), 121–157.

De Smet, H. (2005). A corpus of Late Modern English texts. *International Computer Archive of Modern and Medieval English*, *29*, 69–82. http://www.ovs0.com/PDF/40336372.pdf

Deptuchowa, E., Jasińska, K., Klapper, M., Kołodziej, D., Frodyma, M. i Leńczuk, M. (2019). Korpus Polszczyzny do 1500 r. Referat przedstawiony na seminarium naukowym „Ku integracji korpusów diachronicznych języka polskiego", Warszawa.

Derwojedowa, M., Kieraś, W., Skowrońska, D. i Wołosz, R. (2014). Korpus polszczyzny XIX wieku od mikrokorpusu do korpusu średniej wielkości. *Prace Filologiczne*, *65*, 251–256.

Dressler, W. U. (1997). *On productivity and potentiality in inflectional morphology*. Montreal.

Eder, M. (2013). Mind your corpus: Systematic errors in authorship attribution. *Literary and Linguistic Computing*, *28*(4), 603–614. http://llc.oxfordjournals.org/content/28/4/603

Eder, M. (2014). Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii. *Teksty Drugie*, *2*, 90–105.

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities*, *32*(1), 50–64.

Eder, M. (2018). Words that have made history, or modeling the dynamics of linguistic changes. W: *Digital Humanities 2018: Book of Abstracts* (ss. 362–365). Mexico City. https://dh2018.adho.org/en/abstracts/

Eder, M. i Górski, R. L. (2016). Historical linguistics' new toys, or stylometry applied to the study of language change. W: *Digital Humanities 2016: Conference Abstracts* (ss. 182–184). Kraków: Uniwersytet Jagielloński i Uniwersytet Pedagogiczny. http://dh2016.adho.org/abstracts/398

Eder, M. i Górski, R. L. (w druku). Stylistic fingerprints, POS tags and inflected languages: A case study in Polish. *Journal of Quantitative Linguistics*.

Eder, M. i Rybicki, J. (2013). Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, *28*(2), 229–236.

Eder, M., Klapper, M. i Kołodziej, D. (2015). Dawna polszczyzna i nowe technologie: Testowanie metod przetwarzania języka naturalnego na materiale polskiego piśmiennictwa od średniowiecza po wiek XX. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *71*, 189–202.

Eder, M., Rybicki, J. i Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal*, *8*(1), 107–121. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Ellegård, A. (1953). *The auxiliary "do": The establishment and regulation of its use in English*. Stockholm: Almquist & Wiksell.

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Rybicki, J. i Byszuk, J. (2018). Attributing authorship in the noisy digitized correspondence of Jacob and Wilhelm Grimm. *Frontiers in Digital Humanities*, *5*, 4. https://doi.org/10.3389/fdigh.2018.00004

Gaeta, L. i Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, *44*(1), 57–89. https://doi.org/10.1515/LING.2006.003

Górski, R. L. (2016). Jeszcze raz o produktywności formantów przymiotnikowych. *Prace Filologiczne*, *68*, 111–128.

Górski, R. L. i Król, M. (2018). The Polish adverbial perfect participle: A corpus-based study. W: B. Szymanek i W. Guz (red.), *Canonical and non-canonical structures in Polish* (ss. 55–69). Lublin: Wydawnictwo KUL.

Górski, R. L. i Łaziński, M. (2012). Reprezentatywność i zrównoważenie korpusu. W: A. Przepiórkowski, M. Bańko, R. L. Górski, i B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego* (ss. 25–36). Warszawa: PWN.

Gries, S. T. i Hilpert, M. (2008). The identification of stages in diachronic data: Variability--based neighbour clustering. *Corpora*, *3*(1), 59–81.

Gruszczyński, W., Adamiec, D. i Ogrodniczuk, M. (2013). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.). Prezentacja projektu badawczego. *Polonica*, *33*, 309–316.

Hastie, T., Friedman, J. i Tibshirani, R. (2001). *The elements of statistical learning*. New York: Springer. https://doi.org/10.1007/978-0-387-21606-5

Hilpert, M. (2008). *Germanic future constructions: A usage-based approach to language change*. John Benjamins.

Hirst, G. i Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, *22*(4), 405–417.

Hjelmslev, L. (1942). Langue et parole. *Cahiers Ferdinand de Saussure*, *2*, 29–44.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, *41*(2), 174–203.

James, G., Witten, D., Hastie, T. i Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer.

Jasińska, K. i Kołodziej, D. (2019). The Lexical Database of the Medieval Polish Language – on the investigation of the Old Polish inflection. Referat przedstawiony na konferencji ICHLL-10, Leeuwarden.

Karasiowa, A. i Mayenowa, M. R. (1978). Oboczność in(n)y|inszy w historii języka polskiego. W: *Tekst. Język. Poetyka* (ss. 279–297). Wrocław: Ossolineum.

Kestemont, M., Karsdorp, F. i Düring, M. (2014). Mining the twentieth century's history from the Time Magazine corpus. W: *Abstract Book of EACL 2014: The 14th Conference of the European Chapter of the Association for Computational Linguistics* (s. 62).

Klemensiewicz, Z. (1965). *Historia języka polskiego*. Warszawa: PWN.

Kleszczowa, K. i Mika, T. (2018). Wyzwania badawcze i metodologiczne lingwistyki historycznej. Refleksje po dyskusji. W: M. Pastuch i M. Siuciak (red.), *Historia języka w XXI wieku. Stan i perspektywy* (ss. 639–652). Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Koppel, M. i Schler, J. (2004). Authorship verification as a one-class classification problem. W: *Proceedings of the twenty-first international conference on machine learning* (ss. 1–7). New York: ACM. http://dl.acm.org/citation.cfm?id=1015448

Kowalska, A. (1978). Rozwój nowych form słowa posiłkowego: jestem, jesteś, jesteśmy, jesteście. *Poradnik Językowy*, *9*, 377–384.

Köhler, R. (2015). Linguistic modelling of sequential phenomena. W: A. Tuzzi, G. K. Mikros i J. Macutek (red.), *Sequences in language and text* (ss. 109–123). Walter de Gruyter.

Krążyńska, Z. (1993). Synonimiczne funkcje przyimków do i k(u) w języku staropolskim. W: M. Basaj i Z. Zagórski (red.), *Munera linguistica Ladislao Kuraszkiewicz dedicata* (ss. 163–170). Wrocław: Ossolineum.

Krążyńska, Z. (2001). *Staropolskie konstrukcje z przyimkami*. Cz. 2: *Po, przez, prze, mi(e)mo, nad, wz, pod, przed, za, o, miedzy + acc*. Poznań: WiS.

Krążyńska, Z., Mika, T. i Słoboda, A. (2015). *Składnia średniowiecznej polszczyzny*. Cz. 1: *Konteksty – metody – tendencje*. Poznań: Rys.

Kroch, A. S. (1989). Function and grammar in the history of English: Periphrastic "do". W: R. W. Fasold i D. Schriffin (red.), *Language change and variation* (ss. 132–172). Amsterdam–Philadelphia: John Benjamins.

Król, M., Derwojedowa, M., Górski, R. L., Gruszczyński, W., Opaliński, K., Potoniec, P., Kieraś W., Woliński M. i Eder, M. (2019). Narodowy Korpus Diachroniczny Polszczyzny. Projekt. *Język Polski*, *99*, 92–101.

Kubát, M. i Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339–349. https://doi.org/10.1080/09296174.2013.830552

Kuraszkiewicz, W. (1953). *Pochodzenie polskiego języka literackiego w świetle wyników dialektologii historycznej*. Wrocław: Ossolineum.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K. i Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny współczesnej* (t. 1-2). Kraków: IJP PAN.

Kytö, M. (2011). Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, *11*(2), 417–457.

Le, X., Lancashire, I., Hirst, G. i Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, *26*(4), 435–461.

Leopold, E. (2005). Das Piotrowski-Gesetz. W: R. Köhler, G. Altmann i R. Piotrowski (red.), *Quantitative Linguistik – Quantitative Linguistics. Ein internationales Handbuch* (ss. 627–633). Berlin–New York: de Gruyter.

Lindquist, H. i Mair, C. (red.). (2004). *Corpus approaches to grammaticalization in English*. Amsterdam–Philadelphia: John Benjamins.

Mair, C. (2006). *Twentieth-century English: History, variation and standardization*. Cambridge: Cambridge University Press.

McEnery, T. i Wilson, A. (2001). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Michalska, P. (2013). *Status staropolskich oboczności wyrazowych w polszczyźnie doby średniopolskiej*. Poznań: Wydawnictwo PTPN.

Moisl, H. (2014). *Cluster analysis for corpus linguistics*. Berlin: Mouton de Gruyter.

Mosteller, F. i Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Stanford: CSLI Publications.

Motyl, A. (2014). *Normalizacja fleksji werbalnej w zakresie kategorii czasu w dobie średniopolskiej*. Poznań: Wydawnictwo PTPN.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Ogura, M. (1993). The development of periphrastic "do" in English: A case of lexical diffusion in syntax. *Diachronica*, *10*(1), 51–85.

Osiewicz, M. (2012). Wpływ zecera na ukształtowanie graficznojęzykowe tekstu drukowanego. Uwagi wstępne do analizy „Ksiąg o Gospodarstwie" z 1549 r. *LingVaria*, *14*(2), 65–76.

Ostaszewska, D. (red.). (2002). *Polszczyzna XVII wieku: Stan i przeobrażenia*. Katowice: Śląsk.

Pawłowski, A. (2016). Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish. W: *Digital Humanities 2016: Conference Abstracts* (ss. 311–313). Kraków: Uniwersytet Jagielloński i Uniwersytet Pedagogiczny.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimization. *TASK Quarterly*, *11*(1–2), 151–167.

Pihan-Kijasowa, A. (1992). Z dziejów kształtowania się norm polszczyzny literackiej XVII wieku w zakresie leksyki. W: J. Zieniukowa (red.), *Procesy rozwojowe w językach słowiańskich* (ss. 125–136). Warszawa: Slawistyczny Ośrodek Wydawniczy.

Piotrowskaja, A. i Piotrowski, R. (1974). Matematičeskie modeli v diachronii i tekstoobrazovanii. W: R. Piotrowski (red.), *Statistika reči i avtomatičeskij analiz teksta* (ss. 361–400). Leningrad: Nauka.

Piotrowski, M. (2012). *Natural language processing for historical texts*. San Rafael, CA: Morgan & Claypool.

Pisarkowa, K. (1984). *Historia składni języka polskiego*. Wrocław: Ossolineum.

Plecháč, P., Bobenhausen, K. i Hammerich, B. (2018). Versification and authorship attribution: A pilot study on Czech, German, Spanish, and English poetry. *Studia Metrica et Poetica*, *5*(2), 29–54.

Przepiórkowski, A., Bańko, M., Górski, R. L. i Lewandowska-Tomaszczyk, B. (red.). (2012). *Narodowy Korpus Języka Polskiego*. Warszawa: PWN.

Rabiega-Wisniewska, J. (2008). Wpływ fleksji na derywację – dyskusja podziału morfologii. *LingVaria*, *6*(2), 41–60.

Rosemeyer, M. (2015). How usage rescues the system: Persistence as conservation. W: A. Aria, G. G. Marco i K. Göz (red.), *Variation in language: System- and usage-based approaches* (ss. 289–316). Berlin–Boston: de Gruyter. https://doi.org/10.1515/9783110346855

Rospond, S. (1950). *Studia nad językiem polskim XVI wieku: Jan Seklucjan, Stanisław Murzynowski, Jan Sandecki-Małecki, Grzegorz Orszak*. Wrocław: Wrocławskie Towarzystwo Naukowe.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator: Stylometry in translation. W: M. P. Oakes i M. Ji (red.), *Quantitative methods in corpus-based translation studies* (ss. 231–248). Amsterdam: John Benjamins.

Rydén, M. (1979). *An introduction to the historical study of English syntax*. Stockholm: Almqvist & Wiksell.

Sang, E. T. K. (2016). Improving part-of-speech tagging of historical text by first translating to modern text. W: *Computational history and data-driven humanities* (ss. 54–64). Dublin: Springer.

Schöch, C. (2017). Topic modeling genre: An exploration of French classical and Enlightenment drama. *Digital Humanities Quarterly*, *11*(2), 1–53.

Stachowski, K. (2013). The influx rate of Turkic glosses in Hungarian and Polish post/mediaeval texts. W: R. Köhler i G. Altmann (Eds.), *Issues in quantitative linguistics 3* (ss. 100–116). Lüdenscheid: RAM-Verlag.

Stachowski, K. (w druku). Piotrowski-Altmann law: State of the art, ss. 1–6.

Stamatatos, E., Fakotakis, N. i Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, *26*(4), 471–495.

Stamou, C. (2008). Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, *23*(2), 181–199.

Stefanowitsch, A. i Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Szmrecsanyi, B. (2015). About text frequencies in historical linguistics: Disentangling environmental and grammatical change. *Corpus Linguistics and Linguistic Theory*, *12*(1), 153–171. https://doi.org/10.1515/cllt-2015-0068

Sławski, F. (1952). *Słownik etymologiczny języka polskiego*. Kraków: TMJP.

Taszycki, W. (1946). Staropolskie formy czasu przeszłego robiłech, robilichmy. W: *Sprawozdania Polskiej Akademii Umiejętności* (t. 46, ss. 7–10).

Taylor, A. i Kroch, A. S. (1994). The Penn-Helsinki Parsed Corpus of Middle English. University of Pennsylvania.

Tibshirani, R., Hastie, T., Narasimhan, B. i Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, *99*(10), 6567–6572. https://doi.org/10.1073/pnas.082099299

Twardzik, W. (red.). (2005). *Opis źródeł Słownika staropolskiego*. Kraków: Lexis.

Twardzik, W. i Górski, R. L. (2003). Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie. W: S. Gajda (red.), *Językoznawstwo w Polsce. Stan i perspektywy* (ss. 155–157). Opole: Wydawnictwo Uniwersytetu Opolskiego.

Tweedie, F. i Baayen, H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323–252. https://doi.org/10.1023/A:1001749303137

Urbańczyk, S. (1953). *Rola wielkich pisarzy odrodzenia na tle innych czynników kształtujących język literacki*. Warszawa: Polska Akademia Nauk.

Vulanović, R. (2007). Fitting periphrastic "do" in affirmative declaratives. *Journal of Quantitative Linguistics*, *14*(2–3), 111–126.

Vulanović, R. i Baayen, H. (2007). Fitting the development of periphrastic "do" in all sentence types. W: P. Grzybek i R. Köhler (Eds.), *Exact methods in the study of language and text: Dedicated to Gabriel Altmann on the occasion of his 75th birthday* (ss. 679–688). Berlin: de Gruyter.

Waldenfels, R. von. (2012). *The grammaticalization of "give" + infinitive: A comparative study of Russian, Polish, and Czech*. Bergen: de Gruyter.

Wierzbicka, A. (1966). *System składniowo-stylistyczny prozy polskiego renesansu*. Warszawa: PIW.

Winter, T. N. (1999). Roberto Busa, S.J., and the invention of the machine-generated concordance. *The Classical Bulletin*, *75*(1), 3–20.

Woliński, M., Głowińska, K. i Świdziński, M. (2011). A preliminary version of Składnica – a treebank of Polish. W: *Proceedings of the 5th Language & Technology Conference*. Poznań.

Zeldes, A. (2012). *Productivity in argument selection: From morphology to syntax*. Berlin: de Gruyter.

Zipf, G. (1949). *Human behavior and the principle of least effort. An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Summary

The present book discusses the possible applications, as well as limitations, of corpus and quantitative methods in the field of historical linguistics, with special attention paid to the history of the Polish language. The authors examine several language changes, well-known to the previous generations of scholars, in order to obtain a more detailed picture of language evolution, but also aim at detecting language phenomena whose existence was not reported in literature using machine-learning methods of classification. In other words, the book combines both the corpus-based and the corpus-driven approaches.

Chapter 1 provides a general overview of the advances of corpus and quantitative linguistics in examining the history of a language. Understandably, the vast majority of the studies conducted focuses on the history of English, which is the most well-resourced language both in terms of the available corpora and when it comes to natural language processing tools.

Chapter 2 discusses the research material used in the present study, namely an annotated diachronic corpus of the Polish language, covering the timespan 1380–1850, compiled of existing diachronic corpora and supplemented by several texts scraped from the internet or OCR-ed from printed sources, or manually transcribed from early modern prints. Different text normalization and modernization strategies applied to particular texts, in order to research the corpus in a consistent way over the centuries, are also discussed in this chapter. In its final form, the diachronic corpus consists of over 12 million words. Even if inherently opportunistic, the corpus tries to meet the requirement of representativeness, and to provide an even coverage of texts over the centuries, so as to reduce the timespans not represented by any text.

Chapter 3 is devoted to modeling a few of the language changes in Polish which took place in the 15th–18th centuries. These include isolated changes such as *więtszy* > *większy*, *abo* > *albo*, *barzo* > *bardzo*, *wszytek* > *wszystek*, morphological changes of the superlative marker (*na-* > *naj-*), verbal inflection (*-bych* > *-bym*, *-bychmy* > *-byśmy*, and *-łech* > *-łem*), and a phonological change (*-ir-* > *-ier-*). Since the process of replacement of an older (or recessive) form by a new (or innovative) one is prolonged in time and gradual, yet never linear, it can be best modelled by logistic regression. It is a mathematical model, designed to capture a dynamic change between two states or two phases of a given phenomenon, and it

seems to describe language changes quite well. In fact, most of the changes we have chosen for our study could be modelled with reasonable accuracy. A special case is *abo* > *albo*, which turned out to be a reversed (or not fully accomplished) change. It can be modelled either with polynomial logistic regression (showing relatively high accuracy) or, to a lesser degree, by a combination of two independent logistic models.

When modeling any diachronic process, the researcher has to divide the corpus into chronologically ordered subcorpora of an arbitrarily-set size: such a decision affects (or might affect) the final results. Obviously, bigger yet fewer subcorpora (e.g., 10 units covering 50 years each) provide smoother results, while smaller yet denser subcorpora (e.g. 50 units per 10 years) lead to fine-grained outcomes. To examine the degree of a model's stability despite the changing size of the subcorpora, the fit of the logistic regression and the input parameters was systematically tested. This has proven that 20 years is a minimal size of a subcorpus yielding credible data. Still, the size of the subcorpus affected the goodness of fit to a very limited extent in the case of those changes, where the observed and expected values were close to each other. And reversely, where the values observed were far removed from the expected ones, the goodness of fit increased with the timespan of the subcorpus.

Chapter 4 discusses different methods of automatic text classification, including multidimensional scaling, bootstrap consensus networks, supervised binary classifiers, and so forth. In the first place, however, it introduces a new method of finding turning points in the history of a language. The underlying assumption is that, although the language evolves continuously, there are certain moments when this evolution accelerates. Here, a corpus is a collection of chronologically-ordered texts. If there is a turning point in the evolution of language, then the texts written before this date should be more similar to each other rather than to those created after the assumed turning point and vice versa. The corpus is divided into two subcorpora, one preceding, the other following the hypothetical date of maximal change; for the sake of convenience we call them *ante* and *post*, respectively. An unsupervised classification is then conducted in order to attribute each text either to the *ante* or *post* class. The entire procedure is repeated several times, with each iteration shifting the date dividing the corpus into two subcorpora by a fixed number of years. The highest accuracy over the iterations indicates that the corpus is divided in two most heterogenous subcorpora, or – in other words – that the texts written before and after that date show the biggest difference. This, in turn, suggests that such a date indicates a major change in the course of the language's evolution.

Another approach to detecting major turning points is a variant of the hierarchical cluster analysis method. In its classical flavor, it sequently merges most similar subcorpora, finally grouping them into two clusters. In the variant used here, it is exclusively the neighboring corpora that are allowed to merge. Consequently, the

two top clusters cover subcorpora preceding and following a certain date. The two most diverse adjacent subcorpora are the one directly before and the one directly after that date, which can also be seen as the moment of maximal change in the phenomenon observed. In this study, prepositions were used as a discriminator. The most significant change in the relative frequency of prepositions occurred in the 16th century, which is commonly assumed as the beginning of the Middle Polish period.

Chapter 5 is a case study on the history of a selected grammatical category, namely the adverbial past participle. Since the frequency of this category underwent a substantial change, the question arises what was the underlying factor responsible for the change. Two answers were considered; either it was mere linguistic fashion which made the speakers overuse the form in the 17th century, or alternatively, the form became more productive. In the first case a similar set of verbs would be used with higher frequency, in the second there would be a much higher number of types, which would indicate that the number of word-types has increased significantly, presumably because the semantic restrictions imposed on coining these participles became somewhat looser. Productivity was estimated only with quantitative criteria. The data shows that there are some changes in productivity, however they are less significant than the changes in frequency. This, in turn, supports both claims: the form did become somewhat fashionable, but this consequently led to higher productivity. The drop of frequency was followed by a much more restricted drop of productivity.