Instrukcja korzystania z wyszukiwarki do *Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII wieku (do 1772 r.)*

Spis treści

Wprowadzenie

Niniejsza instrukcja opisuje sposoby korzystania z zasobów zgromadzonych w *Elektronicznym Korpusie Tekstów Polskich z XVII i XVIII w. (do 1772 r.)* za pomocą wyszukiwarki MTAS. Dokument ten stanowi zmodyfikowaną wersję instrukcji opisującej sposób działania wyszukiwarki MTAS, opracowanej przez Witolda Kierasia na potrzeby *Korpusu tekstów polskich z lat 1830–1918*. Z kolei podstawą tej instrukcji była *Ściągawka do Narodowego Korpusu Języka Polskiego*, czyli instrukcja użytkowania wyszukiwarki Poliqarp, wykorzystującej, podobnie jak MTAS, język zapytań znany pod nazwą *Corpus Query Language* (CQL). *Ściągawka* została opracowana przez Adama Przepiórkowskiego, a następnie poprawiona i rozszerzona przez Jakuba Wilka i Aleksandra Buczyńskiego. Jej pełna wersja znajduje się w repozytorium wyszukiwarki Poliqarp. Modyfikacje wprowadzone do pierwotnej wersji instrukcji uwzględniają różnice w języku zapytań oraz specyfikę korzystania z korpusu historycznego. Za zgodą wszystkich wyżej wymienionych autorów niniejsza wersja dokumentu zostaje udostępniona na zasadach licencji Creative Commons BY-SA.

1. Teksty w korpusie — warstwa transliteracji, transkrypcji i anotacji

Teksty w korpusie dostępne są w dwóch formach ortograficznych — w transliteracji i transkrypcji. W wypadku tekstów pozyskanych bezpośrednio z XVII- i XVIII-wiecznych

druków lub rękopisów transliteracja oddaje ortografię podstawy. Drobne różnice dotyczą jedynie niektórych cech (typo)graficznych, np. nie uwzględniamy różnych form grafemów *s* i *r*, a ligatury oddajemy za pomocą osobnych liter (np. *ß* jako *sz*). W tekstach pozyskanych z wydań późniejszych (filologicznych) wersja transliterowana odpowiada dokładnie temu, co znajduje się w wydaniu. Skutkiem tego jest np. występowanie w pewnych tekstach (pochodzących z wydań dziewiętnastowiecznych) grafemu *é*, który w zasadzie nie pojawiał się w drukach barokowych. Wszystkie teksty poddano automatycznej transkrypcji. W transkrypcji zostały zachowane niektóre cechy fonetyczne i morfologiczne charakterystyczne dla średniopolszczyzny, np. forma imiesłowu przysłówkowego uprzedniego zapisana w oryginale *donioźszy* pojawia się w transkrypcji w postaci *doniósszy* (a nie *doniósłszy*). W tekstach zarówno transliterowanych, jak i transkrybowanych zachowana została pisownia wielką lub małą literą zgodnie z oryginałem (lub podstawą).

Wszystkie teksty zostały poddane automatycznej anotacji morfosyntaktycznej, tzn. poszczególnym jednostkom tekstu (segmentom) zostały przypisane znaczniki określające ich formę podstawową (lemat), klasę gramatyczną oraz odpowiednie wartości kategorii gramatycznych. Przez segmenty rozumiemy zazwyczaj słowa tekstowe (ciągi liter ograniczone spacjami lub znakami interpunkcyjnymi), ale w niektórych wypadkach w obrębie słowa wyróżniamy kilka segmentów. Przede wszystkim wyszukiwarka ignoruje oryginalną segmentację tekstów barokowych, tzn. słowa zapisane niezgodnie ze współczesną ortografią mają segmentację uwspółcześnioną, np. w ciągu *niefrasować* zostaną wyróżnione dwa segmenty: [*nie*] i [*frasować*], a w ciągu *nioczym* '*o niczym*' — trzy segmenty: [*ni*], [*o*], [*czym*]. Ponadto, analogicznie jak w NKJP, jako odrębne segmenty traktowane są tzw. ruchome końcówki osobowe czasowników, np. [*łgał*][*eś*], [*długo*][*śmy*], partykuły *by*, *-ż(e)*, *-li* i *-ć* (*-ci*), np. [*przyszedł*][*by*], [*pragnę*][*ć*], poprzyimkowa nieakcentowana forma zaimka ON, np. [*do*][*ń*] oraz w niektórych wypadkach poszczególne człony wyrazu zawierającego łącznik, np. [*melancholiczno*][-][*choleryczny*], [*puku*][-][*huku*]. Za segmenty uznajemy także znaki interpunkcyjne.

2. Zestaw znaczników morfosyntaktycznych

Każdy znacznik morfosyntaktyczny jest ciągiem wartości rozdzielonych dwukropkami. Przykładowo segmentowi *textem* (postać transliterowana) zostały przypisane następujące wartości: `tekst:subst:sg:inst:m`. Pierwsza wartość (`tekst`) to forma podstawowa (lemat) zapisana w transkrypcji, druga (`subst`) określa klasę gramatyczną, następne zaś to wartości odpowiednich dla tej klasy kategorii gramatycznych. Repertuar klas i kategorii gramatycznych oparty jest na zbiorze klas i kategorii (tagsecie) stosowanym w analizatorze morfologicznym Morfeusz, został on jednak dostosowany do języka średniopolskiego.

2.1. Klasy gramatyczne

W *Elektronicznym Korpusie Tekstów Polskich z XVII i XVIII w.*, podobnie jak w NKJP, klasy gramatyczne są oparte na pojęciu fleksemu, będącym pojęciem węższym od terminu leksem. O ile tradycyjnie ujmowane leksemy mogą zawierać formy, którym przysługują różne kategorie gramatyczne[1], o tyle fleksemy stanowią zbiory tylko tych form, które można scharakteryzować za pomocą tych samych kategorii gramatycznych.

[1] Przykładowo w obrębie leksemu czasownikowego można wyróżnić m.in. osobowe formy czasownika (odmienne przez osoby i liczby), bezokolicznik (nieodmienny), a w niektórych ujęciach także odsłownik (odmienny przez przypadki i liczby).

Tradycyjne części mowy	**Fleksem**	**Przykład**	**Symbol**	**Charakterystyka formy podstawowej**	**Forma podstawowa**
rzeczowniki	rzeczownik	*woda*, *drzwi*	`subst`	M. l. poj. (*pl. tant.* — l. mn.)	WODA, DRZWI
liczebniki	liczebnik główny	*dwa*, *pięciu*	`num`	M. rodz. `mnanim`	DWA, PIĘĆ
	liczebnik zbiorowy	*dwoje*	`numcol`	M. rodz. `mnanim`	DWA
	liczebnik przymiotnikowy	*drugi*, *dwojaki*	`adjnum`	M. l. poj. rodz. `m`	DRUGI, DWOJAKI
	liczebnik przysłówkowy	*dwakroć*, *dwojako*	`advnum`	jedyna forma fleksemu	DWAKROĆ, DWOJAKO
przymiotniki	przymiotnik	*polski*, *dobry*	`adj`	M. l. poj. rodz. `m` st. równego odm. złoż.	POLSKI, DOBRY
	przymiotnik odm. niezłożona	*zdrów*, *polsku*	`adjb`	M. l. poj. rodz. `m` st. równego odm. złoż.	ZDROWY, POLSKI
	przymiotnik przyprzymiotnikowy	*polsko-*	`adja`	M. l. poj. rodz. `m` st. równego odm. złoż.	POLSKI
przysłówki	przysłówek	*dobrze*, *barzo*	`adv`	forma stopnia równego	DOBRZE, BARZO
zaimki	zaimek nietrzecioos.	*ja*, *ty*, *my*, *wy*	`ppron12`	mianownik	JA, TY, MY, WY
	zaimek trzecioos.	*on*	`ppron3`	mianownik l. poj.	ON
	zaimek SIEBIE	*siebie*	`siebie`	biernik	SIEBIE
czasowniki	forma nieprzeszła	*czytam*	`fin`	bezokolicznik	CZYTAĆ
	forma przyszła BYĆ	*będę (tam)*	`bedzie`	bezokolicznik	BYĆ
	pseudoimiesłów	*czytał*	`praet`	bezokolicznik	CZYTAĆ
	rozkaźnik	*czytaj*	`impt`	bezokolicznik	CZYTAĆ
	bezosobnik	*czytano*	`imps`	bezokolicznik	CZYTAĆ
	bezokolicznik	*czytać*	`inf`	bezokolicznik	CZYTAĆ
	odsłownik	*czytanie*	`ger`	bezokolicznik	CZYTAĆ
	imiesłów przysłówk. współczesny	*czytając*	`pcon`	bezokolicznik	CZYTAĆ
	imiesłów przysłówk. uprzedni	*(prze)czytawszy*	`pant`	bezokolicznik	(PRZE)CZYTAĆ
	imiesłów przymiot. czynny	*czytający*, *bolątszy*	`pact`	bezokolicznik	CZYTAĆ, BOLEĆ
	imiesłów przymiot. czynny odm. niezł.	*będęcy*, *jadęcy*	`pactb`	bezokolicznik	BYĆ, JECHAĆ
	imiesłów przymiot. bierny	*czytany*	`ppas`	bezokolicznik	CZYTAĆ
	imiesłów przymiot. bierny odm. niezł.	*umęczon*	`ppasb`	bezokolicznik	UMĘCZYĆ
	imiesłów przeszły	*osiwiały*	`ppraet`	bezokolicznik	OSIWIEĆ
	forma BYĆ — wykładnik cz. przyszłego	*będę (czytać)*	`fut`	bezokolicznik	BYĆ
	forma BYĆ — wykładnik cz. zaprzeszłego	*był (czytał)*	`plusq`	bezokolicznik	BYĆ
	aglutynant BYĆ	*-(e)m*, *-(e)śmy*	`aglt`	bezokolicznik	BYĆ
	aglutynant aoryst. BYĆ	*-(e)ch*, *-(e)chmy*	`agltaor`	bezokolicznik	BYĆ
	czasownik typu WINIEN	*winien*, *powinien*	`winien`	forma l. poj. rodz. `m`	WINIEN, POWINIEN
	predykatyw	*trzeba*	`pred`	jedyna forma fleksemu	TRZEBA
przyimki	przyimek	*na*, *przeze*	`prep`	forma niewokaliczna	NA, PRZEZ
spójniki	spójnik współrz.	*oraz*, *i*, *więc*	`conj`	jedyna forma fleksemu	ORAZ, I, WIĘC
	spójnik podrz.	*że*, *ponieważ*	`comp`	jedyna forma fleksemu	ŻE, PONIEWAŻ
partykuły	kublik	*nie*, *-li*, *-że*	`qub`	niewokaliczna forma fleksemu	NIE, -LI, -Ż
wykrzykniki	wykrzyknik	*ach*, *dlaboga*	`interj`	jedyna forma fleksemu	ACH, DLABOGA
inne	skrót	*r.*	`brev`	forma hasłowa rozwinięcia skrótu	ROK
	człon wyrażenia	*kolwiek*	`frag`	jedyna forma fleksemu	KOLWIEK
	znak interpunkcyjny	*:*, *!*, *?*	`interp`	jedyna forma fleksemu	:, !, ?
	ciało obce	*item*	`xxx`	forma fleksemu występująca w tekście	ITEM
	zapis cyfr. (notacja arabska)	*1*, *225*	`dig`	jedyna forma fleksemu	1, 225
	zapis cyfr. (notacja rzymska)	*MDCCLIII*	`romandig`	jedyna forma fleksemu	MDCCLIII
	wyraz nieodmienny o niejasnej funkcji		`ignndm`	jedyna forma fleksemu	
	wyraz o niejasnej funkcji i niejasnej lematyzacji		`ign`	forma fleksemu występująca w tekście	

Tabela 1. Klasy gramatyczne

	liczba	przypadek	rodzaj	osoba	stopień	aspekt	zanegowanie	akcentowość	poprzyim.	aglutyn.	wokal.	kropk.
rzeczownik	⊕	⊕	⊙									
liczebnik główny	⊙	⊕	⊕									
liczebnik zbiorowy	⊙	⊕	⊙									
liczebnik przymiotnikowy	⊕	⊕	⊕		⊕							
liczebnik przysłówkowy					⊕							
przymiotnik	⊕	⊕	⊕		⊕							
przymiotnik odm. niezłożona	⊙	⊕	⊕		⊙							
przysłówek					⊕							
zaimek nietrzecioosobowy	⊙	⊕	⊕	⊙				⊕				
zaimek trzecioosobowy	⊙	⊕	⊕	⊙				⊕	⊕			
zaimek SIEBIE		⊕										
forma nieprzeszła	⊕			⊕		⊙						
forma przyszła BYĆ	⊕			⊕		⊙						
pseudoimiesłów	⊕		⊕			⊙				⊕		
rozkaźnik	⊕			⊕		⊙						
bezosobnik						⊙						
bezokolicznik						⊙						
odsłownik	⊕	⊕	⊙			⊙	⊕					
im. przys. współczesny						⊙						
im. przys. uprzedni						⊙						
imiesłów przymiotnikowy czynny odm. złożona	⊕	⊕	⊕		⊕	⊙	⊕					
imiesłów przymiotnikowy czynny odm. niezłożona	⊙	⊙	⊙		⊙	⊙	⊕					
imiesłów przymiotnikowy bierny odm. złożona	⊕	⊕	⊕		⊕	⊙	⊕					
imiesłów przymiotnikowy bierny odm. niezłożona	⊙	⊕	⊕		⊙	⊙	⊕					
imiesłów przeszły	⊕	⊕	⊕		⊕	⊙	⊕					
forma BYĆ cz. przyszłego	⊕			⊕		⊙						
forma BYĆ cz. zaprzeszłego	⊕		⊕			⊙					⊕	
aglutynant BYĆ	⊕			⊕		⊙					⊕	
aglutynant aorystyczny BYĆ	⊕			⊕		⊙					⊕	
czasownik typu WINIEN	⊕		⊕			⊙						
przyimek		⊙									⊕	
kublik											⊕	
skrót												⊕

Tabela 2. Charakterystyka morfosyntaktyczna klas gramatycznych

Tabela 1 zawiera listę wszystkich klas gramatycznych przyjętych w niniejszym tagsecie (w zestawieniu z tradycyjnymi częściami mowy) wraz z informacją o ich formach podstawowych oraz o ich symbolach używanych w korpusie[2]. W stosunku do zestawu klas gramatycznych wyróżnionych w NKJP zostały tu zastosowane następujące modyfikcje:

- rezygnacja z traktowania form deprecjatywnych rzeczownika jako odrębnych fleksemów (różnica między formami typu *królowie* i *króle* została oddana za pomocą różnych wartości kategorii rodzaju — por. p. 2.2);
- wyróżnienie (w oparciu o kryterium semantyczne) dwóch dodatkowych fleksemów liczebnikowych: liczebnika przymiotnikowego (`adjnum`), np. *dwojaki*, pełniącego funkcję przymiotnika, i liczebnika przysłówkowego (`advnum`), np. *dwojako*, pełniącego funkcję przysłówka;
- zastąpienie dwóch fleksemów — przymiotnika poprzyimkowego (`adjc`) i przymiotnika predykatywnego (`adjp`) — fleksemem o nazwie *przymiotnik w odmianie niezłożonej* (`adjb`), zawierającym formy typu *polsku*, *zdrów*, *gwałtownę*;
- wyróżnienie użyć czasownika BYĆ w funkcji słowa posiłkowego czasów złożonych: przyszłego (`fut`), np. *będę* (*jechał*), i zaprzeszłego (`plusq`), np. *był* (*jechał*);
- wyróżnienie aglutynantu aorystycznego (`agltaor`), obejmującego dawne końcówki osobowe czasownika, typu *-(e)ch* i *-(e)chmy*;
- powiększenie zbioru imiesłowów przymiotnikowych o imiesłów przeszły (`ppraet`), np. *osiwiały*, a także o imiesłowy w odmianie niezłożonej: czynny (`pactb`), np. *będęcy*, i bierny (`ppasb`), np. *umęczon*.

Tabela 2 zawiera przybliżoną charakterystykę morfosyntaktyczną wyróżnionych w korpusie klas gramatycznych (z pominięciem tych klas, którym nie przysługują żadne kategorie gramatyczne). Symbol $\oplus$ oznacza, że dla danej klasy gramatycznej dana kategoria gramatyczna jest morfologiczna (fleksemy należące do tej klasy „odmieniają się" przez tę kategorię), zaś symbol $\odot$ oznacza, że dana kategoria jest słownikowa (wszystkie formy fleksemu należącego do tej klasy mają tę samą wartość tej kategorii[3]).

2.2. Kategorie gramatyczne

Tabela 3 przedstawia repertuar kategorii gramatycznych oraz ich wartości używanych w korpusie. W porównaniu z tagsetem NKJP występują tu następujące zmiany:

- rezygnacja z kategorii akomodacyjności form liczebników;
- dodatkowa wartość kategorii liczby — liczba podwójna (`du`);
- dodatkowa wartość kategorii aspektu — aspekt podwójny (`biasp`), przypisywany czasownikom dwuaspektowym (np. *abdykować*) oraz tym, których aspekt jest niemożliwy do ustalenia ze względu na niewystępowanie w korpusie kontekstów diagnostycznych;
- dwie dodatkowe wartości kategorii rodzaju — rodzaj przymnogi osobowy (`p1`) i przymnogi nieosobowy (`p2`) — przypisywane rzeczownikom *plurale tantum*;
- inne zmiany w kategorii rodzaju: rodzaj męski uogólniony (`m`), który ma trzy podrodzaje — męski żywotny 1 (`manim1`), np. *panowie*, *ptacy*; męski żywotny 2 (`manim2`), np. *pany*, *ptaki*; męski nieżywotny (`mnanim`), np. *stół*.

[2] Dwie ostatnie klasy: `ignndm` i `ign` mogą pojawić się wyłącznie w korpusie anotowanym ręcznie. W taki sposób zostały oznaczone formy, których funkcji i/lub postaci lematu anotatorzy nie byli w stanie rozpoznać.

[3] Wyjątkiem jest tutaj kategoria rodzaju w odniesieniu do rzeczowników — poszczególnym formom rzeczownika mogą zostać przypisane różne wartości kategorii rodzaju pod warunkiem, że są niesprzeczne (np. `m` i `manim1`) — patrz p. 2.2.

Liczba		
pojedyncza	`sg`	*rzeka*
podwójna	`du`	M. B. W. (*dwa*) *szczyta*, *męża*; (*dwie*) *ręce*, *świecy*; (*dwie*) *ście*, *plecy*; D. Msc. (*dwu*) *panu*, *królu*; *kopu*, *niedzielu*; *latu*, *pokoleniu*; C.N. (*dwiema*) *mężoma*, *zakonoma*; *niewiastama*, *rzeczoma*; *latoma*, *plecoma*
mnoga	`pl`	*rzeki*
Przypadek		
mianownik	`nom`	*woda*
dopełniacz	`gen`	*wody*
celownik	`dat`	*wodzie*
biernik	`acc`	*wodę*
narzędnik	`inst`	*wodą*
miejscownik	`loc`	*wodzie*
wołacz	`voc`	*wodo*
Rodzaj		
męski (uogólniony)	`m`	*aktor*, *baran*, *dzban*
męski rzeczowy	`mnanim`	*stół*
męski żywotny 1	`manim1`	*aktorzy*, *wilcy*
męski żywotny 2	`manim2`	*baranki*, *babsztyle*
żeński	`f`	*stuła*
nijaki	`n`	*dziecko*, *okno*, *co*
przymnogi osobowy	`p1`	*(jadą owi) Państwo*, *królestwo (jedli)*
przymnogi nieosobowy	`p2`	*arcaby*
Osoba		
pierwsza	`pri`	*piszę*, *piszewa*, *piszemy*
druga	`sec`	*piszesz*, *piszeta*, *piszecie*
trzecia	`ter`	*pisze*, *piszeta*, *piszą*
Stopień		
równy	`pos`	*cudny*
wyższy	`com`	*cudniejszy*
najwyższy	`sup`	*nacudniejszy*
Aspekt		
niedokonany	`imperf`	*iść*
dokonany	`perf`	*zajść*
podwójny	`biasp`	*abdykować*
Zanegowanie		
niezanegowana	`aff`	*pisanie*, *czytanego*
zanegowana	`neg`	*niepisanie*, *nieczytanego*
Akcentowość		
akcentowana	`akc`	*niego*, *jego*, *tobie*
nieakcentowana	`nakc`	*go*, *-ń*, *ci*
zneutralizowana	`zneut`	*one*, *im*, *je*
Poprzyimkowość		
poprzyimkowa	`praep`	*niego*, *-ń*
niepoprzyimkowa	`npraep`	*jego*, *go*
Aglutynacyjność		
nieaglutynacyjna	`nagl`	*niósł*
aglutynacyjna	`agl`	*niosł-*
Wokaliczność		
wokaliczna	`wok`	*-em*
niewokaliczna	`nwok`	*-m*
Kropkowalność		
z następującą kropką	`pun`	*r*
bez następującej kropki	`npun`	*zł*

Tabela 3. Kategorie gramatyczne

Największa modyfikacja w obrębie kategorii rodzaju wiąże się jednak nie tyle ze zmianą w zakresie znaczników, ile z przyjęciem odmiennych niż w NKJP procedur przypisywania wartości rodzaju poszczególnym formom. Ze względu na częsty brak odpowiednich kontekstów diagnostycznych, które pozwoliłyby określić rodzaj całego fleksemu rzeczownikowego, przyjęliśmy zasadę, że rodzaj jest przypisywany poszczególnym formom i określany z takim stopniem dokładności, na jaki pozwala kontekst, w którym dana forma występuje. Przykładowo formie *alarmami* zostanie przyporządkowana wartość rodzaju `0` (tzw. rodzaju zneutralizowanego, który jest równoważny alternatywie wszystkich rodzajów przyjętych w tagsecie), gdyż na podstawie postaci narzędnika liczby mnogiej nie można określić, czy forma ta reprezentuje rzeczownik rodzaju męskiego (ALARM), żeńskiego (ALARMA) czy nijakiego (ALARMO)[4]. Natomiast forma *alarmę* otrzyma wartość (`f`), ponieważ taka postać biernika umożliwia precyzyjne określenie rodzaju.

Podobnie wygląda dystrybucja wartości rodzaju męskiego i jego trzech podrodzajów. Wartości odpowiednich podrodzajów przypisujemy tylko tym formom, które pozwalają określić żywotność lub nieżywotność rzeczownika, czyli formom w mianowniku liczby mnogiej lub w bierniku (obu liczb). W pozostałych przypadkach formy rzeczownikowe otrzymają wartość rodzaju męskiego uogólnionego (`m`). A zatem forma (*ci*) *tygrysowie* zostanie scharakteryzowana jako `manim1`, a forma (*nie ma*) *tygrysów* — jako `m`.

3. Automatyczna anotacja — dwie wersje korpusu

Jak było powiedziane, teksty zgromadzone w *Elektronicznym Korpusie Tekstów Polskich z XVII i XVIII w.* zostały poddane automatycznej anotacji morfosyntaktycznej. W tym celu zastosowano niezależnie od siebie dwa narzędzia (tagery): Concraft i Toygger. Na skutek tego powstały dwie wersje korpusu, które mogą się od siebie różnić na poziomie lematyzacji i przypisanych poszczególnym formom znaczników morfosyntaktycznych. Domyślnie w wyszukanych wynikach pojawiają się interpretacje Toyggera, zaś interpretacje Concrafta wyświetlają się w dymku zawierającym podstawowe informacje o szukanej formie. Można jednak zadać pytanie, w którym zostaną sformułowane warunki odnoszące się tylko do Concrafta albo do obu tagerów z osobna, np. pytanie o wszystkie segmenty, które Toygger uznał za rzeczownik, a Concraft za reprezentanta innej klasy gramatycznej (por. p. 5.7).

4. Interfejs wyszukiwarki MTAS

Interfejs wyszukiwarki zawiera pasek wyszukiwania, w którym wpisuje się zapytanie skonstruowane zgodnie ze składnią zapytań programu MTAS (por. p. 5). Po prawej stronie paska znajdują się przyciski umożliwiające wpisanie liter, które nie występują we współczesnej polskiej ortografii, ale mogą się pojawiać w warstwie transliteracyjnej przeszukiwanych tekstów. Okno KORPUS powyżej paska wyszukiwania pozwala określić, czy chcemy przeszukiwać pełny, 13,5-milionowy korpus, zaanotowany automatycznie (KORBA AUTOMATYCZNA) czy półmilionowy podkorpus zaanotowany ręcznie (KORBA RĘCZNA).

Poniżej paska wyszukiwania znajdują się trzy przyciski dające dodatkowe możliwości pracy z wyszukiwarką. Przycisk KONSTRUKTOR ZAPYTAŃ daje dostęp do prostej nakładki na wyszukiwarkę, która ułatwia zdefiniowanie warunków określających cechy segmentu lub sekwencji segmentów występujących w zapytaniu, np. postać segmentu, formę hasłową, klasę gramatyczną, wartości kategorii gramatycznych. Poszczególne warunki w obrębie segmentu mogą być łączone operatorami *oraz* (koniunkcja) i *lub* (alternatywa). Zbudowane za pomocą konstruktora zapytanie pojawi się w pasku wyszukiwania, dzięki czemu można zweryfikować jego poprawność.

[4] W języku średniopolskim funkcjonowały wszystkie te rzeczowniki.

Przycisk ODRZUĆ OBCE SEGMENTY spowoduje dodanie do zapytania ograniczenia, dzięki któremu wśród wyników wyszukiwania nie pojawią się segmenty obcojęzyczne równokształtne z polskimi, np. łaciński wyraz *do* ‘daję’.

Przycisk METADANE umożliwia wprowadzanie do zapytania warunków ograniczających jego zasięg, na przykład do tekstów danego autora lub tekstów powstałych w danym przedziale czasowym (por. p. 6). Przycisk DODAJ OGRANICZENIE pozwala na łączenie poszczególnych warunków, można więc ograniczyć badanie np. do pism politycznych powstałych na terenie Małopolski w pierwszej połowie XVII w.

Wyniki wyszukiwania mogą wyświetlać się w postaci albo transliterowanej, albo transkrybowanej (uwspółcześnionej). Sposób prezentacji wyników wybieramy w oknie WARSTWA WYŚWIETLANIA. Na liście wyników, podobnie jak w NKJP, wyszukiwana sekwencja segmentów opatrzonych znacznikami zostaje wyróżniona kolorem, a po nakierowaniu na nią kursora pokazuje się dymek z podstawowymi informacjami o lokalizacji wyszukanego cytatu, z dokładnością do numeru strony książki będącej podstawą transliteracji. Po kliknięciu na wybrany wynik na dole strony pokazuje się szerszy kontekst wraz ze szczegółowymi metadanymi tekstu. W oknie LICZBA WYNIKÓW NA STRONĘ można określić, ile wyszukiwanych pozycji ma być widocznych jednocześnie.

5. Język zapytań wyszukiwarki MTAS

Jak już było powiedziane, składnia zapytań w programie MTAS została oparta na języku zapytań o nazwie *Corpus Query Language* (CQL) wykorzystywanym m.in. w znanym z NKJP Poliqarpie. Należy jednak zwrócić uwagę na drobne różnice, ponieważ mogą one wpływać na poprawność formułowanych zapytań.

Niniejsza instrukcja nie uwzględnia wszystkich możliwości wyszukiwarki, częściowo ze względu na jej skrótowy charakter, a częściowo dlatego, że nie wszystkie zapytania będą miały sens w odniesieniu do korpusu XVII i XVIII wieku (mogą być natomiast użyteczne w tych korpusach, w których uwzględniono znakowanie warstw innego typu, np. nazw własnych, nadrzędników składniowych, sensu słów itp.). Podstawowa dokumentacja wyszukiwarki znajduje się na jej stronie internetowej.

5.1. Proste zapytania o segmenty i formy podstawowe

Zapytania wpisujemy w nawiasach kwadratowych według schematu: [atrybut = „wartość atrybutu”]. W najprostszych pytaniach o kształt tekstowy segmentu atrybutem będzie `translit` — jeśli chcemy wyszukać segment w postaci transliterowanej lub `orth` — jeśli poszukujemy segmentu w postaci transkrybowanej (uwspółcześnionej). Wartością obu tych atrybutów będzie poszukiwany ciąg liter, np.

```
[translit="seym"]
[orth="sejm"]
```

W wypadku zapytań o kształt segmentów w warstwie uwspółcześnionej można pominąć nawiasy kwadratowe oraz nazwę atrybutu. Zatem poniższe zapytania:

```
 [orth="sejm"]
 sejm
```

zwrócą takie same wyniki.

Domyślnie rozróżniana jest wielkość liter, a więc poniższe trzy zapytania:

```
[orth="sejm"]
[orth="Sejm"]
[orth="SEJM"]
```

dadzą różne wyniki. Aby znaleźć podane słowo niezależnie od wielkości poszczególnych liter, należy użyć atrybutu `orth` z rozszerzeniem `_lc` (ang. *lower case*):

```
[orth_lc="sejm"]
```

Wynikiem wszystkich powyższych zapytań będzie jedynie ciąg liter podany jako wartość atrybutu. Aby znaleźć wszystkie formy rzeczownika SEJM, należy użyć następującego zapytania:

```
[base="sejm"]
```

Wartością atrybutu `base` jest forma podstawowa szukanego fleksemu, a zatem mianownik liczby pojedynczej dla rzeczowników czy bezokolicznik dla wszystkich fleksemów czasownikowych. Formy podstawowe zapisane są wyłącznie w postaci uwspółcześnionej, zatem zapytanie:

```
[base="sejm"]
```

zwróci wystąpienia różnych form fleksyjnych rzeczownika SEJM zapisane według różnych konwencji ortograficznych (np. *Sejm*, *seymu*, *sejmem*). Z kolei zapytanie:

```
[base="seym"]
```

nie zwróci żadnego wyniku. Aby znaleźć wszystkie formy rzeczownika SEJM zapisane przez *y*, należy dołączyć dodatkowy warunek na segment opisujący jego postać w warstwie transliterowanej (por. p. 5.4).

5.2. Wyrażenia regularne

W zapytaniach o segmenty i formy podstawowe (a także w innych zapytaniach, które zostaną opisane niżej) można używać standardowych wyrażeń regularnych wykorzystujących znaki specjalne, takie jak: `!`, `?`, `*`, `.`, `|`, `[`, `]`, `(`, `)` oraz liczby naturalne pisane cyframi arabskimi, np. `0` czy `21`. Poniżej kilka przykładów zastosowania takich wyrażeń:

1. `[translit="(sejm|seym)"]`
   znak `|` oznacza alternatywę dwóch wyrażeń (całość należy dodatkowo ująć w nawiasy okrągłe), a zatem zapytanie to może zostać użyte do znalezienia wszystkich wystąpień segmentów, które w warstwie transliteracyjnej mają postać *sejm* lub *seym*;
2. `[orth="[Ss]ejm"]`
   nawiasy kwadratowe oznaczają alternatywę znaków, a zatem zapytanie to może zostać użyte do znalezienia wszystkich segmentów o postaci *sejm* niezależnie od wielkości litery początkowej;
3. `[translit="komm?endant"]`
   znak zapytania oznacza opcjonalność znaku lub ujętego w nawiasy okrągłe wyrażenia bezpośrednio przed nim, a zatem w wyniku zadania tego zapytania znalezione zostaną segmenty *komendant* i *kommendant*;
4. `[orth="k.żdy"]`
   kropka oznacza dowolny znak, a zatem wynikiem tego zapytania będą segmenty *każdy* i *kożdy*;
5. `[orth="owa.*"]`
   gwiazdka oznacza dowolną liczbę wystąpień znaku lub ujętego w nawiasy okrągłe wyrażenia bezpośrednio przed nią, a zatem zapytanie to pozwoli znaleźć segmenty zaczynające się na *owa*, np. *owa*, *owaką*, *owakimi*, *owad*, *owada*, *owalną*, *owałaszyć*;
6. `[orth="do{1,2}koła"]`
   konstrukcja typu `n,m` oznacza od `n` do `m` wystąpień znaku lub ujętego w nawiasy okrągłe wyrażenia bezpośrednio przed nią, a zatem zapytanie to pomoże znaleźć segmenty

zaczynające się od litery *d*, po której następuje ciąg od 1 do 2 liter *o*, a następnie ciąg *koła*, a więc zarówno *dokoła*, jak i *dookoła*;

7. `[orth="(pra){2,}.*"]`
   konstrukcja typu `n,` oznacza co najmniej `n` wystąpień znaku lub ujętego w nawiasy okrągłe wyrażenia bezpośrednio przed nią, a zatem zapytanie to może posłużyć do znalezienia segmentów, w których ciąg *pra* występuje przynajmniej 2 razy z rzędu, np. *praprababa*, *prapraprapradziad*. Specyfikacje segmentów muszą pasować do całych segmentów, stąd konieczność umieszczenia po prawej stronie ciągu `(pra){2,}` wyrażenia `.*`, pasującego do dowolnego ciągu znaków.

UWAGA: Jeśli zadajemy pytanie o znak interpunkcyjny, który jest jednocześnie znakiem specjalnym używanym w wyrażeniach regularnych, należy poprzedzić go znakiem \, np.

```
[base="\."]
```

5.3. Łączenie atrybutów

W zapytaniach o segmenty lub formy podstawowe segmentów (a także w innych zapytaniach, które zostaną opisane niżej) można określić kilka atrybutów, łącząc je operatorem koniunkcji `&` lub alternatywy `|`. W takich zapytaniach może się także pojawić operator negacji `!`. Poniżej kilka przykładów takich zapytań:

1. `[base="sejm" & translit="seym.*"]`
   jako wynik otrzymamy wystąpienia wszystkich form rzeczownika SEJM zapisane przez *y*;
2. `[orth="minę" & base="mina"]`
   w wyniku tego zapytania znalezione zostaną te segmenty, które są jednocześnie segmentem *minę* i formą rzeczownika MINA, a więc wyłącznie te wystąpienia segmentu *minę*, które są interpretowane jako formy rzeczownika MINA (a nie czasownika MINĄĆ);
3. `[orth="minę" & !base="minąć"]`
   w wyniku tego zapytania znalezione zostaną te segmenty, które są jednocześnie segmentem *minę* i nie są formą czasownika MINĄĆ, a zatem zapytanie to jest równoważne poprzedniemu;
4. `[orth="minę" | base="mina"]`
   w tym przypadku znalezione zostaną te segmenty, które są albo segmentem *minę* (niezależnie od tego, czy jest to forma fleksemu MINA czy MINĄĆ), albo dowolną formą rzeczownika MINA, np. *mina*, *miny*, *minami*.

5.4. Zapytania o kilka pozycji

Jeśli chcemy wyszukać kilka sąsiadujących ze sobą segmentów lub form podstawowych, każdą pozycję zapisujemy w oddzielnych nawiasach kwadratowych, np.:

```
[base="sejm"][base="walny"]
```

Oznaczając za pomocą pustych nawiasów kwadratowych dowolny segment, możemy znaleźć formy, które nie sąsiadują ze sobą bezpośrednio, np.:

```
[orth="się"][][][base="bać"]
```

W wyniku tego zapytania zostaną znalezione takie ciągi, jak *się nic nie boi* czy *się o to boję*. Użycie wyrażeń regularnych pozwala na określenie minimalnej i maksymalnej odległości pomiędzy szukanymi formami, np.:

```
[orth="się"][]{2,4}[base="bać"]
```

W wyniku tego zapytania zostaną znalezione segmenty *się* oraz segmenty odpowiadające formom czasownika BAĆ przedzielone dwoma, trzema lub czterema segmentami, a zatem

oba ciągi uzyskane w wyniku poprzedniego zapytania, a także na przykład ciąg *się już niczego nie boją*.

5.5. Zapytania o znaczniki morfosyntaktyczne

W zapytaniach można również określać wartości klasy gramatycznej (za pomocą atrybutu `pos`, ang. *part of speech*), a także wartości poszczególnych kategorii gramatycznych, np. przypadka czy rodzaju. Wartościami atrybutu `pos` są symbole nazw klas gramatycznych znajdujące się w tabeli 1. Atrybuty poszczególnych kategorii gramatycznych i ich wartości wymienione są w tabeli 4.

Kategoria	**Atrybut**	**Możliwe wartości**
liczba	`number`	`sg du pl`
przypadek	`case`	`nom gen dat acc inst loc voc`
rodzaj	`gender`	`m manim1 manim2 mnanim f n p1 p2`
osoba	`person`	`pri sec ter`
stopień	`degree`	`pos com sup`
aspekt	`aspect`	`imperf perf biasp`
zanegowanie	`negation`	`aff neg`
akcentowość	`accentability`	`akc nakc zneut`
poprzyimkowość	`post-prepositionality`	`npraep praep`
aglutynacyjność	`agglutination`	`agl nagl`
wokaliczność	`vocalicity`	`nwok wok`
kropkowalność	`fullstoppedness`	`pun npun`

Tabela 4. Atrybuty kategorii gramatycznych i ich wartości

W zapytaniach tego typu również można używać wyrażeń regularnych. Możliwe jest zadanie na przykład następujących zapytań:

1. `[number="sg"]`
   znalezione zostaną wszystkie formy w liczbie pojedynczej,
2. `[pos="subst" & number="sg"]`
   znalezione zostaną formy rzeczownikowe w liczbie pojedynczej,
3. `[pos="subst" & !gender="f"]`
   znalezione zostaną formy rzeczownikowe rodzaju męskiego, nijakiego lub przymnogiego,
4. `[number="sg" & case="nom|acc" & gender="n"]`
   znalezione zostaną pojedyncze mianownikowe lub biernikowe formy nijakie.

O klasy gramatyczne i kategorie gramatyczne można także pytać łącznie, używając do tego atrybutu `tag`. Na przykład, aby znaleźć wszystkie rzeczowniki żeńskie w mianowniku liczby pojedynczej, można zadać następujące zapytanie:

```
[tag="subst:sg:nom:f"]
```

Aby zadać pytanie o pełną interpretację, czyli łącznie o formę podstawową, klasę gramatyczną i przysługujące jej kategorie gramatyczne, należy określić wartości dwóch atrybutów: `base` i `tag`, np.:

```
[base="baba" & tag="subst:sg:nom:f"]
```

W wyniku zadania tego zapytania znalezione zostaną wszystkie wystąpienia rzeczownika BABA w mianowniku liczby pojedynczej.

5.6. Ograniczenie zapytania do zdania lub akapitu

Teksty zawarte w korpusie zostały podzielone na zdania i akapity. Informację tę można wykorzystać w zapytaniach, na przykład ograniczając dopasowanie do jednego zdania. Aby ograniczyć zasięg zapytania, należy dopisać do zapytania słowo kluczowe `within`, a po nim `<s/>` lub `<p/>`, w zależności od tego, czy zasięg ma być ograniczony do zdania (ang. *sentence*) czy do akapitu (ang. *paragraph*). Ilustruje to następujący przykład zapytania o zdania, w których forma *się* występuje za formą leksemu BAĆ, w odległości co najmniej jednego i nie więcej niż dziesięciu segmentów:

```
[base="bać"][!orth="się"]{1,10}[orth="się"] within <s/>
```

Dodatkowo można również na elementy `<s/>` i `<p/>` nałożyć pewne warunki dotyczące tego, czy zawierają segmenty innego typu. Przykładowo, za pomocą następującego zapytania można znaleźć wszystkie wystąpienia czasownika BYĆ występującego w funkcji słowa posiłkowego czasu przyszłego złożonego ograniczone do zdań zawierających formę bezokolicznika:

```
[pos="fut"] within (<s/> containing [pos="inf"])
```

Wśród wyników będą oczywiście również takie zdania, w których czas przyszły został utworzony z formy pseudoimiesłowu, a bezokolicznik pełni w zdaniu inną funkcję gramatyczną.

5.7. Zapytania o interpretacje Concrafta

Jak zaznaczono w p. 3., wyszukane formy mają domyślnie przypisane interpretacje tagera Toygger. Jeśli chcemy wyszukać formy z interpretacją Concrafta, używamy nazw atrybutów poszerzonych o `_c`, a więc `base_c`, `pos_c`, `tag_c`, `number_c`, `case_c` itd. Istnieje także możliwość porównywania interpretacji obu tagerów. W tym celu w jednym zapytaniu należy określić atrybuty odnoszące się do obu interpretacji z osobna. Przykładowo, jeśli chcemy wyszukać wszystkie segmenty, którym Concraft przypisał interpretację rzeczownikową, a Toygger dowolną inną, zadamy następujące pytanie:

```
[pos_c="subst" & !pos="subst"]
```

Podobnie jak w przypadku innych wyszukiwań interpretacje Toyggera pojawią się przy znalezionych formach, a interpretacje Concrafta będą widoczne w dymku po nakierowaniu kursora na daną formę.

6. Ograniczenie zapytania za pomocą metadanych

Wszystkie teksty w korpusie zostały opatrzone bogatymi metadanymi, czyli informacjami o autorstwie, tytule, miejscu wydania, gatunku, temacie itd. Wyszukiwarka umożliwia ograniczanie wyników do wybranych tekstów za pomocą tych metadanych. W oknie METADANE można wybrać następujące pola:

- Skrót tekstu (identyfikator tekstu)
- Data (wydania lub powstania tekstu)
- Autor
- Tytuł
- Miejsce (wydania lub powstania tekstu)
- Region
- Typ mowy (proza lub wiersz)
- Rodzaj

- Gatunek
- Tematyka
- Poetyka żartu
- Drukarnia
- Tłumacz
- Typ źródła (teksty oryginalne — pochodzące z rękopisów i wydań XVII-XVIII-wiecznych, lub teksty uwspółcześnione — pochodzące z późniejszych wydań filologicznych)

W oknie Ograniczenie pojawiają się różne sposoby określania metadanych, zdefiniowane odmiennie dla poszczególnych pól. Ograniczenia można wprowadzać za pomocą następujących wyrażeń:

1. Wyrażenie „zawiera” — dostępne dla pól Skrót tekstu, Autor, Tytuł, Miejsce, Drukarnia, Tłumacz. W tym przypadku wielkość liter nie ma znaczenia. Można korzystać z wyrażeń regularnych. Wyszukiwarka znajduje teksty zawierające w określonym polu podany ciąg liter, np.:
   - Tytuł zawiera *akademia* — otrzymujemy tekst o tytule *Akademia dziecinna*.
   - Tytuł zawiera *akad dziec* — wynik jest taki sam, jak powyżej.
   - Tytuł zawiera *akad* — wynikiem będą teksty *Akademia dziecinna* i *Notyfikacja o niniejszym królewskiej akademii rycerskiej...*.
2. Wyrażenie „nie zawiera” — działa analogicznie do wyrażenia „zawiera”. Wyszukiwarka pomija teksty zawierające w określonym polu podany ciąg liter, np.:
   - Tytuł nie zawiera *akademia* — otrzymujemy wszystkie teksty z wyjątkiem tych, które w tytule mają słowo *akademia*.
3. Wyrażenie „=” — dostępne dla pól Skrót tekstu, Data, Autor, Tytuł. W tym przypadku istotna jest wielkość liter. Można korzystać z wyrażeń regularnych, np.:
   - Skrót tekstu = AkDziec — otrzymujemy tekst oznaczony skrótem *AkDziec*.
   - Skrót tekstu = .*Dzie.* — otrzymujemy teksty oznaczone skrótami zawierającymi ciąg liter *Dzie*, a więc *AkDziec*, *KwiatDzieje* itp.
   - Autor = Zbigniew Morsztyn — otrzymujemy teksty autorstwa Zbigniewa Morsztyna.
   - Autor = .*Morsztyn — otrzymujemy teksty, których autorami byli Hieronim Morsztyn, Jan Andrzej Morsztyn, Stanisław Morsztyn lub Zbigniew Morsztyn.
   - Autor = (Zbigniew Morsztyn)|(Stanisław Morsztyn) — otrzymujemy teksty, których autorami byli albo Zbigniew Morsztyn, albo Stanisław Morsztyn. W przypadku gdy łączymy znakiem alternatywy kilka członów wielowyrazowych, każdy z nich należy wziąć w nawias okrągły.
   - Data = 1633 — otrzymujemy teksty, które były wydane (lub napisane, w przypadku rękopisów) w roku *1633*.
   - Data = 16.* — otrzymujemy teksty, które były wydane (lub napisane, w przypadku rękopisów) w latach *1601–1699*.
4. Wyrażenie „$\neq$” — działa analogicznie do wyrażenia „=”. Wyszukiwarka pomija teksty, których odpowiednie pola zostały dokładnie określone, np.:
   - Skrót tekstu $\neq$ *AkDziec* — otrzymujemy wszystkie teksty z wyjątkiem *Akademii Dziecinnej*.
5. Wyrażenie „<” — dostępne tylko dla pola Data. Wyszukiwarka znajduje teksty, które zostały wydane (lub napisane, w przypadku rękopisów) przed określoną datą, np.:
   - Data < 1635 — otrzymujemy teksty o datach określonych dokładnie (np. *1630*) i w przybliżeniu (np. *przed 1630*).

6. Wyrażenie „>” — dostępne tylko dla pola DATA. Wyszukiwarka znajduje teksty, które zostały wydane (lub napisane, w przypadku rękopisów) po określonej dacie, np.:
   - DATA > 1670 — otrzymujemy teksty o datach określonych dokładnie (np. *1680*) i w przybliżeniu (np. *po 1690*).
7. Wyrażenie „z przedziału” — dostępne tylko dla pola DATA. Wyszukiwarka znajduje teksty, które zostały wydane (lub napisane, w przypadku rękopisów) w określonym przedziale czasowym. Wymaga podania dwóch dat oddzielonych myślnikiem, np.:
   - DATA z przedziału 1633–1670 — otrzymujemy teksty o datach określonych dokładnie (np. *1654*) i w przybliżeniu (np. *między 1635 a 1644*).

Dla pól, które mają ograniczoną liczbę wyborów (np. REGION), wystarczy wybrać pożądane wartości z rozwijanej listy.

Poszczególne ograniczenia można łączyć za pomocą przycisku DODAJ OGRANICZENIE.

Opracowanie: Włodzimierz Gruszczyński i Renata Bronikowska