Korpusomat — narzędzie do tworzenia przeszukiwalnych korpusów języka polskiego

Witold Kieraś **Łukasz Kobyliński** Maciej Ogrodniczuk

Instytut Podstaw Informatyki PAN

III Konferencja DARIAH-PL – Poznań – 9.11.2016

Dlaczego warto zajmować się lingwistyką korpusową?

Korpus to systematycznie wybrany zbiór tekstów, wykorzystywanych w analizach lingwistycznych, przechowywanych najczęściej w formie elektronicznej, często uzupełniony dodatkowymi warstwami anotacji.

Przykłady zastosowań analiz korpusowych

- obliczanie częstości wystąpień słów, fraz i kolokacji,
- badanie najczęstszych kontekstów wystąpień słów lub fraz,
- badanie zmian języka w czasie, przy wykorzystaniu korpusów tekstów historycznych,
- badanie rzeczywistego wykorzystania języka przez jego użytkowników (korpusy dziedzinowe, korpusy obcojęzyczne).

FIND SAMPLE: 100 200 500 1000
PAGE: << < 1 / 231 > >>

CLICK FOR MORE CONTEXT . [?]

1	FU4	W_fict_drama	A	B	C	off with your clothes. PAMELA: unwillinglyI'll get undressed if you lock the **door** and let me have the keys in my own hand. MRS. JEWKES:
2	FU4	W_fict_drama	A	B	C	go to the bottom of the elm walk. I will steal out of the **door** unperceived. She puts on gloves and picks up her fan. MRS. JEWKES
3	FU4	W_fict_drama	A	B	C	for me and I beg to withdraw. LADY DAVERS: Jackey, shut the **door**, my young lady and I must not have done so soon. Where's
4	FU4	W_fict_drama	A	B	C	will not ask you who is of your party... BELVILLE exits, slamming the **door**. I believe I have shed as many tears as would drown by baby.
5	CH1	W_newsp_tabloid	A	B	C	. Andrew, now 29, was 15 that summer when he knocked at the **door** and introduced himself.' Denis Heymer, Frankie's manager, answered and said
6	CH1	W_newsp_tabloid	A	B	C	smash-hit album Use Your Illusion 1 and 11, which features Knocking On Heaven's **Door** and November Rain. PLUS... we have 100 copies of a new EP,
7	CH1	W_newsp_tabloid	A	B	C	and slippery steps. # 5) # If a child can open the front **door**, fit an extra lock. # Sitting room # 1) # Use heavy
8	CH1	W_newsp_tabloid	A	B	C	child to lock himself in. Preferably, fit a bolt high up on the **door**. # 5) # Turn down the temperature of your hot water. Then
9	CH1	W_newsp_tabloid	A	B	C	Lewis Bronze,' and we like them to have a girl or boy next **door** image.' So BBC bosses have to be ultra careful about who they hire
10	CH1	W_newsp_tabloid	A	B	C	tall man in a vest, braces and crumpled suit is stooped next to a **door**, demonstrating that he has no more notion of how a Savoy room key works
11	CH1	W_newsp_tabloid	A	B	C	about being his wife, wearing big hats, being chauffeur-driven and waltzing through the **door** of Number 10 if he got to be Prime Minister.' She liked to
12	CH1	W_newsp_tabloid	A	B	C	were only her private secretary and the ever-present detective. Diana dashed to the front **door** wearing the kind of understated clothes appropriate for meeting w
13	CH1	W_newsp_tabloid	A	B	C	white top and a black and white striped skirt. Sandra was waiting at the **door**. She asked:' Would you like to come up to the top of
14	CH1	W_newsp_tabloid	A	B	C	these men have this need to control?' In a small adjoining room next **door** a group of women who act as counsellors and administrators were waiting to meet her
15	CH1	W_newsp_tabloid	A	B	C	' But we'll be treating my daughter and our four grandchildren who live next **door**.' Today's game -- Page 25 # THE LIMIT # RICK SKY #
16	CH1	W_newsp_tabloid	A	B	C	Mail mountain bike. I'll pin Harry Prosser's great picture on my front **door** to give our old postman the idea of how it should be done. --
17	CH1	W_newsp_tabloid	A	B	C	gang suddenly burst in and demanded all the ticket money from the guy on the **door**.' They were firing machine guns into the air. It was like a
18	CH1	W_newsp_tabloid	A	B	C	we have all been reaching for our brollies and in some cases sandbagging the front **door** over the past few weeks. Because a team of National Aeronautical Space
19	CH1	W_newsp_tabloid	A	B	C	topped the album charts earlier this month.' The worst moment was when the **door** flew open. I thought I was going to be sucked out. I've
20	CH1	W_newsp_tabloid	A	B	C	that windy weather is on the way. Or the pine cone hanging by his **door**. He checks it each morning to see whether it is going to rain.
21	CH1	W_newsp_tabloid	A	B	C	found him in the kitchen, grabbed his arm and ran off through a side **door**. No one knew why. Lord Charles and his bride seemed happy enough.

NKJP

Narodowy Korpus Języka Polskiego

Poliqarp search engine for NKJP data

Query
Settings
File a bug
Help

Query: drzwi

Corpus: manually annotated corpus v1.2 (1.2M segments)

Search

Results

Found 196 results so far

Displaying results 1—10

1.	zabezpieczenia pasażerów przed przycięciem przez	**drzwi** [drzwi:subst:pl:acc:n]	(czujnik jest umieszczony w
2.	Trzynacha. Odsunął się od	**drzwi** [drzwi:subst:pl:gen:n]	i zapalił światło. Ciemny
3.	do pokoju, zostawił jednak	**drzwi** [drzwi:subst:pl:acc:n]	otwarte na oścież. Wpadł
4.	i frasunku. Gdy już	**drzwi** [drzwi:subst:pl:nom:n]	zamknęły się za ostatnim,
5.	chwili ruch się uczynił od	**drzwi** [drzwi:subst:pl:gen:n]	, stuk licznych kroków i
6.	wy na to? Gdy	**drzwi** [drzwi:subst:pl:nom:n]	zapadły, ujrzał się Kazimierz
7.	pomagając sobie nogą, zatrzasnęła	**drzwi** [drzwi:subst:pl:acc:n]	służbowego mieszkania. Lewicki wystartował
8.	to mogli przecież zadzwonić do	**drzwi** [drzwi:subst:pl:gen:n]	, a nie od razu
9.	wdzianko z odblaskami. Zza	**drzwi** [drzwi:subst:pl:gen:n]	mieszkania numer sto piętnaście dobiegł
10.	samochodu. Trudno było otworzyć	**drzwi** [drzwi:subst:pl:acc:n]	. Podjęto próbę wydostania się

Dlaczego warto tworzyć korpusy tekstowe?

Przykłady istniejących korpusów tekstowych

- Narodowy Korpus Języka Polskiego,
- British National Corpus,
- Penn Treebank,
- ale też: Słownik Warszawski, Korpus Języka Młodzieży, ...

Według jakiego klucza można utworzyć korpus?

- wg dziedziny, np. teksty medyczne, ekonomiczne, prawnicze,
- wg autora, np. Stanisław Lem,
- wg epoki, np. korpus polszczyzny XVIII w.,
- ...

Korpusomat

Czym jest Korpusomat?

Narzędzie (serwis internetowy), służące do tworzenia własnych korpusów tekstowych, automatycznie anotowanych w warstwie morfosyntaktycznej.

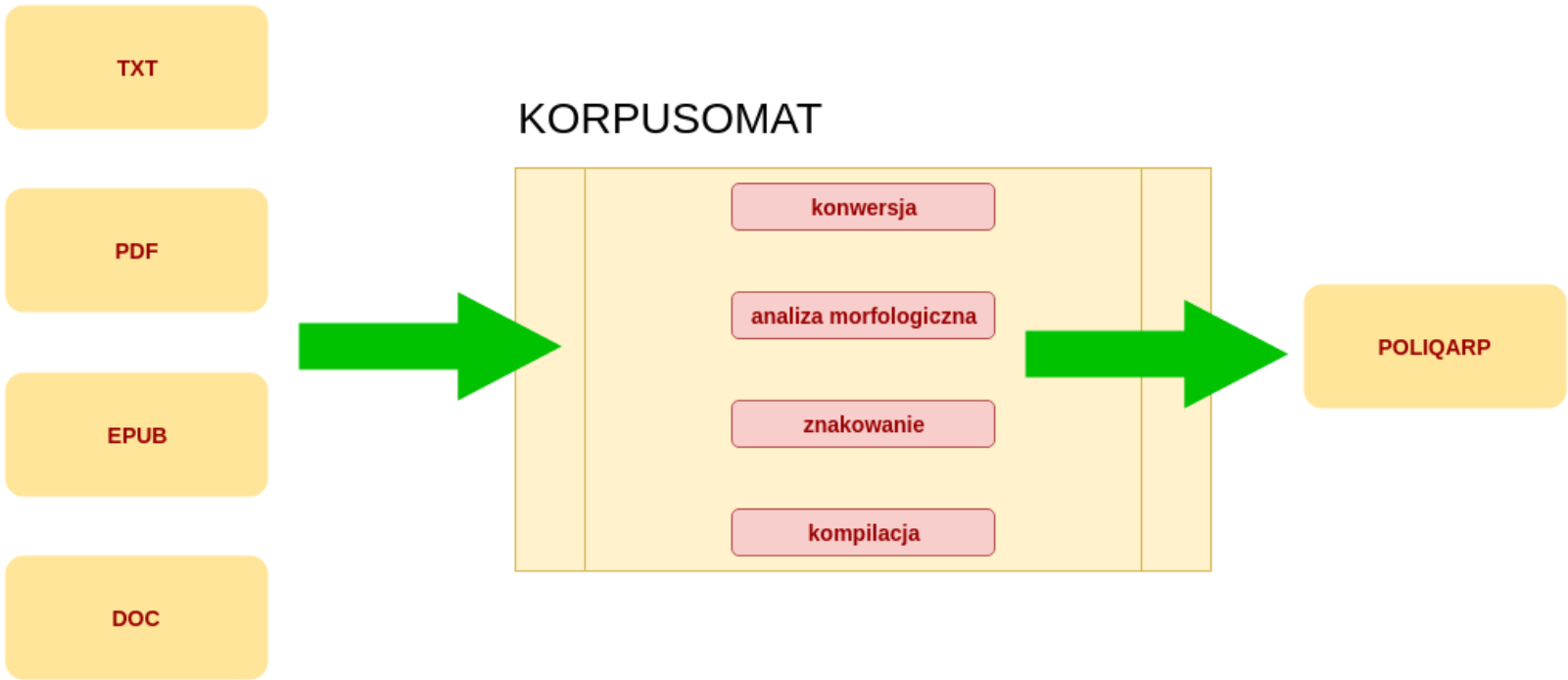
Motywacja

- analizy korpusowe są cennym narzędziem wspierającym pracę lingwistów, leksykografów, tłumaczy, studentów i nauczycieli,
- istniejące narzędzia są:
  - związane z istniejącymi korpusami, bez możliwości wykorzystania własnych danych,
  - trudne do wykorzystania przez osoby nietechniczne,
  - niedostosowane do języka polskiego,
  - komercyjne/płatne.

Idea Korpusomatu

Idea Korpusomatu

- tworzenie korpusu nie wymaga specjalistycznej wiedzy,
- korpus można utworzyć z dowolnego zbioru własnych zasobów,
- instalacje na własnym komputerze są ograniczone do wyszukiwarki korpusowej.



Korpusomat - działanie

Etapy przetwarzania

- konwersja formatów binarnych na format tekstowy,
- konwersja kodowania tekstu do UTF-8,
- analiza morfologiczna tekstu (za pomocą analizatora Morfeusz i słownika SGJP),
- znakowanie morfosyntaktyczne (za pomocą tagera Concraft),
- tworzenie binarnej postaci korpusu, do przeszukiwania oprogramowaniem Poliqarp.

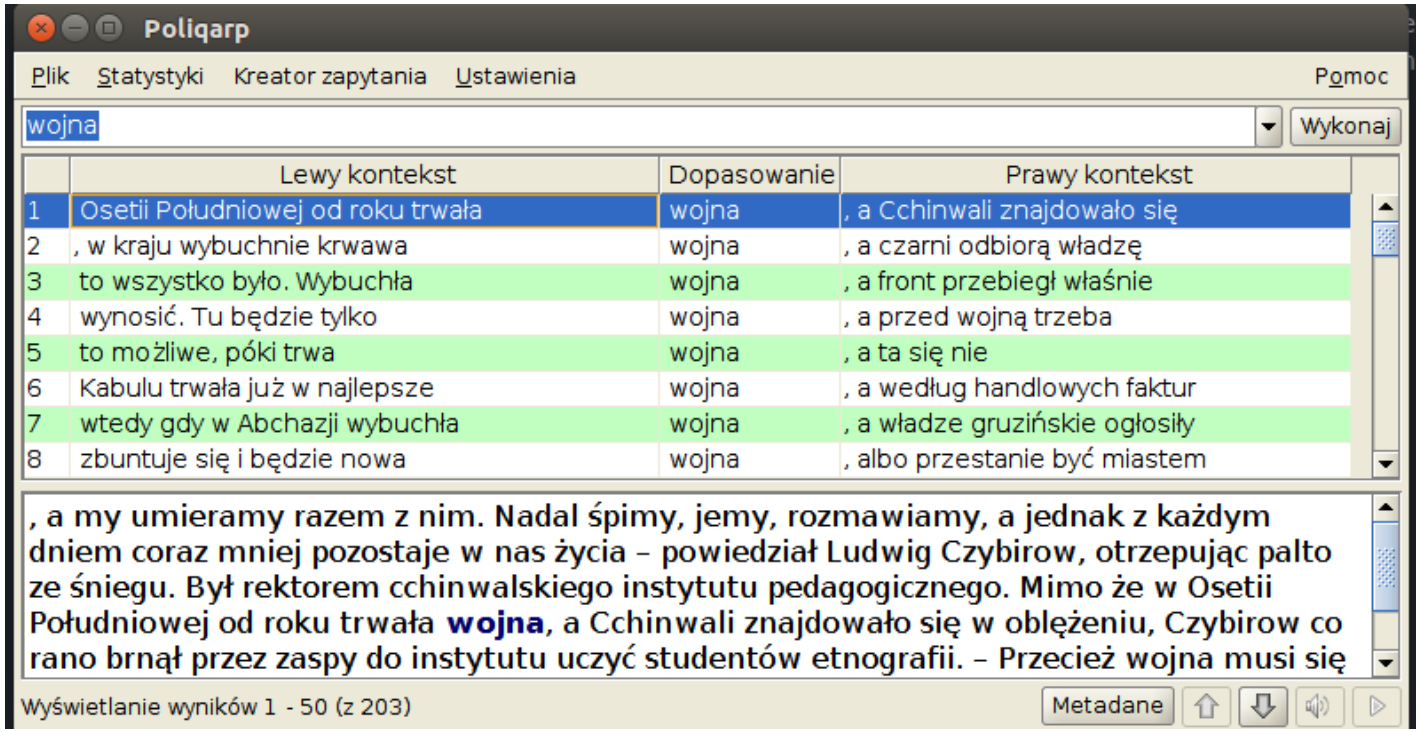
Demo

http://korpusomat.nlp.ipipan.waw.pl

DEMO

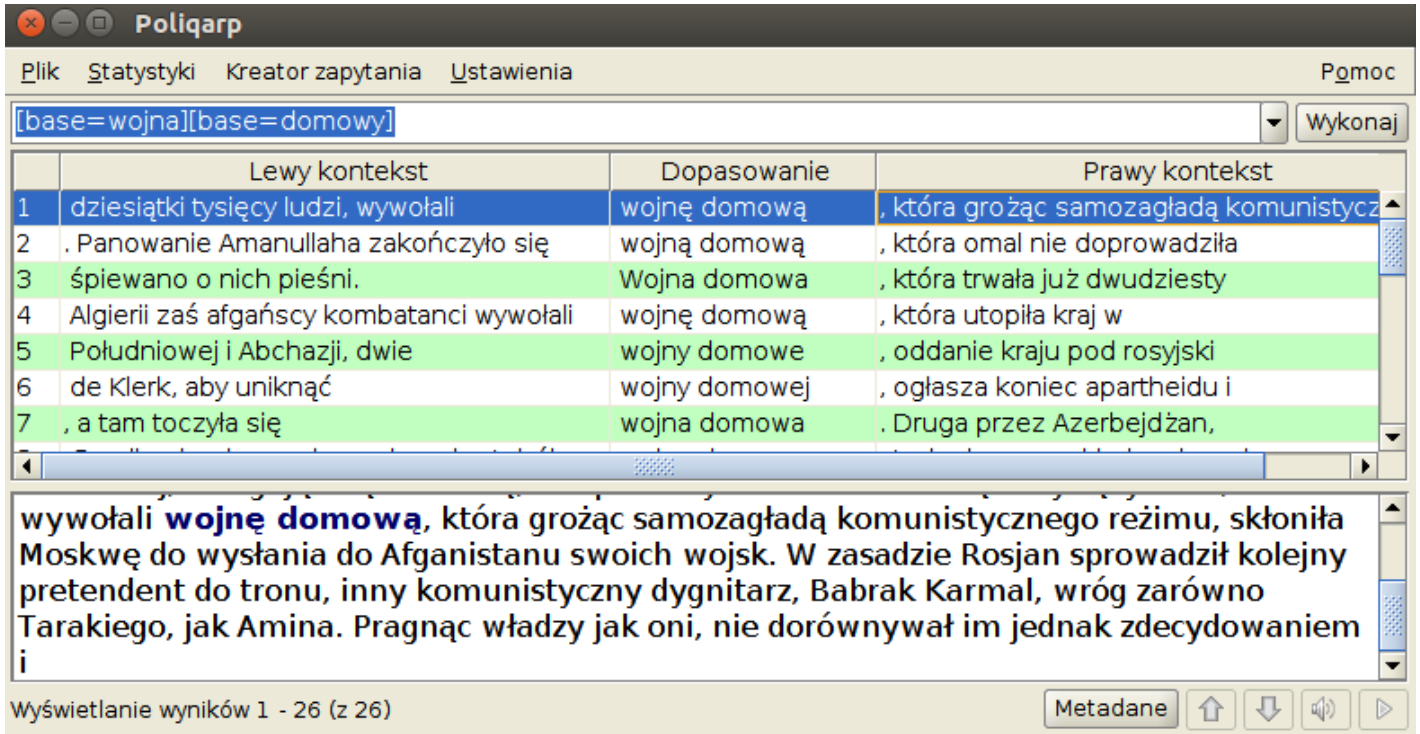
Przykład analizy językowej

Konteksty rzeczownika wojna



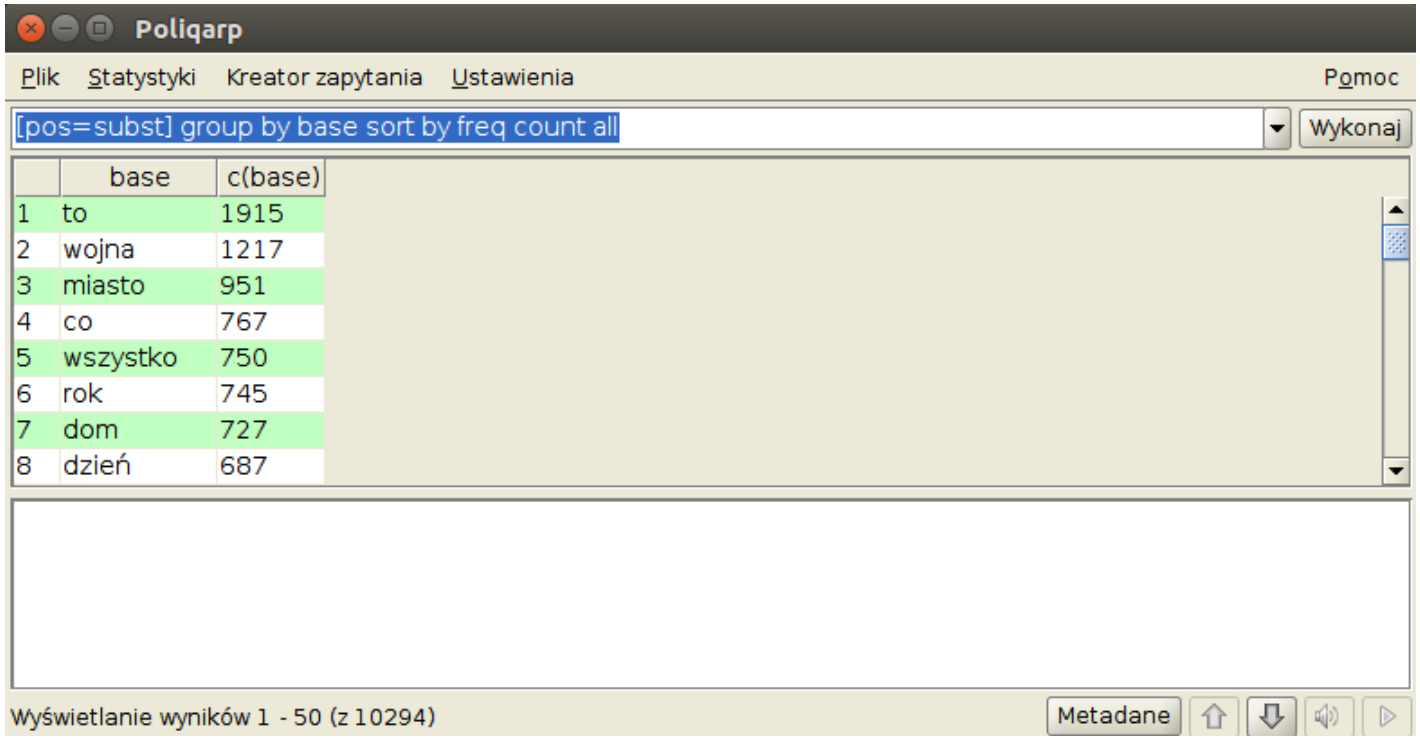
Przykład analizy językowej

Konteksty wszystkich form frazy wojna domowa



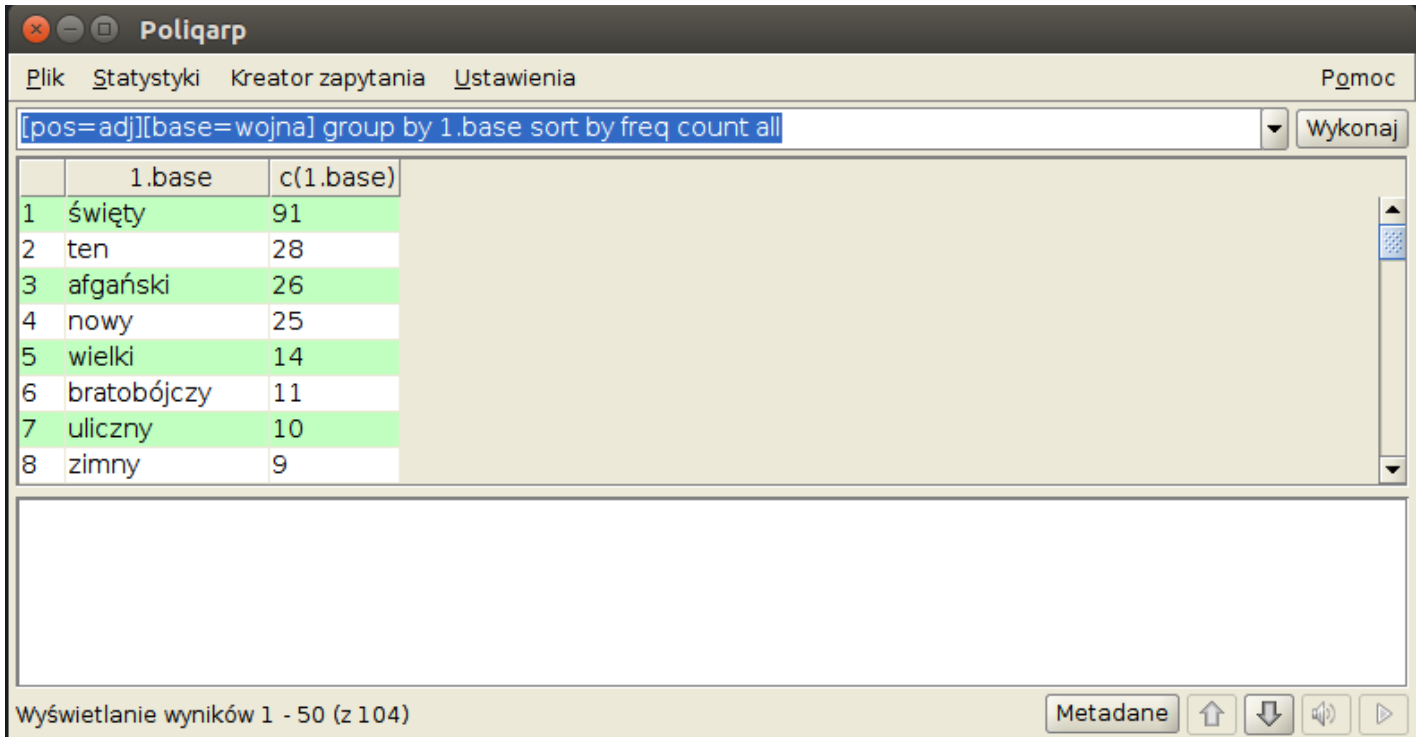
Przykład analizy statystycznej

Lista frekwencyjna rzeczowników



Przykład analizy statystycznej

Lista frekwencyjna przymiotników w lewym kontekście



Dalsze plany

Nowe możliwości

- pobieranie tekstów ze wskazanych adresów internetowych (web-scraping),
- masowe ładowanie wielu tekstów z plików lub Internetu,
- konfiguracja własnej struktury metadanych,
- interfejs webowy do Poliqarpa,
- wykorzystanie Morfeusza2 i alternatywnych słowników morfologicznych.

Sugestie mile widziane!

Dziękujemy!

Dziękujemy za uwagę.