

# 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).



SEARCH

FREQUENCY

CONTEXT

HELP

FIND SAMPLE: [100](#) [200](#) [500](#) [1000](#)

PAGE: &lt;&lt; &lt; 1 / 231 &gt; &gt;&gt;

CLICK FOR MORE CONTEXT .

 [?]

1	FU4	W_fict_drama	A B C	off with your clothes. PAMELA: unwillingly! I'll get undressed if you lock the <b>door</b> 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 <b>door</b> 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 <b>door</b> , 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 <b>door</b> . 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 <b>door</b> 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 <b>Door</b> 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 <b>door</b> , 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 <b>door</b> . # 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 <b>door</b> 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 <b>door</b> , 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 <b>door</b> 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 <b>door</b> 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 <b>door</b> . 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 <b>door</b> 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 <b>door</b> .' 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 <b>door</b> 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 <b>door</b> .' 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 <b>door</b> 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 <b>door</b> 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 <b>door</b> . 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 <b>door</b> . No one knew why. Lord Charles and his bride seemed happy enough.



# NARODOWY KORPUS JĘZYKA POLSKIEGO

## Poliquarp search engine for NKJP data

QUERY  
SETTINGS  
FILE A BUG  
HELP

Query:

Corpus:

## Results

Found 196 results so far

Displaying results 1—10

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

## Czym jest Korpusomat?

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

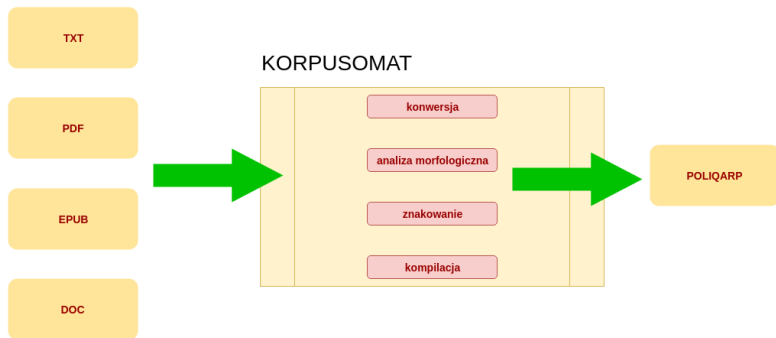
## Motywacja

- 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.



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

DEMO

# Przykład analizy językowej

## Konteksty rzeczownika wojna

The screenshot shows the Poliqarp application window. The search term 'wojna' is entered in the search bar. The results are displayed in a table with three columns: 'Lewy kontekst', 'Dopasowanie', and 'Prawy kontekst'. The first row is highlighted in blue, and the remaining rows are highlighted in green. Below the table, a text snippet is shown, containing the word 'wojna' in bold. The interface includes a menu bar with 'Plik', 'Statystyki', 'Kreator zapytania', and 'Ustawienia', and a toolbar with 'Wykonaj' and 'Pgmoc'.

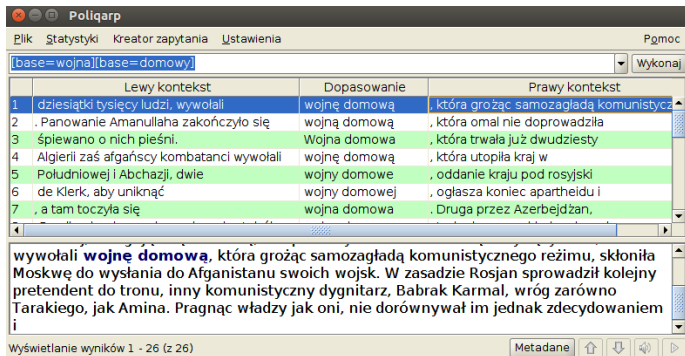
	Lewy kontekst	Dopasowanie	Prawy kontekst
1	Osetii Południowej od roku trwała	wojna	, a Cchinwali znajdowało się
2	, w kraju wybuchnie krwawa	wojna	, a czarni odbiorą władzę
3	to wszystko było. Wybuchła	wojna	, a front przebiegł właśnie
4	wynosić. Tu będzie tylko	wojna	, a przed wojną trzeba
5	to możliwe, póki trwa	wojna	, a ta się nie
6	Kabulu trwała już w najlepsze	wojna	, a według handlowych faktur
7	wtedy gdy w Abchazji wybuchła	wojna	, a władze gruzińskie ogłosiły
8	zbuntuje się i będzie nowa	wojna	, albo przestanie być miastem

, a my umieramy razem z nim. Nadal śpimy, jemy, rozmawiamy, a jednak z każdym dniem coraz mniej pozostaje w nas życia – powiedział Ludwig Czybirow, otrzepując palto ze śniegu. Był rektorem cchinwalskiego instytutu pedagogicznego. Mimo że w Osetii Południowej od roku trwała **wojna**, a Cchinwali znajdowało się w oblężeniu, Czybirow co rano brnął przez zasypy do instytutu uczyć studentów etnografii. – Przecież wojna musi się

Wyświetlanie wyników 1 - 50 (z 203)

# Przykład analizy językowej

## Konteksty wszystkich form frazy wojna domowa



The screenshot shows the Poliqarp search interface. The search query is "[base=wojna][base=domowy]". The results are displayed in a table with three columns: "Lewy kontekst", "Dopasowanie", and "Prawy kontekst".

	Lewy kontekst	Dopasowanie	Prawy kontekst
1	dziesiątki tysięcy ludzi, wywołali	wojnę domową	, która grożąc samozagładą komunistycz
2	. Panowanie Amanullaha zakończyło się	wojną domową	, która omal nie doprowadziła
3	śpiewano o nich pieśni.	Wojna domowa	, która trwała już dwudziesty
4	Algierii zaś afgańscy kombatanci wywołali	wojnę domową	, która utopila kraj w
5	Południowej i Abchazji, dwie	wojny domowe	, oddanie kraju pod rosyjski
6	de Klerk, aby uniknąć	wojny domowej	, ogłasza koniec apartheidu i
7	, a tam toczyła się	wojna domowa	. Druga przez Azerbejdżan,

Below the table, a detailed context is shown for the first result:

wywołali **wojnę domową**, która grożąc samozagładą komunistycznego reżimu, skłoniła Moskwę do wysłania do Afganistanu swoich wojsk. W zasadzie Rosjan sprowadził kolejny pretendent do tronu, inny komunistyczny dygnitarz, Babrak Karmal, wróg zarówno Tarakiego, jak Amina. Pragnąc władzy jak oni, nie dorównywał im jednak zdecydowaniem i

Wyświetlanie wyników 1 - 26 (z 26)

# Przykład analizy statystycznej

## Lista frekwencyjna rzeczowników

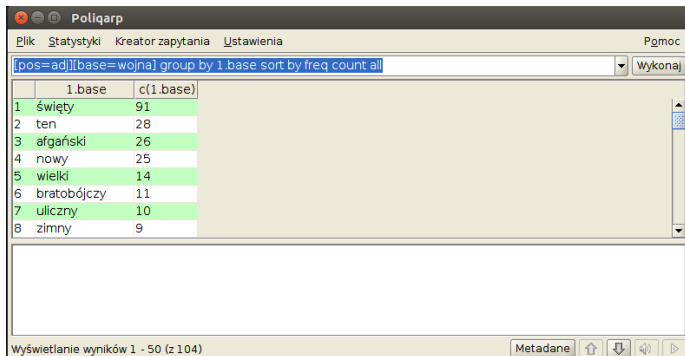


The screenshot shows the Poliqarp application interface. At the top, there are menu items: Plik, Statystyki, Kreator zapytania, Ustawienia, and Pomoc. Below the menu is a search bar containing the query `[pos=subst] group by base sort by freq count all` and a 'Wykonaj' button. The main area displays a table with two columns: 'base' and 'c(base)'. The table contains 8 rows of data, with the first column numbered 1 through 8. Below the table, there is a status bar indicating 'Wyświetlanie wyników 1 - 50 (z 10294)' and a 'Metadane' button with several icons.

	base	c(base)
1	to	1915
2	wojna	1217
3	miasto	951
4	co	767
5	wszystko	750
6	rok	745
7	dom	727
8	dzień	687

# Przykład analizy statystycznej

## Lista frekwencyjna przymiotników w lewym kontekście



The screenshot shows the Poliqarp application window. The title bar reads "Poliqarp". The menu bar includes "Plik", "Statystyki", "Kreator zapytania", "Ustawienia", and "Pomoc". The main input field contains the query: "[pos=adj][base=wojna] group by 1.base sort by freq count all". A "Wykonaj" button is located to the right of the input field. Below the input field is a table with two columns: "1.base" and "c(1.base)". The table contains the following data:

	1.base	c(1.base)
1	święty	91
2	ten	28
3	afgański	26
4	nowy	25
5	wielki	14
6	bratobójczy	11
7	uliczny	10
8	zimny	9

At the bottom of the window, it says "Wyświetlanie wyników 1 - 50 (z 104)". There are also buttons for "Metadane" and navigation icons.

# 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.