WITOLD KIERAŚ*, **MARCIN WOLIŃSKI****, **BARTŁOMIEJ NITOŃ*****

INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK, WARSZAWA

Nowe wielowarstwowe znakowanie lingwistyczne zrównoważonego Narodowego Korpusu Języka Polskiego[1]

Słowa kluczowe: korpus, przetwarzanie języka naturalnego, znakowanie fleksyjne, znakowanie składniowe.

doi: http://dx.doi.org/10.31286/JP.101.2.5

1. Wprowadzenie

W tym roku minęło dziesięć lat od udostępnienia pierwszej, wstępnej wersji Narodowego Korpusu Języka Polskiego (NKJP), osiem zaś – od zakończenia projektu i udostępnienia wersji ostatecznej (Przepiórkowski i in. 2012). Od tamtego czasu korpus nie był aktualizowany ani pod względem zawartości, ani też pod względem technicznym. Jest wciąż największym – bo jedynym – korpusem referencyjnym polszczyzny, czyli reprezentującym użycia wielu zróżnicowanych rejestrów i typów funkcjonalnych języka polskiego. Jednocześnie przez cały ten czas NKJP służył i wciąż służy jako podstawa materiałowa badań zarówno *stricte* językoznawczych, jak i prac z zakresu inżynierii lingwistycznej. Korpus jest również podstawą redagowanych obecnie słowników, choć ze względu na brak nowych tekstów w coraz mniejszym stopniu może stanowić źródło wiedzy o słownictwie najnowszym. Najbardziej znanym przykładem takiego słownika jest WSJP PAN, choć na NKJP opierają się także mniejsze słowniki specjalistyczne, na przykład słownik walencyjny Walenty (Przepiórkowski i in. 2017).

W niniejszym artykule prezentujemy ten sam niezmieniony zbiór tekstów składający się na NKJP, ale w nowym opracowaniu technicznym, na które składa się znakowanie fleksyjne oraz znakowanie jednostek nazewniczych opracowane za pomocą współczesnych narzędzi informatycznych, zupełnie nową warstwę znakowania składniowego pochodzącą z parsera zależnościowego, a także nową wyszukiwarkę korpusową pozwalającą na zindeksowanie wszystkich powyższych warstw i odnoszenie się do nich w zapytaniach korpusowych.

* wkieras@ipipan.waw.pl; ORCID: 0000-0002-8062-5881

** wolinski@ipipan.waw.pl; ORCID: 0000-0002-7498-1484

*** bartek.niton@gmail.com; ORCID: 0000-0003-3306-7650

1 Prace związane z opisanym w niniejszym artykule nowym opracowaniem korpusu NKJP300M zostały dofinansowane w ramach środków infrastruktury badawczej CLARIN-PL, pochodzących z Ministerstwa Nauki i Szkolnictwa Wyższego. Współautorstwo obejmuje wszystkie etapy pracy nad artykułem, a wkład każdego z autorów w jego powstanie jest równy. Autorem opracowania programistycznego opisywanych prac jest Bartłomiej Nitoń.

2. NKJP, jego znakowanie lingwistyczne, wyszukiwarki korpusowe

NKJP został udostępniony w trzech wariantach różniących się znakowaniem lingwistycznym. W najmniejszym podkorpusie, znanym jako NKJP1M, a mającym objętość około 1,2 mln segmentów, wszystkie elementy znakowania lingwistycznego zostały wprowadzone lub zweryfikowane przez człowieka. Większy zrównoważony korpus NKJP300M ma objętość około 300 mln segmentów, a największy, tzw. oportunistyczny, NKJP1800M – około 1,8 mld segmentów. W tych wariantach znakowanie zostało wykonane za pomocą narzędzi automatycznych tworzonych na bazie wzorcowych danych NKJP1M.

Podstawowym elementem we wszystkich wariantach NKJP jest warstwa znakowania morfosyntaktycznego, zawierająca opis form fleksyjnych. Została ona opracowana za pomocą analizatora fleksyjnego Morfeusz, a następnie ręcznie ujednoznaczniona i uzupełniona w NKJP1M, automatycznie natomiast ujednoznaczniona – w NKJP300M i NKJP1800M. Oprócz tego w analogicznym trybie opracowano warstwy słów składniowych, grup składniowych i jednostek nazewniczych, a w NKJP1M – eksperymentalną warstwę znaczeń słów (Przepiórkowski i in. 2012).

Od początku swego istnienia NKJP był udostępniany równolegle za pomocą dwóch różnych wyszukiwarek korpusowych: Poliqarp i Pelcra. Poliqarp w zapytaniach korpusowych pozwala się odwołać do postaci ortograficznej słowa, jego formy hasłowej (lematu) oraz jego interpretacji fleksyjnej (znacznika). Ciekawą cechą tej wyszukiwarki jest możliwość odwołania się zarówno do ujednoznacznionych, jak i do nieujednoznacznionych wyników analizy fleksyjnej. Wyszukiwarka Pelcra umożliwia wyszukiwanie wyłącznie według form ortograficznych i hasłowych, ale za to ma ona istotną przewagę w postaci możliwości wyszukiwania kolokacji. Obie wyszukiwarki mają swoje ograniczenia, przede wszystkim dają użytkownikowi dostęp jedynie do warstwy fleksyjnej, a pomijają pozostałe warstwy znakowania, które opracowano w projekcie.

Z dzisiejszej perspektywy ograniczenia techniczne związane z przeszukiwaniem wielu warstw korpusów zniknęły, pojawiły się mianowicie wyszukiwarki przystosowane do danych wielowarstwowych. Jedną z nich jest holenderska wyszukiwarka MTAS (Brouwer i in. 2017) wykorzystana w kilku polskich korpusach (np. w największym korpusie polszczyzny dawnej KorBa – Gruszczyński i in. 2020), jak również w aplikacji do tworzenia korpusów o nazwie Korpusomat (Kieraś i in. 2018). Nie ma zatem technicznych przeszkód, by te istniejące w plikach źródłowych korpusu warstwy zindeksować i udostępnić użytkownikom. Przeciwko takiemu podejściu przemawia jednak czas, który upłynął od momentu stworzenia tych warstw anotacji, i postęp, jaki się dokonał w dziedzinie przetwarzania języka naturalnego. O ile ręcznie anotowany korpus NKJP1M wciąż służy jako podstawowy dla polszczyzny zasób do trenowania narzędzi statystycznych używanych w celu rozwiązywania bardzo różnorodnych zadań z zakresu inżynierii lingwistycznej, o tyle same narzędzia stworzone w projekcie NKJP są już w znacznej większości przestarzałe.

W niektórych wypadkach rozwój dziedziny doprowadził do zmiany podejścia do problemów analizy języka. Tak jest choćby w wypadku analizy składniowej – w NKJP powstały warstwy słów składniowych oraz grup składniowych. Pierwsza z nich reprezentuje jednostki tekstowe, które mogą (choć nie muszą) przekraczać granicę słowa ortograficznego (od spacji

do spacji), ale na poziomie składniowym są interpretowane jako całostki (takim wyrażeniem jest np. *po polsku* składające się z dwóch słów ortograficznych, ale pełniące funkcję przysłówkową, choć żaden z jego członów nie jest przysłówkiem). Na podstawie słów składniowych zbudowano z kolei warstwę grup składniowych będących przykładem tzw. płytkiej analizy składniowej, polegającej na identyfikowaniu maksymalnych ciągów słów składniowych skupionych wokół jednego centrum. Ograniczono się przy tym jedynie do grup niewerbalnych. Podejście to wynikało z przekonania (popartego doświadczeniem), że ówczesny stan rozwoju inżynierii lingwistycznej nie pozwalał – w odniesieniu do języka takiego jak polski – na stworzenie efektywnego głębokiego, czyli budującego pełne struktury składniowe wypowiedzeń, analizatora składniowego zwracającego interpretacje dla dowolnych wypowiedzeń i działającego z zadowalającą dokładnością. Współcześnie jednak tego typu narzędzia są powszechnie dostępne, choć najbardziej rozpowszechnione i najskuteczniejsze z nich nie budują struktur składnikowych (frazowych), ale zależnościowe, a zatem efektem ich działania na żadnym etapie nie jest grupa składniowa w takim sensie, jak ją opisano w anotacji składniowej NKJP. A zatem ostatnia dekada przyniosła nie tylko postęp techniczny, ale i zmianę paradygmatu[2]. W naturalny sposób zatem należy odświeżyć myślenie o informacji składniowej w NKJP.

W stworzonej w NKJP anotacji jednostek nazewniczych (Savary i in., 2012) sprawa ma się podobnie, choć jest nieco prostsza – współcześnie istniejące narzędzia oznaczające automatycznie jednostki nazewnicze działają z dużo większą dokładnością niż rozwiązanie zastosowane przy automatycznym znakowaniu NKJP, dlatego zasadna wydaje się jej aktualizacja. Nie zmienił się jednak sam schemat anotacji stworzony i opisany w projekcie NKJP.

Wreszcie ostatnia z istniejących w NKJP warstw to warstwa rozróżniania znaczeń słów. Zarówno anotacja tej warstwy, jak i próby stworzenia narzędzia do automatycznego rozróżniania znaczeń słów na podstawie kontekstu miały jednak charakter wybitnie eksperymentalny, czego sami twórcy nie ukrywali. W ramach tworzenia ręcznie znakowanych danych wzorcowych wybrano zaledwie 106 częstych wieloznacznych słów polskich, których różne znaczenia opisano na podstawie *Innego słownika języka polskiego* (ISJP) i za pomocą których następnie ręcznie oznaczono ich wystąpienia w milionowym podkorpusie NKJP. Można więc uznać, że dane z tej warstwy są jedynie zalążkowe. Pomimo upływu czasu problem rozróżniania znaczenia słów (ang. *word sense disambiguation*) dla języka polskiego nie został rozwiązany nawet w częściowo zadowalający sposób. Dlatego zdecydowaliśmy się w niniejszej pracy pominąć tę warstwę anotacji. Nie wykluczamy jednak, że w przyszłości korpus zostanie zaktualizowany również o tego typu informację.

Kierując się przedstawionymi racjami, postanowiliśmy odświeżyć znakowanie korpusu NKJP300M, uwzględniając warstwę morfosyntaktyczną, warstwę jednostek nazewniczych i nowo wprowadzoną warstwę drzew zależnościowych. Bezpośrednią inspiracją do wybrania takich poziomów anotacji i zastosowania konkretnych rozwiązań technicznych był Korpusomat,

2 Nie znaczy to oczywiście, że wcześniej nie istniały zależnościowe analizatory składniowe czy tym bardziej teoretyczne opisy składni języków naturalnych w aparacie zależnościowym. Nie znaczy to również, że współcześnie nie istnieją i nie są rozwijane analizatory składnikowe. W dziedzinie przetwarzania języka naturalnego nastąpiła jednak bardzo wyraźna zmiana, dzięki której w stosunkowo krótkim czasie rozwiązania oparte na opisie zależnościowym przeszły od niszy do głównego nurtu.

który aktualnie oferuje automatyczne znakowanie takich właśnie warstw. Podstawowym pomysłem na aktualizację NKJP było zatem możliwie silne zbliżenie go pod względem technicznym do Korpusomatu, dzięki czemu użytkownicy uzyskaliby podobne funkcje przeszukiwania zarówno w dużym referencyjnym korpusie, jak i we własnych korpusach specjalistycznych. Opracowana wersja NKJP300M została upubliczniona pod adresem https://nkjp.nlp.ipipan.waw.pl/.

3. Warstwa znakowania morfosyntaktycznego

Powszechnie używanym analizatorem morfologicznym dla języka polskiego jest Morfeusz (Woliński 2014; Kieraś, Woliński 2017), którego dane pochodzą z SGJP. W NKJP zastosowano system znaczników zbliżony do tego, którym operuje Morfeusz SGJP, jednak z nim nietożsamy. Między systemami znakowania Morfeusza i NKJP istnieje wiele szczegółowych różnic zarówno w samym zestawie znaczników, jak i w sposobie interpretowania niektórych z nich. Odświeżenie korpusu uznaliśmy za dobrą okazję, by te różnice zniwelować i wykorzystując niemal dekadę doświadczeń ze znakowaniem fleksyjnym polskich tekstów, zaproponować pewne modyfikacje opisu. Przy czym modyfikacje są obustronne – ostateczny zestaw znaczników nie jest tożsamy ani z oryginalnym tagsetem Morfeusza (choć Morfeusz od pewnego już czasu go stosuje), ani z tagsetem NKJP.

3.1. Znakowanie kategorii rodzaju

Najbardziej jaskrawą różnicą między NKJP a Morfeuszem SGJP był opis systemu rodzajowego. W obu systemach rodzaj jest rozumiany jako kategoria wprowadzająca podział zbioru leksemów rzeczownikowych na podstawie obserwacji łączliwości form tych leksemów z formami leksemów innych klas gramatycznych. W NKJP przyjęto opracowany przez Witolda Mańczaka zestaw pięciu rodzajów: męskoosobowego, męskozwierzęcego, męskonieżywotnego, nijakiego i żeńskiego, oznaczanych w korpusie odpowiednio jako: m1, m2, m3, n, f. Podstawą wyróżnienia takich rodzajów jest obserwacja łączliwości rzeczowników z formami przymiotnikowymi w bierniku liczby pojedynczej i mnogiej (Mańczak 1956). Morfeusz pierwotnie stosował system dziewięciu rodzajów Zygmunta Saloniego (Saloni 1974), w którym dodatkowo rozróżnia się dwa rodzaje nijakie: rzeczowników łączących się z tzw. liczebnikami zbiorowymi (n1) i z tzw. liczebnikami głównymi (n2). Oprócz tego Z. Saloni wyodrębnił trzy rodzaje przymnogie, określone na podstawie braku łączliwości rzeczowników *plurale tantum* z formami liczby pojedynczej innych leksemów. Wśród nich są: rodzaj przymnogi męskoosobowy (p1; np. GENERAŁOSTWO), rodzaj przymnogi niemęskoosobowy łączący się z liczebnikami zbiorowymi (p2; np. DRZWI) i rodzaj przymnogi niemęskoosobowy niełączący się w ogóle z liczebnikami (p3; np. WYMIOCINY).

Jedną z zalet systemu Z. Saloniego jest to, że pozwala opisać łącznie tzw. liczebniki główne i zbiorowe jako jednolite leksemy: formy *pięciu* i *pięcioro* należą do tego samego leksemu PIĘĆ, a różnią się rodzajem. Wadą w zastosowaniach praktycznych jest jednak szczegółowość. Ze względu na uzgodnienia każda forma przymiotnikowa i czasownikowa musi być oznaczona konkretną spośród dziewięciu wartością rodzaju, mimo że formy te nie wykazują tak szczegółowych zróżnicowań. Można ten system także skrytykować na gruncie teoretycznym: nie ma

podstaw, by uznać, że zróżnicowanie dotyczące jedynie dwóch klas gramatycznych, rzeczowników i liczebników, powinno być częścią kategorii obejmującej więcej klas. Wprowadzenie rodzajów przymnogich jest z kolei niespójne z traktowaniem kategorii rodzaju jako niezależnej od kategorii liczby (por. Woliński 2019: 33). Dlatego w przyjętym przez nas systemie postulujemy posługiwanie się rodzajem w rozumieniu W. Mańczaka oraz wprowadzenie nowej kategorii, roboczo nazwanej przyrodzajem, która przysługuje jedynie rzeczownikom i liczebnikom.

Przyjrzyjmy się wartościom klasy gramatycznej (fleksemu), rodzaju i przyrodzaju przypisywanym wybranym formom w każdym z systemów:

	SYSTEM Z. SALONIEGO	**ORYGINALNE ZNAKOWANIE NKJP**	**NOWE ZNAKOWANIE**
dwoje	num:n|l	numcol:n	num:n:col
dzieci	subst:n|l	subst:n	subst:n:col
dwa	num:n2	num:n	num:n:ncol
okna	subst:n2	subst:n	subst:n:ncol
zmęczone	adj:n|l	adj:n	adj:n
dziecko	subst:n|l	subst:n	subst:n:col
spało	praet:n|l	praet:n	praet:n

System stosowany w NKJP wyróżnia liczebniki zbiorowe za pomocą wartości numcol klasy gramatycznej. Formy rzeczownikowe *dzieci* i *okna* są jednak w tym systemie opatrzone tym samym znacznikiem subst:n, a więc nie jest w żaden sposób odnotowana niemożliwość połączeń *dwoje okien* lub *dwa dzieci*. W nowym systemie znakowania odpowiednia cecha jest różnicowana wartością przyrodzaju: col lub ncol, która podlega uzgodnieniu między rzeczownikiem a liczebnikiem. Przy znakowaniu rodzajami Z. Saloniego fragmentu *zmęczone dziecko spało* trzeba oznaczyć formę przymiotnikową i przeszlik szczegółową wartością rodzaju n|, mimo że w tym wypadku wystarczyłoby znakowanie zaproponowane w NKJP. W nowym znakowaniu wartością rodzaju wszystkich trzech form jest n, a tylko forma rzeczownika niesie wartość przyrodzaju col, która jednak w tym zdaniu, pod nieobecność liczebnika, z niczym się nie uzgadnia. Tak więc nowe znakowanie daje tę samą dokładność opisu co system Z. Saloniego, wprowadzając jednak szczegółową informację tylko tam, gdzie jest ona potrzebna.

W nowym systemie również rodzaje przymnogie Z. Saloniego są reprezentowane z wykorzystaniem kategorii przyrodzaju, przy czym zrezygnowaliśmy z reprezentowania najmniej wyrazistego rozróżnienia między rodzajami p2 a p3. O ile bowiem faktycznie trudno liczyć wymiociny, to w SGJP jest wiele leksemów, na przykład nazw miejscowości *plurale tantum*, które mają przypisany rodzaj p2/p3, co wypada odczytać jako niemożność rozstrzygnięcia,

czy dany obiekt daje się liczyć. Wiele obiektów daje się liczyć z trudem, co w naszym odczuciu ma bardziej charakter uwarunkowań semantycznych niż gramatycznych, rodzajowych.

Leksemom rzeczownikowym p1 według Z. Saloniego przypisujemy jak w NKJP wartość m1 rodzaju (łączą się one z formami męskoosobowymi czasowników) oraz dodajemy specjalną wartość pt przyrodzaju. Leksemom p2 i p3 przypisujemy rodzaj n (taka decyzja jest niesprzeczna z ich łączliwością z przymiotnikami i czasownikami, a dwuliczbowe leksemy nijakie są jedynymi, które miewają łączliwość z liczebnikami zbiorowymi) oraz opatrujemy je znacznikiem pt. Wartość pt należy na potrzeby uzgodnień uznawać za identyczną z col, formy tych leksemów łączą się bowiem z formami liczebników zbiorowych (*dwoje państwa*, *dwoje skrzypiec*). Osobny symbol pozwala jednak na wyszukanie w korpusie wystąpień rzeczowników *plurale tantum*, co było niemożliwe w znakowaniu NKJP, a wydaje się interesujące dla badacza.

3.2. Inne zmiany w znakowaniu

We współczesnym paradygmacie przymiotnika oprócz bloku form, które można scharakteryzować wartościami kategorii liczby, przypadka i rodzaju, występują pewne formy nietypowe. Niektóre przymiotniki tworzą formę, która może wystąpić jedynie po przyimku *po* w konstrukcjach takich jak *po polsku*, *po angielsku*, i nie jest homonimiczna z żadną inną formą danego przymiotnika. W oryginalnym systemie znaczników Morfeusza zaproponowano znacznik adjp na oznaczenie tych form (jeżeli przymiotnik nie tworzy tej formy, w kontekście po *po* używana jest forma celownika męskiego). W znakowaniu NKJP1M znacznik adjp był stosowany również na oznaczenie innych form nietypowych, w szczególności seryjnej formy przymiotnikowej występującej po przyimku wymagającym dopełniacza w kontekstach takich jak *z polska*, *z wolna*, *od dawna*. Formy te są zawsze homonimiczne z mianownikiem żeńskim, dlatego nie były wyróżniane w Morfeuszu (ani w SGJP). Przyjęte w NKJP znakowanie wypada uznać za niefortunne: tym samym znacznikiem adjp jest bowiem oznaczana zarówno forma *polsku*, jak i forma *polska* (po *z*), a konteksty wystąpień tych form są wyraziście różne. Dlatego w bieżącej wersji Morfeusza stosujemy oznaczenia bardziej szczegółowe: formy typu *polsku* otrzymują znacznik adjp:dat, formy zaś typu *[z] polska* – znacznik adjp:gen. Warto też zauważyć, że w ten sposób grupujemy w jednym fleksemie skostniałe formy o podobnym pochodzeniu historycznym, czyli wywodzące się z tzw. niezłożonej (rzeczownikowej) odmiany przymiotników.

Za nietrafną uznaliśmy także przyjętą w NKJP konwencję, że wszystkie formy liczebników mają mnogą wartość liczby. W nowym znakowaniu przyjmujemy za SGJP, że formy liczebnikowe takie jak *pół* i *ćwierć* łączą się z formami liczby pojedynczej rzeczowników.

Zapisy cyfrowe w znakowaniu NKJP były opatrywane znacznikiem num (liczebnik) lub adj (przymiotnik) i charakteryzowane wartościami liczby, przypadka i rodzaju. Ponieważ zapis cyfrowy charakteryzuje się pełnym synkretyzmem, można to było wykonać jedynie, jeśli odpowiednie wartości wynikały z kontekstu. W wielu wypadkach znakujący musiał podejmować decyzje arbitralne. W nowym znakowaniu przyjęliśmy, że nie warto wprowadzać takiej komplikacji. Liczby w tekście zapisane cyframi arabskimi są obecnie oznaczane znacznikiem dig bez dalszej charakterystyki. Podobnie zapisy z użyciem cyfr rzymskich są oznaczane znacznikiem romandig.

Drobną zmianę systemu znaczników stanowi wprowadzenie nowego fleksemu numcomp. Reprezentuje on formę liczebników występującą w złożeniach, na przykład *pięcio*. Potrzeba użycia tego znacznika pojawia się rzadko, kiedy taka forma występuje samodzielnie z łącznikiem: *cztero- lub pięciodrzwiowym*. W przykładzie tym forma *cztero* zostanie oznaczona symbolem numcomp, zaś *pięciodrzwiowym* zostanie w całości uznane za formę przymiotnika PIĘCIODRZWIOWY.

Za kosmetyczną można uznać zmianę symbolu dwóch fleksemów istniejących. Pierwsza zmiana dotyczy symbolu qub, wprowadzonego początkowo na oznaczenie klasy resztkowej[3], gromadzącej leksemy nieodmienne niesklasyfikowane dokładniej. W miarę rozwoju słownika Morfeusza przeprowadzono dokładniejszą klasyfikację, a w rezultacie obecnie klasa ta obejmuje w przeważającej większości jednostki partykułowe w sensie SGJP (choć są tu również wyjątki, jak SIĘ, którego różnorodne funkcje w tagsecie nie są rozróżniane). Stąd zmiana jej symbolu na part. Druga to zmiana nieczytelnej etykiety burk na frag, która odpowiada klasie jednostek stanowiących w SGJP człony (fragmenty) wyrażeń o charakterze frazeologicznym.

Zmieniona została zasada lematyzacji wieloczłonowych nazw własnych. Anotatorzy NKJP1M przypisywali obu członom nazw takich jak *Morze Czarne* lematy pisane wielką literą (wprowadzając tym samym trudne do obronienia leksemy będące przymiotnikami pisanymi wielką literą). W nowym znakowaniu przyjęliśmy, że tego rodzaju nazwy składają się z wyrazów pospolitych i jako takie powinny być lematyzowane. Lemat pisany wielką literą przysługuje więc jedynie leksemom stanowiącym samodzielne nazwy własne i takim, których formy są używane wyłącznie jako część nazwy własnej. Wszystkie przymiotniki mają lemat pisany małymi literami.

Zmianie uległa też koncepcja znakowania segmentów obcych – słów nienależących do języka polskiego oraz symboli i innych zapisów niebędących słowami. Wedle zasad znakowania korpusu NKJP1M segmenty należące do cytatów obcojęzycznych były oznaczane symbolem xxx (nieznanym Morfeuszowi, którego słownik obejmuje tylko słowa polskie). Wedle instrukcji znakowania symbol ten przysługuje jedynie segmentom „niewchodzącym w bezpośrednie oddziaływania z segmentami polskimi w tekście". Jeżeli takie oddziaływanie dawało się stwierdzić, wszystkim segmentom obcym przypisywano znacznik wynikający z kontekstu. Tak więc we fragmencie „[...] publikowany przez tygodnik Flight International [...]" segmentom *Flight* oraz *International* przypisano interpretację rzeczownikową subst:sg:nom:m3 (rodzaj m3 był przyjmowany na zasadzie konwencji). Rozwiązanie to cechowało się znaczną nieoczywistością w rozpoznawaniu „bezpośrednich oddziaływań" oraz arbitralnością przypisywania cech gramatycznych, ponieważ często kontekst determinuje tylko niektóre z nich. Dlatego w zmienionym znakowaniu postanowiliśmy stosować znacznik xxx do wszelkich segmentów obcych z jednym wyjątkiem: jeżeli fragment obcy występuje w zdaniu w kontekście narzucającym przypadek, stosowany jest nowy znacznik xxs:‹przypadek› – jego użycie nie wiąże się z koniecznością dorozumiewania się wartości innych kategorii, zwłaszcza wartości rodzaju. Tak więc w nowym znakowaniu segmenty *Flight International* opatrzono symbolem xxs:nom, sygnalizującym, że element obcy wystąpił jako tzw. mianownik nazywający.

3 Symbol qub miał być skrótem od nazwy klasy „kublik", a ta z kolei nawiązaniem do kubła na śmieci.

3.3. Zastosowane narzędzia

Warstwę znakowania morfosyntaktycznego opracowano za pomocą nowych wersji narzędzi informatycznych. Do analizy fleksyjnej zastosowano Morfeusza (ściślej: Morfeusza 2 SGJP). Znaczącą zmianę widać w stosowanym przez program słowniku. O ile wersja Morfeusza zastosowana do znakowania w roku 2012 pozostawiła w korpusie NKJP300 4,4% segmentów nierozpoznanych, to program z bieżącą wersją słownika „nie zna" jedynie 1,6% segmentów w tym korpusie. Oznacza to, że w odświeżonym znakowaniu interpretację słownikową otrzymało nieomal 8,5 mln segmentów wcześniej nierozpoznanych.

Do ujednoznaczniania interpretacji fleksyjnych w nowym znakowaniu użyto tagera Concraft (Waszczuk i in. 2018) z modelem wytrenowanym na zaktualizowanej wersji korpusu NKJP1M. Dokładność ujednoznacznienia (ang. *accuracy*) można oszacować na 92% (dla poprzednio używanego tagera dokładność wynosiła ok. 89%). Należy zaznaczyć, że są dziś dostępne rozwiązania osiągające jeszcze wyższą jakość ujednoznaczniania. Za wykorzystaniem tagera Concraft przemawiała jednak efektywność oraz moce obliczeniowe dostępne do wykonania opisywanych prac. Ponadto Concraft jako jedyny tager może pracować bezpośrednio na niejednoznacznych grafach fleksyjnych generowanych przez Morfeusza.

4. Warstwa jednostek nazewniczych

Drugą warstwą automatycznego znakowania jest warstwa jednostek nazewniczych. Do prezentowanej w niniejszym artykule wersji NKJP300M nie wprowadzono żadnych zmian koncepcyjnych w stosunku do oryginalnego projektu opisanego w artykule Savary i in. (2012), który stanowi też oficjalną dokumentację schematu znakowania tej warstwy. Ograniczyliśmy się jedynie do zastosowania nowszego narzędzia do oznaczania jednostek nazewniczych w tekście (Marcińczuk i in. 2018) i – co najważniejsze – do zindeksowania tej warstwy w wyszukiwarce, by można było tworzyć odwołujące się do niej zapytania korpusowe.

Jednostki nazewnicze w ujęciu NKJP to klasa szersza niż tradycyjne nazwy własne. Zaliczają się do nich jednostki jedno- lub wielowyrazowe i tekstowo ciągłe. W języku zapytań korpusu oznaczone są znacznikiem <ne /> (od ang. *named entity*) – nawias kątowy oznacza, że jednostka opisana takim znacznikiem może się składać z więcej niż jednego segmentu. Na jednostki nazewnicze składają się:

– nazwy osób (persName), w tym ich imiona (forename), nazwiska (surname) i pseudonimy (addName);

– nazwy geopolityczne (placeName), w tym nazwy osiedli (settlement), dzielnic (district), regionów (region), państw i obszarów odrębnych, choć niekoniecznie samodzielnych politycznie (country) oraz organizacji międzypaństwowych (bloc) – do tej klasyfikacji włączone są również nazwy mieszkańców tych jednostek, czyli tradycyjne etnonimy, a także przymiotniki od nich derywowane;

– nazwy organizacji (orgName);

– nazwy geograficzne (geogName);

– określenia dat i czasu (date oraz time).

Jak widać, jest to klasyfikacja szeroka. Do poszczególnych jej klas można się odnosić w zapytaniach korpusowych za pomocą wyżej wspomnianego znacznika <ne />. Na przykład:

```
<ne="persName.surname" />
```

zwróci w wynikach wyłącznie nazwiska. Zapytanie tego typu można łączyć z warunkami dotyczącymi na przykład kształtu segmentów wchodzących w skład jednostki nazewniczej. Powyższe zapytanie można więc rozszerzyć do postaci:

```
<ne="persName.surname" /> containing [base="-"]
```

dzięki czemu wyniki zostaną ograniczone do nazwisk dwuczłonowych z dywizem.

5. Warstwa składniowa

Trzecią warstwą automatycznego znakowania jest informacja składniowa pochodząca z parsera zależnościowego COMBO (Rybak, Wróblewska 2018). Parser dla każdego segmentu w wypowiedzeniu określa, który z pozostałych segmentów jest jego bezpośrednim nadrzędnikiem składniowym oraz jakiego typu relacja te segmenty łączy. W efekcie parser dla każdego wypowiedzenia tworzy drzewo opisujące jego strukturę zależnościową: każdy segment w wypowiedzeniu ma dokładnie jeden nadrzędnik (nadrzędnikiem centrum całego wypowiedzenia jest konwencjonalny element ROOT).

Drzewa składniowe są strukturami hierarchicznymi, których pełne przeszukiwanie, tzn. nieograniczone bezpośrednią relacją nadrzędności, ale uwzględniające pełną ścieżkę łączącą dwa wierzchołki drzewa z potencjalnie dowolnie wieloma wierzchołkami pośrednimi, wymaga mechanizmów rekurencyjnych. Do tego jednak wyszukiwarki korpusowe ogólnego zastosowania nie są przystosowane. Istnieją co prawda wyspecjalizowane wyszukiwarki do tzw. banków rozbiorów składniowych, ale one nie są z kolei przystosowane do wielowarstwowego znakowania innego rodzaju. W efekcie zatem uwzględnienie informacji składniowej wymaga pewnych kompromisów, które co prawda uniemożliwiają rekurencyjne przeszukiwanie drzew składniowych, ale mimo to rozszerzają możliwości przeszukiwania korpusu. Takim kompromisem jest indeksowanie przy każdym segmencie szczegółowych informacji tylko o jego bezpośrednim nadrzędniku. Te informacje to: forma hasłowa nadrzędnika (atrybut head.base), klasa fleksyjna nadrzędnika (head.pos), typ relacji zależnościowej łączącej oba segmenty (deprel), lewo- lub prawostronne umiejscowienie bezpośredniego nadrzędnika względem segmentu (head.position), odległość liczona w segmentach od bezpośredniego nadrzędnika (head.distance)[4].

Informacja o typie relacji zależności pochodzi ze schematu anotacji Polskiego Banku Zależnościowego (Wróblewska 2014), na który składają się trzydzieści dwie etykiety wskazujące rodzaje relacji zależności składniowej między formami wyrazowymi w wypowiedzeniu. Wśród nich są na przykład relacja subj łącząca formę finitywną z podmiotem, relacja obj

4 Podobne rozwiązanie zastosowano również w wyszukiwarce Kontext Czeskiego Korpusu Narodowego. Zindeksowano w nim typ relacji zależności oraz cechy morfosyntaktyczne bezpośredniego nadrzędnika składniowego, a także niektóre inne cechy specyficzne dla praskiej koncepcji opisu składniowego (Klyueva, Straňák 2016), choć już nie pozycję i odległość bezpośredniego nadrzędnika w linearnym porządku wypowiedzenia.

łącząca formę finitywną z dopełnieniem bliższym oraz relacja comp_inf łącząca formę finitywną z wymaganym przez nią członem w bezokoliczniku[5].

Aby zilustrować użyteczność przedstawionego wyżej znakowania składniowego, posłużymy się przykładami zaczerpniętymi z artykułu poświęconego możliwym zastosowaniom korpusu w badaniu gramatyki, pochodzącego z tomu podsumowującego projekt NKJP (Górski 2012). Autor prezentuje w nim przykłady zjawisk gramatycznych, które można badać metodami ilościowymi za pomocą wyszukań korpusowych w oparciu o warstwę znakowania fleksyjnego, czyli jedynego dostępnego dotąd w NKJP. Wśród nich znalazły się wyszukania wystąpień konstrukcji w stronie biernej oraz analitycznych form czasu przyszłego czasowników niedokonanych. W wypadku strony biernej proponowane zapytanie korpusowe w wyszukiwarce Poliqarp ma postać:

```
[pos="ppas"]  [!pos="(fin|praet|imps|impt)"]{,2}  [base="(być|zostać)"]|
[base="(być|zostać)"]  [!pos="(fin|praet|imps|impt)"]{,2}  [pos="ppas"]
```

Zapytanie jest symetryczne, tzn. pierwsza część zapytania uwzględnia sytuację, gdy słowo posiłkowe BYĆ lub ZOSTAĆ występuje w wypowiedzeniu po imiesłowie, druga – odwrotnie, tzn. gdy słowo posiłkowe w konstrukcji poprzedza imiesłów. Dodatkowo zapytanie uwzględnia również sytuację, gdy słowo posiłkowe i imiesłów nie występują bezpośrednio obok siebie, ale być może są przedzielone jednym lub dwoma segmentami niebędącymi formami czasownikowymi. Ogólniej zaś w zapytaniu tego typu trzeba brać pod uwagę nieciągłość i swobodny szyk konstrukcji składniowej lub analitycznej formy fleksyjnej oraz wyeliminować przynajmniej część potencjalnych nadmiarowych dopasowań. Proponowane zapytanie nie chroni zresztą przed wszystkimi fałszywymi alarmami, nie uwzględnia na przykład sytuacji, w której potencjalne słowo posiłkowe i imiesłów bierny znajdują się w różnych składnikach zdania współrzędnie złożonego, a zatem nie mogą wspólnie tworzyć konstrukcji biernej, choć również o taki przypadek można rozszerzyć powyższe zapytanie. Zapytanie to nie zwróci jednak wypowiedzeń, w których odległość między oboma członami konstrukcji jest większa niż trzy segmenty. Okno można oczywiście poszerzyć, ale za cenę większej liczby fałszywych dopasowań wśród wyników.

Dużo łatwiej znaleźć konstrukcje bierne (i inne podobne zjawiska) za pomocą warstwy informacji składniowej. Zapytanie:

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

zwróci wszystkie takie wypowiedzenia, w których forma leksemu BYĆ lub ZOSTAĆ jest bezpośrednim podrzędnikiem składniowym formy imiesłowu biernego[6]. Można rozszerzyć zapytanie, by dodatkowo uwzględniało tylko konstrukcje, w których oba słowa łączy relacja typu aux

5 Pełny wykaz typów relacji zależności stosowanych w schemacie Polskiego Banku Zależnościowego można znaleźć m.in. na stronie http://zil.ipipan.waw.pl/PDB/DepRelTypes (dostęp: 2 grudnia 2020).

6 Jak zostało to wcześniej wyjaśnione, w znakowaniu składniowym korzystamy ze schematu znakowania przyjętego w Polskim Banku Zależnościowym (PBZ), który powstał niezależnie od projektu NKJP, i nie wprowadzamy w nim żadnych zmian. W szczególności w PBZ przyjęto, że w konstrukcji biernej słowo posiłkowe jest podrzędnikiem imiesłowu biernego, a nie odwrotnie, czyli inaczej niż na przykład w *stricte* powierzchniowych opisach polskiej składni.

(deprel="aux"), co powinno zwiększyć dokładność wyników. Nie ma przy tym znaczenia szyk ani nieciągłość konstrukcji – użytkownik nie musi pamiętać o uwzględnieniu tego typu sytuacji. Mimo to, jeśli zajdzie taka potrzeba, wciąż może mieć kontrolę nad szykiem i odległością linearną obu członów konstrukcji dzięki atrybutom head.distance i head.position. W naturalny sposób wyszukania ograniczone też zostaną do słów znajdujących się w jednym wypowiedzeniu, ponieważ relacje zależności składniowych nie wykraczają poza granice wypowiedzenia.

W podobny sposób można odnaleźć w korpusie wiele innych konstrukcji gramatycznych oraz klasyfikować uzyskane wyniki ze względu na ich szyk i nieciągłość. Zachęcamy czytelnika do eksperymentowania z zapytaniami wykorzystującymi warstwę składniową korpusu.

6. Plany na przyszłość

NKJP wymaga przede wszystkim rozszerzenia o teksty najnowsze reprezentujące w miarę możliwości wszystkie gatunki i style funkcjonalne polszczyzny. Duży reprezentatywny, zrównoważony i aktualny korpus jest bowiem dla języka o statusie takim jak polski absolutnie podstawowym zasobem – mającym zastosowanie zarówno w badaniach czysto lingwistycznych, jak i z zakresu przetwarzania języka i sztucznej inteligencji, a nawet w przemyśle informatycznym. Rozszerzenie podstawy materiałowej NKJP to jednak przedsięwzięcie o zupełnie innej skali niż prace zaprezentowane w niniejszym artykule. Mimo to uważamy utrzymanie tego korpusu na poziomie technicznym i użytkowym odpowiadającym współczesnym standardom za zadanie równie ważne.

Chcielibyśmy w najbliższej przyszłości dbać o techniczną dostępność i przystępność NKJP zwłaszcza dla użytkowników indywidualnych, prowadzących na nim swoje badania. Dlatego planujemy w miarę możliwości uaktualniać dostępne warstwy znakowania automatycznego, by podnosić jego jakość w miarę rozwoju nowych narzędzi informatycznych, a być może również rozszerzyć je o nowe warstwy. Warto w tym celu zadbać o utrzymanie równoległości automatycznego znakowania w NKJP300M i w Korpusomacie, dzięki czemu łatwiej można porównywać wyniki w samodzielnie zebranym materiale i w korpusie referencyjnym.

Planujemy ponadto udostępnienie w ramach tego samego serwisu internetowego ręcznie znakowanego NKJP1M w wersji ze zaktualizowanym tagsetem morfosyntaktycznym, wzbogaconego o warstwę informacji składniowej wraz z wizualizacją drzew zależnościowych.

Bibliografia

Brouwer M., Brugman H., Kemps-Snijders M. 2017: *MTAS: A Solr/Lucene based Multi Tier Annotation Search solution*, [w:] *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, Linköping University Electronic Press, Linköpings Universitet, s. 19–37.

Górski R.L. 2012: *Zastosowanie korpusów w badaniu gramatyki*, [w:] Przepiórkowski i in. (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa, s. 291–301.

Gruszczyński W., Adamiec D., Bronikowska R., Wieczorek A. 2020: *Elektroniczny korpus tekstów polskich z XVII i XVIII w. – Problemy teoretyczne i warsztatowe*, „Poradnik Językowy", z. 8, s. 32–51.

Kieraś W., Kobyliński Ł., Ogrodniczuk M. 2018: *Korpusomat – a tool for creating searchable morphosyntactically tagged corpora*, „Computational Methods in Science and Technology", vol. 24, s. 21–27.

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

Klyueva N., Straňák P. 2016: *Improving corpus search via parsing*, [w:] N. Calzolari i in. (red.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, European Language Resources Association, Portorož, s. 2862–2866.

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

Marcińczuk M., Kocoń J., Gawor M. 2018: *Recognition of Named Entities for Polish-Comparison of Deep Learning and Conditional Random Fields Approaches*, [w:] M. Ogrodniczuk, Ł. Kobyliński (red.), *Proceedings of the PolEval 2018 Workshop*, Institute of Computer Science, Polish Academy of Science, Warszawa, s. 63–73.

NKJP: Narodowy Korpus Języka Polskiego (online: http://nkjp.pl).

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

Przepiórkowski A., Hajnicz E., Andrzejczuk A., Patejuk A., Woliński M. 2017: *Walenty: gruntowny składniowo-semantyczny słownik walencyjny języka polskiego*, „Język Polski" XCVII, s. 30–47.

Rybak P., Wróblewska A. 2018: *Semi-supervised neural system for tagging, parsing and lemmatization*, [w:] *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Brussels, s. 45–54.

Saloni Z. 1974: *Kategoria rodzaju we współczesnym języku polskim*, [w:] Urbańczyk S. i in. (red.), *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, Ossolineum, Wrocław, s. 41–75.

Savary A., Chojnacka-Kuraś M., Wesołek A., Skowrońska D., Śliwiński P. 2012: *Anotacja jednostek nazewniczych*, [w:] A. Przepiórkowski i in. (red.), *Narodowy Korpus Języka Polskiego*, Wydawnictwo Naukowe PWN, Warszawa.

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

Waszczuk J., Kieraś W., Woliński M. 2018: *Morphosyntactic Disambiguation and Segmentation for Historical Polish with Graph-Based Conditional Random Fields*, [w:] P. Sojka i in. (red.), *Text, Speech, and Dialogue. TSD 2018*, Lecture Notes in Computer Science 11107, Springer, s. 188–196.

Woliński M. 2014: *Morfeusz reloaded*, [w:] N. Calzolari i in. (red.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation. LREC 2014*, European Language Resources Association, Reykjavík, s. 1106–1111.

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*, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa.

WSJP PAN: *Wielki słownik języka polskiego PAN*, red. P. Żmigrodzki (online: https://wsjp.pl).

Summary

New multilayer linguistic annotation of the balanced National Corpus of Polish

Keywords: corpus, natural language processing, morphosyntactic annotation, syntactic annotation.

The article describes the well-known and widely used National Corpus of Polish in a new setup. The update consists of the annotation scheme modification in the morphosyntactic layer (especially in its parts related to the grammatical gender), as well as adding new layers of annotation: the syntactic layer and the named entities layer. All three layers are indexed in the MTAS corpus search engine and can be referenced in CQL corpus queries.