MIROSŁAW BAŃKO* | UNIWERSYTET WARSZAWSKI

Badanie dystrybucji stylistycznej za pomocą wyszukiwarki PELCRA w Narodowym Korpusie Języka Polskiego

Słowa kluczowe: dystrybucja stylistyczna, Narodowy Korpus Języka Polskiego, wyszukiwarka PELCRA.

doi: https://doi.org/10.31286/JP.00968

Liczba narzędzi cyfrowych do przetwarzania języka polskiego stale rośnie (zob. ich wykaz na stronach CLIP[1]; Ogrodniczuk, Przepiórkowski 2013), ale wciąż do najczęściej wykorzystywanych należą wyszukiwarki Narodowego Korpusu Języka Polskiego (NKJP): wyszukiwarka PELCRA[2] (Pęzik 2012), wyszukiwarka Instytutu Podstaw Informatyki PAN, znana też pod nazwą Poliqarp[3] (Przepiórkowski 2004), i dwie mniej popularne: wyszukiwarka MTAS[4] (Kieraś i in. 2021) oraz późniejsza wersja wyszukiwarki PELCRA (nazwiemy ją tu PELCRA 2[5]), oferująca nowe funkcjonalności. Z nielicznymi wyjątkami różnice między tymi narzędziami nie będą tu przedmiotem uwagi.

Celem artykułu jest opis działania funkcji Profil wbudowanej w wyszukiwarkę PELCRA, ze szczególnym uwzględnieniem słabo udokumentowanych cech, pominiętych lub mylnie opisanych w plikach pomocy dostępnych online, w artykule na temat PELCR-y (Pęzik 2012) i w publikacjach poświęconych NKJP (Przepiórkowski i in. 2012; Bańko, Górski 2014). Ponieważ funkcja Profil służy do badania dystrybucji stylistycznej wyrazu, tzn. jego częstości w tekstach różnego typu, w artykule odnoszę się do pytania, czym są profile stylistyczne generowane przez PELCR-ę, a więc jak rozumieć zawarte w nich informacje. Bodźcem do zajęcia się tą kwestią były błędy w interpretacji profili stylistycznych popełniane przez studentów w pracach dyplomowych z zakresu porównawczej analizy synonimów. W artykule wyjaśniam, jak można porównać dystrybucję stylistyczną wyrazów synonimicznych. Ponadto pokazuję na przykładach, że podstawową wiedzę o dystrybucji wyrazu – nie tylko stylistycznej, ale i dziedzinowej – może dać kwerenda w internetowym katalogu bibliotecznym.

* m.banko@uw.edu.pl; ORCID: 0000-0002-5396-4327

1 http://clip.ipipan.waw.pl/ (dostęp: 8 października 2023).

2 http://www.nkjp.uni.lodz.pl/ (dostęp: 8 października 2023).

3 http://www.nkjp.pl/poliqarp/ (dostęp: 8 października 2023).

4 https://nkjp.nlp.ipipan.waw.pl/query_corpus/ (dostęp: 8 października 2023).

5 http://pelcra.clarin-pl.eu/NKJP/ (dostęp: 8 października 2023).

Pomysł, by w publikacji naukowej zająć się narzędziem znanym i używanym od kilkunastu lat, może się wydawać dziwaczny, uzasadniają go jednak luki, niejasności i nieścisłości w jego opisach i w interfejsie, a także łatwość, z jaką można wpaść w pułapkę mylnej interpretacji wyników. Niniejszym artykułem chcę pomóc początkującym użytkownikom funkcji Profil oraz włączyć się w dyskusję nad przyszłym kształtem NKJP i oprogramowaniem przydatnym do jego obsługi.

Choć funkcja Profil działa zarówno w pełnym NKJP (ok. 1,5 miliarda słów), jak i w podkorpusie zrównoważonym (ok. 240 milionów słów), przykłady omawiane niżej dotyczą tylko podkorpusu zrównoważonego, ponieważ kwerendy wykonane w nim dają bardziej miarodajne wyniki.

Jak działa funkcja Profil

Funkcja Profil informuje o częstości występowania słowa lub innego szukanego obiektu[6] w poszczególnych rodzajach tekstów, opiera się więc na klasyfikacji tekstów w NKJP. Ponieważ klasyfikacja ta służy także innym operacjom w korpusie – np. konkordancje i kolokacje można ograniczyć do wybranego typu tekstów – jej podstawy omówiono w publikacji dotyczącej NKJP (Górski, Łaziński 2012). Ukazana tam klasyfikacja nie w pełni jednak pokrywa się z wdrożoną w korpusie, np. jako odrębny typ ujmuje listy, choć kategorii takiej nie uwzględnia wyszukiwarka PELCRA (uwzględnia natomiast PELCRA 2). Przydatne są w wymienionej publikacji wyjaśnienia, jakie teksty należą do poszczególnych typów, np. co kryje się za nazwą *qmow* lub *nd* (w artykule na temat PELCR-y, zob. Pęzik 2012, a także w pomocy online można znaleźć tylko rozwiązania skrótów). Cenna jest też informacja, że kategoria *lit* (literatura piękna) nie jest sumą kategorii *lit_poezja*, *lit_proza* i *lit_dramat*, gdyż tylko niektóre utwory prozatorskie weszły do kategorii *lit_proza*, pozostałe zaś utworzyły kategorię *lit*. Dzięki funkcji Profil, która podaje liczbę słów w poszczególnych typach tekstów, można się upewnić, że cztery kategorie literackie wymienione wyżej są rzeczywiście rozłączne. Ale dlaczego teksty pisane prozą znalazły się w dwóch zbiorach, z których jeden ma mylącą nazwę – tego autorzy czytelnikom nie wyjaśnili.

Funkcja Profil nie informuje wyłącznie, ile razy szukane słowo występuje w tekstach różnego typu. Takie dane byłyby niemiarodajne, gdyż udział poszczególnych typów tekstów w NKJP jest nierówny, np. największy segment stanowi publicystyka (prawie 130 mln słów), a najmniejszy – teksty dramaturgiczne (niespełna 49 tys. słów)[7]. Aby zneutralizować efekt nierównej struktury korpusu, funkcja Profil oblicza frekwencję słowa względem wielkości każdego segmentu, czyli podaje ją w przeliczeniu na milion słów w danym typie. Statystyki dla kwerendy *zabójstwo*** (dwie gwiazdki gwarantują uwzględnienie wszystkich form odmiany wyrazu) podano w tabeli 1. Frekwencję w przeliczeniu na milion słów określa ostatnia kolumna.

6 Wyszukiwane mogą być słowa (określone w całości lub w części), sekwencje słów (także nieciągłe i o dowolnym szyku), leksemy, jak również kombinacje tych elementów. W niniejszym artykule dla uproszczenia będę pisał o wyszukiwaniu słów, mając na myśli różne obiekty, przeważnie leksemy.

7 Architektura NKJP jest efektem kompromisu między dostępnością tekstów a próbą pogodzenia dwóch innych kryteriów, zob. szczegóły w artykule: Górski, Łaziński 2012.

Tabela 1. Statystyki dotyczące leksemu *zabójstwo*

#	Typ	Liczba wystąpień (A)	Słowa w kategorii (B)	A/(B/1M)
1	publ	4,493	129,475,805	34.701
2	lit	666	30,230,684	22.031
3	fakt	398	14,874,971	26.756
4	qmow	203	23,323,854	8.704
5	net_interakt	149	9,218,718	16.163
6	nd	80	8,680,759	9.216
7	lit_proza	70	4,512,240	15.513
8	nklas	58	2,245,020	25.835
9	net_nieinterakt	45	3,717,246	12.106
10	inf_por	31	8,078,242	3.837
11	urzed	9	3,450,662	2.608
12	konwers	2	1,774,144	1.127
13	lit_dramat	2	48,374	41.345
14	lit_poezja	0	74,793	0
Łącznie:		**6,206**	**239,705,512**	**25.89**

Źródło: wyszukiwarka PELCRA w NKJP.

Tabelę 1 przeniosłem prawie bez zmian z okna wyszukiwarki, zmieniłem jedynie wielkość czcionki. Widoczne są niedogodności, zarówno natury redakcyjnej (np. liczby powinny być wyrównane w kolumnach do cyfry najmniej znaczącej lub do przecinka dziesiętnego), jak i merytorycznej (np. zastosowano anglosaski zapis liczb, z kropką zamiast przecinka dziesiętnego i z przecinkiem rozdzielającym trójelementowe grupy cyfr, co może utrudniać dalszą obróbkę liczb, np. w arkuszu kalkulacyjnym, jeśli zawartość tabeli zostanie do niego skopiowana).

Z nagłówka tabeli 1 wynika, że frekwencję słowa wyrażono liczbą jego wystąpień w kolejnych typach tekstów. Nie jest to ścisłe: podane liczby nie dotyczą wcale wystąpień, lecz akapitów, w których dane słowo występuje raz lub więcej niż raz. Na przykład liczba wystąpień leksemu *zabójstwo* w kategorii *lit_proza* zgodnie z tabelą 1 miałaby wynosić 70, podczas gdy kwerenda wykonana wyszukiwarką PELCRA przynosi 73 wyniki. Ogólnie rzecz biorąc, PELCRA podaje liczbę akapitów zawierających szukane słowo, liczbę wystąpień trzeba natomiast ustalić samodzielnie, patrząc na numer ostatniego wiersza konkordancji. W wypadku słów częstych wymaga to podniesienia limitu wyników ze 100 (wartość domyślna) do 10 000 (wartość maksymalna), a w wypadku słów bardzo częstych, gdy podniesienie limitu nie wystarcza, kwerendę trzeba dzielić na części, np. najpierw objąć nią teksty wydane do roku 2000, a potem pozostałe. W kontekście działania funkcji Profil ważniejsze od tych utrudnień jest jednak to, że wartości frekwencji względnej w ostatniej kolumnie tabeli 1 nie odnoszą się do wystąpień

słowa w przeliczeniu na milion słów w danym typie tekstów, tylko do liczby akapitów w przeliczeniu na milion słów.

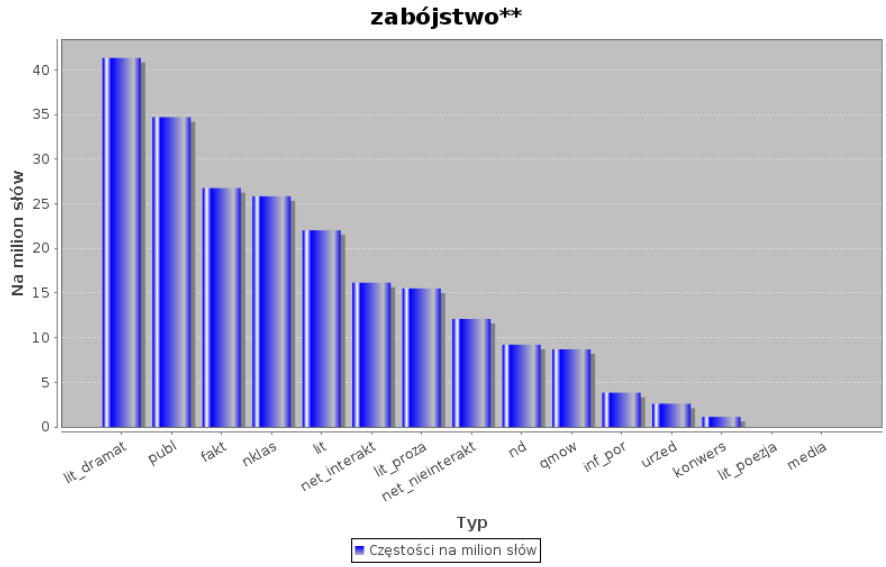
Nasuwają się dwa pytania. Czy frekwencję słowa lepiej mierzyć liczbą jego wystąpień, liczbą akapitów, w których to słowo występuje, czy jeszcze inaczej (np. wyszukiwarka PELCRA 2 wyraża frekwencję liczbą zdań)? Oraz czy różnice między tymi miarami wpływają na wyniki działania funkcji Profil? Pierwsze pytanie pozostaje poza zakresem tego artykułu, warto jednak zanotować, że dla słów o wysokiej częstości liczba akapitów może znacznie ustępować liczbie wystąpień, natomiast dla słów rzadkich obie miary powinny dawać podobne wyniki. Co więcej, jeśli słowo występuje wielokrotnie w zaledwie kilku akapitach, a poza nimi wcale lub prawie wcale, liczba akapitów może być nawet lepszą miarą jego frekwencji niż liczba wystąpień. Mimo to standardową miarą frekwencji słów w tekście pozostaje liczba wystąpień. O tym, że w wyszukiwarce PELCRA jest inaczej, zadecydował wzgląd na szybkość jej działania, czynnik szczególnie istotny w wypadku aplikacji internetowej[8].

Na drugie pytanie – w jakim stopniu przyjęcie liczby akapitów zamiast liczby wystąpień słowa jako miary jego frekwencji wpływa na wyniki działania funkcji Profil – nie da się odpowiedzieć bez badań empirycznych. Można jedynie przypuszczać, że ponieważ dla funkcji Profil istotniejsze są różnice we frekwencji słowa w poszczególnych rodzajach tekstów niż jego frekwencja bezwzględna, przyjęcie liczby akapitów jako miary frekwencji nie powinno deformować profilu słowa w zakresie tych typów tekstów, w których jego frekwencja względna jest najwyższa.

Prócz tabel funkcja Profil dostarcza poglądowych wykresów pokazujących, w których typach tekstów dane słowo występuje najczęściej w przeliczeniu na milion słów, w których rzadziej, w których jeszcze rzadziej itd. Wykres 1 przedstawia profil stylistyczny leksemu *zabójstwo*. Kolejność i wysokość słupków są zgodne z wartościami podanymi w ostatniej kolumnie tabeli 1.

Zwraca uwagę obecność tekstów dramaturgicznych na pierwszym miejscu w profilu *zabójstwa*. Z tabeli 1 wynika, że należą do nich tylko dwa wystąpienia tego leksemu (ściślej biorąc, dwa akapity, ale przypadkiem liczba akapitów pokrywa się tu z liczbą wystąpień). Jest to liczba zbyt mała, by uważać ją za miarodajną, ale wystarczająco duża, by w relacji do skromnego segmentu tekstów dramaturgicznych w NKJP wysunąć go na pierwsze miejsce na wykresie. Interpretując wyniki, teksty dramaturgiczne w profilu stylistycznym *zabójstwa* trzeba po prostu zignorować, podobnie jak teksty konwersacyjne (typ *konwers*, tylko dwa akapity, por. tab. 1), i tak samo należy postąpić w analizie dystrybucji innych słów, gdy za jakimś słupkiem na wykresie kryje się zaledwie kilka akapitów. Przykłady tego rodzaju nie są wcale rzadkie, zwracano już na nie uwagę (Piotrowski, Grabowski 2013: 66–68). Skalę problemu można by ograniczyć przez zwiększenie NKJP, ale nawet w kilkukrotnie większym korpusie znajdą się wyrazy o tak małej liczbie wystąpień w danym rodzaju tekstów, że w analizie ich dystrybucji rodzaj ten trzeba będzie pominąć.

8 Decyzja o przyjęciu akapitu jako jednostki indeksowania korpusu została podjęta kilkanaście lat temu, z uwzględnieniem mocy ówczesnych komputerów i ówczesnej technologii. Profesorowi Piotrowi Pęzikowi dziękuję za wyjaśnienie tej kwestii, jak również za inne uwagi na temat niniejszego artykułu.



Wykres 1. Profil stylistyczny leksemu *zabójstwo*

Źródło: wyszukiwarka PELCRA w NKJP.

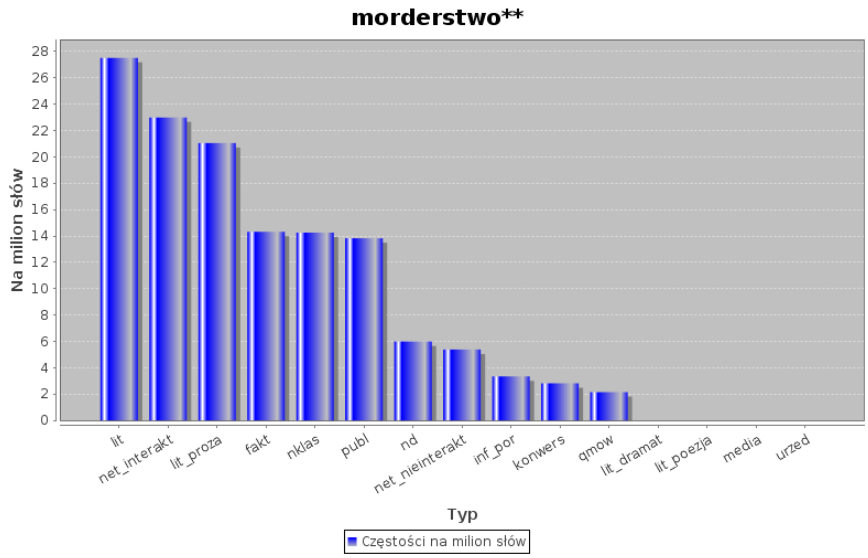
Co mówią wyniki funkcji Profil

Tu dochodzimy do pytania, czym właściwie jest profil stylistyczny słowa. Można powiedzieć, że informuje on o pewnego rodzaju preferencjach w zakresie jego użycia. Widoczne na wykresie 1 *zabójstwo* jest w przeliczeniu na milion słów najczęściej spotykane w tekstach publicystycznych, nieco rzadziej w literaturze faktu, jeszcze rzadziej np. w literaturze pięknej, a relatywnie bardzo rzadko w tekstach informacyjno-poradnikowych. Słowo *relatywnie* jest tu istotne, mniej bowiem liczą się wartości bezwzględne (wysokość słupków) niż relacje między nimi (kolejność słupków i różnice wysokości między nimi).

Profil *zabójstwa* porównajmy z profilem *morderstwa* (ponownie z uwzględnieniem wszystkich form odmiany; zob. wykres 2 na następnej stronie). Ten drugi leksem występuje najczęściej w beletrystyce, nieco rzadziej w internetowych tekstach interaktywnych, zdecydowanie rzadziej zaś w literaturze faktu i publicystyce, które zajmowały pierwsze miejsca w profilu *zabójstwa*.

Porównanie wykresów 1 i 2 naprowadza na istotną różnicę między *zabójstwem* i *morderstwem*. Pierwsze słowo, najczęściej obecne w tekstach o charakterze sprawozdawczym, ma charakter oficjalny i neutralny. Drugie, najczęstsze w wypowiedziach nacechowanych subiektywizmem, łatwo podlega wartościowaniu i służy przekazywaniu negatywnych emocji. Do takich wniosków uprawniają też inne obserwacje, m.in. nieobecność *morderstwa* w polskim języku prawnym i prawniczym (np. *Kodeks karny* zna tylko *zabójstwo*) i bardziej negatywne konteksty *morderstwa* w porównaniu z *zabójstwem* (np. wśród pierwszych dziesięciu prepozycyjnych przymiotników znalezionych za pomocą kolokatora w wyszukiwarce PELCRA *morderstwo* ma więcej określeń nacechowanych negatywnie). Jak widać, funkcja Profil może być użyteczna

w porównawczej analizie synonimów, ale bez wyraźnej świadomości, o czym właściwie informuje, łatwo popełnić błąd w interpretacji wyników.



Wykres 2. Profil stylistyczny leksemu *morderstwo*

Źródło: wyszukiwarka PELCRA w NKJP.

Szczególnie często, jak się przekonałem, w pracach studenckich pojawia się mylny wniosek, że jeśli jedno słowo ma w swoim profilu na samym początku np. teksty publicystyczne, a drugie np. teksty literackie, to pierwsze z tych słów częściej od drugiego występuje w publicystyce, a drugie częściej od pierwszego w beletrystyce. Tak może być (i tak akurat jest w wypadku *zabójstwa* i *morderstwa*, niezależnie od tego, czy brać pod uwagę wartości bezwzględne, czy przeliczone na milion słów w poszczególnych rodzajach tekstów), lecz wcale tak być nie musi. Zilustrujmy to przykładem analogicznym i bliższym życiu: przypuśćmy, że osoba X trzyma swoje oszczędności głównie na lokatach bankowych i tylko małą część inwestuje w akcje, podczas gdy osoba Y czyni na odwrót – kupuje akcje i tylko niewielką część swoich oszczędności odkłada w banku. Z tego oczywiście nie wynika, że osoba X ma więcej pieniędzy na lokatach od osoby Y, a Y ma więcej w akcjach niż X.

Profile stylistyczne mogą być instruktywne w porównawczej analizie wyrazów synonimicznych, ale trzeba pamiętać, że każdy profil powinien być analizowany osobno. Profile nie informują, jak proporcja między frekwencją jednego synonimu a frekwencją drugiego zależy od rodzaju tekstu, lecz z jaką częstością w przeliczeniu na milion słów dane słowo występuje w tekstach różnych rodzajów.

Względna frekwencja synonimów w zależności od rodzaju tekstów

Aby zbadać, jak zmienia się proporcja między frekwencją wyrazów synonimicznych w zależności od typu tekstów, nie trzeba analizować wykresów generowanych przez funkcję Profil.

Wystarczy porównać liczbę wystąpień jednego i drugiego synonimu w poszczególnych rodzajach tekstów, ewentualnie liczbę akapitów z tymi słowami, czyli wykorzystać dane z pierwszej kolumny liczbowej tabel dostarczanych przez funkcję Profil, takich jak tabela 1.

W tabeli 2 zestawiono statystyki pobrane z dwóch tabel wygenerowanych przez funkcję Profil dla leksemów *zabójstwo* i *morderstwo*. Typy tekstów ułożono w porządku od najliczniejszych (tzn. obejmujących najwięcej słów) do najmniej licznych. Z porównania łącznej liczby akapitów wynika, że w całym zrównoważonym podkorpusie NKJP *zabójstwo* przeważa nad *morderstwem* w stosunku około 1,86. W niektórych rodzajach tekstów proporcja ta rośnie (*zabójstwo* zwiększa swoją przewagę nad *morderstwem*), a w innych maleje (*zabójstwo* traci na rzecz *morderstwa*). Przewaga *zabójstwa* nad *morderstwem* najbardziej wzrasta w tekstach quasi-mówionych (są to głównie protokoły obrad Sejmu RP), w mniejszym stopniu w publicystycznych i internetowych nieinteraktywnych. Maleje zaś najsilniej w tekstach konwersacyjnych, w internetowych interaktywnych i w literackich, a więc takich, które sprzyjają wyrażaniu ocen subiektywnych i emocjonalnych. Obserwacje te potwierdzają, że *zabójstwo* to termin oficjalny i neutralny emocjonalnie, w przeciwieństwie do *morderstwa*.

Tabela 1. Zależność względnej frekwencji leksemów *zabójstwo* i *morderstwo* od typu tekstów

TYP	ZABÓJSTWO	MORDERSTWO	PROPORCJA
publ	4493	1790	2,51
lit	666	832	0,80
qmow	203	50	4,06
fakt	398	213	1,87
net_interakt	149	212	0,70
nd	80	52	1,54
inf_por	31	27	1,15
lit_proza	70	95	0,74
net_nieinterakt	45	20	2,25
urzed	9	0	
nklas	58	32	1,81
konwers	2	5	0,40
lit_poezja	0	0	
lit_dramat	2	0	
Łącznie	6206	3328	**1,86**

Źródło: wyszukiwarka PELCRA, opracowanie własne.

Niektóre komórki tabeli 2 są puste, gdyż obliczenie proporcji wymagałoby dzielenia przez zero. Ponieważ jednak w takich wypadkach zarówno frekwencja *zabójstwa*, jak i *morderstwa* wyraża się liczbą zaledwie jednocyfrową, nie ma to większego wpływu na całość wyników.

Jak wiadomo, NKJP został zamknięty w roku 2010 i odczuwa się potrzebę jego uzupełnienia o nowe teksty. Trwają dyskusje o jego przyszłym kształcie (np. Ogrodniczuk i in. 2017), warto więc rozważyć udostępnienie w nowej wersji wyszukiwarki PELCRA narzędzia, które automatycznie wyliczałoby proporcje takie jak w tabeli 2 i prezentowałoby je obrazowo w formie wykresu. Byłoby ono czymś analogicznym do bazy HASK[9] (Pęzik 2013), która ułatwia porównanie kolokacji dwóch wyrazów synonimicznych, generując poglądowe grafy i prawie gotowe listy różnicowe (można je dopracować np. za pomocą filtrowania wyników w arkuszu Excel).

Katalog biblioteczny jako źródło informacji o dystrybucji słów

Podkorpus zrównoważony NKJP obejmuje czternaście typów tekstów, które widać w tabelach 1 i 2 oraz na wykresach wyżej. W pełnym korpusie można znaleźć ponadto teksty typu *media*, czyli medialne mówione. Nawet z nimi jednak typologia tekstów NKJP jest stosunkowo prosta i nie zawsze ujawnia różnice w dystrybucji stylistycznej badanych słów. W takich wypadkach na ogół nie pomaga też uwzględnienie podziału tekstów na tzw. kanały (książki, prasa, internet itp.). Choć PELCRA oprócz profili omówionych wyżej udostępnia profile oparte na kanałach, są one zwykle mniej instruktywne (i dlatego zostały pominięte w tym artykule).

Gdy profile stylistyczne PELCR-y zawodzą, pewien wgląd w dystrybucję wyrazów, stylistyczną lub dziedzinową, może dać kwerenda w internetowym katalogu bibliotecznym, ograniczona do tytułów publikacji. Na przykład w katalogu Biblioteki Narodowej[10] słowo *składniki* występuje głównie w tytułach tekstów z zakresu biotechnologii i dietetyki, podczas gdy synonimiczne *komponenty* są obecne w tytułach publikacji bardziej zróżnicowanych tematycznie. Podobnie *śmigłowce* występują przede wszystkim w tytułach publikacji z zakresu wojskowej techniki lotniczej, *helikoptery* natomiast mają szerszy zakres użycia. Ze słowem *helikopter* można się spotkać m.in. w tytułach tekstów literackich, i to nawet przeznaczonych dla dzieci, co znaczy, że jest ono bliższe codziennej polszczyźnie, podczas gdy *śmigłowiec* należy raczej do terminologii specjalistycznej.

Obserwacje takie mają, rzez jasna, ograniczoną przydatność i trudno ująć je liczbowo, ponieważ katalog biblioteczny nie służy celom badawczym i nie wspiera analizy ilościowej. Z tego powodu nie proponuję tu żadnego schematu badawczego, żadnej procedury poszukiwań wyniku w tytułach książek. Kwerendy w katalogu bibliotecznym mają charakter pomocniczy, ale wynikające z nich wnioski zyskują na wiarygodności, gdy są zgodne z obserwacjami opartymi na innych źródłach (jak np. w wypadku *helikoptera* i *śmigłowca*, por. analizę tej pary synonimów w artykule: Bańko 2013).

9 http://pelcra.pl/hask_pl/ (dostęp: 8 października 2023).

10 https://katalogi.bn.org.pl/discovery/search?vid=48OMNIS_NLOP:48OMNIS_NLOP (dostęp: 8 października 2023).

Podsumowanie

Profile stylistyczne PELCR-y są wykorzystywane zarówno w publikacjach naukowych (np. Kasza 2014), jak i w pracach studenckich i doktoranckich. Tym ważniejsze jest, aby stosować je ze świadomością ich przeznaczenia i ograniczeń. W niniejszym artykule zwrócono uwagę na pewne ich cechy, które nie zostały udokumentowane w publikacjach dotyczących PELCR-y bądź zostały opisane w sposób nieprecyzyjny i mylący. Wskazano także na błędy popełniane w interpretacji wyników dostarczanych przez funkcję Profil i zaproponowano prostą metodę porównania dystrybucji stylistycznej wyrazów synonimicznych, którą warto zaimplementować w nowej wersji wyszukiwarki PELCRA.

Ocena użyteczności innych wyszukiwarek NKJP w analizie dystrybucji stylistycznej słów nie należała do zakresu tego artykułu, pobieżne obserwacje wskazują jednak, że choć PELCRA 2 wyraża frekwencję słów nie liczbą akapitów, lecz liczbą zdań, profile przez nią generowane (nazywane tu fasetami, ang. *facets*) na ogół nie różnią się od profili PELCR-y w zakresie tych typów tekstów, w których badane słowo ma najwyższą frekwencję względną.

Wyniki działania funkcji Profil są czasem mało diagnostyczne z powodu – jak można przypuszczać – nie dość szczegółowej i nie zawsze trafnej klasyfikacji tekstów w NKJP. Zagadnienie to znalazło się jednak poza obszarem niniejszych rozważań.

W artykule zwrócono za to uwagę, że pewien wgląd w zakres użycia słowa może dać kwerenda w internetowym katalogu bibliotecznym, ale wymaga ona weryfikacji na podstawie innych źródeł.

Bibliografia

Bańko M. 2013: *Helikopter, śmigłowiec*, http://www.approval.uw.edu.pl/helikopter (dostęp: 8 października 2023).

Bańko M., Górski R. 2014: *Praktyczny przewodnik po korpusie języka polskiego*, [w:] M. Hebal-Jezierska (red.), *Praktyczny przewodnik po korpusach języków słowiańskich*, Wydział Polonistyki Uniwersytetu Warszawskiego, Warszawa, s. 11–28.

CLIP: Computational Linguistics in Poland (online: http://clip.ipipan.waw.pl/, dostęp: 8 października 2023).

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

Kasza M. 2014: *O możliwości wykorzystania automatycznej analizy źródeł w NKJP przy wyborze kwalifikatora dla jednostki słownikowej*, „Język Polski" XCIV, z. 5, s. 399–416.

Kieraś W., Woliński M., Nitoń B. 2021: *Nowe wielowarstwowe znakowanie lingwistyczne zrównoważonego Narodowego Korpusu Języka Polskiego*, „Język Polski" CI, z. 2, s. 59–70.

NKJP: Narodowy Korpus Języka Polskiego (online: http://nkjp.pl, dostęp: 8 października 2023).

Ogrodniczuk M., Derwojedowa M., Łaziński M., Pęzik P. 2017: *Narodowy Korpus Języka Polskiego – co dalej?*, „Prace Filologiczne" LXXI, s. 237–245.

Ogrodniczuk M., Przepiórkowski A. 2013: *CLIP – portal internetowy łączący projekty, narzędzia, zespoły związane z komputerowym przetwarzaniem języka polskiego*, „Język Polski" XCIII, z. 3, s. 242.

Pęzik P. 2012: *Wyszukiwarka PELCRA dla danych NKJP*, [w:] A. Przepiórkowski, M. Bańko, R. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 253–273.

Pęzik P. 2013: *Paradygmat dystrybucyjny w badaniach frazeologicznych. Powtarzalność, reprodukcja i idiomatyzacja*, [w:] P. Stalmaszczyk (red.), *Metodologie językoznawstwa. Ewolucja języka, ewolucja teorii językoznawczych*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, s. 141–160.

Piotrowski T., Grabowski Ł. 2013: *Interpretacja danych frekwencyjnych z korpusów językowych: opis pewnych problemów (na kilku przykładach z życia wziętych)*, [w:] W. Chlebda (red.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*, Wydawnictwo Uniwersytetu Opolskiego, Opole, s. 59–71.

Przepiórkowski A. 2004: *Korpus IPI PAN: wersja wstępna*, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.

Przepiórkowski A., Bańko M., Górski R., Lewandowska-Tomaszczyk B. 2012: *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 253–273.

Summary

Studying stylistic distribution using the PELCRA search engine in the National Corpus of Polish

Keywords: stylistic distribution, National Corpus of Polish, PELCRA search engine.

The article discusses poorly or incorrectly documented features of the Profile function used to examine stylistic distribution by means of the PELCRA search engine in the National Corpus of Polish. Attention is paid to errors made by users in interpreting the results provided by the Profile function. The use of the function to compare the stylistic distribution of synonymous words is shown. Online library catalogs are also included as a complementary source of information about the stylistic distribution of words.