WITOLD KIERAŚ[Ia], MAŁGORZATA MARCINIAK[IIa], MAREK ŁAZIŃSKI[IIIa,b], MARCIN WOLIŃSKI[IVa], KRYSTYNA BOJAŁKOWSKA[Vc], WIKTOR EŹLAKOWSKI[VIa], ŁUKASZ KOBYLIŃSKI[VIIa], DOROTA KOMOSIŃSKA[VIIIa], KATARZYNA KRASNOWSKA-KIERAŚ[IXa], MICHAŁ RUDOLF[Xa], ALEKSANDRA TOMASZEWSKA[XIa], JOANNA WOŁOSZYN[XIIa], NATALIA ZAWADZKA-PALUEKTAU[XIIIa]

[a] INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK, WARSZAWA

[b] UNIWERSYTET WARSZAWSKI

[c] UNIWERSYTET MIKOŁAJA KOPERNIKA W TORUNIU

Korpus Współczesnego Języka Polskiego. Dekada 2011–2020

Słowa kluczowe: Korpus Współczesnego Języka Polskiego, znakowanie lingwistyczne, rozbiory składniowe.

doi: https://doi.org/10.31286/JP.001055

1. Od pierwszych korpusów referencyjnych w Polsce do Korpusu Współczesnego Języka Polskiego

W Polsce zaczęto budować duże referencyjne korpusy językowe już w latach 90. XX wieku. Wtedy powstały Korpus PELCRA Uniwersytetu Łódzkiego, korpus stworzony przez instytuty Polskiej Akademii Nauk (Przepiórkowski 2004) oraz korpus Redakcji Słowników Wydawnictwa Naukowego PWN. Każdy z wymienionych zasobów zawierał ponad sto milionów słów. W 2007 roku cztery instytucje utworzyły konsorcjum Narodowego Korpusu Języka Polskiego (dalej: NKJP) (Przepiórkowski i in. (red.) 2012) pod przewodnictwem Instytutu Podstaw Informatyki PAN. Jako efekt kilkuletniego projektu w roku 2011 udostępniono korpus

[I] wkieras@ipipan.waw.pl; ORCID: 0000-0002-8062-5881
[II] malgorzata.marciniak@ipipan.waw.pl; ORCID: 0000-0002-0953-758X
[III] marek.lazinski@ipipan.waw.pl; ORCID: 0000-0001-5718-4435
[IV] marcin.wolinski@ipipan.waw.pl; ORCID: 0000-0002-7498-1484
[V] Krystyna.Bojalkowska@umk.pl; ORCID: 0000-0001-5672-4751
[VI] wiktor.ezlakowski@ipipan.waw.pl; ORCID: 0000-0002-5107-8890
[VII] lukasz.kobylinski@ipipan.waw.pl; ORCID: 0000-0003-2462-0020
[VIII] dorota.komosinska@ipipan.waw.pl; ORCID: 0000-0002-2611-1214
[IX] katarzyna.krasnowska@ipipan.waw.pl; ORCID: 0000-0002-7052-0568
[X] michal@rudolf.waw.pl; ORCID: 0000-0002-3115-9087
[XI] aleksandra.tomaszewska@ipipan.waw.pl; ORCID: 0000-0001-6379-3034
[XII] joanna.woloszyn@ipipan.waw.pl; ORCID: 0000-0002-8923-414X
[XIII] natalia.zawadzka-paluektau@ipipan.waw.pl; ORCID: 0000-0003-4969-2039

zrównoważony o wielkości 300 milionów słów oraz wielki korpus niezrównoważony o wielkości 1,8 miliarda słów wraz z dwiema wyszukiwarkami: Poliqarp (przygotowaną przez IPI PAN) oraz PELCRA (powstałą na UŁ). Pierwsza z wymienionych umożliwia dokładne wyszukiwanie morfosyntaktyczne z klasyfikacją i dezambiguacją (ujednoznacznieniem) form fleksyjnych oraz części mowy, druga pozwala na szybkie, łatwe, intuicyjne wyszukiwanie słów i leksemów, kolokacji oraz szeregów czasowych.

NKJP miał być w założeniu aktualizowany i rozwijany, o czym mówiono wprost w poświęconej mu publikacji: „Aby pozostał największym narzędziem referencyjnym dla rzeszy użytkowników, NKJP musi być ustawicznie uzupełniany, ulepszany, nadzorowany i monitorowany" (Lewandowska-Tomaszczyk i in. 2012: 9). Niestety z różnych przyczyn (głównie formalnych i finansowych) ta wizja się nie ziściła i przygotowany w ramach NKJP zbiór, obejmujący teksty powstałe do 2010 roku, „od zakończenia prac trwa w postaci zamrożonej" (Ogrodniczuk 2017: 24). Przez kolejne lata jedynym korpusem, w którym można było śledzić najnowsze słownictwo i zjawiska językowe, był korpus monitorujący tekstów internetowych, w tym serwisów prasowych, MoncoPL – korpus aktualny, ogromny i wciąż rosnący, ale niezrównoważony gatunkowo i z ograniczonymi możliwościami wyszukiwania gramatycznego (Pęzik 2020).

Przygotowania korpusu pokazującego stan polszczyzny współczesnej podjął się w roku 2021 IPI PAN. Zadanie udało się zrealizować w ramach projektu POIR.04.02.00-00-D006/20 – Cyfrowa infrastruktura badawcza dla humanistyki i nauk o sztuce DARIAH-PL, a jego wynikiem jest Korpus Współczesnego Języka Polskiego obejmujący teksty z lat 2011–2020 (dalej: KWJP). Po trzech latach intensywnego pozyskiwania tekstów wśród wydawców książek i prasy korpus został udostępniony pod koniec 2023 roku. Jego zasoby podzielone są na dwie rozłączne części: korpus zrównoważony (100 mln słów) oraz korpus niezrównoważony (ponad 1,4 mld słów), ten drugi jest dostępny dla zarejestrowanych użytkowników. Dodatkowo korpus niezrównoważony jest podzielony na podkorpusy ze względu na klasyfikację zawartych w nich tekstów.

KWJP może być w pewnym sensie traktowany jako kontynuacja NKJP, ponieważ obejmuje następny odcinek czasowy w procesie monitorowania rozwoju polszczyzny od początku XX wieku. Jest to jednak niezależny korpus, który pokazuje polszczyznę kolejnego dziesięciolecia, a w procesie zrównoważenia uwzględnia specyfikę historyczną tego okresu (zob. rozdz. 2). Zakładamy, że cechy wyszukiwanych słów – ich częstość, kolokacje i konotacje – są takie jak u przeciętnych użytkowników języka.

2. Struktura Korpusu Współczesnego Języka Polskiego

2.1. Dobór tekstów

Dobór tekstów w korpusie jest zawsze kompromisem między tym, co twórcy korpusu chcieliby w nim umieścić, a tym, co faktycznie mogą umieścić. Trzeba jednak podkreślić, że w znacznej większości wypadków spotkaliśmy się z dużą życzliwością i zrozumieniem ze strony wydawców prasy i książek, jak również indywidualnych autorów. Wśród darczyńców korpusu (lista znajduje się na stronie w zakładce „Teksty") są zarówno najwięksi wydawcy książek i prasy w Polsce, jak i niewielkie oficyny wydające książki lub prasę z określonych nisz tematycznych

lub regionalnych. Dzięki temu możemy stwierdzić, że KWJP dość dobrze odzwierciedla zróżnicowanie polskiego piśmiennictwa dekady 2011–2020.

KWJP zawiera mniejszy zakres typów tekstów niż NKJP. Podjęliśmy decyzję o gromadzeniu tekstów podlegających redakcji, co spowodowało wyłączenie z korpusu tekstów mówionych oraz internetowych (tego rodzaju zasoby stanowią ok. 15 procent NKJP). Argumentem za takim rozwiązaniem jest odrębność tych odmian języka, pociągająca za sobą konieczność opracowania niezależnych zasobów i potoków przetwarzania. Do KWJP nie włączono więc tekstów *stricte* internetowych – forów, blogów, wpisów z sieci społecznościowych itp. W NKJP mowa jest reprezentowana przez stosunkowo niewielki podkorpus konwersacji (wyłącznie transkrypcje) oraz znacznie większy podkorpus tekstów określanych jako quasi-mówione, będących przede wszystkim zbiorem zapisów obrad Sejmu, Senatu oraz parlamentarnych komisji śledczych. W latach 2011–2018 w IPI PAN przygotowano obszerny specjalistyczny Korpus Dyskursu Parlamentarnego (Ogrodniczuk 2018) gromadzący wszystkie transkrypcje obrad obu izb parlamentu RP, poczynając od 1919 roku. Nie zdecydowaliśmy się na włączenie zawartych w nim transkrypcji pochodzących z lat 2011–2020 do KWJP, by nie powielać zasobów, które są już dostępne. Jednocześnie, w ramach tej samej infrastruktury DARIAH-PL, w IJP PAN powstał korpus języka mówionego z tego okresu[1], który zawiera ponad tysiąc godzin nagrań dźwiękowych (głównie pochodzących z serwisów internetowych) wraz z ich transkrypcją.

Ze względu na to, że zakres czasowy tekstów w KWJP obejmuje jedynie ubiegłą dekadę, pozyskiwaliśmy teksty wyłącznie w postaci elektronicznej, co jednak nie znaczy, że ich opracowanie sprowadzało się tylko do automatycznej konwersji. Część tekstów dostępna była jedynie w postaci plików PDF, co wymagało półautomatycznej konwersji i ręcznej korekty (w wypadku książek) lub całkowicie ręcznego opracowania (prasa). Do korpusu włączono tylko pełne teksty książek napisanych oryginalnie po polsku i wydanych pierwotnie w dekadzie 2011–2020, choć nie wszystkie książki pozyskano w postaci pierwszych wydań. Wtedy w metadanych odnotowywana jest zarówno data pierwszego wydania, jak i data wydania włączonego do korpusu. Druga z nich może wskazywać na datę po roku 2020, ale są to przypadki rzadkie. Jeśli chodzi o prasę, staraliśmy się pozyskiwać całe numery, lecz postać pozyskanych danych nie zawsze na to pozwalała (problem ten dotyczy jedynie niewielkiej części tytułów prasowych). Numery pozyskane do korpusu zrównoważonego wybierano „półlosowo", tzn. losowo w obrębie poszczególnych lat reprezentowanych w archiwum danego periodyku. Chodziło o to, by możliwie jak najbardziej zróżnicować tematykę tekstów w obrębie danego tytułu prasowego. Przy opracowywaniu numerów czasopism pomijano stopki redakcyjne, ogłoszenia, reklamy, podpisy pod zdjęciami i ilustracjami, bardzo krótkie (najczęściej jednozdaniowe) notki, a także teksty tłumaczone (jeśli dało się je zidentyfikować).

Zbierając prasę regionalną, dbaliśmy o to, by pochodziła ze wszystkich regionów Polski. Pozyskaliśmy 130 różnych tytułów z 44 miast. Staraliśmy się również zapewnić równowagę

1 https://korpusmowy.ijp.pan.pl/.

płci w korpusie zrównoważonym. Włączono do niego 291 książek napisanych przez kobiety (23 mln słów) i 379 napisanych przez mężczyzn (29 mln słów). W tej statystyce pomijamy prace zbiorowe.

2.2. Klasyfikacje

Korpus zrównoważony składa się niemal wyłącznie z książek i tekstów prasowych. Podobnie jak w NKJP każdy tekst ma przypisaną wartość klasyfikacji gatunkowej oraz kanału dystrybucji, wartości tych klasyfikacji są jednak nieco inne.

Po pierwsze, klasyfikacja gatunkowa została wyraźnie uproszczona i składają się na nią trzy wartości: fikcja, fakt i publicystyka. Na fikcję składają się przede wszystkim książki literackie (powieści i zbiory opowiadań) reprezentujące możliwie różne gatunki oraz (w ograniczonym zakresie) czasopisma literackie, publikujące w większości opowiadania. Gatunek ten odpowiada dość dobrze tekstom oznaczonym jako *lit* w klasyfikacji NKJP. Gatunek faktu skupia szeroko pojęte teksty niefikcjonalne – od reportaży, przez dzienniki, biografie, poradniki, aż po teksty popularnonaukowe i naukowe, a także urzędowe, a zatem teksty, które w NKJP zebrano pod kilkoma wartościami klasyfikacji: *fakt* (typowa literatura faktu), *nd* (teksty naukowo-dydaktyczne), *urzed* (teksty urzędowe), *inf-por* (teksty informacyjno-poradnikowe). Inaczej niż w NKJP do kategorii fakt zaliczamy też prasę tematyczną, np. ekonomiczną, sportową, popularnonaukową, pszczelarską. Kryterium wyboru nie jest przy tym zgodność z faktami lub naukowym stanem wiedzy, ale domniemana intencja autorów tekstów, dlatego do tego gatunku zostały również zaliczone czasopisma z zakresu astrologii czy zjawisk paranormalnych. Wreszcie gatunek publicystyka skupia wyłącznie typową prasę informacyjno-publicystyczną (głównie dzienniki i tygodniki). Tak określone trzy ogólne gatunki tekstów stanowią odpowiednio 30 procent (fikcja), 35 procent (fakt) i 35 procent (publicystyka) docelowego korpusu zrównoważonego. Taki bliski równemu rozkład tekstów między gatunkami sprawia, że łatwo można oszacować, czy wyszukiwane przez użytkownika słowo lub wyrażenie jest nadreprezentowane w którymś z gatunków; do celów badawczych warto wszakże posiłkować się informacją o częstości względnej w poszczególnych podkorpusach.

Po drugie, klasyfikacja kanału dystrybucji ograniczona jest do dwóch typów: książek i prasy, przy czym kanał prasowy podzielony jest na dzienniki, tygodniki, miesięczniki oraz inne. Jest tylko jedna niewielka grupa tekstów, którym przypisano jako wartość kanału dystrybucji *internet*. To próbka orzeczeń sądów różnych instancji reprezentujących jeden z typów tekstów urzędowych. W pozostałych wypadkach w podziale na kanał książkowy i prasowy kierowaliśmy się klasyfikacją biblioteczną, czyli numerami ISBN (książki) i ISSN (prasa), nawet wówczas, gdy dana publikacja ma postać wyłącznie elektroniczną. Według klasyfikacji kanału dystrybucji książki stanowią blisko 55 procent korpusu zrównoważonego, prasa zaś – nieco ponad 45 procent (dzienniki niespełna 19 procent, tygodniki nieco ponad 12 procent, miesięczniki 9,5 procent, pozostałe 5,5 procent). Zaklasyfikowane do kanału internetowego orzeczenia sądów powszechnych stanowią 0,3 procent korpusu.

Warto też dodać, że książki w korpusie mają dodatkową klasyfikację tematyczną, którą stanowi bardzo krótki opis wyjaśniający, z jakiego gatunku literackiego (np. kryminał, science fiction, reportaż, biografia) lub z jakiej dziedziny (historia, socjologia, pszczelarstwo) pochodzą konkordancje wyszukane w wyniku danego zapytania. Przypomina ona nieco klasyfikację gatunkową w katalogu Biblioteki Narodowej i spełnia w korpusie podobną funkcję informacyjną. Wartości z tej klasyfikacji nie są jednak zamkniętym zbiorem, w związku z czym grupowanie na jej podstawie wyników zapytań korpusowych nie ma większego sensu.

2.3. Korpus niezrównoważony

Podobnie jak w NKJP oprócz korpusu zrównoważonego udostępniamy również korpus niezrównoważony, czyli oportunistyczny, skupiający większość danych, które udało się nam zgromadzić w trakcie budowania KWJP. Tego typu korpusy mogą być użyteczne w badaniach słownictwa rzadkiego lub nowego, przydają się też m.in. leksykografom jako źródło przykładów użyć konkretnych słów i wyrażeń. Nie powinny być wszakże w większości wypadków źródłem informacji statystycznych do badań ilościowych.

Przytłaczającą część tego zbioru stanowi ogólnopolska prasa codzienna i tygodniowa. Z powodów praktycznych korpus ten został podzielony na pięć części ze względu na kryterium gatunku i kanału dystrybucji tekstów. Pierwsza część to korpus fikcji i faktu (łącznie ok. 200 mln segmentów), na który składają się przede wszystkim czasopisma tematyczne i literackie, w znacznie zaś mniejszym stopniu książki. Pozostałe dane stanowi publicystyka prasowa podzielona na cztery podkorpusy. Największy z nich to korpus dzienników ogólnopolskich (ok. 600 mln segmentów), kolejny to korpus dzienników regionalnych (ok. 290 mln segmentów), wreszcie korpus tygodników (ok. 200 mln segmentów) i innych periodyków (35 mln segmentów). Całkowity rozmiar korpusu niezrównoważonego przekracza 1,4 miliarda segmentów, czyli jest on ponad czternaście razy większy od korpusu zrównoważonego. Inaczej niż w NKJP korpusy zrównoważony i niezrównoważony są rozłączne, można jednak wyszukiwać w zbiorze łączącym wszystkie korpusy.

2.4. Próbka losowa wielkości pół miliona słów KWJP½M

Wzorując się na zasobach udostępnionych w bazie NKJP, przygotowaliśmy też korpus krótkich próbek o nazwie KWJP½M. Z korpusu zrównoważonego wylosowane zostały fragmenty o długości 40–60 słów w taki sposób, by każda książka i tytuł prasowy były reprezentowane w przybliżeniu w takiej samej proporcji jak w oryginalnym korpusie. Korpus ten w zamyśle jest odpowiednikiem korpusu NKJP1M, choć nie zawiera żadnej warstwy znakowania (ręcznego czy automatycznego), a jedynie metadane. Dane uwzględnione w KWJP½M mogą służyć do celów dydaktycznych lub badań niewymagających dostępu do pełnych tekstów. Omawiany podkorpus został umieszczony w repozytorium[2] w postaci plików JSON zawierających pełne metadane źródła oraz listę wylosowanych z niego próbek.

2 https://github.com/ipipan/kwjp100-varia/.

3. Wielopoziomowe znakowanie korpusu

Użyteczność korpusu językowego dla badaczy zależy od zakresu wprowadzonego doń znakowania lingwistycznego oraz od możliwości odwoływania się do tego znakowania w wyszukiwarce korpusowej. Teksty objęte prawami autorskimi nie mogą być udostępniane w całości, więc wyszukiwarka korpusowa pozostaje jedyną publicznie dostępną formą interakcji z tekstami korpusu.

Najpowszechniej wykorzystywanym elementem znakowania gramatycznego jest hasłowanie (lematyzacja) tekstu, pozwalające szukać w tekstach dowolnych form odmiany zadanych leksemów. Jednak informacja dodana w KWJP jest dużo bogatsza. W korpusie uwzględniono następujące typy informacji (tzw. warstwy znakowania lingwistycznego):

- segmentację,
- hasłowanie i znakowanie morfosyntaktyczne,
- jednostki nazewnicze,
- rozbiory składniowe.

W zrównoważonym korpusie NKJP dostępne były do wyszukiwania dwie pierwsze warstwy. W jego odświeżonej wersji (Kieraś i in. 2021) uzupełniono znakowanie o oznaczenie jednostek nazewniczych oraz drzewa zależnościowe. Elementem całkowicie nowym w KWJP są hybrydowe rozbiory składniowe, łączące informację zależnościową i składnikową.

3.1. Segmentacja

Pierwszym etapem interpretacji językoznawczej tekstu jest podział na zdania i segmenty. Jest to przetwarzanie tak podstawowe, że często się o nim zapomina, myśląc o znakowanym korpusie. Lecz w wypadku automatycznego przetwarzania tekstów i tutaj mogą wkraść się błędy.

Zasady podziału na zdania we współczesnym tekście pisanym są względnie niekontrowersyjne[3] i jest to procedura dość prosta. Na poziomie technicznym podział na zdania komplikuje to, że musi on zostać wykonany na tekście przed podjęciem jakiejkolwiek interpretacji językoznawczej. Wszystkie bowiem narzędzia używane w przetwarzaniu KWJP zakładają pracę na poszczególnych zdaniach z osobna. Podział na zdania wykonywany jest prostym narzędziem regułowym. W związku z tym w korpusie mogą się trafić błędne podziały zdań, np. w wypadku zbiegu nietypowego skrótu z kropką z następującym dalej słowem pisanym wielką literą.

Podział na jednostki podlegające opisowi morfosyntaktycznemu – nazywane segmentami (por. Przepiórkowski i in. (red.) 2012) – jest funkcją samego opisu morfosyntaktycznego. Zasady wyróżniania segmentów przejmujemy z NKJP (i wcześniejszych korpusów) bez żadnych modyfikacji. W szczególności oznacza to, że segment nigdy nie zawiera odstępu oraz że słowa (ciągi znaków literowych między odstępami) bywają dzielone na wiele segmentów. Dotyczy to na przykład form 1. i 2. os. czasu przeszłego czasowników, które za NKJP są traktowane jako złożone z dwóch segmentów (pseudoimiesłowu i aglutynantu, np. *czytał+em*).

3 Niektóre sytuacje trzeba rozstrzygnąć arbitralnie, jak bowiem traktować wielozdaniową wypowiedź przytoczoną, np. *Tak. Masz rację – powiedział Janek*.

3.2. Znakowanie morfosyntaktyczne, lematyzacja i jednostki nazewnicze

Celem lematyzacji i znakowania morfosyntaktycznego jest zinterpretowanie segmentów jako form wyrazowych przez przypisanie ich do leksemów języka polskiego oraz wskazanie pełnionej przez daną formę funkcji gramatycznej. Ta ostatnia jest oznaczana w korpusie za pomocą systemu symbolicznych znaczników wyrażających wartości poszczególnych kategorii gramatycznych, którymi charakteryzowana jest dana forma.

Znakowanie KWJP opiera się na systemie opracowanym dla NKJP, jednak nie jest z nim identyczne. Stosujemy mianowicie zmodyfikowaną wersję znakowania, opisaną w artykule Witolda Kierasia, Marcina Wolińskiego i Bartłomieja Nitonia (2021). Jest to zarazem system wykorzystywany przez bieżące wersje analizatora fleksyjnego Morfeusz (Kieraś, Woliński 2017).

Najbardziej widoczną różnicą między NKJP a obecnym ujęciem jest sposób opisu liczebników zbiorowych i ich łączliwości z rzeczownikami. Znakowanie NKJP odróżniało liczebniki główne i zbiorowe jako osobne klasy gramatyczne (num i numcol), ale nie zdawało sprawy z ich łączliwości z rzeczownikami. NKJP operuje bowiem pięcioma wartościami rodzaju gramatycznego (m1, m2, m3, f, n) za Witoldem Mańczakiem (1956), w odróżnieniu od bardziej szczegółowego systemu Zygmunta Saloniego (1974) traktującego liczebniki jako jednolite leksemy, w których obrębie formy główne i zbiorowe różnią się rodzajem. System zaproponowany przez W. Kierasia, M. Wolińskiego i B. Nitonia (2021) wprowadza dodatkową kategorię gramatyczną o wartościach col (zbiorowy) i ncol (główny), roboczo nazwaną przyrodzajem. Kategoria ta przysługuje wyłącznie rzeczownikom i liczebnikom, co zapewnia ekonomię opisu. Dodatkowa wartość pt przyrodzaju pozwala oznaczyć w korpusie wystąpienia rzeczowników *plurale tantum*, a możliwość ich wyszukania (lub wykluczenia z wyników wyszukiwania) wydaje się nam przydatna badawczo.

Oprócz tego w nowym systemie uporządkowano traktowanie nietypowych form odmiany przymiotników, uproszczono interpretację wyrazów obcych i zapisów cyfrowych oraz zmieniono sposób lematyzacji członów nazw własnych. Szczegółowy opis zmian można znaleźć w przywoływanym artykule W. Kierasia, M. Wolińskiego i B. Nitonia (2021).

We wcześniejszych projektach korpusowych, w szczególności przy znakowaniu NKJP w 2012 oraz 2021 roku, automatyczne znakowanie fleksyjne było wykonywane dwuetapowo. Najpierw ustalano słownikowo wszystkie możliwe interpretacje fleksyjne danego segmentu (w tym możliwe lematy) bez uwzględnienia kontekstu. Tę pracę wykonywał analizator fleksyjny Morfeusz (Woliński 2006) oparty na SGJP. Następnie za pomocą narzędzi statystycznych spośród możliwych interpretacji wybierana była jedna najlepiej pasująca do danego kontekstu.

Obecnie wyższą jakość znakowania zapewniają tagery wykorzystujące sztuczne sieci neuronowe i pretrenowane modele językowe. W tym schemacie przetwarzania nie ma komponentu słownikowego – sieć neuronowa uczy się bezpośrednio znaczników fleksyjnych i reguł lematyzacji. Oznacza to, że w razie pomyłek możliwe jest pojawienie się lematów zupełnie nieprawdopodobnych fleksyjnie. W przetwarzaniu KWJP zastosowano więc dodatkowy etap, na którym lematy są korygowane słownikowo, jeśli tylko dana forma występuje w słowniku

Morfeusza SGJP. Znakowanie fleksyjne i lematyzację KWJP wykonano za pomocą programu Hydra (zob. dalej).

W ramach prac nad NKJP podkorpus NKJP1M został ręcznie oznakowany jedno- i wielowyrazowymi jednostkami nazewniczymi (Savary i in. 2012). Przyjęta taksonomia obejmuje nazwy osób, nazwy geograficzne, nazwy organizacji, wyrażenia czasowe, por. ich szczegółowy opis w przywoływanej pracy. Do oznaczenia jednostek nazewniczych w KWJP użyto programu PolDeepNer 2[4] z modelem wytrenowanym na danych NKJP1M.

3.3. Składnia w korpusie

Jako sposób reprezentacji struktur składniowych w KWJP wybrano hybrydowe drzewa, łączące powiązania typowe dla składni zależnościowej i dla opisu składnikowego.

Składnia zależnościowa w centrum uwagi stawia oddziaływania zachodzące między poszczególnymi formami wyrazowymi w wypowiedzeniu. W ujęciu strukturalnym zdaje sprawę przede wszystkim z oddziaływań konotacyjnych i akomodacyjnych (zob. Saloni, Świdziński 1998: 108–289).

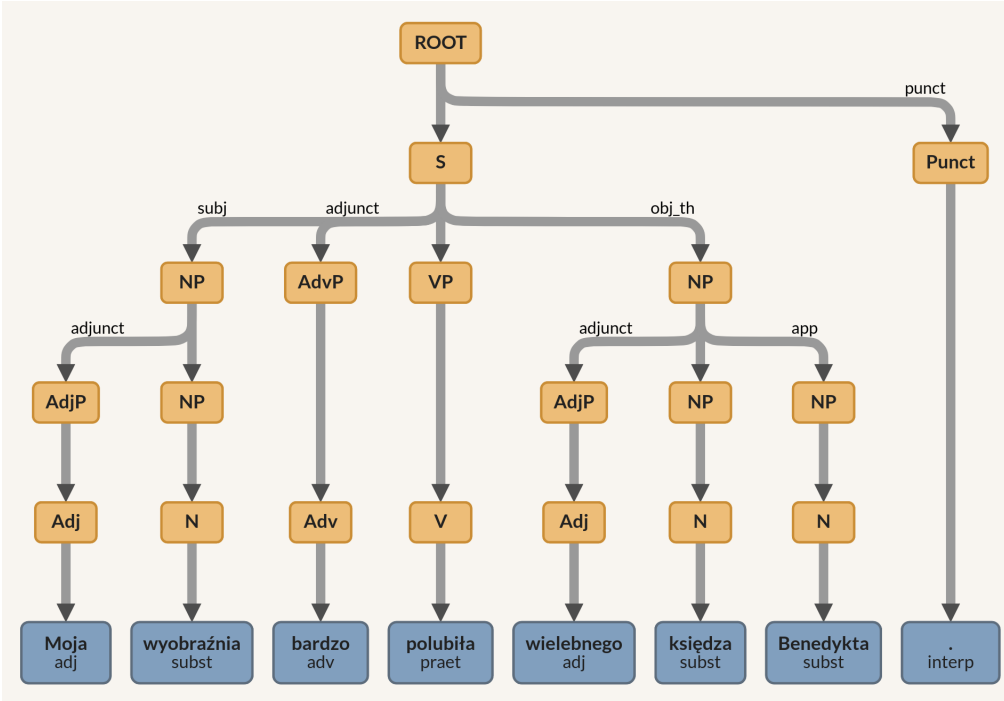
Opis składnikowy (frazowy) koncentruje się na tym, jak krótsze fragmenty wypowiedzenia mogą być zestawiane w dłuższe. Sposobem wyrażenia tej struktury jest drzewo składników bezpośrednich (Saloni, Świdziński 1998: 49–82).

W KWJP struktury składniowe są reprezentowane za pomocą hybrydowych drzew, które ilustruje przykład zamieszczony na rysunku 1. Przedstawioną strukturę można rozumieć jako drzewo składników bezpośrednich z przypisanymi etykietami, wyrażającymi typ składnika (S – zdanie, NP – fraza nominalna, VP – fraza werbalna, AdvP – fraza przysłówkowa, N – forma rzeczownikowa, V – forma czasownikowa itd.). Liśćmi drzewa są formy fleksyjne. Przedstawiona wizualizacja przekazuje również informację o centrach składniowych: składnik centralny jest umieszczany bezpośrednio pod rodzicem w drzewie. W przykładzie należy więc na przykład rozumieć to tak, że dla całego wypowiedzenia (oznaczonego symbolem ROOT) składnikiem centralnym jest zdanie S, którego centrum stanowi fraza werbalna, a jej centrum to forma czasownika POLUBIĆ.

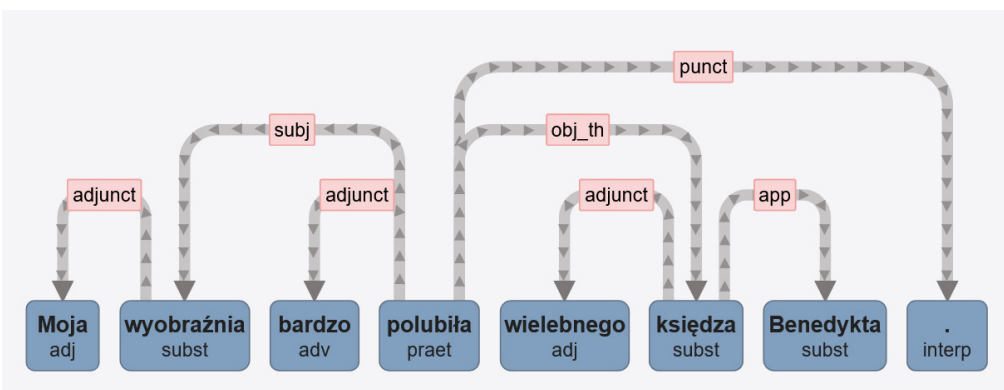
Niecentralne gałęzie drzewa są opatrywane etykietami oznaczającymi relacje zależnościowe, przy czym możliwe są dwa ich odczytania. W drzewie na rysunku 1 etykieta subj została umieszczona na krawędzi łączącej wierzchołek S, znajdujący się nad segmentem *polubiła*, z wierzchołkiem NP nad segmentem *wyobraźnia*. Wskazuje to, że w drzewie zależnościowym istnieje krawędź etykietowana subj od segmentu *polubiła* do *wyobraźnia*. Druga interpretacja głosi, że fraza nominalna *Moja wyobraźnia* występuje w roli subj jako składnik bezpośredni zdania S. W tym odczytaniu zależności zachodzą między składnikami (jednostkami składniowymi), a nie segmentami (formami wyrazowymi).

Oprócz wersji drzewa pokazującego widok hybrydowy (jak na rys. 1) można też osobno wyświetlić wydobyte z niego drzewo zależnościowe, obrazujące relacje zachodzące między segmentami danego zdania (zob. rys. 2).

4 https://github.com/CLARIN-PL/PolDeepNer2.



Rysunek 1. Hybrydowe drzewo struktury składniowej dla zdania *Moja wyobraźnia bardzo polubiła wielebnego księdza Benedykta.* (Tokarczuk 2020, *Czuły narrator*, Wydawnictwo Literackie, KWJP)



Rysunek 2. Drzewo zależnościowe struktury składniowej

Struktury składniowe w KWJP są wynikami analizy automatycznej wykonanej za pomocą parsera Hydra (Krasnowska-Kieraś, Woliński 2024)[5]. Program ten wykorzystuje pretreno-

5 Wersja demonstracyjna parsera dostępna jest pod adresem http://constituency.nlp.ipipan.waw.pl/. Pozwala ona zobaczyć interpretacje dla dowolnych zdań wpisywanych przez użytkownika.

wany model językowy HerBERT (Mroczkowski i in. 2021). Źródłem wiedzy składniowej są dla algorytmu struktury powstałe z połączenia drzew zależnościowych z Polish Dependency Bank (Wróblewska 2014) oraz drzew składnikowych z centrami ze Składnicy frazowej (Woliński 2019).

Ze względu na wspólną historię tych dwóch zasobów zawarte w nich drzewa dość łatwo dają się łączyć w spójne struktury. Zmiany wymagała w zasadzie wyłącznie konwencja zapisu konstrukcji biernych. Jako lepiej pasującą do powierzchniowoskładniowego opisu w korpusie wybrano wersję stosowaną w Składnicy, czyli zasadę, że w konstrukcji biernej forma imiesłowu biernego jest podrzędnikiem posiłkowej formy finitywnej czasownika (forma finitywna zawsze jest nadrzędnikiem). Tym samym struktury zależnościowe w KWJP różnią się w tym względzie od Polish Dependency Bank oraz odświeżonego NKJP (gdzie to imiesłów bierny jest centrum konstrukcji biernej).

Etykiety zależnościowe są zgodne z koncepcją wypracowaną przez Alinę Wróblewską (2014 z późn. zm.), zrezygnowano wszakże z niektórych rozróżnień między podtypami etykiety adjunct[6].

Struktura drzew składnikowych jest zgodna z systemem wypracowanym dla Składnicy frazowej (Woliński 2019). Jako etykiety składników postanowiono jednak używać symboli o proweniencji angielskiej (S – *sentence*, NP – *noun phrase*, AP – *adjectival phrase* itd.), które powinny budzić żywsze skojarzenia niż tworzone *ad hoc* polskojęzyczne symbole w Składnicy.

Należy zaznaczyć, że drzewa składnikowe i zależnościowe generowane przez parser są wynikiem działania modelu statystycznego. Z tego powodu zdarzające się błędne lub nieadekwatne reprezentacje składniowe są nieuniknione. Warto też podkreślić, że warstwa znakowania składniowego w większym stopniu niż pozostałe elementy opisu lingwistycznego ma charakter eksperymentalny – zakładamy możliwość zmian w kolejnych wydaniach korpusu.

4. Wyszukiwanie w korpusie

Podstawowym narzędziem korzystania z KWJP jest wyszukiwarka korpusowa wykorzystująca język zapytań CQL (*Corpus Query Language*). Inaczej niż w NKJP, zamiast formułować zapytania w składni CQL, można skorzystać z konstruktora pozwalającego na wybór wartości parametrów z listy. W oknie wyszukiwania można również wpisać dowolne słowo – jeśli będzie to forma hasłowa (np. *kot*), wynikiem takiego zapytania będą konkordancje zawierające dowolne formy fleksyjne danego leksemu. Jeśli zaś będzie to forma niehasłowa (np. *kotem*), to wyszukane zostaną tylko wystąpienia tej konkretnej formy.

Istnieje możliwość łączenia warstw znakowania w zapytaniach. Jeśli interesuje nas, jakie nazwy geograficzne są używane po przyimku NA wymagającym miejscownika, wystarczy zadać następujące zapytanie:

```
[lemma="na" & case="loc"] <ne="geogName"/>
```

6 Pełną listę stosowanych etykiet można znaleźć w konstruktorze zapytań wyszukiwarki korpusowej.

W wyniku uzyskujemy między innymi: *na Azorach*, *na Krecie*, *na Dolnym Śląsku*, *na Krakowskim Przedmieściu*, *na Saharze*, *na Bałkanach*, *na Węgrzech*.

Celem niniejszego artykułu nie jest wyczerpujące opisanie możliwości języka zapytań w odniesieniu do wszystkich wymienionych wyżej warstw znakowania. Skupimy się jedynie na znakowaniu składniowym i pokażemy na przykładach, jak można korzystać z anotacji zależnościowej i składnikowej w korpusie.

4.1. Wyszukiwanie struktur składniowych

Wyszukiwarka korpusowa KWJP nie jest wyszukiwarką przeznaczoną do pracy na strukturach drzewiastych, w związku z czym w języku zapytań nie ma operatorów wyrażających relacje zależnościowe między segmentami ani wyrażających relację bycia składnikiem bezpośrednim. Korpus KWJP został jednak zaindeksowany w wyszukiwarce w taki sposób, aby możliwe były zapytania o lokalne fragmenty struktury zależności oraz o poszczególne typy składników. Dla znalezionych zdań można wyświetlić kompletną strukturę drzewiastą (rysunek 1 jest przykładem wizualizacji drzewa składniowego w wyszukiwarce).

Za przykład wykorzystania warstwy zależnościowej niech posłuży ponownie (por. Kieraś i in. 2021) kwestia wyszukiwania konstrukcji biernych. Zacznijmy od poszukiwania form imiesłowu biernego oznaczonych rolą zależnościową pd, która w znakowaniu PDB odpowiada tradycyjnemu pojęciu orzecznika[7]:

```
[pos="ppas" & deprel="pd"]
```

Warto wykluczyć z wyszukiwania czasowniki BYĆ i ZOSTAĆ, które są najczęstszymi nadrzędnikami imiesłowów biernych w tego typu związkach. Pozwoli to na stwierdzenie, jakie inne czasowniki mogą wystąpić jako formy finitywne w konstrukcjach biernych.

```
[pos="ppas" & deprel="pd" & !head.base="(być|zostać)"]
```

Dalsza eksploracja prowadzi do grupy czasowników BYWAĆ, ZOSTAWAĆ oraz POZOSTAĆ, POZOSTAWAĆ.

```
[pos="ppas" & deprel="pd" & !head.base="(by(wa)?ć|(po)?zosta(wa)?ć)"]
```

Wyniki tego zapytania zawierają m.in. czasowniki: STANOWIĆ, WYDAWAĆ SIĘ, ZDAWAĆ SIĘ, STAĆ SIĘ i STAWAĆ SIĘ. Zapytanie wyszukujące konstrukcje z czasownikami: BYĆ, BYWAĆ, ZOSTAĆ, ZOSTAWAĆ mogłoby wyglądać następująco:

```
[pos="ppas" & deprel="pd" & head.base="(by(wa)?ć|zosta(wa)?ć)"]
```

7 Jest to inne zapytanie niż w artykule W. Kierasia, M. Wolińskiego i B. Nitonia (2021) ze względu na wspomnianą wcześniej zmianę w stosunku do PDB.

Wyszukiwanie składników (fraz) wykonuje się za pomocą wyrażenia <c/>, które dopasowuje się do dowolnego składnika (*constituent*). Składniki zadanego typu, np. frazy zdaniowe CP, można wyszukać w następujący sposób:

```
<c="CP"/>
```

Za pomocą operatora containing można sformułować zapytanie odwołujące się do zagnieżdżenia składników. Trzeba jednak pamiętać, że wyraża on zawieranie na dowolnym poziomie struktury, a nie powiązania jednostki składniowej z jej bezpośrednimi składnikami. W związku z tym następujące wyrażenie wyszukuje frazy nominalne zawierające gdziekolwiek w swojej strukturze frazę zdaniową:

```
<c="NP"/> containing <c="CP"/>
```

Wśród wyników zobaczymy na przykład frazy: *pisarzy, których dzieła były odrzucane przez cenzurę*; *cisza, że w uszach dzwoni*; *miejscu gdzie autor osiadł po wyjeździe z Polski po zakończeniu stanu wojennego*; ale także frazę: *twórczości pisarzy, których dzieła były odrzucane przez cenzurę* zawierającą zdanie względne na niższym poziomie zagnieżdżenia składników.

Nieco trudniejsze jest ograniczenie wyników zapytania do fraz względnych. Następujący wariant wymaga, aby w obrębie frazy nominalnej NP znalazł się taki ciąg: miejsce rozpoczęcia frazy zdaniowej (<c="CP"> w odróżnieniu od pełnej frazy <c="CP"/>), przecinek i dowolna forma leksemu KTÓRY:

```
<c="NP"/> containing (<c="CP"> [orth=","] [base="który"])
```

Zapytanie to łączy warunki wybierające segmenty w nawiasach prostokątnych z warunkami dotyczącymi fraz w nawiasach kątowych. Innym przykładem tego rodzaju jest zapytanie wyszukujące wypowiedzenia składające się z dokładnie dwóch segmentów:

```
[]{2} fullyalignedwith <s/>
```

Operatorów typu containing można użyć w zapytaniu wielokrotnie. Konieczne jest wtedy zastosowanie nawiasów, aby jawnie wskazać, do jakich elementów operatory się odnoszą. Oto wariant poprzedniego zapytania wyszukujący krótkie wypowiedzenia, złożone z 4–6 segmentów. Został on uzupełniony o warunek, że wypowiedzenie musi zawierać finitywną formę czasownika CZYTAĆ oraz jawnie wyrażony podmiot (wymagamy obecności segmentu, który pełni funkcję składniową subj):

```
(([]{4,6} containing [base="czytać" & pos="(fin|praet)"]) containing [deprel="subj"])
fullyalignedwith <s/>
```

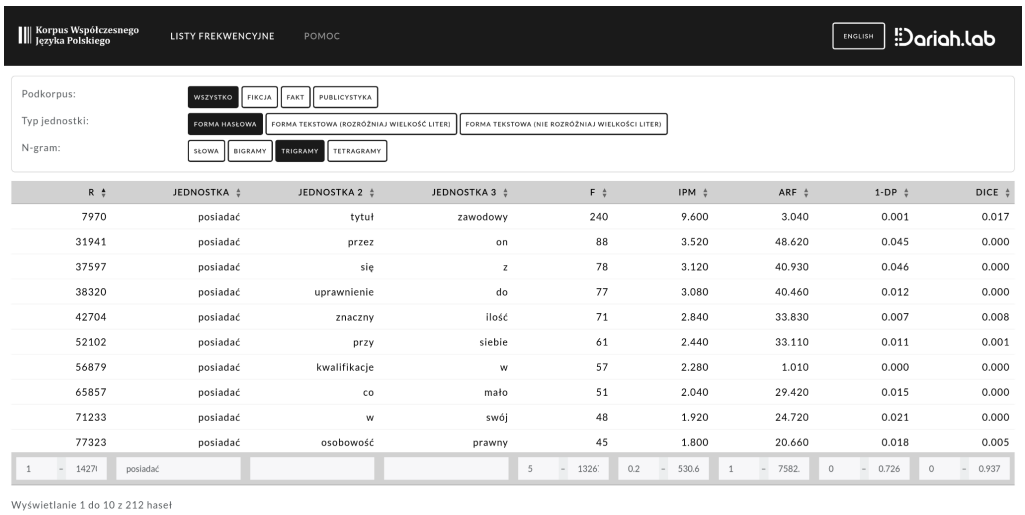
4.2. Informacja frekwencyjna

Najbardziej podstawową informacją, jaką operuje językoznawstwo korpusowe, jest częstość występowania jednostek tekstowych w badanym korpusie. Użytkownik KWJP zainteresowany prostą informacją, ile razy występuje w nim dany wyraz czy dana konstrukcja, lub który chciałby porównać częstość w poszczególnych latach czy typach tekstów, może skorzystać z informacji wyświetlanej przy każdym wyniku wyszukiwania. Pokazywane są wówczas częstość bezwzględna wyszukiwanej frazy oraz jej częstość na milion słów.

Oprócz możliwości samodzielnego sprawdzenia w ten sposób częstości danego wyrazu udostępniamy również przygotowane wcześniej listy frekwencyjne słów i *n*-gramów w korpusie zrównoważonym z podziałem na podzbiory gatunkowe, co umożliwia sprawdzenie częstości nie tylko pojedynczego słowa lub wyrażenia, ale też innych jednostek o takiej samej lub podobnej częstości. Listy zostały wzbogacone o wartości miar dyspersji oraz skorygowanej częstości. Listy frekwencyjne zostały zebrane na podstawie automatycznej lematyzacji i znakowania morfosyntaktycznego, mogą zatem zawierać błędy. Na listach znajdują się wyłącznie słowa zapisane literami alfabetu łacińskiego (z diakrytami), ewentualnie z dywizem.

Listy utworzono w kilku wariantach. Po pierwsze, w podziale na trzy główne podzbiory gatunkowe korpusu: fikcję, fakt i publicystykę prasową. Po drugie, ze względu na zamieszczone na nich formy: hasłowe oraz tekstowe, te drugie dodatkowo w podziale na formy rozróżniające wielkość liter i nierozróżniające wielkości liter. Zarówno listy pojedynczych słów, jak i listy *n*-gramów są ograniczone do jednostek, które wystąpiły co najmniej 5 razy w całym korpusie (liczby wystąpień w poszczególnych podzbiorach gatunkowych mogą być mniejsze).

Każda lista składa się z kilku kolumn (zob. rys. 3). Kolumna *R* zawiera rangi jednostek, czyli kolejne liczby porządkowe na liście w porządku częstości. Kolumna *Jednostka* zawiera słowa lub *n*-gramy, a na liście form hasłowych znajduje się dodatkowo kolumna *POS* zawierająca klasę gramatyczną (fleksem), którą przypisano do danej formy hasłowej – na liście mogą być zatem hasła homonimiczne przypisane do różnych klas. Kolumna *F* zawiera częstość jednostki w korpusie, kolumna *IPM* (*Items per Million*) zaś częstość względną przeliczoną na milion słów. Kolumna *ARF* (*Average Reduced Frequency*) (Savický, Hlaváčová 2002; Hlaváčová 2006) zawiera wartość miary tzw. skorygowanej frekwencji o takiej nazwie, której celem jest zredukowanie zwykłej częstości słów występujących w korpusie w bliskich skupiskach, np. w jednym lub kilku tekstach, w przeciwieństwie do słów dość równomiernie rozłożonych w całym korpusie, dla jakich wartość ARF będzie stosunkowo bliska zwykłej częstości (*F*). Z kolei kolumna *1-DP* zawiera wartość miary dyspersji DP (*Deviation of Proportions*) (Gries 2008, 2020) przeskalowanej w taki sposób, by słowa o względnie równomiernej dystrybucji w korpusie miały wartości bliskie 1, a słowa o bardzo nierównomiernej dystrybucji – bliskie 0. Przykładowo przymiotnik SPACJALNY ma w korpusie 151 wystąpień, ale wszystkie znajdują się w jednej książce. Z tego powodu jego wartość częstości skorygowanej (*ARF*) została zredukowana aż do 1,170, czyli o dwa rzędy wielkości, wartość *1-DP* wynosi zaś 0,001, co również wskazuje na bardzo nierównomierne rozmieszczenie w korpusie.



R	JEDNOSTKA	JEDNOSTKA 2	JEDNOSTKA 3	F	IPM	ARF	1-DP	DICE
7970	posiadać	tytuł	zawodowy	240	9.600	3.040	0.001	0.017
31941	posiadać	przez	on	88	3.520	48.620	0.045	0.000
37597	posiadać	się	z	78	3.120	40.930	0.046	0.000
38320	posiadać	uprawnienie	do	77	3.080	40.460	0.012	0.000
42704	posiadać	znaczny	ilość	71	2.840	33.830	0.007	0.008
52102	posiadać	przy	siebie	61	2.440	33.110	0.011	0.001
56879	posiadać	kwalifikacje	w	57	2.280	1.010	0.000	0.000
65857	posiadać	co	mało	51	2.040	29.420	0.015	0.000
71233	posiadać	w	swój	48	1.920	24.720	0.021	0.000
77323	posiadać	osobowość	prawny	45	1.800	20.660	0.018	0.005
1 – 1427(	posiadać			5 – 1326	0.2 – 530.6	1 – 7582.	0 – 0.726	0 – 0.937

Rysunek 3. Początek listy frekwencyjnej trigramów, których pierwszy segment jest formą czasownika POSIADAĆ

Listy *n*-gramów zawierają też dodatkowo kolumnę *Dice* zawierającą wartość tzw. współczynnika Dice'a interpretowanego jako miara siły współwystępowania dwóch lub większej liczby słów. Współczynnik ten osiąga maksymalną wartość 1 dla słów, które występują w tekstach wyłącznie obok siebie (nie występują w innych kontekstach).

Wszystkie opisane powyżej kolumny można filtrować za pomocą pól znajdujących się w ich dolnej części: kolumny o wartościach liczbowych (*R, F, IPM, ARF, 1-DP*) za pomocą zakresów wartości, kolumnę *POS* za pomocą menu zawierającego wszystkie wartości kategorii gramatycznych z tagsetu, kolumnę *Jednostka* natomiast za pomocą dowolnego podciągu liter szukanego słowa. W polu tym można korzystać z wyrażeń regularnych.

Na koniec omówimy przykład korzystania z informacji zawartej w listach frekwencyjnych. Rysunek 3 przedstawia górną część listy frekwencyjnej trigramów, z których pierwszy ma być formą czasownika POSIADAĆ, pozostałe natomiast dowolnymi innymi słowami. Wśród kilkunastu najczęstszych połączeń zwracają uwagę trójki *posiadać tytuł zawodowy* oraz *posiadać kwalifikacje w* – w obu wypadkach wartość częstości skorygowanej (*ARF*) jest kilkanaście razy mniejsza niż ich częstość bezwzględna (*F*), tymczasem w wypadku pozostałych połączeń te wartości różnią się mniej więcej dwu- lub trzykrotnie. Wartość *ARF* niemal zawsze jest wyraźnie mniejsza niż częstość bezwzględna, ponieważ słowa czy połączenia wyrazowe nigdy nie są w korpusie rozłożone idealnie równomiernie, ale tak duża różnica między zwykłą częstością i *ARF* świadczy o tym, że zdecydowana większość wystąpień tych połączeń jest skupiona w jednym lub kilku tekstach. Na to samo również wskazują wyraźnie niższe niż w wypadku pozostałych połączeń sąsiadujących na liście wartości miary dyspersji *1-DP*. Proste zapytanie w korpusie połączone z grupowaniem wyników ze względu na tytuł tekstu źródłowego wskazuje, że ogromna większość wystąpień obu połączeń pochodzi z jednego tekstu – opublikowanego w Dzienniku Ustaw rozporządzenia ministra. Tego typu zbitki o wysokiej frekwencji

w korpusie są charakterystyczne dla języka urzędowego, ale mogą też wystąpić w tekstach specjalistycznych (np. w książkach naukowych) lub stanowić wykładnik bardzo specyficznego stylu konkretnego autora.

Podsumowanie

Celem niniejszego artykułu było przekazanie czytelnikowi zwięzłej informacji na temat Korpusu Współczesnego Języka Polskiego 2011–2020. Mamy nadzieję, że opis zasobu oraz możliwości jego wykorzystania sprawią, że korpus stanie się przyjaznym narzędziem w pracy lingwistów. Duża część informacji oraz opcji KWJP jest dostępna dla użytkowników zalogowanych do systemu. Uzyskanie dostępu rejestrowanego jest czynnością prostą i natychmiastową, wystarczy bowiem zgodzić się na zasady korzystania z zasobu i założyć konto w systemie. W wypadku wykorzystania korpusu w badaniach i publikacjach naukowych prosimy o cytowanie korpusu w następujący sposób: M. Marciniak, W. Kieraś, K. Bojałkowska, P. Borkowski, M. Borys, W. Eźlakowski, W. Guz, Ł. Kobyliński, D. Komosińska, K. Krasnowska-Kieraś, M. Łaziński, M. Miernecka, B. Nitoń, M. Ogrodniczuk, M. Rudolf, A. Tomaszewska, M. Woliński, J. Wołoszyn, B. Wójtowicz, A. Wróblewska, N. Zawadzka-Paluektau, *Korpus Współczesnego Języka Polskiego: 2011–2020*, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2023. URL: https://kwjp.pl. Korpus nie powstałby bez współpracy dziesiątek wydawców, redakcji prasowych, bibliotek i instytucji kultury. Pamiętajmy o nich, korzystając z korpusu i cytując przykłady z pełnym adresem bibliograficznym (autor, tytuł, wydawca, rok) oraz akronimem KWJP.

Bibliografia

Gries S.T. 2008: *Dispersions and adjusted frequencies in corpora*, „International Journal of Corpus Linguistics", no. 13(4), s. 403–437.

Gries S.T. 2020: *Analyzing dispersion*, [w:] M. Paquot, S.T. Gries (red.), *A practical handbook of corpus linguistics*, Springer, Cham, s. 99–118.

Hlaváčová J. 2006: *New approach to frequency dictionaries – Czech example*, [w:] *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genua, Włochy, s. 373–378.

Kieraś W., Woliński M. 2017: *Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego*, „Język Polski" XCVII, z. 1, s. 75–83.

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

Krasnowska-Kieraś K., Woliński M. 2023: *Constituency parsing with spines and attachments*, [w:] J. Mikyška, C. de Mulatier, M. Paszynski, V.V. Krzhizhanovskaya, J.J. Dongarra, P.M. Sloot (red.), *Computational Science – ICCS 2023. ICCS 2023. Lecture Notes in Computer Science,* vol. 14073, Springer, Cham, s. 191–205.

Krasnowska-Kieraś K., Woliński M. 2024: *Parsing headed constituencies*, [w:] *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turyn, Włochy, ELRA and ICCL, s. 12633–12643.

Lewandowska-Tomaszczyk B., Bańko M., Górski L.R., Łaziński M., Pęzik P., Przepiórkowski A. 2012: *Narodowy Korpus Języka Polskiego. Geneza i dzień dzisiejszy*, [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN SA, Warszawa, s. 3–10.

Mańczak W. 1956: *Ile rodzajów jest w polskim?*, „Język Polski" XXXVI, z. 2, s. 116–121.

Marciniak M., Kieraś W., Bojałkowska K., Borkowski P., Borys M., Eźlakowski W., Guz W., Kobyliński Ł., Komosińska D., Krasnowska-Kieraś K., Łaziński M., Miernecka M., Nitoń B., Ogrodniczuk M., Rudolf M., Tomaszewska A., Woliński M., Wołoszyn J., Wójtowicz B., Wróblewska A., Zawadzka-Paluektau N. 2023: *Korpus Współczesnego Języka Polskiego*, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa, URL: https://kwjp.pl.

MoncoPL: Wyszukiwarka korpusowa Monco (online: http://monco.frazeo.pl/, dostęp: 3 października 2024).

Mroczkowski R., Rybak P., Wróblewska A., Gawlik I. 2021: *HerBERT: Efficiently pretrained transformer-based language model for Polish*, [w:] *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kijów, Ukraina, Association for Computational Linguistics, s. 1–10.

NKJP: Narodowy Korpus Języka Polskiego (online: www.nkjp, dostęp: 3 października 2024).

Ogrodniczuk M. 2017: *Lingwistyka komputerowa dla języka polskiego: dziś i jutro*, „Język Polski" XCVII, z. 1, s. 18–28.

Ogrodniczuk M. 2018: *Polish Parliamentary Corpus*, [w:] D. Fišer, M. Eskevich, F. de Jong (red.), *Proceedings of the LREC 2018 Workshop ParlaCLARIN. Creating and using Parliamentary Corpora*, European Language Resources Association (ELRA), Paryż, s. 15–19.

Pęzik P. 2020: *Budowa i zastosowania korpusu monitorującego MoncoPL*, „Forum Lingwistyczne", nr 7(7), s. 133–150.

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

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

Saloni Z. 1974: *Klasyfikacja gramatyczna leksemów polskich*, „Język Polski" LIV, z. 1, s. 3–13 oraz LIV, z. 2, s. 93–101.

Saloni Z., Świdziński M. 1998: *Składnia współczesnego języka polskiego*, wyd. 4 zmienione, Wydawnictwo Naukowe PWN, Warszawa.

Savary A., Chojnacka-Kuraś M., Wesołek A., Skowrońska D., Śliwiński P. 2012: *Anotacja jednostek nazewniczych*, [w:] A. Przepiórkowski, M. Bańko, R.L. Górski, B. Lewandowska-Tomaszczyk (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 129–167.

Savický P., Hlaváčová J. 2002: *Measures of word commonness*, „Journal of Quantitative Linguistics", no. 9, s. 215–231.

SGJP: Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska, *Słownik gramatyczny języka polskiego*, wyd. 3 online, Warszawa 2015 (online: http://sgjp.pl, dostęp: 3 października 2024).

Woliński M. 2006: *Morfeusz – a practical tool for the morphological analysis of Polish*, [w:] M.A. Kłopotek, S.T. Wierzchoń, K. Trojanowski (red.), *Intelligent Information Processing and Web Mining, Advances in Soft Computing*, Springer-Verlag, Berlin, s. 503–512.

Woliński M. 2019: *Automatyczna analiza składnikowa języka polskiego*, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

Wróblewska A. 2014: *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*, rozprawa doktorska, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.

Summary

The Corpus of Contemporary Polish. Decade 2011–2020

Keywords: The Corpus of Contemporary Polish, linguistic annotation, syntactic trees.

The goal of this article is to provide readers with a concise overview of the Corpus of Contemporary Polish 2011–2020. We start by presenting the characteristics of acquired texts, detailing their sources and the structure of the balanced corpus. We then proceed to describe the layers of linguistic annotation, which include: segmentation, lemmatization, morphosyntactic tagging, named entities, and syntactic parsing (represented by trees that illustrate the constituent structure of utterances and dependency relations within them). Finally, we discuss how users can access the information within each annotation layer. The article includes useful webpage addresses that allow the user to fully take advantage of the gathered linguistic material.