ARTYKUŁY I ROZPRAWY

MACIEJ EDER*, RAFAŁ L. GÓRSKI**

INSTYTUT JĘZYKA POLSKIEGO POLSKIEJ AKADEMII NAUK, KRAKÓW

Typologia tekstów oparta na miarach kwantytatywnych: studium korpusowe o zróżnicowaniu polszczyzny[1]

Słowa kluczowe: stylistyka, genologia językoznawcza, językoznawstwo korpusowe, metody wielowymiarowe.

doi: http://dx.doi.org/10.31286/JP.99.3.1

Wprowadzenie

Polonistyczna genologia lingwistyczna może się pochwalić wieloma niewątpliwymi osiągnięciami (ważne podsumowania tego dorobku zawierają m.in. prace Gajda (red.) 1995; Malinowska i in. (red.) 2013), również w zakresie ilościowej charakterystyki tekstów (Mikołajczak 1990; Ruszkowski 1997; Górski, Łaziński 2010 itp.). Niemniej jednak wciąż brakuje typologii, która byłaby oparta nie tyle na funkcjach, jakie tekst spełnia w komunikacji językowej, ile na jego cechach gramatycznych[2]. W niniejszym artykule przedstawiamy wyniki badań nad taką właśnie ilościową charakterystyką różnic gramatycznych pomiędzy z góry przyjętymi typami tekstów. Punktem wyjścia jest typologia zaproponowana przez Rafała Ludwika Górskiego i Marka Łazińskiego (2012) dla NKJP. Typologia ta została lekko zmodyfikowana przez podział pewnych typów na podtypy. Tak więc jest to typologia NKJP o zwiększonej ziarnistości.

W niniejszej pracy stosujemy podejście kwantytatywne, zadajemy zatem pytanie, czy policzalne cechy języka (np. częstości słów, konstrukcji składniowych, cech fleksyjnych, a nawet interpunkcji czy, powiedzmy, długości akapitów) mają rozkład na tyle podobny w obrębie jakiejś grupy tekstów, że możemy wiarygodnie, to znaczy na przyjętym poziomie istotności statystycznej, wyodrębnić ową grupę tekstów spomiędzy innych tekstów, na przykład z całego korpusu.

W obrębie metod kwantytatywnych odwołujemy się również do analizy wielowymiarowej (w kilku różnych wariantach), czyli analizy wielu cech językowych j e d n o c z e ś n i e.

* maciej.eder@ijp.pan.pl; ORCID: 0000-0002-1429-5036

** rafal.gorski@ijp.pan.pl; ORCID: 0000-0003-4727-2639

1 Wkład autorski: Maciej Eder (50%) – współautorstwo koncepcji badania, statystyczne opracowania materiału, stworzenie skryptów i rysunków; Rafał L. Górski, autor korespondencyjny (50%) – współautorstwo koncepcji badania, wybór cech językowych, przygotowanie korpusu.

2 Pomijamy tu prace Stanisława Mikołajczaka (1990) i Marka Ruszkowskiego (1997) dotyczące charakterystyki ilościowej wybranych cech składniowych.

Istotę takiego podejścia unaocznić może następujący (wymyślony naprędce) przykład: wolno sądzić, że w języku mówionym znajdzie się znacznie więcej form czasownikowych w 2. os. niż w języku pism urzędowych. Z kolei w utworach literackich powinniśmy oczekiwać bogatszej składni (a więc liczniejszych wskaźników zespolenia: *że*, *który*, *wprawdzie*, *ale* itd.) niż w pismach urzędowych, ale jednocześnie możemy się spodziewać mniejszej liczby słów cztero- czy pięciosylabowych (oraz dłuższych). Metody wielowymiarowe starają się odpowiedzieć na następujące pytanie: czy jednoczesne porównanie frekwencji form 2. os. czasownika, liczby spójników oraz liczby słów wielosylabowych przyporządkuje badane teksty do trzech (spodziewanych) grup? W rzeczywistych studiach kwantytatywnych liczba badanych cech języka jest oczywiście większa: na ogół mierzy się frekwencje kilkunastu, kilkudziesięciu, a nawet kilkuset zmiennych jednocześnie.

Metody wielowymiarowe stosowane są z powodzeniem w językoznawstwie kwantytatywnym (pomijamy tutaj pierwotne zastosowania w naukach ścisłych, w naukach technicznych, w medycynie itd.). Chyba najczęściej podejmowanym problemem badawczym jest pytanie o autorstwo tekstu anonimowego przez porównanie frekwencji kilkudziesięciu (czasem kilkuset) wyrazów: jeśli dysponujemy anonimowym tekstem oraz korpusem porównawczym, to przez takie właśnie zbadanie frekwencji jesteśmy w stanie wskazać w korpusie porównawczym grupę tekstów, do których owo anonimowe dzieło jest najbardziej podobne. Autor tej grupy tekstów jest też najprawdopodobniej autorem tekstu anonimowego. Powiedzmy wszakże z naciskiem: jedyną rzeczą, którą porównujemy, są różnice we frekwencji słów lub innych cech językowych, nie zaś sam system językowy, który zasadniczo jest wspólny dla współczesnych sobie autorów. Można by to określić następująco: różni autorzy mają wspólny słownik i gramatykę, ale korzystają z nich w sposób dla siebie charakterystyczny.

O tym, że jest to metoda skuteczna, przekonuje duża liczba tak zwanych kontrolowanych eksperymentów przeprowadzanych na tekstach, co do których autorstwa nie ma wątpliwości: jeśli weźmiemy 10 powieści Tadeusza Dołęgi-Mostowicza i 10 powieści Marii Rodziewiczówny i z nimi porównamy tekst *Braci Dalcz i S-ki*, to okaże się, że jest on w dużo większym stopniu podobny do powieści pierwszego ze wspomnianych pisarzy. Zauważmy jednak, że w praktyce historyka literatury pytanie o anonimowe autorstwo jest rzadkie. Metoda ta znajduje wszakże zastosowanie poza ściśle rozumianą atrybucją, na przykład w przekładoznawstwie, gdzie daje odpowiedź na pytanie, czy autor oryginału przebija się przez tekst tłumaczony (Rybicki, Heydel 2013), w socjologii literatury, gdzie pozwala zobaczyć, czy i do jakiego stopnia język pisarek różni się od języka pisarzy (Weidman, O'Sullivan 2017), w językoznawstwie, gdzie umożliwia kwantytatywne spojrzenie na problem zróżnicowania języków o dużym stopniu pokrewieństwa, takich jak serbski, chorwacki i bośniacki (Waldenfels, Eder 2016) czy angielski w odmianie amerykańskiej, irlandzkiej i brytyjskiej (Jockers 2013), wreszcie w stylistyce językoznawczej (Biber 1988).

Warto podkreślić, że przywołane powyżej studia, odnajdujące grupy podobnych tekstów i separacje między grupami, odwołują się do w gruncie rzeczy tych samych lub bardzo podobnych „cech języka". Są to z jednej strony cechy leksykalne, a więc frekwencje słów, z drugiej zaś strony cechy gramatyczne, takie jak frekwencje różnych części mowy (Hirst, Feiguina 2007)

albo fragmenty drzewek (a więc w istocie podstruktury składniowe, zob. van Cranenburgh 2012, 2016), cechy prozodyczne oraz interpunkcyjne, wreszcie cechy mniej uchwytne, takie jak analiza sentymentu (tj. nasycenia słownictwem nacechowanym emocjonalnie). Problem, o którym tu mowa, polega na tym, że sygnał autorski, chronologiczny, gatunkowy i każdy inny kryje się w tym samym uniwersum „cech języka" i siłą rzeczy wszystkie one wchodzą we wzajemne interferencje. Wynika to po części z tego, że każdy tekst jest tworem wielowarstwowym, na którego ostateczny kształt językowy oraz stylistyczny wpływają nie tylko uniwersalne prawa języka, ale też wymogi gatunku, językowy „smak" epoki, a nawet wiek autora czy przebyte przez niego choroby, jak o tym przekonuje studium na temat zmian stylistycznych w twórczości Agathy Christie (Le i in. 2011). Więcej nawet, ta sama cecha językowa, na przykład nieco częstsze użycie angielskiego rodzajnika *the*, może zarówno odbijać zróżnicowanie odmiany amerykańskiej i brytyjskiej (Jockers 2013), jak i pokazywać różnice między językiem kobiet i mężczyzn (Pennebaker 2011).

Zróżnicowanie niektórych rejestrów wydaje się oczywiste i nie trzeba żadnego aparatu matematycznego czy statystycznego, by tego dowieść: polszczyznę mówioną i pisaną rozpozna każdy chyba użytkownik języka, nawet na małej próbce tekstu. Nieco bardziej wprawnego ucha wymaga rozróżnienie języka rozpraw naukowych od tekstu ustaw, ale nadal jest to zadanie stosunkowo łatwe. Ale czy równie łatwe będzie rozróżnienie języka publicystyki od literatury faktu? Mało tego, prędzej czy później musi pojawić się pytanie, do jakiego stopnia istniejące typologie rejestrów polszczyzny, które za punkt wyjścia obierają raczej funkcjonowanie tekstu we wspólnocie komunikacyjnej, odzwierciedlają rzeczywiste zróżnicowanie formalne tekstów. Oczywiście to zróżnicowanie jest obecne i na poziomie leksykalnym, i gramatycznym, przy czym to pierwsze jest bardziej widoczne. W niniejszej pracy jednak całkowicie abstrahujemy od warstwy leksykalnej, co naturalnie nie znaczy, że jej nie dostrzegamy. Koncentrujemy się na gramatyce właśnie dlatego, że zróżnicowanie pod względem leksykalnym było już przedmiotem badań. Dodajmy zresztą, że kwalifikatory stylistyczne w słownikach stanowią również pewien opis zróżnicowania leksyki poszczególnych rejestrów języka.

W tym miejscu sformułujmy wprost hipotezę badawczą: zakładamy, że typy tekstów wyróżnionych przez R.L. Górskiego i M. Łazińskiego (2012) na potrzeby NKJP różnią się rozkładem ilościowym rozmaitych cech gramatycznych. Jeśli hipoteza jest prawdziwa, będziemy w stanie wymodelować (odnaleźć) różnice między poszczególnymi rejestrami polszczyzny za pomocą takich znaczników stylu, jak dystrybucja poszczególnych przypadków, liczba zaimków osobowych czy stopniowanie przymiotników i przysłówków. W niniejszym studium pod uwagę bierzemy w sumie 60 cech.

Materiał badawczy i metoda

Korpus tekstów

Opisane w niniejszym studium eksperymenty zostały przeprowadzone na zbiorze 1190[3] tekstów pozyskanych losowo z NKJP (Przepiórkowski i in. (red.) 2012). Liczba tekstów przynależnych

3 Nie przywiązujemy wagi do magii okrągłych liczb, zależało nam jedynie na tym, by tekstów było ponad tysiąc, stąd liczba akurat 1190 tekstów jest dość przypadkowa.

do poszczególnych typów jest dość nierówna, między 10 dla mówionych medialnych a 207 dla powieści. Z każdego tekstu wyciągnięto pierwsze 10 tysięcy słów, zaokrąglonych do pełnego zdania, tak że każda próbka jest jednakowej długości.

Za NKJP przyjęliśmy następujące typy tekstów[4], oznaczane w samym korpusie kodami podanymi w nawiasie:

1) literatura piękna (lit), 207 próbek;
2) literatura faktu (fakt), 202 próbki;
3) publicystyka i krótkie wiadomości prasowe (publ), 262 próbek; na ten zbiór składa się 150 tekstów z dzienników (w niniejszym opracowaniu oznaczane jako publDz), 81 tekstów z periodyków wydawanych raz w tygodniu lub rzadziej (publPer) oraz 31 tekstów z prasy lokalnej (publReg);
4) typ naukowo-dydaktyczny (nd), 171 próbek, w tym 55 reprezentujących nauki humanistyczne, 50 społeczne oraz 66 nauki ścisłe, przyrodnicze i techniczne;
5) typ informacyjno-poradnikowy (inf-por), 40 próbek;
6) książka niebeletrystyczna niesklasyfikowana (nklas), 45 próbek;
7) inne teksty pisane: teksty urzędowo-kancelaryjne (urzed), 93 próbki[5];
8) teksty mówione konwersacyjne (konwers), 23 próbki;
9) teksty mówione medialne (media), 10 próbek;
10) teksty quasi-mówione (qmow – w praktyce są to stenogramy parlamentarne), 92 próbki.

W wypadku kategorii 1, 2, 4–6 próbki pochodzą z pojedynczych tekstów, w kategoriach 3 oraz 8–10 próbki losowano z pojedynczego pliku NKJP, który zawierał zazwyczaj większą liczbę tekstów o objętości często znacznie mniejszej niż 10 tysięcy słów, tak więc na pojedynczą próbkę składa się szereg artykułów (pochodzących wszakże z jednego tytułu prasowego), wypowiedzi bądź aktów prawnych. W wypadku publicystyki na próbkę składały się artykuły pochodzące z jednego tytułu prasowego. Kryteria podziału i procedura przypisywania tekstu do poszczególnych typów zostały szczegółowo omówione w pracy R.L. Górskiego i M. Łazińskiego (2012).

Na potrzeby niniejszego studium wprowadziliśmy podział prasy na trzy kategorie: codzienną, wydawaną rzadziej, a także lokalną. Podział na prasę codzienną i pozostałą prasę był niejawnie przyjęty w NKJP – mimo że oba typy tekstów należały do jednej kategorii, proporcje pomiędzy nimi były z góry zadane i odzwierciedlały proporcje ich czytelnictwa. Nie dbano natomiast w trakcie tworzenia korpusu o zrównoważenie między prasą lokalną a tą o zasięgu ogólnopolskim, niemniej z metadanych źródeł NKJP da się wyczytać, które teksty przynależą do tych dwu kategorii. Również książkę publicystyczną uznaliśmy za osobny typ tekstu, mimo że w NKJP nie był on wyróżniony. Znów jednak daje się łatwo wyróżnić taką

4 NKJP wyróżnia ponadto dwa rodzaje tekstów internetowych, tj. interaktywne strony WWW (fora, czaty, listy dyskusyjne itp.) i statyczne strony WWW. Nie uwzględniliśmy tego typu tekstów, jako że jest on zupełnie nowy i pomijany w dotychczasowej literaturze. Ponadto obawialiśmy się, że statyczne strony WWW są kategorią mocno heterogeniczną i jako taka zaburzy ogólny obraz prezentowanych wyników.

5 Uwzględniliśmy jedynie ten podtyp kategorii „Inne teksty pisane", ponieważ pozostałe teksty są zarówno nieliczne, jak i – co ważniejsze – bardzo krótkie.

kategorię jako część wspólną publicystyki (typ tekstu) i książki (kanał tekstu); kategorię tę reprezentowały 32 próbki.

W klasyfikacji tematycznej Biblioteki Narodowej pojawia się kategoria „szkice" (Górski, Łaziński 2012), w istocie dość bliska temu, co zwykliśmy nazywać esejem, choć z pewnością nie tożsama. Z reguły – z punktu widzenia typologii NKJP – są to „książki niebeletrystyczne niesklasyfikowane". Zdecydowaliśmy się wprowadzić tę nieistniejącą w NKJP kategorię szkiców, którą reprezentuje 13 próbek, do naszych badań, gdyż chcieliśmy się przekonać, czy rzeczywiście stanowi ona stylistycznie odrębną klasę. Zaznaczamy wszakże, że w tym wypadku kierowaliśmy się decyzjami bibliografów z BN. Oczywiście mieszamy tutaj ugruntowaną typologię językoznawczą z typologią innego rodzaju, raczej utylitarną niż opartą na teoretycznych roztrząsaniach, o czym trzeba pamiętać, gdy przejdziemy do referowania wyników naszych badań.

Uważny czytelnik dostrzeże znaczne różnice między liczbą próbek reprezentujących daną kategorię, co może mieć pewien wpływ na nasze wyniki. Należy jednak wziąć pod uwagę, że dostępność poszczególnych typów tekstów jest zróżnicowana. Gdyby każdy typ miał być reprezentowany przez jednakową liczbę próbek, ich łączna liczba musiałaby być znacząco mniejsza, bo dostosowana do najbardziej marginalnego typu.

Cechy językowe (znaczniki stylu)

Wybór odpowiednich znaczników stylu do analizy nie jest rzeczą prostą. W tym miejscu należy jednak przeprowadzić bardzo ważne rozróżnienie. Otóż kategoryzacja tekstu, a więc automatyczne przypisanie tekstu do jakiejś z góry zadanej kategorii, jest dość rutynowym zadaniem w lingwistyce komputerowej, tym bardziej że ma ona duże znaczenie praktyczne. Z inżynierskiego punktu widzenia istotna jest tylko skuteczność (dokładność) klasyfikacji, same zaś cechy, które miałyby różnicować teksty, są o tyle tylko istotne, o ile pozwalają lepiej wykonać zadanie. Dla językoznawcy z kolei jest to zagadnienie ciekawe poznawczo z punktu widzenia ogólnej wiedzy o tym, jak działa język, dlatego też równie ważne jest to, co rejestry języka różnicuje, jak i to, co ich n i e r ó ż n i c u j e.

My przyjęliśmy stanowisko pośrednie: wybraliśmy pewną liczbę cech podejrzewanych o to, że są charakterystyczne dla określonych typów tekstów, na przykład imiesłowy bierne uzupełnione o agensa (*pisany przez autora*), rozbudowane frazy nominalne, zdania podrzędne. Ponadto posłużyliśmy się wszystkimi cechami, których frekwencja daje się bardzo łatwo pozyskać automatycznie z korpusu. Były to przede wszystkim kategorie gramatyczne (części mowy w rozumieniu tagsetu NKJP (por. Woliński 2003), a więc: rzeczownik, czasownik w czasie nieprzeszłym, imiesłów przysłówkowy uprzedni itp.), a także cechy fleksyjne (przypadek, osoba, liczba, stopień przymiotnika), wreszcie cecha czysto ortograficzna, jaką jest przecinek – zauważmy jednak, że przecinki również stanowią pewne (oczywiście odległe) przybliżenie składni. Intuicyjnie należy się spodziewać, że niektóre z tych cech nie mogą w żaden sposób różnicować tekstów, gdyż ich rozkład jest dość równomierny i niezależny od stylu (np. przypadek), niemniej nie ma szczególnego powodu, by je z góry odrzucać. Jednocześnie nie potraktowaliśmy jako cechy różnicującej na przykład strony biernej, choć jej frekwencja (niska bądź wysoka) jest ewidentnie wyróżnikiem stylu (Górski 2008), gdyż nie są to dane,

które można by pozyskać automatycznie. Uwzględniliśmy natomiast taką cechę, jaką jest użycie imiesłowu biernego przed frazą przyimkową z *przez*, ponieważ da się ona wyszukiwać automatycznie z dużą dokładnością. Wyszukiwanie w korpusie jest bowiem zawsze pewnego rodzaju kompromisem między zwrotem (*recall*), a więc pozyskaniem możliwie wielu rzeczywistych przykładów szukanego zjawiska, i precyzją (*precision*), czyli pozyskaniem minimalnej liczby fałszywych przykładów. I tak jeśli zdefiniujemy wzorzec wyszukiwania czasu przyszłego jako *być* w czasie przyszłym i czasownik w czasie przeszłym, oddzielone nie więcej niż 5 segmentami (w składni wyszukiwarki Poliqarp [pos=bedzie][]{0,5}[pos=praet]), to korpus zwróci zarówno rzeczywisty przykład czasu przyszłego:

(1) *będę nie na swój własny obraz tworzył*,

jak i ciąg, który z pewnością nie reprezentuje tej formy fleksyjnej:

(2) *będzie Donald Tusk i jak bardzo wysilił* (co stanowi część ciągu *z jakich armat strzelać będzie Donald Tusk i jak bardzo wysilił się* PSL).

Jeśli zmniejszymy dystans pomiędzy *być* a formą czasu przeszłego, to wyeliminujemy szereg fałszywych przykładów, czyli ciągów, które nie są szukaną formą, takich jak (2), a więc zwiększymy precyzję, ale jednocześnie wyeliminujemy rzeczywiste przykłady czasu przyszłego, takie jak (1), czyli zmniejszymy zwrot.

I jeszcze jedna bardzo istotna uwaga: pewne cechy są ze sobą powiązane, na przykład suma frekwencji przypadków jest zarazem sumą frekwencji odpowiednich części mowy, które odmieniają się przez przypadki[6].

W tym miejscu przypomnijmy pojęcie segmentu wprowadzone w Korpusie Instytutu Podstaw Informatyki Polskiej Akademii Nauk i stosowane w NKJP. Segmentem jest każde słowo ortograficzne, znak interpunkcyjny, a także wykładnik czasu przeszłego i partykuła *by* bez względu na to, czy jest pisana łącznie, czy rozłącznie. Każdy segment jest scharakteryzowany pod względem gramatycznym. I tak na przykład segment *celem* charakteryzują 4 kategorie fleksyjne: część mowy, liczba, przypadek i rodzaj, które przybierają wartości: rzeczownik (oznaczany w korpusie jako *subst*), liczba pojedyncza (*sg*), narzędnik (*inst*) i rodzaj męski rzeczowy (*m3*). Z technicznego punktu widzenia istotne dla nas jest to, że można dzięki temu wyszukać wszystkie segmenty, w których na przykład kategoria przypadka (*cas*) przyjmuje wartość *inst*.

W analizie wzięliśmy pod uwagę następujące kategorie gramatyczne, zakładając, że niektóre z nich mogą się okazać cechami różnicującymi style:

1) liczba segmentów przynależnych do danej części mowy (w rozumieniu tagsetu NKJP); przypomnijmy, że z punktu widzenia tagsetu NKJP każda forma, której przysługuje niepowtarzalny zestaw kategorii fleksyjnych, stanowi odrębną część mowy, nie ma więc czasowników, są

6 Przypomnijmy, że kategoria przypadka jest w NKJP przypisywana również przyimkom.

tylko czasowniki przeszłe, nieprzeszłe, imiesłowy przymiotnikowe czynne itd. – każda z tych kategorii jest traktowana jako odrębna część mowy;

2) liczba segmentów występujących w danym przypadku;

3) przymiotniki: wszystkie przymiotniki ([pos=adj]), a także osobno: przymiotniki w stopniu wyższym ([pos=adj&deg=com]), najwyższym ([pos=adj&deg=sup]) syntetycznym, wyższym i najwyższym opisowym, definiowanym jako sekwencja segmentów [base=bardzo&deg=com] [pos=adj], bądź [base=bardzo&deg=sup] [pos=adj], przymiotniki przyprzymiotnikowe (np. *biało* w *biało-czerwony* [pos=adja]), przymiotniki poprzyimkowe (*polsku* w *po polsku* [pos=adjp]), a także przymiotniki w postpozycji (definiowane jako sekwencja rzeczownik–przymiotnik [pos=subst][pos=adj]);

4) czas przyszły niedokonany (złożony) oparty na bezokoliczniku (definiowany jako sekwencja [pos=bedzie][pos=inf]) i na formie przeszłej ([pos=bedzie][pos=praet]); czas przyszły prosty definiowany jako kombinacja kategorii [pos=fin&aspect=perf];

5) kategoria osoby. Są to: wszystkie segmenty zawierające kategorię *person*, która przybiera wartość *pri* lub *sec*, a więc formy nieprzeszłe czasownika, aglutynanty, zaimki pierwszej i drugiej osoby, oraz osobno zaimki pierwszej i drugiej osoby w mianowniku. Te dwie ostatnie cechy wskazują na to, że mamy do czynienia z tekstem „osobistym", gdzie nadawca jawnie i jednocześnie redundantnie (a nie tylko poprzez fleksję czasownika) wskazuje na siebie i odbiorcę;

6) stopień rozbudowania zdania: frekwencja spójników (*aby*, *albowiem*, *bowiem*, *by*, *iż*, *który*, *że* i *żeby*), przecinków, a także liczba rozbudowanych fraz nominalnych definiowanych jako nieprzerwane sekwencje segmentów z kategorią *pos* o wartości *subst*, wreszcie liczba sekwencji imiesłowów biernych i wyrazu *przez* ([pos=ppas][base=przez]), która jest dobrym przybliżeniem agentywnego passivum (np. *pity przez*). Podobnie frekwencję wyrazu *który* traktujemy przede wszystkim jako dość odległe przybliżenie liczby zdań względnych.

Jeszcze raz z całym naciskiem chcemy podkreślić, że nie ograniczamy się do cech, co do których można się spodziewać – na podstawie już to intuicji, już to dotychczasowych badań – że ich frekwencja jest uzależniona od typu tekstu. I jeszcze jedno: bardzo ważkie pytanie o to, które cechy istotnie różnicują typy tekstów przekracza ramy niniejszego artykułu.

Statystyczne metody wielowymiarowe

W sytuacji, gdy badacz staje przed koniecznością zmierzenia frekwencji wielu cech jednocześnie (np. frekwencji wielu słów albo frekwencji wielu *n*-gramów kategorii gramatycznych), najskuteczniejsze okazują się tak zwane metody wielowymiarowe, które, posługując się matematycznymi operacjami, tworzą przestrzeń o kilkudziesięciu czy nawet kilkuset wymiarach. Tego typu przestrzeń jest bytem czysto abstrakcyjnym, pozwala jednak w bardzo wygodny sposób mierzyć podobieństwa i różnice między analizowanymi danymi liczbowymi. Dzieje się tak dlatego, że w dowolnej przestrzeni geometrycznej można się odwołać do pojęcia *odległości*, które z kolei można w sposób bardzo intuicyjny przekładać na pojęcie *podobieństwa*. W rzeczywistości korpusowej teksty przyjmują postać szeregu liczb (frekwencji słów), za pomocą których można obliczyć wzajemną odległość między tekstami. Im większa geometryczna odległość między tekstami, tym większa różnica stylistyczna (czy, mówiąc szerzej, różnica językowa).

Do najczęściej używanych metod wielowymiarowych należą hierarchiczna analiza skupień (*hierarchical cluster analysis*), analiza głównych składowych (*principal components analysis*) czy skalowanie wielowymiarowe (*multidimensional scaling*)[7]. Pomijamy tutaj cały szereg wyrafinowanych metod tak zwanego nadzorowanego uczenia maszynowego (poświęcimy im osobne studium).

Stosunkowo spora liczba analizowanych tekstów sprawia, że klasyczna metoda wizualizacji danych zwana hierarchiczną analizą skupień może się okazać nie dość czytelna. Z tego powodu użyta została rozszerzona wersja metody, która analizę skupień łączy z teorią grafów i relacje między badanymi tekstami pokazuje w postaci sieci wzajemnych podobieństw stylistycznych (Eder 2014, 2017). Analizy zostały przeprowadzone za pomocą stworzonego specjalnie do przeprowadzenia niniejszych badań skryptu, który został napisany w języku R.

Rezultaty

Omówienie rezultatów zaczniemy od spojrzenia „z lotu ptaka" na cały zbiór. Sieć 1190 tekstów pogrupowanych na podstawie frekwencji 60 cech językowych pokazano na rysunku 1.

Rysunek 1. Sieć podobieństw między tekstami[8]

7 Matematyczne i konceptualne założenia zarówno skalowania wielowymiarowego, jak i analizy skupień zostały drobiazgowo wyjaśnione w klasycznych podręcznikach z zagadnień klasyfikacji i uczenia maszynowego; stosunkowo „łagodne" wprowadzenie do tematyki – zogniskowane zarazem na zastosowaniach *stricte* językoznawczych – daje np. R. Harald Baayen (2008: 118–160) w podręczniku do językoznawstwa kwantytatywnego, w rozdziale na temat klasyfikacji.

8 Autorzy zamieścili kolorowe, łatwiejsze w interpretacji wersje wykresów z niniejszej pracy w Internecie (https://github.com/computationalstylistics/typologia). Tam też znajdują się wszystkie dane liczbowe użyte w niniejszym studium oraz skrypt do wygenerowania wykresów.

Już pierwszy rzut oka na wykres pozwala sformułować kilka wstępnych wniosków. Po pierwsze zatem: grupy tekstów należące do tego samego typu mają tendencję do układania się blisko siebie (sąsiedztwo na sieci oznacza bliskość stylistyczną/językową danych próbek), co potwierdza stylistyczną jednorodność poszczególnych stylów funkcjonalnych. Wyraźnie wyodrębnić można jednorodne skupiska tekstów. I tak teksty naukowo-dydaktyczne (głęboko szare punkty, stanowiące lewą stronę centralnego skupiska), literatura faktu (sam jego środek) i teksty literackie (jasnoszare na lewo) tworzą pewne *continuum*, a wokół nich gromadzą się prasa codzienna (wyodrębnione skupisko na dole wykresu), prasa regionalna (bladoszare „warkocze" wyraźnie odcinające się na lewo), eseistyka i publicystyka książkowa (czarne punkty rozsiane w środkowej części). Interesujące jest rozbicie literatury faktu na dwie grupy: większość skupia się pośrodku, część jednak jest wyraźnie odseparowana od reszty, tworząc długi „warkocz" skierowany ku lewej górnej części wykresu. Wprawdzie położenie skupisk na wykresie (góra czy dół, lewo czy prawo) nie niesie żadnej informacji (algorytm nie „rozróżnia" kierunków), ale stopień ich separacji już tak. Ciekawe są zatem grupy tekstów, które wyraźnie oddzielają się od głównej części sieci; nie mniej interesujące są jednak i te zbiory, w których n i e n a s t ę p u j e wyraźna separacja pod względem gatunku.

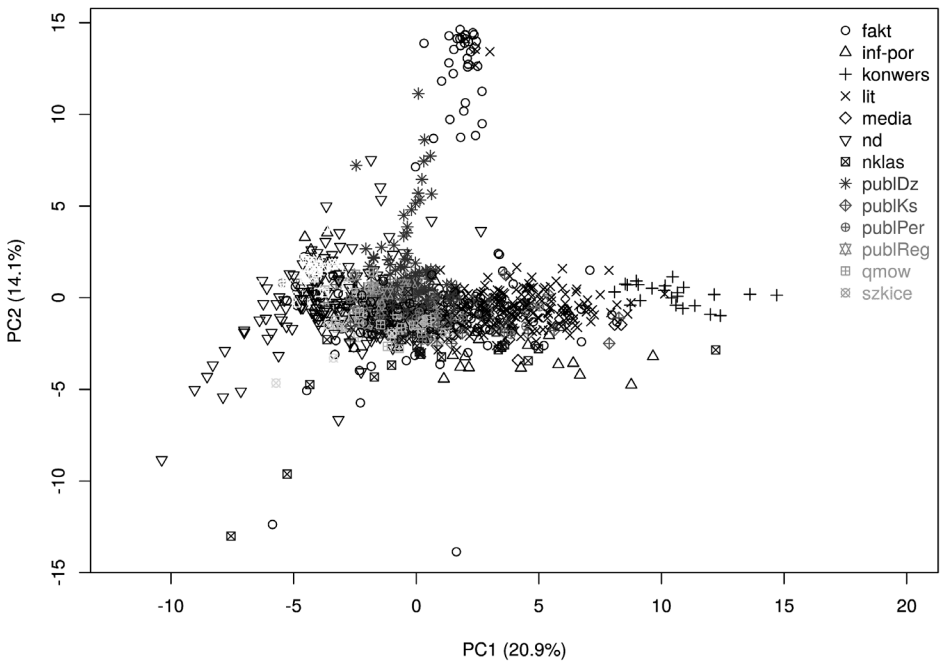
Po drugie, niektóre ze skupisk wykazują wyraźną tendencję odśrodkową: teksty prawnicze (po lewej), część literatury faktu (punkty na górze sieci), przemówienia sejmowe (odrębne skupiska punktów pośrodku góry wykresu), podczas gdy inne skupiają się w centralnej części sieci. Widać wreszcie, po trzecie, że w centrum sieci pojawia się spora liczba tekstów niegrupujących się w wyraźne skupiska. Są to opisane wyżej szczegółowo „szkice".

Warto też zauważyć, że punkty symbolizujące teksty naukowo-dydaktyczne tworzą co prawda gęstą konstelację, ale jednocześnie znaczna ich część jest rozproszona wśród innych typów tekstów. Sugeruje to, że tylko część tekstów, które referują odkrycia naukowe, nosi wszystkie wykładniki tego stylu i należy przypuszczać, że są to teksty z zakresu nauk ścisłych i społecznych.

Trudny do zinterpretowania, ale nie mniej przez to ciekawy, jest łańcuszek tekstów prasowych, który wychodzi z głównego skupiska prasy (na dole sieci) i zmierza w kierunku małej galaktyki (u góry) zawierającej głównie literaturę faktu. Jedną z najciekawszych jednak obserwacji jest stopniowe przejście od tekstów prawniczych (po lewej) przez dyskurs naukowy, potem literaturę faktu, literaturę piękną aż do polszczyzny mówionej (po prawej). Być może głównym czynnikiem odpowiedzialnym za owo stopniowe przejście między kolejnymi podzbiorami jest frekwencja form czasownikowych 1 os. oraz 2 os. (bardzo częstych w języku mówionym, w ogóle nieobecnych w tekstach prawniczych), choć zapewne i inne czynniki miały tu znaczenie.

W analizach wielowymiarowych chodzi przede wszystkim o to, by odnaleźć separację między tekstami lub grupami tekstów. Im ostrzejsze są granice między czytelnie zarysowanymi zbiorami, tym mocniejszy dowód, że cechy językowe użyte w bieżącej analizie są mocnymi predyktorami odpowiedzialnymi za zróżnicowanie tekstów.

Rysunek 2 przedstawia wykres analizy głównych składowych. Próbki grupują się może nieco mniej wyraźnie niż na przedstawionej powyżej sieci (rys. 1), ale mimo to i tutaj widać jednolite obszary gatunkowe – i są to obszary niemal takie same jak powyżej.



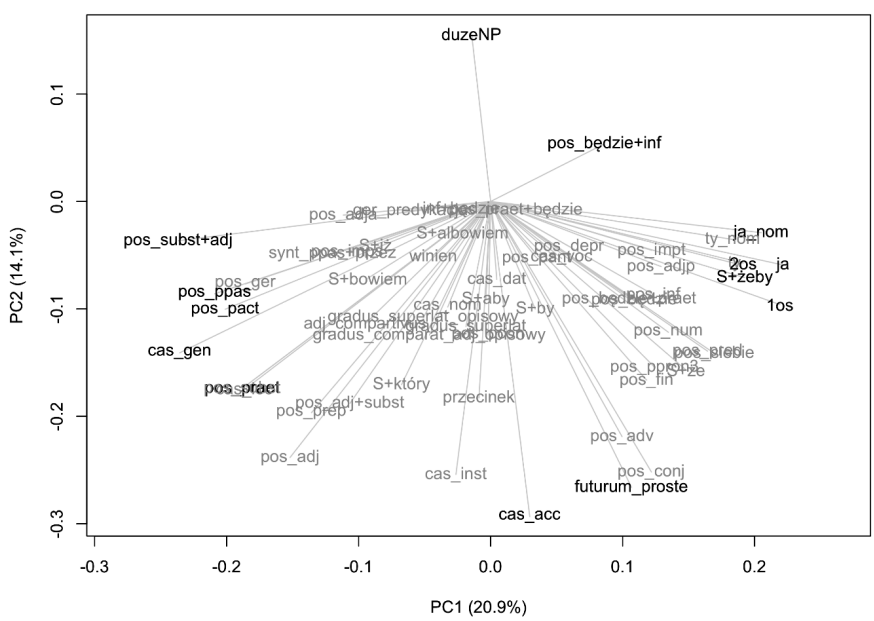
Rysunek 2. Analiza głównych czynników

Analiza głównych składowych została tak skonstruowana, by najważniejsza separacja ułożyła się horyzontalnie, czyli wzdłuż pierwszej głównej składowej (czytelnika zainteresowanego matematycznymi założeniami metody odsyłamy do wspomnianych wyżej podręczników), od tego zatem należy zacząć omawianie wyników. Otóż widzimy tutaj niemal dokładnie taki sam obraz: idąc od lewej, mamy prawodawstwo, naukę (rozmytą), literaturę piękną zmieszaną z gatunkami prasowymi i z mowami parlamentarnymi (nieco u dołu), dalej literaturę faktu i wreszcie teksty mówione medialne.

W tym miejscu językoznawca zadaje sobie pytanie, które cechy najbardziej różnicują typy tekstów. Poczyniliśmy założenie, że za separację tekstów na czytelne podgrupy odpowiada wprawdzie wiele cech jednocześnie, ale przecież niektóre z nich są silniejszymi predyktorami niż inne. Na przykład intuicja podpowiada, że różnica między językiem mówionym i pisanym będzie w większym stopniu zasadzała się na wyznacznikach składniowych niż, powiedzmy, na częstości użycia narzędnika.

Rysunek 3 przedstawia siłę dyskryminacyjną użytych w badaniu zmiennych (cech językowych). Wykres ów należy czytać paralelnie z rysunkiem 2, na którym zostały pokazane

właściwe próbki tekstowe, rozsunięte od środka wykresu tak mocno, jak mocno ujawnia się w nich działanie poszczególnych cech językowych zaznaczonych na rysunku 3. Jak widać, owe cechy mają nie tylko różną siłę dyskryminacyjną (najsilniejsze zaznaczono kolorem czarnym), ale też różny kierunek. I tak można z niego wyczytać, że zaimki pierwszej osoby („ja"), zaimki pierwszej osoby w mianowniku („ja_nom"), pierwsza osoba zaimka i czasownika („1os"), druga osoba zaimka i czasownika („2os"), zaimek drugoosobowy w mianowniku („ty_nom") czy zdania podrzędne ze spójnikiem *żeby* („S_zeby") będą występowały mniej więcej w podobnym natężeniu w różnych tekstach i na ogół ich duże nasycenie będzie szło w parze z niskim nasyceniem sekwencji fraz nominalnych z przymiotnikiem w postpozycji („pos_subst.adj"), imiesłowów przymiotnikowych biernych („pos_ppas"), imiesłowów przysłówkowych czynnych („pos_pact") czy dopełniaczy („cas_gen"), czy też rzeczownikami („pos_subst"). Zauważmy, że o ile pierwszy zestaw cech (tj. nasycenie zaimkami, w tym redundantnymi zaimkami w mianowniku, zdania ze spójnikiem *żeby*) na rysunku 3 lokuje się tam, gdzie na rysunku 2 znalazły się teksty mówione konwersacyjne, o tyle drugi zestaw (imiesłowy, dopełniacze, przymiotniki w postpozycji) znajduje się tam, gdzie część tekstów naukowych, czyli odpowiednio po prawej i lewej stronie. Są to cechy raczej przewidywalne, mniej natomiast oczywiste jest powiązanie literatury faktu z rozbudowanymi frazami nominalnymi (co prawda – przypomnijmy – rozumianymi dość wąsko, jako nieprzerwane ciągi co najmniej czterech rzeczowników – „duże_NP"), a zwłaszcza z czasem przyszłym złożonym opartym na bezokoliczniku („pos_bedzie.inf").



Rysunek 3. Analiza głównych czynników (ładunki)

Podsumowanie

W niniejszym artykule staraliśmy się pokazać, że typ tekstu nakłada autorowi dość sztywny gorset stylistyczny, któremu to gorsetowi większość autorów się poddaje. Stwierdzenie to jest i oczywiste, i zgodne z dotychczasowymi ustaleniami. To, co nowe w naszych badaniach, to przede wszystkim ich skala. Daje nam ona pewność, że większość tekstów lokuje się wśród innych przedstawicieli swojego typu; pozwala też stwierdzić, że istnieją nieliczne teksty, które nie wpasowują się w szablon. Jest to więc głos w dyskusji, czy wyznaczniki gatunkowe są pewną idealizacją tworzoną przez badaczy, ale nie ma czystych gatunkowo tekstów, czy też, przeciwnie, owe wyznaczniki rzeczywiście istnieją. Nasze badania wzmacniają raczej tę drugą tezę.

Dodajmy w tym miejscu jeszcze jedno bardzo ważne stwierdzenie: niektóre z cech wynikają nie tyle z dostosowania się do wzorców gatunkowych, ile zwyczajnie z tematyki lub konsytuacji tekstu – tak jest niewątpliwie w wypadku pierwszej i drugiej osoby bądź w konwersacjach czy też w wypadku czasu gramatycznego.

Wreszcie możemy stwierdzić, które typy są do siebie stosunkowo bardziej podobne. I tak na przykład teksty prawne są bardzo podobne do pewnej części tekstów naukowo-dydaktycznych. Z kolei teksty informacyjno-poradnikowe są od tych ostatnich całkowicie odrębne.

Na koniec wszakże podkreślmy raz jeszcze, że nasze wyniki opierają się na cechach policzalnych i powtarzalnych, te zaś pozwalają z dość dużą precyzją wskazać, które teksty i w jakiej liczbie wykazują podobieństwo do pozostałych przedstawicieli swojego typu.

Bibliografia

Baayen H. 2008: *Analysing linguistic data: a practical introduction to statistics using R*, Cambridge University Press, Cambridge.

Biber D. 1988: *Variation across speech and writing*, Cambridge University Press, Cambridge.

Cranenburgh van A. 2012: *Literary authorship attribution with phrase-structure fragments*, [w:] *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, June, 2012, Montréal, Canada, Association for Computational Linguistics, s. 59–63 (online: https://www.aclweb.org/anthology/W12-2508, dostęp: 1 czerwca 2019).

Cranenburgh van A. 2016: *Rich statistical parsing and literary language*, University of Amsterdam, Amsterdam.

Eder M. 2014: *Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii*, „Teksty Drugie", nr 2, s. 90–105.

Eder M. 2017: *Visualization in stylometry: cluster analysis using networks*, „Digital Scholarship in the Humanities", Vol. 32(1), s. 50–64 (online: https://doi.org/10.1093/llc/fqv061, dostęp: 1 czerwca 2019).

Gajda S. (red.) 1995: *Przewodnik po stylistyce polskiej*, Uniwersytet Opolski, Instytut Filologii Polskiej, Opole.

Górski R.L. 2008: *Diateza nacechowana w polszczyźnie. Studium korpusowe*, Lexis, Kraków.

Górski R.L., Łaziński M. 2010: Wzór stylu i wzór na styl. *Zróżnicowanie stylistyczne tekstów Narodowego Korpusu Języka Polskiego*, [w:] M. Milewska-Stawiany, E. Rogowska-Cybulska (red.), *Polskie języki. O językach zawodowych i środowiskowych. Materiały VII Forum Kultury Słowa*, Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk, s. 41–55.

Górski R.L., Łaziński M. 2012: *Typologia tekstów w NKJP*, [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 13–23.

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

Jockers M.L. 2013: *Macroanalysis: digital methods and literary history*, University of Illinois Press, Urbana.

Le X., Lancashire I., Hirst G., 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", Vol. 26(4), s. 435–461.

Malinowska E., Nocoń J., Żydek-Bednarczuk U. (red.) 2013: *Style współczesnej polszczyzny. Przewodnik po stylistyce polskiej*, Towarzystwo Autorów i Wydawców Prac Naukowych „Universitas", Kraków.

Mikołajczak S. 1990: *Składnia tekstów naukowych. Dyscypliny humanistyczne*, Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań.

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

Przepiórkowski A., Bańko M., Górski R.L., Lewandowska-Tomaszczyk B. (red.) 2012: *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 13–23.

Ruszkowski M. 1997: *Główne tendencje syntaktyczne w polskiej prozie artystycznej dwudziestolecia międzywojennego*, Wyższa Szkoła Pedagogiczna im. Jana Kochanowskiego w Kielcach, Kielce.

Rybicki J., Heydel M. 2013:. *The stylistics and stylometry of collaborative translation: Woolf's 'Night and Day' in Polish*, „Literary and Linguistic Computing", Vol. 28(4), s. 708–717.

Waldenfels von R., Eder M. 2016: *A stylometric approach to the study of differences between standard variants of Bosnian/Croatian/Serbian, or: is the Hobbit in Serbian more Hobbit or more Serbian?*, „Russian Linguistics", Vol. 40(1), s. 11–31.

Weidman S.G., O'Sullivan J. 2017: *The limits of distinctive words: Re-evaluating literature's gender marker debate*, „Digital Scholarship in the Humanities", Vol. 33(2), s. 374–390.

Woliński M. 2003: *System znaczników morfosyntaktycznych w korpusie IPI PAN*, „Polonica" XXII–XXIII, s. 39–55.

Summary

Polish text types in a quantitative approach: a corpus based study on diversity of Polish

Keywords: stylistics, text typology, corpus linguistics, multivariate methods.

The article seeks to answer the question whether it is possible to establish a typology of Polish texts based exclusively on their grammatical features. An additional aim was to find whether the typology adopted in the National Corpus of Polish (NCP), based on purely extra-linguistic criteria, groups together texts that are grammatically similar.

The study was conducted on a corpus of 1190 texts randomly chosen from the NCP. For each text the frequency of some 60 grammatical features was counted, such as the number words belonging to a part of speech, occurring in a particular case, person or tense etc. With these data Bootstrap Consensus Network analysis as well as multidimensional scaling was conducted. The results show that most members of a text type cluster together showing similarity one to another. Moreover, the typology of texts adopted in the NCP gains additional support.