Piotr Pęzik
Uniwersytet Łódzki
https://orcid.org/0000-0003-0019-5840
piotr.pezik@uni.lodz.pl

Forum Lingwistyczne
nr 7, 2020
ISSN 2450-2758 (wersja elektroniczna)

DOI: http://doi.org/10.31261/FL.2020.07.11

Budowa i zastosowania korpusu monitorującego MoncoPL

Design and Applications of MoncoPL as a Monitor Corpus of Polish

Abstract: This paper introduces the methodology of compiling and maintaining MoncoPL, a large monitor corpus of web-based Polish. Furthermore, an overview of the search engine of the same name is provided to show how the size and composition of the corpus, currently reaching over 5.6 billion word tokens, facilitates research on distributional properties of rare words, neologisms and phraseological units. Finally, the article exemplifies some advantages of using a densely-sampled diachronic corpus for the purposes of observing frequency trends and cycles of various constructions in online media discourse.

Key words: MoncoPL, monitor corpus, Polish, diachronic corpora

Abstrakt: Artykuł przedstawia budowę i zastosowania korpusu monitorującego polszczyzny MoncoPL oraz główne funkcje wyszukiwarki o tej samej nazwie. Omówiona zostaje rola referencyjna korpusu oraz jego użycia w identyfikacji neosemantyzmów występujących w komunikacji internetowej. Wielkość indeksu MoncoPL (5,6 mld słów w połowie 2019 r.) pozwala zaobserwować subtelne zjawiska językowe, takie jak łączliwość frazeologiczna rzadkich słów czy wariantywność frazemów. Z kolei wysoka częstość próbkowania danych umożliwia badanie trendów i cykliczności występowania różnorodnych konstrukcji językowych w internetowych rejestrach dyskursu medialnego.

Słowa kluczowe: MoncoPL, korpus monitorujący polszczyzny, korpusy diachroniczne

1. Wstęp

Korpusy językowe stanowią dziś podstawę badań lingwistycznych, leksykograficznych, a także prac informatycznych z dziedziny komputerowego przetwarzania języka naturalnego. Szczególne znaczenie dla empirycznych badań języka mają korpusy określane mianem referencyjnych, takie jak *Narodowy Korpus Języka Polskiego* (NKJP) (Przepiórkowski i in., 2012). Są to zazwyczaj liczące co najmniej kilkaset milionów słów zbiory wielu odmian, rejestrów i typów funkcjonalnych tekstów, skompilowane według określonych z góry kryteriów z myślą o szerokich zastosowaniach praktycznych i teoretycznych. Korpusy monitorujące można uznać za specjalny typ korpusów referencyjnych. Ich struktura i zawartość mają w założeniu umożliwiać ciągłą obserwację najnowszych zmian i trendów w języku. W odróżnieniu od nieaktualizowanych korpusów diachronicznych, które również składają się z tekstów próbkowanych z podokresów historycznych, a także inaczej niż w przypadku niezrównoważonych diachronicznie korpusów referencyjnych, korpusy monitorujące mają strukturę otwartą i są stale uzupełniane o nowe próbki reprezentowanych w nich odmian

języka. Korpus monitorujący, który przestaje być aktualizowany, pozostaje jedynie zamkniętym korpusem diachronicznym, a jego funkcja referencyjna ulega nieuchronnemu osłabieniu w miarę upływu czasu. Podczas gdy ogólne korpusy referencyjne opisuje się stałą wartością ich wielkości, na przykład liczbą słów, zdań czy też tekstów w nich zawartych, podstawowym parametrem ilościowym korpusu monitorującego jest tempo przyrostu (ang. *rate of flow*, por. Sinclair, 1996), mierzone zazwyczaj w segmentach słów przypadających na jednostkę czasu. Do niedawna tempo przyrostu zawartości korpusów monitorujących wynosiło nie więcej niż kilkadziesiąt milionów słów na rok, na przykład popularny korpus monitorujący angielszczyzny amerykańskiej COCA powiększył się w latach 2010–2017 z około 400 do 560 mln segmentów wyrazowych (Davies, 2010). W epoce cyfrowej dostępność niektórych typów danych językowych jest tak duża, że współczesne korpusy monitorujące można powiększać w tempie milionów słów dziennie. Przykładowo, korpus monitorujący angielszczyzny globalnej NOW powiększa się o 140–160 mln słów miesięcznie[1].

Prawdopodobnie jedynym publicznie dostępnym, stale aktualizowanym korpusem monitorującym polszczyzny jest obecnie MoncoPL[2] i to właśnie temu zasobowi oraz narzędziom do jego przeszukiwania dostępnym w serwisie *monco.frazeo.pl* poświęcony jest ten artykuł. Pierwsza część tekstu przedstawia metodologię budowy i aktualizacji MoncoPL, główne funkcje jego wyszukiwarki, w tym składnię zapytań, funkcje agregowania wyników według metadanych oraz moduł ekstrakcji kolokacji. W dalszej części omówione są przykłady zastosowań korpusu monitorującego w badaniu neosemantyzmów oraz identyfikacji kolokacji występujących w internetowych rejestrach polszczyzny[3].

2. Indeks MoncoPL

Podstawowym źródłem danych indeksowanych w korpusie MoncoPL są kanały RSS[4] udostępniane przez ponad 1500 serwisów informacyjnych, branżowych oraz platform blogowych, z których około 600 było aktywnych w lipcu 2019 roku. Poza źródłami dynamicznymi do indeksu MoncoPL zostały jednorazowo włączone niektóre dostępne na otwartej licencji korpusy, z których największym jest *Polski Korpus Sejmowy* (Ogrodniczuk, 2017). Aktualną listę największych źródeł korpusu można znaleźć w widoku *Narzędzia -> Statystyki -> Indeks*. Średni przyrost danych wynosi aktualnie około 1,64 mln słów, czyli około 87 tys. zdań dziennie. Liczba ta waha się z odchyleniem standardowym = 409 tys. słów, w wyraźnych cyklach tygodniowych, wynikających z dużo niższej podaży artykułów w serwisach informacyjnych w soboty i niedziele w porównaniu z pozostałymi dniami tygodnia. Wykres przyrostu danych od 2010 roku w odstępach miesięcznych ukazano na rysunku 1.

[1] Zob. online: https://www.english-corpora.org/now/; data dostępu: 15.06.2020.

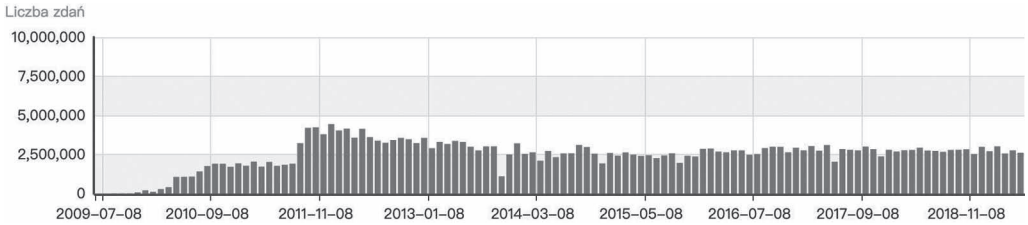
[2] Ponieważ jedynym narzędziem umożliwiającym korzystanie z korpusu jest dostępna pod adresem monco.frazeo.pl wyszukiwarka, nazwa MoncoPL jest używana w tym artykule zamiennie jako określenie witryny wyszukiwarki i samego korpusu.

[3] O zastosowaniach MoncoPL w dydaktyce polonistycznej piszą Beata Duda i Karolina Lisczyk (Duda, Lisczyk, 2018: 148–149).

[4] RSS (ang. *RDF Site Summary*) to oparty na języku XML format używany do udostępniania list nagłówków artykułów publikowanych w serwisach WWW.

Poza okresem rozruchu na przełomie lat 2010/2011 i dwoma epizodami awarii systemu zbierania danych indeks MoncoPL powiększał się w tempie około 2,5 mln zdań (47 mln słów) miesięcznie. W połowie 2019 roku korpus osiągnął chwilową wielkość 5,644 mld segmentów wyrazowych w ponad 300 mln jednostek zdaniowych.

Ogromną większość tekstów wchodzących w skład korpusu stanowią artykuły publikowane w internetowych serwisach informacyjnych. Ponieważ za ekstrakcję danych tekstowych ze stron WWW odpowiada generyczny algorytm, część artykułów jest indeksowanych wraz z sekcjami komentarzy, co uwidacznia się w niektórych wynikach. Takie nie do końca zamierzone zróżnicowanie można uznać za pewną zaletę, jako że styl anonimowych komentarzy znacząco się różni od rejestru artykułów, nawet jeżeli są one z nimi powiązane tematycznie. To z kolei pozwala znaleźć w korpusie przykłady form językowych, które w artykułach występują sporadycznie.



Rys. 1. Przyrost danych w indeksie MoncoPL liczony w zdaniach na miesiąc

Równomierne próbkowanie danych z ziarnistością na poziomie jednego dnia otwiera możliwości badań trendów i cykli frekwencyjnych w danych językowych. Również sam fakt zindeksowania niemal 6 mld segmentów wyrazowych z tekstów opublikowanych głównie po roku 2010, a więc po zakończeniu konstrukcji NKJP, nadaje korpusowi MoncoPL cechy uzupełniającego korpusu referencyjnego polszczyzny, choć ograniczonego tylko do kilku typów funkcjonalnych tekstów. Przykłady tych funkcji i zastosowań omówiono w dalszej części artykułu.

3. Funkcje wyszukiwarki MoncoPL

Ze względu na ograniczenia wynikające z praw autorskich, którymi objęte są teksty zindeksowane w MoncoPL, jedynym narzędziem do udostępniania tego korpusu użytkownikom zewnętrznym jest wyszukiwarka dostępna w serwisie *monco.frazeo.pl*. Funkcje tego narzędzia można podzielić na wyszukiwawcze i agregacyjne. Podstawową funkcją wyszukiwawczą jest generowanie konkordancji dla zapytań korpusowo-bibliograficznych. Do funkcji agregacyjnych można zaliczyć tworzenie tzw. faset[5] dla wyników wyszukiwania oraz oddzielny moduł ekstrakcji kolokacji.

[5] *Faseta* (od ang. *facet*, a pierwotnie od fr. *facette*) to w kontekście systemów wyszukiwawczych termin oznaczający widok wyników oparty na agregacji wyników po jednej metadanej lub większej ich liczbie. Fasetą może być tabela określająca liczbę wystąpień danej frazy w różnych źródłach lub przedziałach czasowych.

3.1. Konkordancje i składnia zapytań

3.1.1. Słowoformy i dokładne frazy

Aby wyszukać dokładne wystąpienia pojedynczych słowoform lub fraz składających się z dwóch lub więcej słowoform, należy wpisać w polu wyszukiwania dosłowną sekwencję wyszukiwanych słowoform i nacisnąć ikonę lupy. Wielkość liter nie ma znaczenia w zapytaniach. Inaczej niż w przypadku niektórych wyszukiwarek internetowych, fraz nie należy wpisywać w cudzysłowach[6]. Na rysunku 2 ukazano ekran wyników wyszukiwania słowoformy *połówka*. W całym korpusie MoncoPL znaleziono 1551 zdań zawierających ten wyraz.

Konkordancje | Podsumowanie | Fasety | Trendy | Pliki

Wyniki od 1 do 10 z 1551. Kryteria wyszukiwania.

Show / hide columns

#				Źródło	Data
1	ektora , 606904195 , ar_kam@wp.pl , www.ar - kam . com.pl	**Połówka**	bliźniaka w bogatym stanie developerskim ! Pow .	pultusk.tio.com.pl	23/2/2012
2	ja komfort Chotomów Jabłonna Legionowo , 797 171 543 7	Połówka	bliźniaka w bogatym stanie developerskim ! Pow .	pultusk.tio.com.pl	23/3/2012
3	Nawet po przecięciu na pół każda	**połówka**	pływałaby osobno .	interia.pl	4/3/2015
4	Bardzo często za sukcesem przedsiębiorcy - marzyciela , stoi ta druga	**połówka**	na nudnym , ale stabilnym etacie .	polki.pl	15/11/2016
5	-	**Połówka**	jednego ze świątecznych dni pozostaje nieobsadzona , w...	dzienniklodzki.pl	24/12/2016
6	wiartki , będzie to „ Ironman na raty " , a w przypadku 1/ 8IM	**połówka**	na raty " .	portel.pl	4/12/2018
7	Tak podane śniadanie sprawi , że Twoja druga	**połówka**	poczuje się naprawdę wyjątkowo .	iswinoujscie.pl	14/2/2019
8	ozwodu prawo decydować , na jaki kolor jego pozostawiona	**połówka**	małżeństwa ma sobie pomalować ściany w jej własnym ...	onet.pl	11/4/2019
9	Pokaźną rekompensatą zdecydowanie jest druga	**połówka**	, w której Klevberg serwuje już nam klasyczny slasher - pr...	trojmiasto.pl	22/6/2019
10	In minus na pewno słaba pierwsza	**połówka**	filmu , mizerny scenariusz , masa nielogiczności i . . . mał...	trojmiasto.pl	22/6/2019

Poprzednia | 1 | 2 | Następna | Pokaż 10 kontekstów

Rys. 2. Konkordancje słowoformy *połówka*. Zob. https://tinyurl.com/yxuorcsw

Konkordancji na danej stronie może być nieco więcej, niż sugerowałaby to wybrana przez użytkownika wartość opcji limitu (w powyższym przykładzie było to 10 zdań). Dzieje się tak w sytuacji, gdy co najmniej jedno z pobranych z indeksu zdań zawiera więcej niż jeden kontekst pasujący do zapytania. Innymi słowy, gdyby wśród pierwszych dziesięciu zdań znalazło się dokładnie jedno zdanie zawierające podwójne wystąpienie wyrazu *połówka*, to na pierwszym ekranie wyników wyświetlonych byłoby 11 wierszy konkordancji. Poza samym dopasowaniem i jego kontekstem domyślnie wyświetlane są kolumny z nazwą źródła (portalu), z którego pochodzi cytat, oraz datą publikacji tekstu. Nazwa źródła jest jednocześnie bezpośrednim odnośnikiem do oryginalnej strony z pełną wersją tekstu. W niektórych przypadkach wskazywana strona może być w chwili wyświetlenia konkordancji niedostępna. Warto pamiętać, że możliwe jest wyświetlenie dodatkowych lub ukrycie aktualnie wyświetlanych kolumn z metadanymi. Do konfiguracji wyświetlanych kolumn służy przycisk znajdujący się nad tabelą konkordancji. Aktualnie, na jednej

[6] Wyniki przykładu zapytania o dosłowne wystąpienie frazy *od wielkiego dzwonu* można zobaczyć pod adresem online: https://tinyurl.com/y3ypz7u5; data dostępu: 17.04.2020.

stronie można wyświetlić do 1 tys. wyników. Przechodzenie między kolejnymi stronami wyników umożliwiają kontrolki u dołu tabeli. Wykryte w bieżących wynikach duplikaty kontekstów są przekreślane i wyszarzane, co pozwala na ich szybką identyfikację. Po dwukrotnym kliknięciu na wiersz z wynikiem ukazuje się nieco większy kontekst wystąpienia wyszukiwanego wyrazu.

Nieco mniej intuicyjne może się wydawać wyszukiwanie słowoform, które są zapisywane w tekstach łącznie, a następnie dzielone przez analizator morfologiczny Morfeusz (Woliński, 2014) na osobne segmenty. Przykładem może być wyszukiwanie niektórych form czasownika *być*, na przykład formy *byłem*, która jest dzielona w indeksie na dwa segmenty *był* i *-em*. Zapytanie o tę formę powinno mieć postać *był em*[7].

3.1.2. Rozszerzenia fleksyjne i ortograficzne

Niezwykle istotnym elementem składni wyszukiwarki dla korpusu polszczyzny jako języka z bogatą fleksją jest możliwość wyszukania wariantów fleksyjnych zadanej formy podstawowej (tzw. lematu). Dwa przykłady takich zapytań dla słów i prostych fraz przedstawiono poniżej:

*ściemniać***
*Wielki*** *Brytania***[8]

Pierwsze z tych zapytań zwraca konteksty zawierające różne formy czasownika *ściemniać*, a drugie – różne warianty frazy *Wielka Brytania*. Warto zauważyć, że w drugim wypadku formą podstawową rozszerzanego fleksyjnie terminu przymiotnika jest forma rodzaju męskiego, tj. *wielki*, mimo iż jej dosłowne wystąpienie nie pokrywa się z intencją zapytania, ze względu na stałe uzgodnienie rodzaju przymiotnika *wielki* z głową[9] frazy *Wielka Brytania*.

Wyniki uzyskane za pomocą automatycznego rozszerzenia fleksyjnego mogą nie być pełne, jako że nie wszystkie formy podstawowe zostały poprawnie rozpoznane w indeksie. Istnieją dwa sposoby zwiększenia pokrycia form fleksyjnych w zapytaniu. Po pierwsze, możliwe jest wymienienie wszystkich wariantów fleksyjnych, które mają być uwzględnione na danej pozycji, na przykład:

fejs|*fejsem*|*fejsa*|*fejsie*

Po drugie, możliwe jest zastosowanie rozszerzenia ortograficznego, które jednak może dać nadmiarowe wyniki, na przykład:

*fejs.**

[7] Zob. online: https://tinyurl.com/yyjhbl9b; data dostępu: 24.06.2020.

[8] Zob. online: https://tinyurl.com/y2skutxc; data dostępu: 24.06.2020.

[9] Termin *głowa* jest używany tu jako odpowiednik ang. terminu *head* (of a phrase).

Możliwe jest ograniczenie generowanych w ten sposób wariantów na danej pozycji za pomocą operatora negacji, na przykład:

fejs.|!fejsbukow.**[10]

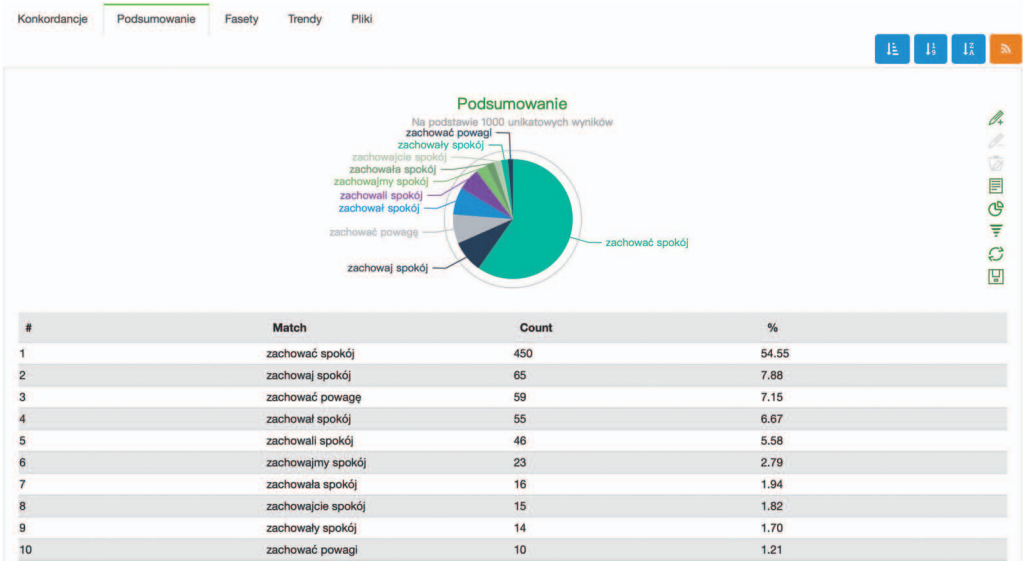
To ostatnie zapytanie zwróci różne formy wyrazowe rozpoczynające się od łańcucha znaków *fejs-* z wyjątkiem form, które zaczynają się od łańcucha *fejsbukow-*. Takie ograniczenie wykluczy więc z wyników wystąpienia przymiotników, co może być efektem pożądanym, jeżeli wyszukujemy jedynie form rzeczownikowych.

3.2. Warianty

Operator wariantu | został już częściowo wprowadzony powyżej. Warto pamiętać, że można go używać również do rozszerzeń leksykalnych. Na przykład, aby znaleźć jednocześnie wystąpienia fraz *zachować spokój* lub *zachować powagę*, można użyć następującego zapytania:

*zachować** spokój**|powaga**

W wynikach zapytań z wariantami lub operatorami rozszerzeń przydatna może być zakładka *Podsumowanie*, w której na podstawie próby o wielkości maksymalnej 1 tys. wystąpień tworzona jest lista frekwencyjna form pasujących do zapytania. Rysunek 3 ukazuje podsumowanie dopasowań do powyższego zapytania, z którego wynika, iż w losowym zbiorze 1 tys. wystąpień najczęściej pojawiała się forma *zachować spokój*, co można wytłumaczyć jej zagnieżdżeniem w dłuższych frazach z czasownikiem modalnym, wyrażającym potrzebę zachowania spokoju, na przykład *należy*, *powinno się*, *trzeba* + *zachować spokój*.



#	Match	Count	%
1	zachować spokój	450	54.55
2	zachowaj spokój	65	7.88
3	zachować powagę	59	7.15
4	zachował spokój	55	6.67
5	zachowali spokój	46	5.58
6	zachowajmy spokój	23	2.79
7	zachowała spokój	16	1.94
8	zachowajcie spokój	15	1.82
9	zachowały spokój	14	1.70
10	zachować powagi	10	1.21

Rys. 3. Widok zagregowanej konkordancji

[10] Zob. online: https://tinyurl.com/yylpyf9h; data dostępu: 12.07.2020.

3.3. Kolejność i odstęp między terminami zapytania

Domyślnie wyszukiwarka MoncoPL próbuje znaleźć dopasowania terminów występujących w indeksie korpusu w takiej kolejności, w jakiej zostały one zdefiniowane w zapytaniu. Aby wyłączyć ten warunek, należy rozwinąć panel opcji i odznaczyć w nim pole *Zachowaj szyk*. Dla przykładowego zapytania:

*kłamać** bezczelnie*

zostaną wtedy zwrócone zarówno wystąpienia frazy *kłamać bezczelnie* oraz jej wariantu pozycyjnego *bezczelnie kłamać*.

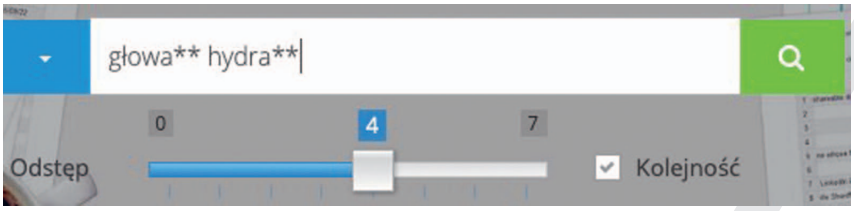
Składnia MoncoPL umożliwia stosunkowo wygodne wyszukiwanie różnego rodzaju związków wyrazowych, na przykład idiomów, formuł konwersacyjnych, kolokacji, a także nieutrwalonych kombinacji wielowyrazowych. Czasem trudne jest określenie z góry dokładnego zbioru, kolejności, a nawet relacji gramatycznych dla wyrazów tworzących dany związek lub frazę. Formułując zapytanie o frazy, możemy zwiększyć pokrycie zwracanego zbioru wyników poprzez odpowiednie ustawienie parametru odstępu. Domyślnie jego wartość wynosi *0*, co oznacza, że między kolejnymi pozycjami dopasowania wyrażanymi terminami zapytania nie może wystąpić żaden wyraz. Dla przykładu, zapytanie:

*głowa** hydra**

z domyślnym ustawieniem odstępu zwróci jedynie konteksty, w których te terminy zapytania występują bezpośrednio obok siebie. Wyrazy te jednak tworzą często figuratywne kolokacje i warianty idiomu opartego na metaforze złożonego, powracającego problemu jako mitycznej Hydry, której głowy odrastają po odcięciu. Po zwiększeniu wartości tego parametru do 4 wśród wyników znajdą się między innymi takie dopasowania:

hydra, która podnosi głowę
hydra, która podnosi głowę
hydra nie ma akurat tylu głów
Hydrze administracji odrastają liczne głowy

Jeszcze większe pokrycie w tym przypadku można uzyskać poprzez odznaczenie opcji *Zachowaj szyk*, jak to ukazano na rysunku 4[11].



Rys. 4. Zwiększenie pokrycia zapytania o frazy poprzez jednoczesne rozluźnienie parametrów kontekstu i szyku terminów

[11] Zob. online: https://tinyurl.com/y2r45яc7; data dostępu: 12.08.2020.

Poza podanymi uprzednio kontekstami w wynikach dla tego zapytania znajdują się również takie dopasowania, w których wyraz *głowa* występuje przed wyrazem *hydra*, na przykład:

głowa komunistycznej hydry
głów biurokratycznej hydry
głowę wymyślonej przez siebie faszystowskiej hydrze
głowa ohydnej hydry

3.4. Zapytania morfosyntaktyczne

Składnia MoncoPL umożliwia definiowanie sekwencyjnych wzorców leksykalno-gramatycznych. Kategorie morfosyntaktyczne można definiować, używając specjalnej składni: *<tag=XX>*, gdzie *XX* to znacznik określający część mowy i inne kategorie gramatyczne słowa zgodnie z tagsetem NKJP[12]. Załóżmy, że interesują nas przymiotniki występujące przed różnymi formami rzeczownika *wiara*. Poniższe zapytanie wymusza dopasowanie wystąpienia przed dowolną formą tego słowa segmentu, który został oznakowany jako przymiotnik:

<tag=ADJ.> wiara***[13]

Przykładowe wyniki tego zapytania ukazuje tabela 1. Pierwszy z wyników pokazuje również, że zapytania morfosyntaktyczne mają charakter czysto pozycyjny i nie uwzględniają relacji składniowych między terminami. Na przykład słowoforma przymiotnikowa *anglosaskich* modyfikuje rzeczownik *krajach* w tym zdaniu, podczas gdy powyższe zapytanie miało raczej na celu znalezienie przymiotników modyfikujących rzeczownik *wiara*.

Tabela 1

Przykładowe wyniki zapytania *<tag=ADJ.*> wiara***

Lp.	Lewy kontekst	Dopasowanie	Prawy kontekst	Źródło
1	W krajach	anglosaskich wiarę	w Jedi zadeklarowało około 500 tysięcy osób…	gazeta.pl
2	… wyraźnie pobrzmiewa tu	archaiczna wiara	żydów, że wszystko, co dobre może…	onet.pl
3	Czym jest	autentyczna wiara	?	katolik.pl
4	Wtedy z	autentyczną wiarą	zaczęliśmy mecz, ale w miarę upływu…	weszlo.com
5	Czy można budować	autentyczną wiarę	na wzruszających piosenkach o Jezusie…	denon.pl

Składnia wyszukiwarki jest kompozycyjna, co oznacza, że jej elementy można łączyć w celu zmaksymalizowania pokrycia lub precyzji zapytania. Na przykład, po wyłączeniu opcji zachowania szyku wyrazów, zapytanie:

<tag=adv.> ściemniać***

[12] Zob. online: http://nkjp.pl/poliqarp/help/ense2.html; data dostępu: 12.08.2020.

[13] Zob. online: https://tinyurl.com/y3lnqer7; data dostępu: 12.08.2020.

zwróci zdania zawierające kombinacje przysłówków i dowolnej formy czasownika *ściemniać* w dowolnej kolejności tych terminów. Tabela 2 przedstawia przykłady zapytań o wyrazy, proste frazy i wzorce leksykalno-gramatyczne:

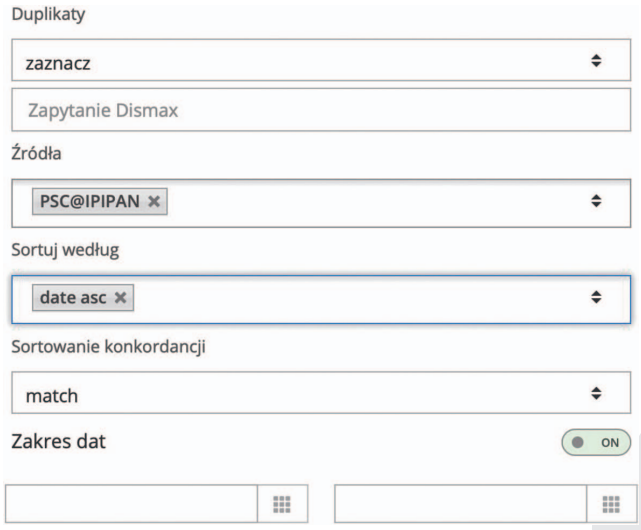
Tabela 2

Przykłady zapytań wykorzystujących różne aspekty składni wyszukiwarki MoncoPL

#	Zapytanie	Odstęp	Szyk	Uwagi
1	*brać** jak leci*	2	tak	Terminy mogą występować w odległości do 2 tokenów od siebie, np. 'brali wszystko jak leci'.
2	*stąpać**|chodzić**| po <tag=adj.*> <tag=subst.*>*	2	tak	Sekwencja *stąpać* + *po* + *przym.* + *rzeczownik*. Do 2 nieokreślonych segmentów między terminami.
3	*koń** ząb** patrzeć**|zaglądać***	4	nie	Różne warianty idiomu.
4	*<tag=fin.*>|<tag=infin.*>|<tag=praet.*> przykład***	2	tak	Wybrane formy czasownikowe (zob. tagset NKJP), po których występuje dowolna forma rzeczownika 'przykład'.
5	*brzęk** <tag=.*gen.*>*	2	tak	Dowolna forma wyrazu 'brzęk', po której występuje dowolny wyraz w dopełniaczu.

3.5. Sortowanie wyników

Wyszukiwarka obsługuje dwie niezależne metody sortowania wyników. *Sortowanie głębokie* to sortowanie wszystkich zindeksowanych zdań, które pasują do zapytania według metadanych takich, jak data publikacji i nazwa źródła. Opcja *Sortowanie konkordancji* pozwala z kolei posortować zbiór wydobytych w danym żądaniu konkordancji według pasujących fragmentów zdań lub ich bezpośrednich kontekstów. Obie opcje sortowania, a także opcje filtrowania wyników według źródeł tekstów dostępne są w rozwijanym pod polem zapytania panelu opcji zaawansowanych. Rysunek 5 ukazuje, w jaki sposób można ograniczyć wyniki wyszukiwania form rzeczownika *demokracja* do jednego źródła, tj. *Polskiego Korpusu Sejmowego*, i jednocześnie posortować je rosnąco według daty publikacji.

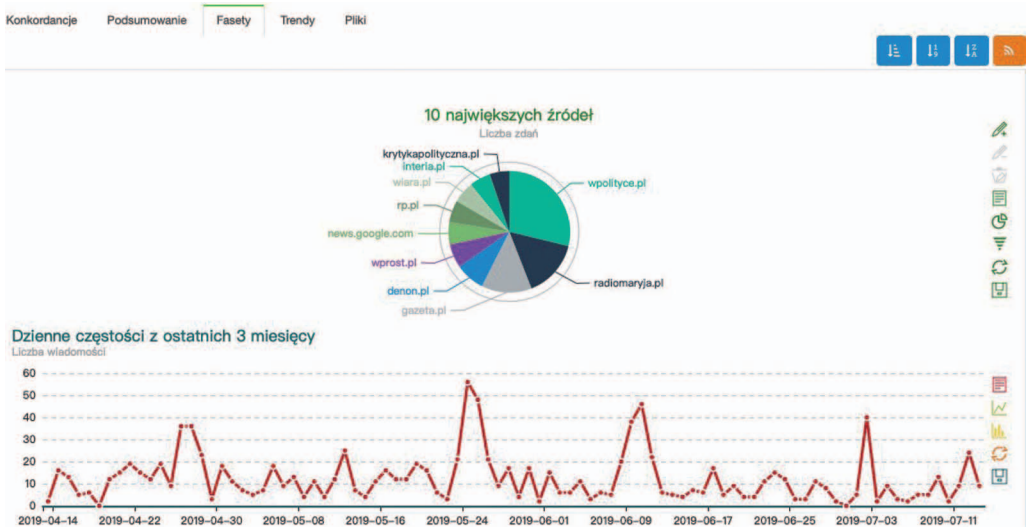


Rys. 5. Zaawansowane opcje wyszukiwania i sortowania

Nazwy źródeł są podpowiadane po wpisaniu dwóch pierwszych liter w polu *Źródła*. W formularzu można też ograniczyć zakres wyszukiwania do konkretnego przedziału czasowego.

3.6. Fasety

Dla każdego zapytania korpusowego obliczane są całkowite częstości dopasowań w różnych kategoriach metadanych. Aktualnie są to źródła oraz przedziały czasowe, w których znaleziono pasujące zdania. Wykresy z faset są dostępne w zakładce *Fasety* w widoku wyników wyszukiwania. Rysunek 6 przedstawia wykresy 10 najczęstszych źródeł dla zapytania *gender.**, które zwraca wystąpienia rzeczownika *gender* oraz jego derywaty, na przykład *genderowy*, *genderyzm* itp. Najczęściej pojawiającymi się źródłami w próbie ponad 33 tys. wystąpień tego typu form okazują się portale *wpolityce.pl* oraz *radiomaryja.pl*.



Rys. 6. Największe źródła i częstości z 3 miesięcy dla zapytania *gender**

3.7. Eksportowanie wyników

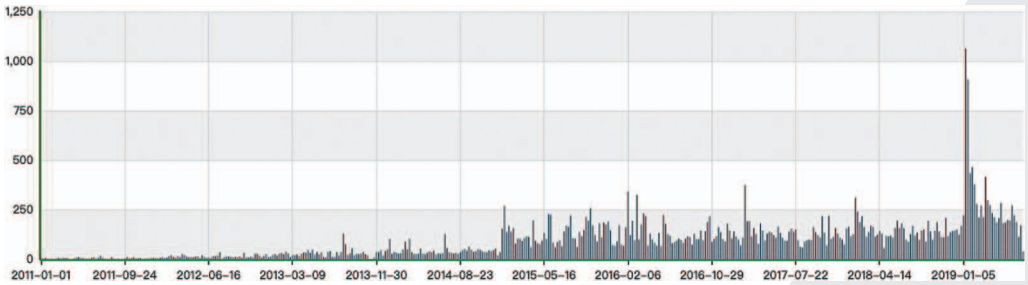
Aby umożliwić użytkownikom bardziej zaawansowaną analizę danych korpusowych, wyszukiwarka MoncoPL pozwala na pobranie w zakładce *Pliki* arkusza kalkulacyjnego zawierającego do 10 tys. wierszy konkordancji dla danego zapytania. W arkuszach wynikowych poza konkordancjami z pełnymi metadanymi znajdują się tabele z fasetami oraz dokładne informacje o rozmiarze i wersji korpusu użytego w chwili wysłania zapytania.

4. Funkcja referencyjna MoncoPL

Można zaryzykować tezę, że po zakończeniu konstrukcji NKJP w 2011 roku, MoncoPL pełni w ograniczonym stopniu funkcję dużego korpusu referencyjnego, przynajmniej w badaniach zmian językowych występujących w polszczyźnie. Zilustrujmy to prostym przykładem neosemantyzmu, jakim jest (a raczej jakiś czas temu był) czasownik *ściemniać* w znaczeniu *kłamać*/*kręcić* oraz jego rzeczownikowy derywat *ściema*. Dzięki dostępności w metada-

nych NKJP daty publikacji każdego tekstu możliwe jest znalezienie najwcześniejszych poświadczeń wystąpienia danego słowa lub frazy w tymże korpusie oraz sprawdzenia zmian w częstości ich występowania na przestrzeni kilkudziesięciu lat. Rzeczownik *ściema* po raz pierwszy pojawił się w danych NKJP w tekście opublikowanym w listopadzie 1998 roku na łamach „Zielonych Brygad" jako głowa dłuższej frazy rzeczownikowej *totalna* ***ściema*** *słupskich policjantów*. W latach 1998–2010 obserwujemy od kilku do kilkudziesięciu wystąpień tego rzeczownika rocznie. Z kolei pierwsze wystąpienie czasownika *ściemniać* w znaczeniu *kłamać/kręcić* pojawia się w NKJP w powieści beletrystycznej z 1999 roku. Można więc przypuszczać, że nowe znaczenie czasownika *ściemniać* pojawiło się na szerszą skalę najpóźniej w drugiej połowie lat 90. XX wieku. Ponieważ jednak ostatnie próbki języka włączone do NKJP pochodzą z roku 2011, w Korpusie Narodowym nie można już zaobserwować dalszych tendencji ani ilościowych, ani też jakościowych w użyciu tych dwóch słów. Z kolei w danych MoncoPL, częściowo prawdopodobnie ze względu na wysoki udział polemicznych i mocno nieformalnych komentarzy pod artykułami w ogólnej puli danych, znajdujemy ponad 10 tys. wystąpień samego rzeczownika *ściema*. Należy zaznaczyć, że sporą zaletą zrównoważonej kompozycji NKJP w tego typu badaniach jest możliwość sprawdzania stopnia przyjęcia neosemantyzmów i neologizmów we wtórnych rejestrach i typach funkcjonalnych tekstów, czyli na przykład ich poświadczeń w publicystyce i w rejestrach literackich, jeżeli ich pierwotnym rejestrem był język mówiony. Możliwości MoncoPL są w tym zakresie bardziej ograniczone.

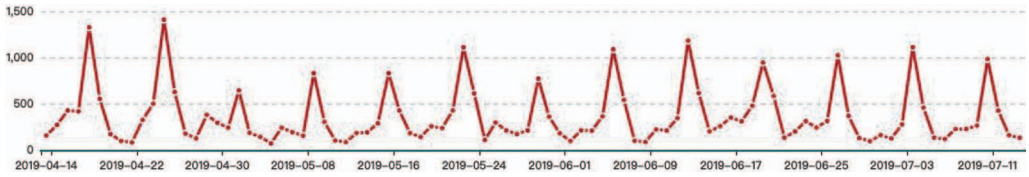
Jeszcze bardziej jaskrawym przykładem luki, jaką wypełnia MoncoPL po zamrożeniu puli tekstów NKJP, są innowacje leksykalne, które w korpusach polszczyzny powstałych przed rokiem 2011 nie były odnotowywane w ogóle lub były bardzo rzadko. Przykładem niech będzie cała rodzina wyrazów i frazemów zbudowanych na bazie morfemu *hejt*, na przykład *hejt*, *hejtować*, *hejter*, *hejterski* itp. W pełnej puli danych NKJP wyrazy takie występują sporadycznie. Rysunek 7 ukazuje stopniowy przyrost częstości ich występowania po roku 2011 aż do połowy roku 2019 na podstawie 40 tys. wystąpień w korpusie MoncoPL. Ogromną popularność wyrazu *hejt* i jego derywatów można tłumaczyć faktem, iż stanowią one swoiste autodeskryptory dyskursu internetowego znamionujące agresję słowną jako jeden z wyróżników języka internautów. Świadczą o tym typowe modyfikatory rzeczownika *hejt* zidentyfikowane za pomocą opisanego w dalszej części artykułu modułu ekstrakcji kolokacji, wśród których znajdziemy przymiotniki takie, jak *internetowy* (363 wystąpienia) *wszechobecny* (26), *zmasowany* (23), *obrzydliwy* (17), *prymitywny* (14), *pisowski* (14) oraz *lewacki* (8). Kolokacje te świadczą o tym, że *hejt* ma jednoznacznie negatywny wydźwięk i jest przypisywany różnym stronom sporów politycznych i ideologicznych.



Rys. 7. Częstości tygodniowe występowania wyrazów zaczynających się od członu *hejt-* w latach 2011–2019 w MoncoPL

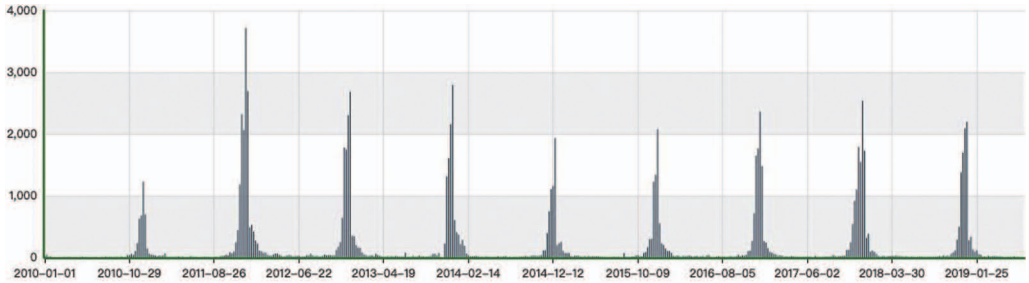
5. Cykle frekwencyjne

Jedną z unikatowych zalet dużego korpusu monitorującego z dzienną ziarnistością próbkowania danych jest możliwość badania cykli frekwencyjnych, w których zazwyczaj występują jednostki leksykalne. Jest to być może mało zbadany, ale bardzo interesujący aspekt użycia słów i skonwencjonalizowanych wyrażeń wielowyrazowych. Przykładem bardzo przewidywalnego i ustalonego cyklu frekwencyjnego jest występowanie nazw dni tygodnia w prasie codziennej i internetowych tekstach informacyjnych. Dla przykładu, rysunek 8 przedstawia częstości dzienne występowania rzeczownika *czwartek* w okresie od kwietnia do lipca 2019 roku. Regularne maksima lokalne tego szeregu czasowego można wyjaśnić tendencją do opisywania daty wydarzenia w prasie codziennej za pomocą nazwy dnia tygodnia, która jest określeniem mniej względnym niż na przykład wyrazy *dziś*, *wczoraj* lub *jutro*.



Rys. 8. Częstości dzienne występowania rzeczownika *czwartek*

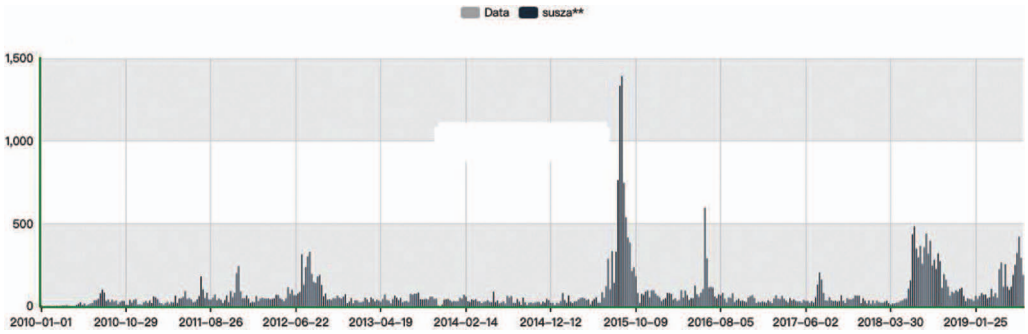
Cykle frekwencyjne wielu słów mają mniej precyzyjnie określone, ale równie niezawodnie występujące maksima. Przykładem może być słowo *choinka*, które w latach 2010–2019 osiągało maksimum częstości w okolicach świąt Bożego Narodzenia, utrzymując wyższą od przeciętnej częstość przez kilka dni. Widać to wyraźnie na rysunku 9, który został wygenerowany z konkordancji tego rzeczownika za pomocą funkcji *Trendy* wyszukiwarki MoncoPL.



Rys. 9. Częstości tygodniowe rzeczownika *choinka w* latach 2010–2019

Terminy oznaczające zjawiska pogodowe lub astronomiczne mogą mieć cykle bardzo regularne z punktowym maksimum (np. *Perseidy*), ale mogą też występować z mniej przewidywalną częstością. Na przykład największe względne częstości występowania rzeczownika *susza* obserwuje się w miesiącach letnich w latach, w których stwierdza się wystąpienie suszy rolniczej na obszarze Polski. Co ciekawe, wykres z rysunku 10 w zasadzie pokrywa się z wystąpieniami tego zjawiska w ostatnich latach, a zwłaszcza z rekordową suszą w roku 2015 i jej brakiem w roku 2017[14].

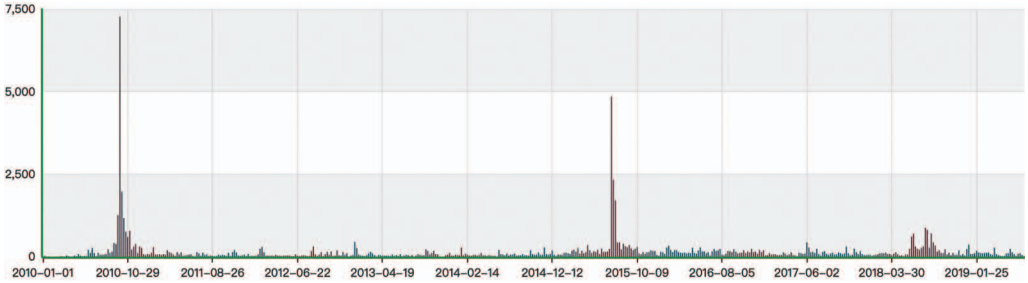
[14] Zob. online: http://www.susza.iung.pulawy.pl; data dostępu: 14.07.2020.



Rys. 10. Częstości tygodniowe występowania rzeczownika *susza* w latach 2010–2019

Chociaż ocena istotności cykliczności jako elementu opisu leksykograficznego nie jest tematem tego artykułu, to jednak warto zaznaczyć, że MoncoPL daje pewne możliwości badań cykli frekwencyjnych na przykład rzeczowników pospolitych w polszczyźnie.

Szeregi czasowe generowane z korpusu MoncoPL pozwalają również zweryfikować pewne intuicje z pogranicza językoznawstwa korpusowego i badań medioznawczych. Na przykład rysunek 11 przedstawia szereg czasowy częstości tygodniowych występowania rzeczownika *dopalacz*. Z szeregu tego wynika, że w polskim dyskursie medialnym wyróżniamy dwa lub trzy epizody szczególnego zainteresowania tematyką *dopalaczy* w ciągu ostatniej dekady, co może stanowić punkt wyjścia do dalszej analizy korpusowo-medioznawczej.



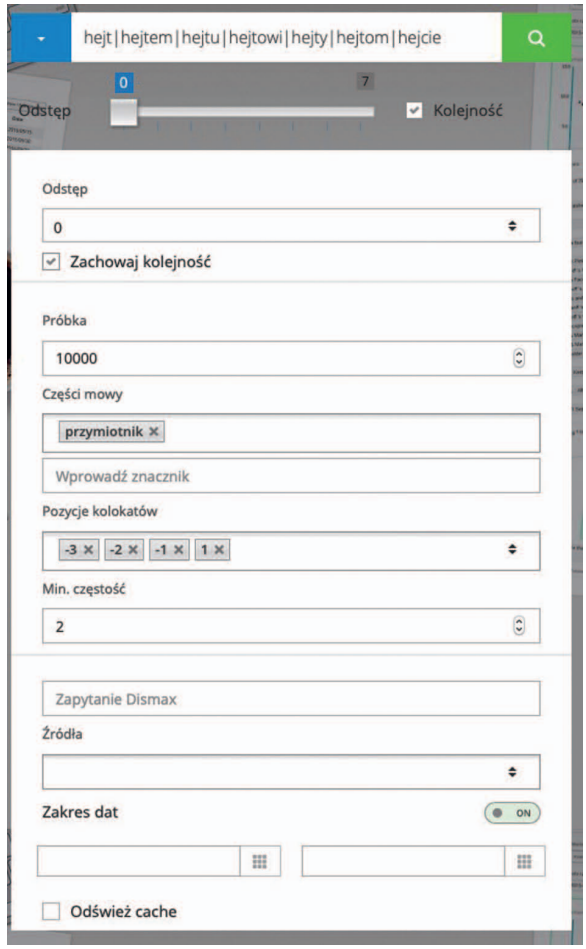
Rys. 11. Częstości tygodniowe występowania rzeczownika *dopalacz* w latach 2010–2019

6. Ekstrakcja kolokacji

Korpusowe narzędzia do ekstrakcji kolokacji, idiomów i różnych innych typów frazemów znajdują szerokie zastosowania w pracach leksykograficznych i językoznawczych. W serwisie MoncoPL moduł ekstrakcji kolokacji można znaleźć w zakładce *Narzędzia -> Kolokacje*. Wykorzystuje on pozycyjny mechanizm identyfikacji kolokacji z możliwością użycia filtrów morfosyntaktycznych. Formularz ekstrakcji kolokacji przedstawia rysunek 12. W głównym polu wyszukiwania należy wpisać zapytanie korpusowe definiujące tzw. ośrodek kolokacji w składni opisanej powyżej. W omawianym przykładzie jest to zapytanie o różne formy rzeczownika *hejt*. Parametr *Próbka* określa maksymalną liczbę zdań, z których wyszukiwarka ma ekstrahować potencjalne kolokaty (czyli regularnie współwystępujące wyrazy) dla

zadanego ośrodka kolokacji. Ograniczenie to wynika z ogromnej liczby danych, jakie musiałyby być przetwarzane dla niektórych ośrodków kolokacji w czasie rzeczywistym. Limit można zwiększać do 50 tys. wystąpień, ale wiąże się to z wydłużonym czasem oczekiwania na wynik zapytania. Dodatkowo, podczas ekstrakcji kolokacji identyfikowane i eliminowane są prawdopodobne duplikaty, czyli identyczne lub niemal identyczne zdania o zadanej minimalnej długości. Po zakończeniu ekstrakcji wyszukiwarka podaje liczbę unikatowych kontekstów (zdań), których użyto do wygenerowania listy kolokacji.

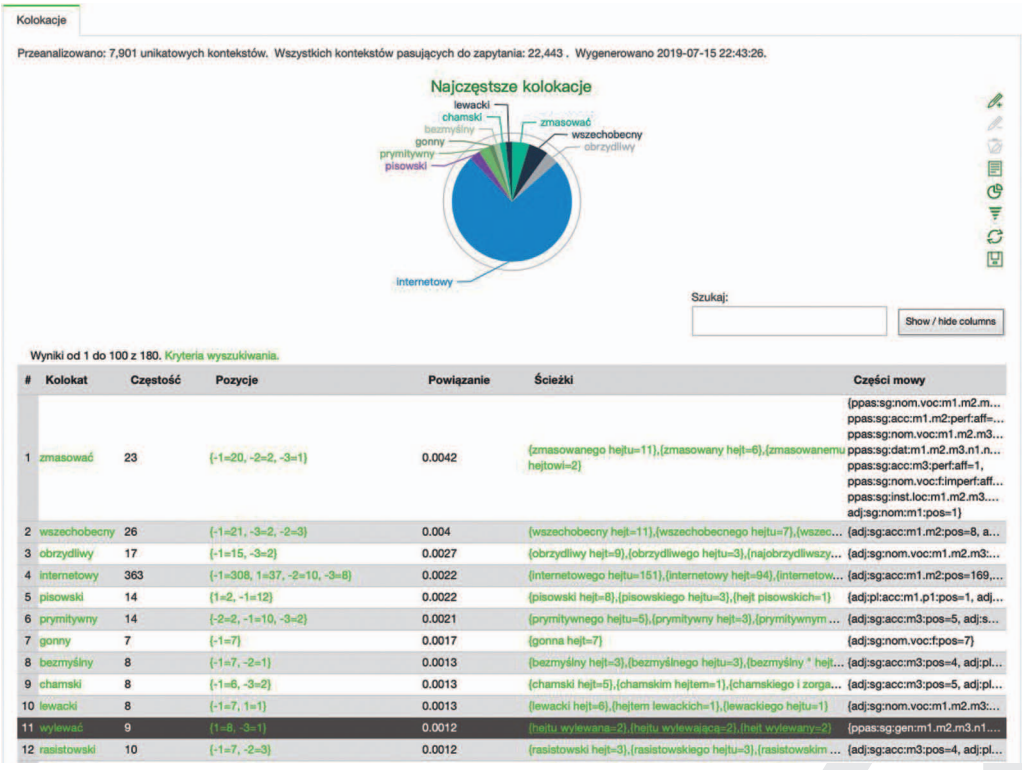
W polu *Części mowy* można określić klasy części mowy kolokatów, które mają być brane pod uwagę w procesie ekstrakcji. Z kolei w polu *Pozycje kolokatów* określamy maksymalną odległość pozycyjną potencjalnych kolokatów od ośrodka kolokacji. W przykładzie z rysunku 12 pod uwagę brane są tylko przymiotniki występujące na pozycjach od –3 do +1 w odniesieniu do pozycji rzeczownika *hejt*. Tego rodzaju ograniczenia zwiększają prawdopodobieństwo wystąpienia relacji składniowych między ośrodkiem kolokacji i współwystępującymi z nim wyrazami. W przykładowym zapytaniu zwiększamy szansę na zaobserwowanie modyfikatorów przymiotnikowych rzeczownika *hejt*.



Rys. 12. Formularz modułu ekstrakcji kolokacji

Opcja minimalnej częstości pozwala odciąć okazjonalne współwystąpienia. W polu *Źródła* można określić nazwy źródeł, z których mają być ekstrahowane kolokacje, dzięki czemu możliwe jest identyfikowanie nacechowania pewnych słów i terminów w niektórych serwisach. Dzięki tej opcji dowiadujemy się na przykład, że wśród kolokatów przymiotnikowych rzeczownika *ideologia* występujących w portalu *gazeta.pl* stosunkowo często występują przymiotniki *faszystowska*, *zbrodnicza* (głównie w odniesieniu do nazizmu), *nazistowska* czy *neonazistowska*. Z kolei w portalu *wpolityce.pl* typowymi modyfikatorami przymiotnikowymi rzeczownika *ideologia* są wyrazy: *zbrodnicza* (głównie w odniesieniu do komunizmu), *lewacka*, *marksistowska* i *genderowa*. Może to świadczyć o wyraźnie negatywnej prozodii semantycznej rzeczownika *ideologia*, który używany jest do relatywizacji i deprecjacji konkurencyjnego systemu przekonań.

Formularz wyszukiwania kolokacji umożliwia zawężenie zakresu dat publikacji tekstów, w których występują potencjalne kolokacje. I tak na przykład wśród najczęściej występujących w 2011 roku kolokatów rzeczownika *fala* znajdziemy dopełniaczowe formy rzeczowników *tsunami* i *protesty*[15]. To samo zapytanie z zakresem dat ograniczonym do roku 2018 jako jedno z najczęstszych współwystąpień *fali* zwraca rzeczownik *hejt(u)*[16]. Tego typu zestawienia mogą być przydatne w identyfikacji dominujących lub wręcz powstających w danym czasie sensów wyrazów wieloznacznych.



#	Kolokat	Częstość	Pozycje	Powiązanie	Ścieżki	Części mowy
1	zmasować	23	{-1=20, -2=2, -3=1}	0.0042	{zmasowanego hejtu=11},{zmasowany hejt=6},{zmasowanemu hejtowi=2}	{ppas:sg:nom.voc:m1.m2.m... ppas:sg:acc:m1.m2:perf:aff=... ppas:sg:nom.voc:m1.m2.m3... ppas:sg:dat:m1.m2.m3.n1.n... ppas:sg:acc:m3:perf:aff=1, ppas:sg:nom.voc:f:imperf:aff... ppas:sg:inst.loc:m1.m2.m3.... adj:sg:nom:m1:pos=1}
2	wszechobecny	26	{-1=21, -3=2, -2=3}	0.004	{wszechobecny hejt=11},{wszechobecnego hejtu=7},{wszec...	{adj:sg:acc:m1.m2:pos=8, a...
3	obrzydliwy	17	{-1=15, -3=2}	0.0027	{obrzydliwy hejt=9},{obrzydliwego hejtu=3},{najobrzydliwszy...	{adj:sg:nom.voc:m1.m2.m3:...
4	internetowy	363	{-1=308, 1=37, -2=10, -3=8}	0.0022	{internetowego hejtu=151},{internetowy hejt=94},{internetow...	{adj:sg:acc:m1.m2:pos=169,...
5	pisowski	14	{1=2, -1=12}	0.0022	{pisowski hejt=8},{pisowskiego hejtu=3},{hejt pisowskich=1}	{adj:pl:acc:m1.p1:pos=1, adj...
6	prymitywny	14	{-2=2, -1=10, -3=2}	0.0021	{prymitywnego hejtu=5},{prymitywny hejt=3},{prymitywnym ...	{adj:sg:acc:m3:pos=5, adj:s...
7	gonny	7	{-1=7}	0.0017	{gonna hejt=7}	{adj:sg:nom.voc:f:pos=7}
8	bezmyślny	8	{-1=7, -2=1}	0.0013	{bezmyślny hejt=3},{bezmyślnego hejtu=3},{bezmyślny * hejt...	{adj:sg:acc:m3:pos=4, adj:pl...
9	chamski	8	{-1=6, -3=2}	0.0013	{chamski hejt=5},{chamskim hejtem=1},{chamskiego i zorga...	{adj:sg:acc:m3:pos=5, adj:pl...
10	lewacki	8	{-1=7, 1=1}	0.0013	{lewacki hejt=6},{hejtem lewackich=1},{lewackiego hejtu=1}	{adj:sg:nom.voc:m1.m2.m3:...
11	wylewać	9	{1=8, -3=1}	0.0012	{hejtu wylewana=2},{hejtu wylewającą=2},{hejt wylewany=2}	{ppas:sg:gen:m1.m2.m3.n1....
12	rasistowski	10	{-1=7, -2=3}	0.0012	{rasistowski hejt=3},{rasistowskiego hejtu=3},{rasistowskim ...	{adj:sg:acc:m3:pos=4, adj:pl...

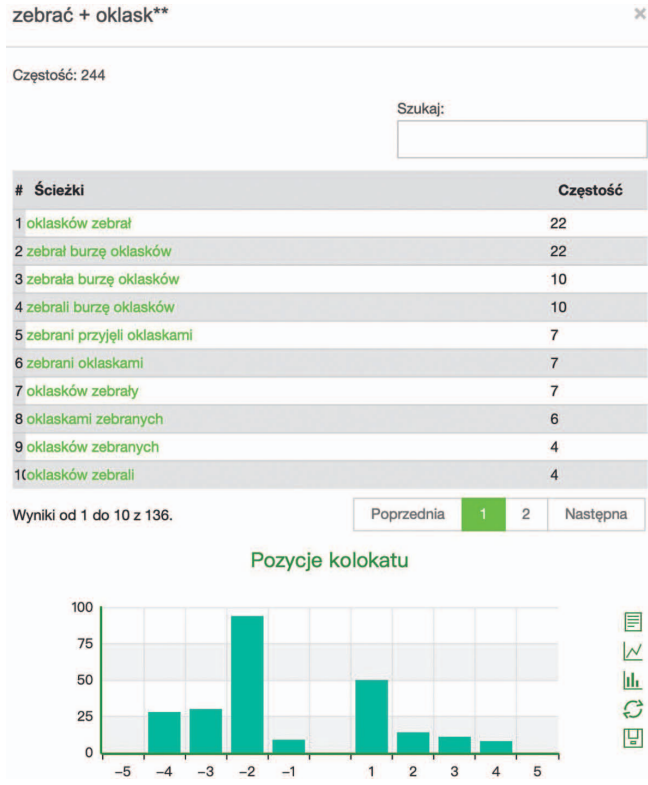
Rys. 13. Wyniki ekstrakcji kolokacji przymiotnikowych rzeczownika *hejt*

[15] Zob. online: https://tinyurl.com/y33bxogw; data dostępu: 14.07.2020.

[16] Zob. online: https://tinyurl.com/y4qc9tds; data dostępu: 14.07.2020.

Rysunek 13 przedstawia główny ekran wyników ekstrakcji kolokacji. U góry tego widoku podana jest informacja o wielkości zdeduplikowanej próby użytej w zapytaniu, a także o całkowitej liczbie wystąpień ośrodka kolokacji w korpusie. Nieco niżej znajduje się wykres kołowy obrazujący rozkład frekwencji 10 najczęstszych kolokacji zaobserwowanych w przeanalizowanej próbie zdań. Potencjalne kolokacje ukazane są w zestawieniu pod wykresem. W przykładowym zestawieniu z rysunku 13 znajduje się łącznie 180 potencjalnych kolokacji przymiotnikowych rzeczownika *hejt*, które w próbie 7901 zdań wystąpiły co najmniej dwukrotnie. Domyślnie wyniki posortowane są malejąco według prostej miary powiązania opartej na współczynniku Dice'a, która promuje nadspodziewanie częste współwystąpienia wyrazów. Klikając na nagłówki kolumn *Częstość* i *Powiązanie*, można zmieniać odpowiednio kolejność wyników w tabeli.

W kolumnie *Pozycje* podane są częstości występowania pozycji, na których wystąpił dany kolokat. W omawianym przykładzie pozwala to stwierdzić, że różne formy przymiotnika *wszechobecny* wystąpiły tylko w prepozycji do rzeczownika *hejt*, w tym 21 razy bezpośrednio przed nim, dwukrotnie jako drugie i raz jako trzecie słowo przed ośrodkiem kolokacji. W kolumnie *Ścieżki* podane są częstości ciągów słowoform (tzw. n-gramów wyrazowych) tworzących kolokacje wraz z wyrazami występującymi między nimi. Takie listy n-gramów pozwalają zaobserwować w niektórych przypadkach typowe formy gramatyczne składników kolokacji, a czasami również dodatkowe wyrazy tworzące dłuższe związki



Rys. 14. Widok szczegółowy n-gramów wyrazowych tworzących kolokację *zebrać + oklaski*

frazeologiczne. Na przykład na liście czasownikowych kolokatów rzeczownika *oklaski* znajdziemy formy dokonane czasownika *zebrać*. Po kliknięciu w ten wynik otwiera się okno dialogowe podobne do tego z rysunku 14, z którego wynika, że ponad 25% (22 + 22 + 10 + 10 z 244) współwystąpień czasownika *zebrać* z rzeczownikiem *oklaski* to kombinacje uwięzione w nieco dłuższej konstrukcji idiomatycznej *zebrać burzę oklasków*. Dodatkowo w oknie dialogowym ścieżek wyświetlany jest wykres pozycji kolokatu w stosunku do ośrodka kolokacji.

Warto zauważyć, że ośrodkiem wyszukiwanych kolokacji może być wyrażenie wielowyrazowe. Przykładem ilustrującym zasadność próby ekstrakcji kolokacji wyrażenia wielowyrazowego może być zapytanie o kolokaty frazy *parasol ochronny*[17], które zwraca wyniki ujęte częściowo w tabeli 3. Na uwagę zasługuje stosunkowo duża dowolność w konstrukcji idiomów opartych na metaforze *roztaczania parasola ochronnego* nad kimś lub czymś. Co ciekawe, poza czasownikiem *roztaczać* w wynikach pojawiają się bardzo często formy synonimicznych czasowników, takich jak *rozpiąć*, *rozpostrzeć*, *rozciągnąć* czy też *rozłożyć*. Tego typu listy okazują się pomocne w weryfikacji wariantywności wielowyrazowego idiomu, co ma pewnie praktyczne zastosowania na przykład w redakcji frazeostylistycznej tekstów.

Tabela 3

Kolokaty czasownikowe frazy *parasol ochronny*

#	Kolokat	Częstość występowania	Powiązanie
1	roztoczyć	138	0.077
2	roztaczać	185	0.0635
3	rozpiąć	94	0.0347
4	rozpostrzeć	61	0.0333
5	rozpinać	18	0.0103
6	rozciągnąć	40	0.0093
7	zwinąć	27	0.0065
8	rozpościerać	14	0.0065
9	zwijać	25	0.0061
10	rozłożyć	63	0.0041

7. Implementacja

Na koniec warto wspomnieć o rozwiązaniach technicznych, bez których nie byłoby możliwe wydajne aktualizowanie i serwowanie tak dużego korpusu monitorującego. Mechanizm wyszukiwania MoncoPL oparty jest na technologii Apache Solr. Format indeksu został dostosowany do potrzeb omówionej tu korpusowej składni zapytań. W szczególności, każdy segment wyrazowy jest reprezentowany w indeksie jako trójelementowa krotka złożona ze słowoformy, lematu i znacznika morfosyntaktycznego. Zapytania w składni MoncoPL są tłumaczone na wyrażenia regularne odpowiadające segmentom słów wraz z ich znacz-

[17] Zob. online: https://tinyurl.com/y39ves7t/; data dostępu: 12.06.2020.

nikami, a następnie mapowane na składnię zapytań logicznych Apache Solr. Zastosowane rozwiązanie jest skalowalne horyzontalnie, dzięki czemu możliwe jest rozproszenie indeksu wyszukiwarki na dodatkowe maszyny na odrębnych węzłach.

8. Plany utrzymania korpusu

Celem podjętym w tym artykule było omówienie budowy i funkcji, ale również wykazanie zasadności utrzymywania dużego korpusu monitorującego polszczyzny, jakim jest w pewnym zakresie MoncoPL. Aktualizacja tego korpusu jest planowana przynajmniej do czasu rozpoczęcia prac nad kolejną wersją NKJP. W przekonaniu autora, docelowo to właśnie korpus narodowy powinien pełnić funkcję monitorującą, również dla tych rejestrów i odmian polszczyzny, których odpowiednia reprezentacja wymaga znacznych nakładów i zasobów.

Literatura

Davies M., 2010: *The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English*. "Literary and Linguistic Computing" XXV, No. 4, s. 447–465. DOI: 10.1093/llc/fqq018.

Duda B., Lisczyk K., 2018: *Narzędzia cyfrowe w polonistycznej dydaktyce akademickiej – zastosowania, możliwości, perspektywy*. „Forum Lingwistyczne" nr 5, s. 143–154.

Ogrodniczuk M., 2018: *Polish Parliamentary Corpus*. In: Fišer D., Eskevich M., Jong F. de, eds.: *Proceedings of the LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*. Paris, s. 15–19.

Przepiórkowski i in., 2009: Przepiórkowski A., Górski R.L., Łaziński M., Pęzik P.: *Recent Developments in the National Corpus of Polish*. "NLP, Corpus Linguistics, Corpus Based Grammar Research", s. 302–309.

Sinclair J., 1996: *EAGLES Guidelines*. Expert Advisory Group on Language Engineering Standards [online: http://www.ilc.cnr.it/EAGLES96/browse.html; data dostępu: 30.06.2020].

Woliński M., 2014: *Morfeusz Reloaded*. In: Calzolari N., Choukri Kh., Declerck Th., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S., eds.: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC. Reykjavík, s. 1106–1111.