WITOLD KIERAŚ*, ŁUKASZ KOBYLIŃSKI**

INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK, WARSZAWA

Korpusomat – stan obecny i przyszłość projektu[1]

Słowa kluczowe: przetwarzanie języka naturalnego, lingwistyka korpusowa, analiza fleksyjna, analiza składniowa, anotacja tekstu.

doi: http://dx.doi.org/10.31286/JP.101.2.4

1. Wprowadzenie

W niniejszym artykule prezentujemy aktualny stan rozwoju Korpusomatu, aplikacji webowej służącej do samodzielnego gromadzenia i indeksowania znakowanych korpusów językowych z tekstów dostarczonych przez użytkownika[2]. Co prawda Korpusomat był już wcześniej prezentowany (Kieraś i in. 2018), od tamtego czasu projekt przeszedł jednak gruntowne zmiany techniczne i koncepcyjne, większość komponentów została wymieniona lub zaktualizowana, a jedyną częścią, która nie uległa zmianie, jest szata graficzna strony.

Pierwotnie Korpusomat miał być jedynie prostym narzędziem, które ułatwia procedurę znakowania fleksyjnego i indeksowania tekstów do formatu czytelnego dla wyszukiwarki korpusowej Poliqarp (Janus, Przepiórkowski 2007), powstałej na użytek korpusu IPI PAN i używanej również (w wersji webowej) w Narodowym Korpusie Języka Polskiego (NKJP). Stosunkowo szybko okazało się jednak, że ograniczenia techniczne Poliqarpa nie nadążają za potrzebami i pomysłami, w szczególności ograniczają możliwości automatycznego znakowania tekstów jedynie do informacji morfosyntaktycznej. Zmiana wyszukiwarki na MTAS (Multi Tier Annotation Search; Brouwer i in. 2017) otworzyła zaś drogę do rozszerzania Korpusomatu o nowe warstwy znakowania automatycznego.

W artykule skupimy się przede wszystkim na części użytkowej aplikacji, czyli możliwości wydobywania informacji o zgromadzonym zbiorze tekstów, które mogą być podstawą analizy i interpretacji *stricte* lingwistycznej. Szczegóły techniczne pozostawiamy w przypisach bibliograficznych odsyłających do konkretnych narzędzi wkomponowanych w potok przetwarzania Korpusomatu.

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

** lkobylinski@ipipan.waw.pl; ORCID: 0000-0003-2462-0020

1 Prace związane z rozwojem aplikacji Korpusomat 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.

2 Aplikacja znajduje się pod adresem http://korpusomat.pl. Można z niej korzystać bez żadnych opłat, wymagane jest jedynie założenie konta w serwisie.

2. Warstwy znakowania automatycznego

Zmiana wyszukiwarki korpusowej pozwoliła na wprowadzenie do budowanych korpusów dowolnej liczby warstw anotacji. Warstwy to różne sposoby klasyfikowania słów w tekście ze względu na przyjęte kryteria. Przykładowo: w warstwie fleksyjnej każdemu segmentowi przypisano takie cechy, jak forma hasłowa (lemat) i interpretacja fleksyjna (znacznik) zgodna z przyjętym wcześniej zestawem etykiet (tagsetem). W poszczególnych warstwach opisowi podlegają albo słowa, albo ich ciągi.

Wielowarstwowy opis tekstu umożliwia precyzyjniejsze przeszukiwanie zindeksowanych dokumentów ze względu na ich cechy językowe. W jednym zapytaniu można się odnosić do różnych warstw i dzięki temu łączyć różne kryteria przeszukiwania korpusu. W dalszej części opiszemy pokrótce istniejące w Korpusomacie warstwy znakowania wraz z prostymi przykładami ich zastosowania. Skupimy się przede wszystkim na nowych warstwach, ponieważ warstwa opisu fleksyjnego była już wcześniej opisywana i działa niemal identycznie jak w wyszukiwarce Poliqarp w NKJP, a także w wielu innych korpusach polskich. Poświęcimy jej zatem tylko krótkie podsumowanie.

2.1. Informacja fleksyjna

Warstwa znakowania fleksyjnego była obecna w Korpusomacie od początku jego istnienia i użytkownikom powinna być dobrze znana z wielu istniejących korpusów dla polszczyzny, w tym z NKJP. Zawiera się w niej podstawowy najdrobniejszy podział tekstu na tzw. segmenty (w przybliżeniu: słowa). W wyniku automatycznej analizy fleksyjnej za pomocą analizatora Morfeusz (Woliński 2014; Kieraś, Woliński 2017) i ujednoznacznienia za pomocą tagera Concraft (Waszczuk i in. 2018) segment zyskuje interpretację – staje się formą fleksyjną z przypisaną formą hasłową i znacznikiem określającym wartości poszczególnych kategorii gramatycznych. Na przykład słowu tekstowemu *śledź* analizator morfologiczny przypisze dwie interpretacje: formy mianownika liczby pojedynczej rzeczownika o formie hasłowej ŚLEDŹ (znacznik: subst:sg:nom:m2) oraz formy czasownikowej trybu rozkazującego drugiej osoby liczby pojedynczej o formie hasłowej ŚLEDZIĆ (znacznik: impt:sg:sec:imperf). Tager spośród tych interpretacji wybierze tę, która w danym kontekście wydaje się bardziej prawdopodobna.

Znacznik składa się z klasy gramatycznej oraz wartości kategorii gramatycznych przypisanych do danej klasy. Klasyfikacja nie opiera się na tradycyjnych częściach mowy, ale na drobniejszym podziale na tzw. klasy fleksemów. Szczegółowy opis tej klasyfikacji nie jest przedmiotem niniejszego artykułu, a różne jej odmiany były już wcześniej opisywane. Schemat ten został pierwotnie zastosowany w korpusie IPI PAN i opisany w artykule Marcina Wolińskiego (2003). Opis późniejszej jego wersji znajduje się między innymi w instrukcji użytkowania NKJP[3], która z kolei stała się podstawą instrukcji języka zapytań Korpusomatu[4] oraz innych instrukcji użytkowania różnych korpusów. Co istotne, na podstawie tej klasyfikacji można formułować zapytania o całe klasy fleksemów, na przykład liczebniki ([pos="num"]), lub formy o konkretnej

3 http://nkjp.pl/poliqarp/help/pl.html.

4 http://korpusomat.pl/query_language_manual.

charakterystyce fleksyjnej, na przykład rodzaju żeńskiego ([gender="f"]), które dodatkowo można łączyć z warunkami dotyczącymi kształtu segmentu (np. [orth="koronawyborów"]) lub jego formy hasłowej (np. [base="epidemiczny"]).

Warto dodać, że w Korpusomacie dostępny jest również atrybut ign o wartościach true i false, dzięki któremu w zapytaniach dotyczących warstwy fleksyjnej można formułować warunki uwzględniające to, czy dane słowo znane było analizatorowi morfologicznemu posługującemu się listą form wyrazowych pochodzących ze *Słownika gramatycznego języka polskiego* (Saloni i in. 2015). Dzięki temu można z wyników wyszukań odfiltrować przypadki, w których tager błędnie „zgadł" znacznik dla słowa spoza słownika, lub odwrotnie – skupić się tylko na takich wynikach, w których występują słowa nieznane analizatorowi, co może mieć zastosowanie w badaniu neologizmów i najnowszego słownictwa nieuwzględnionego w SGJP. Na przykład zapytanie [orth="korona.*" & ign="true"] powinno zwrócić wystąpienia neologizmów związanych z pandemią koronawirusa użytych w zgromadzonych przez użytkownika tekstach, a nieobecnych w SGJP i Morfeuszu.

2.2. Jednostki nazewnicze

Pierwszą – po warstwie fleksyjnej – nową warstwą anotacji jest rozpoznawanie jednostek nazewniczych. Klasyfikacja jednostek nazewniczych pochodzi wprost z projektu NKJP, w którym oznakowano wzorcowy milionowy podkorpus zgodnie z przyjętą i opisaną tam koncepcją (Savary i in. 2012). Kategoria jednostek nazewniczych obejmuje szerszą klasę jednostek leksykalnych niż tylko typowe nazwy własne. W ich skład wchodzą również tradycyjne etnonimy (np. *Szkot*, *Szkotka*) i derywowane od nich przymiotniki (*szkocki*), a także zapisy dat i czasu. Pełna klasyfikacja obejmuje: nazwy osób, w tym człony tych nazw obejmujące imiona, nazwiska i pseudonimy (np. *Jan „Ptaszyn" Wróblewski*), nazwy geograficzne (*Puszcza Białowieska*), nazwy organizacji (*Organizacja Bezpieczeństwa i Współpracy w Europie*) oraz nazwy miejsc lub też tzw. nazwy geopolityczne nazywające dzielnice (*Ursynów*), regiony (*województwo małopolskie*), państwa i kraje o jakimś rodzaju odrębności niebędące niepodległymi państwami (*Nowa Zelandia*, *Republika Górskiego Karabachu*), bloki państw (*NATO*), a także wyrażenia czasowe. Zapytania dotyczące znakowania jednostek nazewniczych można łączyć z warunkami odnoszącymi się do anotacji fleksyjnej, na przykład zapytanie:

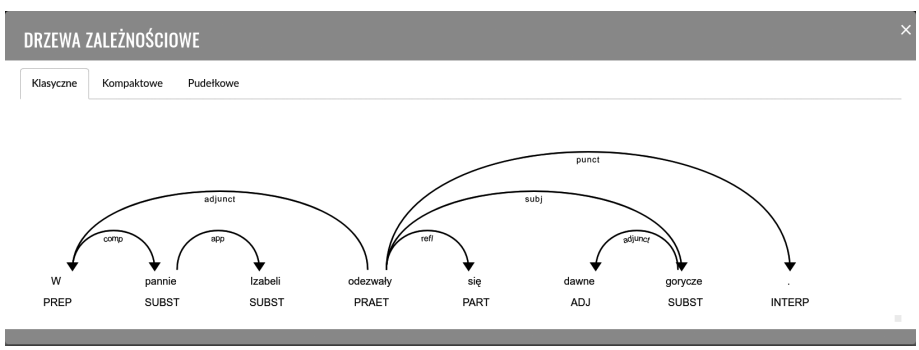
<ne="persName.surname" /> containing [pos="subst" & gender="f"]

znajdzie ciągi słów rozpoznane jako nazwiska osób dodatkowo zawierające co najmniej jeden segment opisany jako rzeczownik w rodzaju żeńskim. Powinno ono zatem znaleźć w tekście nazwiska żeńskie.

W Korpusomacie nie wprowadzaliśmy żadnych zmian do tej klasyfikacji oraz wykorzystaliśmy jedno z istniejących rozwiązań opartych na uczeniu maszynowym (Marcińczuk i in. 2018) do automatycznego oznaczania jednostek nazewniczych w tekstach. Oczywiście, jak zawsze w wypadku automatycznej anotacji, wynik nie jest bezbłędny, ale wystarczająco dobry, by mógł być użyteczny. Przykładowo: za pomocą zapytań odnoszących się do warstwy jednostek nazewniczych można w łatwy sposób uzyskać z tekstu powieści listę wszystkich występujących lub wymienionych w niej postaci.

2.3. Informacja składniowa z parsera zależnościowego

Najnowszą wprowadzoną do Korpusomatu warstwą znakowania jest częściowa informacja składniowa pochodząca z parsera zależnościowego, czyli analizatora składniowego przypisującego zdaniu strukturę składniową w aparacie zależnościowym. Pełna informacja składniowa, czyli możliwość przeszukiwania całych drzew składniowych, wymagałaby zastosowania oddzielnej wyszukiwarki przeznaczonej do przeszukiwania banków rozbiorów składniowych (tzw. *parse banków*). Ponieważ MTAS nie jest wyszukiwarką struktur takich jak drzewa, zindeksowana informacja pochodząca z parsera została ograniczona do cech opisujących relację każdego słowa z jego bezpośrednim nadrzędnikiem składniowym w drzewie. Tymi cechami są: forma hasłowa nadrzędnika, znacznik fleksyjny nadrzędnika, etykieta typu zależności łączącej słowo z jego nadrzędnikiem, lewo- lub prawostronne umiejscowienie nadrzędnika względem danego słowa w porządku linearnym wypowiedzenia oraz liczona w segmentach odległość nadrzędnika od danego słowa. Dzięki takim cechom można łatwo skonstruować zapytania o zjawiska fleksyjne lub składniowe, które trudno byłoby wyrazić wyłącznie za pomocą cech fleksyjnych. Oprócz tego możliwe jest również wyświetlenie pełnego drzewa składniowego dla całego wypowiedzenia. Przykład takiego drzewa znajduje się na rysunku 1.



Rys. 1. Przykładowa wizualizacja rozbioru zależnościowego dla krótkiego zdania

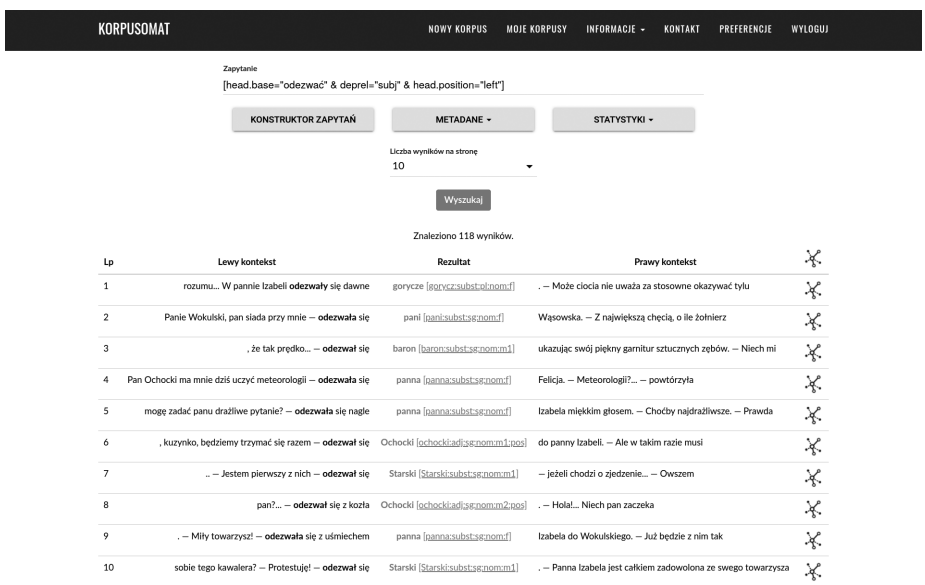
Aby lepiej zrozumieć sens cech zindeksowanych w warstwie składniowej, spójrzmy na przykładowe zdanie z rysunku 1, pochodzące z *Lalki* Bolesława Prusa. Jego centrum składniowym jest forma *odezwały*, od której wychodzą krawędzie łączące ją z:

– *się* morfologicznym (które w niniejszym opisie jest osobnym słowem, choć faktycznie stanowi część formy finitywnej);
– elementem nadrzędnym członu wymaganego (podmiotu), czyli słowem *gorycze*;
– elementem nadrzędnym członu luźnego, czyli przyimkiem *w*;
– znakiem końca (kropką), który zawsze opisuje się jako podrzędnik centrum wypowiedzenia.

Niektóre z tych elementów mają swoje podrzędniki. Każdy element wypowiedzenia (oprócz centrum) ma dokładnie jeden bezpośredni nadrzędnik, dzięki czemu całe wypowiedzenie ma strukturę drzewa. Dodatkowo każda krawędź prowadząca od nadrzędnika do podrzędnika ma etykietę informującą o typie zależności, czyli relacji łączącej oba słowa. Zestaw etykiet

pochodzi z Polskiego Banku Zależnościowego (Wróblewska 2014)[5], na podstawie którego wytrenowano parser COMBO (Rybak, Wróblewska 2018) użyty w Korpusomacie.

Dzięki uwzględnieniu w znakowaniu informacji o cechach i pozycji bezpośredniego nadrzędnika można na przykład zapytać o wszystkie takie elementy, których bezpośrednim nadrzędnikiem jest forma czasownika ODEZWAĆ SIĘ (zapytanie: [head.base="odezwać"]). W przykładowym zdaniu z rysunku 1 jego wynikiem będą cztery wymienione wyżej elementy. Zapytanie można jednak ograniczyć tylko do tych podrzędników, które pełnią funkcję podmiotu ([head.base="odezwać" & deprel="subj"]), wówczas tylko jeden element w przykładowym zdaniu będzie spełniał taki warunek. W całej *Lalce* czasownik ODEZWAĆ SIĘ realizowany wraz z podmiotem pojawia się 127 razy. Charakterystyczną cechą przykładu z rysunku 1 jest szyk – podmiot w porządku linearnym zdania znajduje się po orzeczeniu, a nie przed nim (tzw. szyk VS). Dzięki informacji o pozycji nadrzędnika można dodatkowo sprawdzić, jak częsty jest taki szyk w badanym tekście. W przykładzie z szykiem VS forma finitywna *odezwały* znajduje się na lewo od formy *gorycze* będącej podmiotem. Odpowiednio rozszerzone o dodatkowy warunek wyrażający tę relację linearną zapytanie ([head.base="odezwać" & deprel="subj" & head.position="left"]) odnajdzie wszystkie wystąpienia czasownika ODEZWAĆ SIĘ użyte w realizacji z szykiem VS. Okazuje się, że w *Lalce* takich użyć jest aż 118. Tylko w dziewięciu przypadkach zastosowano szyk kanoniczny SV.



Rys. 2. Wyniki zapytania o przykłady użycia form czasownika ODEZWAĆ SIĘ z realizacją podmiotu po prawej stronie. W lewym kontekście każdego z dopasowań zaznaczono pismem pogrubionym bezpośredni nadrzędnik segmentu wyświetlonego w kolumnie *Rezultat*

5 Schemat znakowania Polskiego Banku Zależnościowego jest propozycją opisu składniowego dla polszczyzny w aparacie zależnościowym. Innym możliwym do zaimplementowania w Korpusomacie rozwiązaniem mógłby być wielojęzyczny schemat Universal Dependencies (Nivre i in. 2016). Uznaliśmy, że będzie on mniej intuicyjny dla polskich użytkowników, ale jeśli zgłoszą oni potrzebę wykorzystania tego schematu, to możliwe jest jego włączenie jako alternatywnego modelu przetwarzania składniowego.

Tego typu zapytania można budować dla form dowolnych leksemów realizujących różne typy zależności. Istotne jest to, że nie muszą one sąsiadować ze sobą w tekście. Oczywiście taka informacja składniowa ma poważne ograniczenia – dotyczy bowiem tylko dwóch segmentów połączonych bezpośrednią relacją. Pomimo to pozwala na wyrażenie w postaci zapytania korpusowego sporej liczby cech gramatycznych oraz konstrukcji składniowych i analitycznych konstrukcji fleksyjnych. Więcej przykładów tego typu zapytań można znaleźć w instrukcji języka zapytań Korpusomatu.

3. Podsumowanie statystyczne korpusów

Oprócz możliwości przeszukiwania zgromadzonych tekstów Korpusomat oferuje również możliwość spojrzenia na ich różnorakie podsumowania statystyczne. Na ekranie statystyk można obejrzeć kilka okien, w których dostępne są różnego rodzaju wyliczenia lub wizualizacje. Nie jest to spójny obraz zbioru tekstów, ale raczej rzut oka na poszczególne aspekty tego zbioru: od listy frekwencyjnej (z możliwością jej filtrowania), przez rozkład słów kluczowych i ekstrakcję terminologii, aż po prostą wizualizację stylometryczną. Celem tego podsumowania jest przede wszystkim podsuwanie możliwych tropów dalszej analizy tekstów, w tym również takich, które wymagają zastosowania innych wyspecjalizowanych narzędzi, jak TermoPL do wydobywania terminologii z tekstów specjalistycznych (Marciniak i in. 2017) czy stylometryczna biblioteka Stylo dla języka programowania R (Eder i in. 2016). Nie jest to zatem gotowa pełnowartościowa analiza kwantytatywna, a jedynie podpowiedź dotycząca tego, jakie potencjalne możliwości może dać właściwa analiza ilościowa. W szczególności na ekranie statystyk prezentowane jest charakterystyczne słownictwo stosowane w korpusie, co pozwala na szybkie podjęcie decyzji o wartości ewentualnej dalszej analizy w tym kierunku z wykorzystaniem dodatkowych możliwości parametryzowania aplikacji TermoPL. Prezentowane jest również w sposób graficzny podobieństwo stylistyczne pomiędzy tekstami korpusu, które może stanowić punkt wyjścia do bardziej zaawansowanych analiz tego rodzaju. Wykorzystywane są metody grupujące i agregujące dane pod różnym kątem, aby zakres tej podstawowej analizy był możliwie szeroki. Dostępne są zatem podejścia rozpatrujące korpus jako całość, biorące pod uwagę poszczególne teksty w oderwaniu od siebie, a także rozkład treści w tekstach z uwzględnieniem ich naturalnej kolejności odczytywania. Na przykład rozkład słów kluczowych w tekście prezentuje w sposób graficzny, jak częste jest użycie poszczególnych słów kluczowych w kolejnych fragmentach tekstu, co pozwala zaobserwować sposób ich rozkładu, na przykład: jednostajny lub skoncentrowany w jednym lub kilku fragmentach tekstu. Wszystkie listy generowane na ekranie statystyk można pobrać i wykorzystać także poza Korpusomatem w celu prowadzenia dalszych analiz.

4. Współdzielenie korpusów

Jedną z istotnych nowości w Korpusomacie jest możliwość współdzielenia korpusów. Każdy korpus zbudowany w aplikacji jest domyślnie zasobem przypisanym wyłącznie do użytkownika, który go stworzył. Możliwe jest jednak współdzielenie korpusów z innymi użytkownikami – wybranymi (przez podanie listy ich identyfikatorów w aplikacji) lub wszystkimi (przez

upublicznienie zasobu). W pierwszym wypadku możliwe są dwa tryby udostępnienia: tylko do odczytu lub z prawem do edycji. Prawa do edycji spowodują, że inni użytkownicy będą mogli modyfikować korpus, w szczególności również dodawać i usuwać z niego teksty, a także pobrać pliki źródłowe korpusu, należy z nich więc korzystać ostrożnie.

Współdzielenie korpusów ma zastosowanie przede wszystkim w małych projektach badawczych, w których kilkoro naukowców pracuje na jednej bazie tekstów, a także w wypadku wykorzystywania Korpusomatu jako wsparcia zajęć uniwersyteckich, na których prowadzący udostępnia studentom przygotowany przez siebie zasób. Upublicznianie korpusów warto rozważyć w odniesieniu do zbioru będącego podstawą materiałową publikacji naukowej, dla której może on stanowić uzupełnienie. Warunkiem upubliczniania korpusów jest oczywiście dysponowanie prawami do rozpowszechniania tekstów w nim zawartych, choć w publicznych korpusach inni użytkownicy nie mogą pobrać tekstów źródłowych ani plików XML Korpusomatu, mogą jedynie przeszukiwać korpus za pomocą zapytań i oglądać konkordancje wraz z ich ograniczonym kontekstem. Takie same ograniczenia dotyczą również korpusów udostępnianych konkretnym użytkownikom tylko do odczytu.

5. Wykorzystanie Korpusomatu

Korpusomat okazał się narzędziem, które jest przydatne nie tylko w swojej podstawowej formie prezentowanej powyżej, ale również w wielu zastosowaniach pokrewnych. Pierwszym oczywistym zastosowaniem było wykorzystanie rdzenia Korpusomatu przy udostępnianiu powstających korpusów językowych, tworzonych w projektach naukowych realizowanych w IPI PAN lub we współpracy z nim. Wśród nich są między innymi Korpus Dyskursu Parlamentarnego[6] (Ogrodniczuk 2018) i Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772)[7] (Gruszczyński i in. 2020). W obu zastosowano różne warstwy normalizacji i anotacji oryginalnego tekstu, ale trzon techniczny oraz interfejs są te same i pochodzą z Korpusomatu.

Mechanizmy wykorzystane w Korpusomacie okazały się również bardzo potrzebne przy rozwiązywaniu nieco odleglejszych problemów, na przykład w analizie Zintegrowanego Rejestru Kwalifikacji[8] prowadzonego przez Instytut Badań Edukacyjnych. Korpusomat został tu wykorzystany nie tylko do ułatwienia wewnętrznej pracy nad rejestrem, a więc do jego pełnotekstowego przeszukiwania uwzględniającego zapytania dotyczące różnych warstw anotacji językowej, ale też przy tworzeniu raportów i podsumowań dających jego redaktorom pełniejszy obraz rejestru. Korpusomatu użyto również jako podstawy do stworzenia aplikacji umożliwiającej automatyczne grupowanie i klasyfikację poszczególnych rekordów Zintegrowanego Rejestru Kwalifikacji za pomocą metod uczenia maszynowego i przetwarzania języka naturalnego na dużym zbiorze tekstów.

6 https://kdp.nlp.ipipan.waw.pl/.

7 https://korba.edu.pl/.

8 https://rejestr.kwalifikacje.gov.pl/.

6. Plany na przyszłość

Po czterech latach od udostępnienia pierwszej wersji aplikacji można uznać, że jest ona projektem dojrzałym, ale będzie dalej rozwijana i ulepszana. Skupimy się przede wszystkim na zwiększaniu wydajności automatycznej anotacji tekstów w miarę przybywania użytkowników, jak również na podnoszeniu jakości tej anotacji w miarę powstawania coraz lepszych narzędzi dla języka polskiego. Dotyczy to przede wszystkim analizy fleksyjnej i składniowej, których jakość w niedalekiej przyszłości może się istotnie poprawić dzięki zastosowaniu coraz nowszych generacji modeli językowych powstających również dla języka polskiego. Aktualnie najlepsze wyniki w prototypowych narzędziach osiągają rozwiązania korzystające z modeli typu BERT[9], są one jednak również dużo bardziej wymagające obliczeniowo, co utrudnia ich wykorzystanie w ogólnodostępnej aplikacji webowej (choć go nie uniemożliwia).

Zespół pracuje również nad kilkoma dodatkowymi funkcjami Korpusomatu. Jedną z nich jest możliwość automatycznego porównywania korpusów ze zbioru użytkownika, polegająca na wydobyciu za pomocą analiz statystycznych takich cech gramatycznych, stylistycznych lub leksykalnych, które możliwie najbardziej dane zbiory różnicują lub pod względem których są one do siebie najbardziej podobne. W wypadku podobieństw leksykalnych pozwoli to na uchwycenie podobieństwa tematycznego tekstów, w odniesieniu zaś do cech gramatycznych i stylistycznych – na wskazanie, które teksty pochodzą z podobnych rejestrów. Innym przygotowywanym rozszerzeniem Korpusomatu jest również włączenie do potoku przetwarzania prostej analizy wydźwięku (ang. *sentiment analysis*), polegającej na przypisaniu wybranym słowom w zgromadzonym korpusie jednej z kilku wartości nacechowania wydźwięku emocjonalnego. Pozwoli to na określenie, które ze zgromadzonych w korpusie tekstów lub które ich fragmenty są najsilniej nacechowane wykładnikami tekstowymi podstawowych emocji.

Prowadzone są również wstępne prace nad udostępnieniem wielojęzycznej wersji Korpusomatu, umożliwiającej przetwarzanie tekstów w językach innych niż polski. Spodziewamy się, że będzie to jedna z poważniejszych zmian w Korpusomacie, zarówno od strony implementacyjnej, a więc architektury aplikacji, jak też od strony użytkownika, który od chwili udostępnienia tego rodzaju funkcji uzyska znacząco szersze spektrum możliwości wykorzystania omawianego oprogramowania, w ramach jednego spójnego interfejsu powiązanego z repozytorium danych badawczych – korpusów tekstowych.

Innym istotnym problemem, którym zajmiemy się w kolejnych odsłonach Korpusomatu, jest kwestia pozyskiwania danych i ich wstępnego przetwarzania przed rozpoczęciem analiz językowych. W szczególności planujemy zaoferować możliwość bardziej zautomatyzowanego pobierania danych z wielu źródeł internetowych, a także wprowadzić funkcję ręcznej korekty tekstu przesłanego do Korpusomatu przed rozpoczęciem przetwarzania. Pozwoli to

9 BERT, czyli *Bidirectional Encoder Representations from Transformers* to rodzina modeli językowych najnowszej generacji stworzona przez firmę Google. Modele tego typu powstają również poza Google. Jednym z takich modeli jest stworzony przez firmę Allegro we współpracy z IPI PAN model HerBERT, którego wykorzystanie planujemy.

na uniknięcie problemów z włączeniem niechcianych partii tekstów do wyników analiz statystycznych czy do wyników przeszukiwania korpusu.

Bibliografia

Brouwer M., Brugman H., Kemps-Snijders M. 2017: *MTAS: A Solr/Lucene based Multi Tier An-notation 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.

Eder M., Rybicki J., Kestemont M. 2016: *Stylometry with R: a package for computational text analysis*, „The R Journal", vol. 8, s. 107–121.

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.

Janus D., Przepiórkowski A. 2007: *Poliqarp: An open source corpus indexer and search engine with syntactic extensions*, [w:] *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, s. 85–88.

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.

Marciniak M., Mykowiecka A., Rychlik P. 2017: *Automatyczne wydobywanie terminologii dziedzinowej z korpusów tekstowych*, „Język Polski" XCVII, s. 64–74.

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.

Nivre J., de Marneffe M., Ginter F., Goldberg Y., Hajič J., Manning Ch., McDonald R., Petrov S., Pyysalo S., Silveira N., Tsarfaty R., Zeman D. 2016: *Universal Dependencies v1: A Multilingual Treebank Collection*, [w:] N. Calzorali i in. (red.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association, Portorož, s. 1659–1666.

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

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), Paris, s. 15–19.

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

Rybak P., Wróblewska A. 2018: *Semi-supervised neural system for tagging, parsing and lemmatization*, [w:] D. Zeman, J. Hajič (red.), *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., Woliński M., Wołosz R., Gruszczyński W., Skowrońska D. 2015: *Słownik gramatyczny języka polskiego*, Warszawa, wyd. 3.

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, s. 129–167.

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

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

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

Summary

Korpusomat – present state and the future of the project

Keywords: natural language processing, corpus linguistics, inflectional analysis, syntactic analysis, text annotation.

The article presents the Korpusomat web application for creating user's own annotated linguistic corpora. The application offers an automatic annotation of texts and the ability to search it based on the annotation of inflectional and syntactic features of words and named entities. All annotation layers are presented along with examples of their application in linguistic analysis. The Korpusomat also offers statistical summaries of the collected data, as well as the possibility of sharing the created corpora with other users.