

Możliwości przeszukiwania Korpusu barokowego

Cele i założenia

Renata Bronikowska
Instytut Języka Polskiego PAN

Ogólne informacje

- Korpus barokowy ma stanowić podkorpus NKJP, co umożliwi przeszukiwanie go za pomocą wyszukiwarki Poliqarp
- Teksty włączone do korpusu reprezentowane są w systemie XML zgodnym ze standardem TEI
- Znakowanie odbywa się w szablonie dokumentu Word umożliwiającym wprowadzanie znaczników za pomocą skrótów klawiaturowych
- Na potrzeby projektu został stworzony program do konwersji plików Worda na format XML

Nowe możliwości przeszukiwania tekstów średniopolskich

- Wariantywność zapisu
- Nieczytelne fragmenty tekstu, niepewne odczytanie
- Fragmenty obcojęzyczne
- Nietranskrybowane fragmenty tekstu
- Fragmenty tekstu inne niż tekst główny
- Numeracja stron

Wariantywność zapisu I

Zapis	Wyszukiwane wyrażenie	Wyniki wyszukiwania
Litera o w górnej frakcji	mego	mego, me ^o
Litery z tyldą na końcu słowa	panem wiatrom	panem, panẽ wiatrom, wiatrõ
Litery kreskowane	sądzić rada	sądzić, sądzić rada, rádá
Litery bez znaków diakrytycznych	który był	który, ktory był, byl
Litera v	uwaga	uwaga, vwaga
Podwojona litera v	powódź	powódź, povvódź
Litery ż i ź	uważać źle	uważać, uważać źle, źle

Wariantywność zapisu II

Słowa doprowadzone do postaci znormalizowanej:

- skróty oznaczone tyldą nad literą:
atramẽt <atrament>
- wyrazy błędnie zapisane: miła <miał>

Wyszukiwane wyrażenie	Wyniki wyszukiwania
atrament	atramẽt / atrament
miła	miła / miał
miał	miła / miał

Nieczytelne fragmenty tekstu

- W wynikach wyszukiwania pojawi się — informacja o niepewności odczytania oraz potwierdzenie poprawności odczytania słowa
 - dobrodziej<?>
 - gorzski<!>
- Znaki nieczytelne w obrębie słowa są zaznaczone odpowiednią liczbą symboli – pozwoli to na wyszukanie słowa o określonej liczbie znaków
 - dobro<.....>

Fragmenty obcojęzyczne

- WOIENNOSC PIESZA Ktorą Łąćinnicy **Pedestrem Militiam** nazywają
- Albo piy/ albo się biy. **Aut bibe, aut abi.**
- Przez Quartę z gory Solmizatia
- [TEKST W JĘZYKU OBCYM]
- Słowo bowiem Greckie [ALFABET NIEŁACIŃSKI] wyrażnie osoby znaczy.

Fragmenty obcojęzyczne w wynikach wyszukiwania

- Fragmenty obcojęzyczne będą wyłączone z przeszukiwania
- Fragmenty obcojęzyczne będą się pojawiały jako kontekst szukanego wyrażenia
- Liczba segmentów we fragmentach obcojęzycznych będzie odpowiadać stanowi faktycznemu
 - [orth=tak] [] [] [orth=mówi] > **ták** **inter cetera** mowi
 - [orth=przed] [] {1,2} [orth=laty] > **przed** dwiema **circiter** láty

Lista języków

- Arabski
- Czeski
- Francuski
- Grecki
- Hebrajski
- Hiszpański
- Litewski
- Łacina
- Niemiecki
- Południowosłowiański
- Ruski
- Skandynawski
- Turecko-tatarski
- Węgierski
- Włoski
- Inny

Nietranskrybowane fragmenty tekstu

- [ILUSTRACJA]
- [WZÓR MATEMATYCZNY]
- [ZAPIS NUTOWY]
- [TEKST W JĘZYKU OBCYM]
- [ALFABET NIEŁACIŃSKI]
- [FRAGMENT NIECZYTELNY]
- [INNA PRZERWA W TEKŚCIE]

Fragmenty nietranskrybowane w wynikach wyszukiwania

- W wynikach wyszukiwania informacja o tekście nietranskrybowanym pojawi się jako tekst w nawiasach kwadratowych
- Fragmenty nietranskrybowane będą traktowane jako jeden segment
 - [orth=kiedy] [] {1, 6} [orth=to] > **kiedy** będą spółem [ZAPIS NUTOWY] dwie noćie takie/ **to** zawsze tá pierwsza trzyma w sobie tákt ieden

Fragmenty tekstu inne niż tekst główny

- Wyrażenie wyszukane w przypisie lub notce marginesowej wyświetli się razem z kontekstem akapitowym

Powietrza rozmańtość.	Powietrze morowe/ ábo iest domorodne/ co się domá sámó vrodzi: ábo zanieśione z inszey kráiny.
----------------------------------	--

- Tekst w przypisach, notkach marginesowych i żywej paginie nie będzie się pojawiał jako kontekst wyszukanego wyrażenia w tekście głównym

Wyodrębnione elementy tekstu głównego

- Elementy strony tytułowej
 - Tytuł publikacji,
 - Podtytuł publikacji,
 - Autor,
 - Język oryginału,
 - Tłumacz,
 - Miejsce wydania,
 - Rok wydania,
 - Drukarnia,
 - Numer wydania,
 - Inny element strony tytułowej
- Nagłówki w tekście
 - Powtórzony tytuł,
 - Tytuł części,
 - Podtytuł części,
 - Streszczenie części,
 - Podpis pod rysunkiem

Numeracja stron

- Każdy segment będzie powiązany z numerem strony – pozwoli to na dokładną lokalizację w źródle wyszukanych wyrażień
- Każdej stronie zostaje przypisany identyfikator, niezależnie od tego, czy znajduje się w przepisywanym tekście
 - Numer strony (1, 2, 3, 4...)
 - Oznaczenie składki (A, Av, A2, A2v...)
 - Strony nieliczbowane (1 nlb., 2 nlb., 3 nlb. ...)

Dziękuję za uwagę