Korpusomat — narzędzie do tworzenia przeszukiwalnych korpusów języka polskiego

Witold Kieraś **Łukasz Kobyliński** Maciej Ogrodniczuk

Instytut Podstaw Informatyki PAN

6.02.2017

Agenda

Wprowadzenie — prezentacja

- Korpusomat — motywacja i sposób działania.
- Parę słów o implementacji.
- Znakowanie a słowniki morfologiczne (Morfeusz1 vs Morfeusz2).

Demonstracja systemu

- Korpusomat — tworzenie własnego korpusu językowego.
- Przykłady analiz.
- Wnioski z dotychczasowych warsztatów i plany na przyszłość.

Jaki problem staramy się rozwiązać?

Lingwistyka korpusowa to kluczowe narzędzie analizy jakościowej i ilościowej języka naturalnego.

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.,
- ...

Przykłady korpusów z ostatnich lat

- Słownik Warszawski,
- Korpus Języka Młodzieży.

Utworzenie korpusów wymagało współpracy z programistami / lingwistami komputerowymi.

Korpusomat

Czym jest Korpusomat?

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

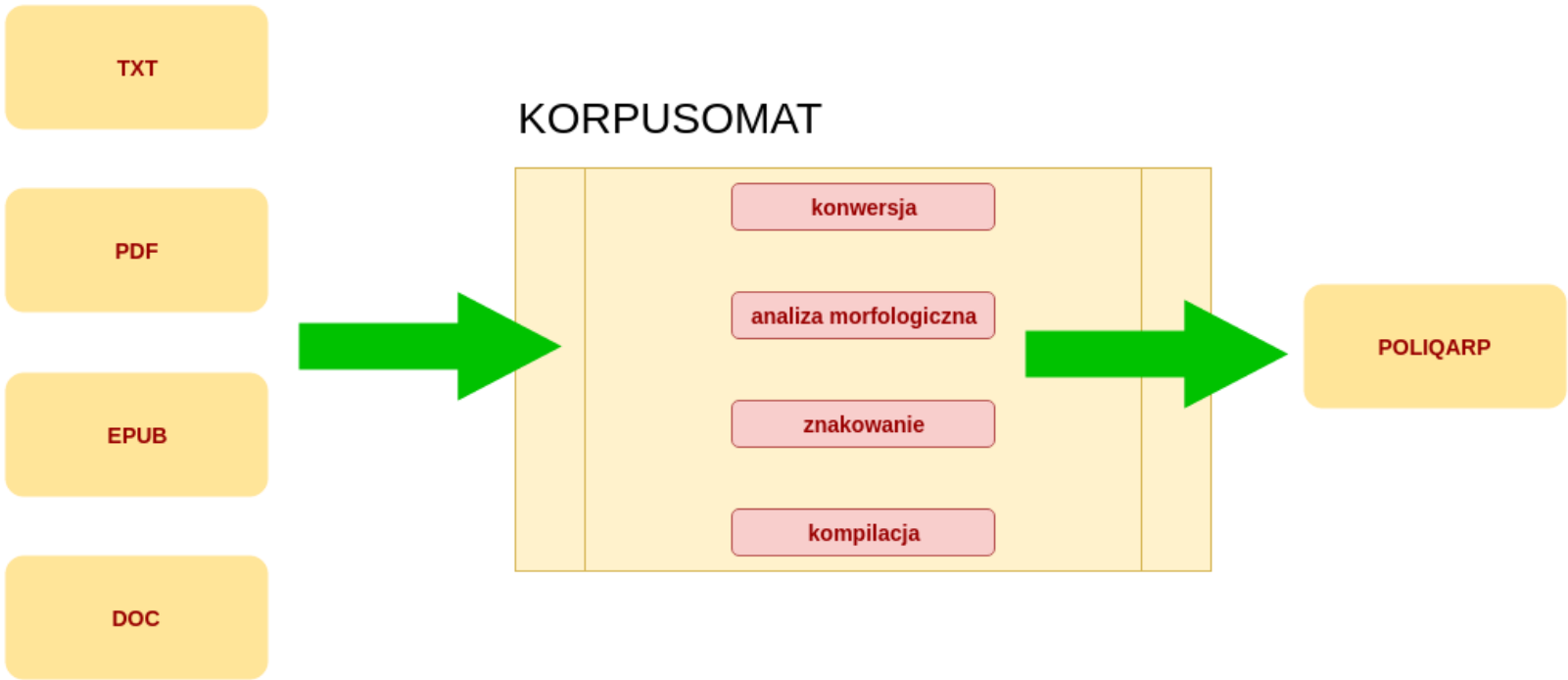
Motywacja

- 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,
  - złożonymi systemami, realizującymi również wiele innych funkcji,
  - 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 — funkcje

Dodatkowe możliwości

- pobieranie tekstów ze wskazanych adresów internetowych (web-scraping),
- masowe ładowanie wielu tekstów z plików (drag-and-drop),
- ładowanie archiwów plików źródłowych (zip),
- autodetekcja metadanych,
- konfiguracja własnej struktury metadanych.

Korpusomat — działanie

Etapy przetwarzania

- ekstrakcja tekstu
  - konwersja formatów binarnych (`Apache Tika, Calibre`),
  - izolacja treści głównej (`newspaper`),
  - konwersja kodowania tekstu do UTF-8 (`enconv`),
  - autodetekcja metadanych,
- segmentacja i analiza morfologiczna tekstu (`Morfeusz, Maca`),
- znakowanie morfosyntaktyczne (`Concraft-pl`),
- tworzenie binarnej postaci korpusu, do przeszukiwania oprogramowaniem `Poliqarp` (`bpng`).

Korpusomat — uwagi implementacyjne

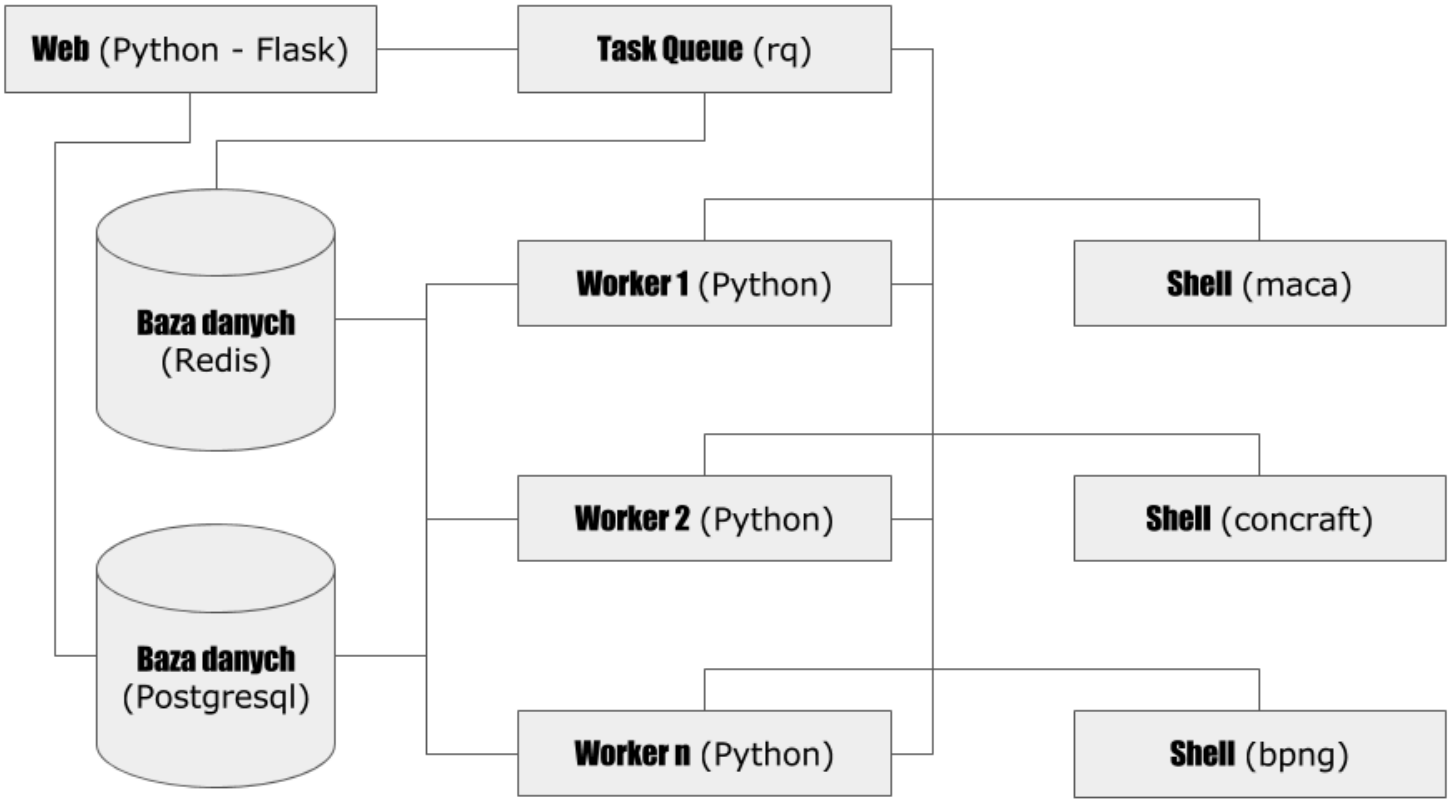
Kluczowe założenia

- aplikacja opiera się na łączeniu wyników istniejących narzędzi,
- powinna być maksymalnie dostępna i łatwa w użyciu dla osób nietechnicznych,
- budujemy prototyp, żeby zebrać informacje od użytkowników,
- minimalne nakłady czasowe i finansowe,
- ale jednocześnie stabilność i pewność działania.

Najważniejsze decyzje projektowe

- aplikacja webowa,
- asynchroniczne przetwarzanie,
- (*) narzędzia wołane z wiersza poleceń,
- (*) przeszukiwanie korpusu w aplikacji desktopowej.

Korpusomat — architektura



Korpusomat — łańcuch przetwarzania języka

Podjęte decyzje

- segmentacja i analiza morfologiczna: wybór pomiędzy `Morfeuszem 1`, a `Morfeuszem 2` pozostawiony użytkownikowi
  - obecne tagery wytrenowane na danych analizowanych `Morfeuszem 1`,
  - pozostałe elementy łańcucha przetwarzania dostosowane do `Morfeusza 1`,
- znakowanie: `Concraft-pl`
  - najwyższa dokładność znakowania wg testów na NKJP 1M wśród pojedynczych tagerów,
- przeszukiwanie: `Poliqarp 1`
  - dla `Poliqarpa 1` istnieje klient desktopowy,
  - `Poliqarp 1` wspiera zapytania statystyczne.

Słownik morfologiczny a tagery

Czy można swobodnie wymieniać słowniki morfologiczne przy znakowaniu tekstu?

- zasadniczo tager powinien być uczony na takich samych danych, które taguje,
- realizacja tego postulatu wymaga przekonwertowania anotacji NKJP 1M oraz pełnego dostosowania tagera Concraft (w trakcie),
- testy eksperymentalne pokazują, że sama podmiana słownika na etapie znakowania może poprawić jakość znakowania, ale wyniki nie są w pełni przewidywalne.

Znakowanie: Morfeusz 1 vs Morfeusz 2

```
co      space                          co      space
    CO      subst:pl:acc:n                         co:c    comp
    CO      subst:pl:dat:n                         co:p1   prep:acc
    CO      subst:pl:gen:n                         co:p1   prep:nom
    CO      subst:pl:inst:n                        co:p2   prep:gen
    CO      subst:pl:loc:n                         co:q    qub
    CO      subst:pl:nom:n                         co:s    subst:sg:acc:n  disamb
    CO      subst:pl:voc:n                         co:s    subst:sg:gen:n
    CO      subst:sg:acc:n  disamb                 co:s    subst:sg:nom:n
    CO      subst:sg:dat:n
    CO      subst:sg:gen:n
    CO      subst:sg:inst:n
    CO      subst:sg:loc:n
    CO      subst:sg:nom:n
    CO      subst:sg:voc:n
    co      comp
    co      prep:acc
    co      prep:nom
    co      qub
    co      subst:sg:acc:n  disamb
    co      subst:sg:gen:n
    co      subst:sg:nom:n
```

Korpusomat w liczbach

Najważniejsze statystyki

- 50+ użytkowników,
- 10M+ przetworzonych segmentów,
- 100+ utworzonych korpusów,
- 2 telefony od użytkowników "na biurko" :)

Demo

DEMO

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

Przykłady analiz — Joseph Conrad

Korpus

- wszystkie utwory Josepha Conrada z Wolnych Lektur (dwie powieści, przygarść opowiadań),
- prawie 400 tys. segmentów.

Zapytanie

```
[pos=subst] group by base sort by freq count all
```

Rezultat

Na liście dać kilka wyraźnie tematyczny (marynistycznych) rzeczowników:

- kapitan (4. miejsce, ponad 1000 wystąpień!)
- statek (8.), morze (21.), pokład (27.)
- okręt (29.), parowiec (30.)

Przykłady analiz — Joseph Conrad

[pos=subst] group by base sort by freq count all

	base	c(base)			
1	to	2407	16	chwila	570
2	człowiek	1198	17	głos	524
3	pan	1073	18	CO	490
4	kapitan	1053	19	życie	484
5	Pan	866	20	dzień	469
6	czas	786	21	morze	453
7	co	774	22	twarz	449
8	oko	770	23	Massy	444
9	statek	739	24	słowo	435
10	głowa	669	25	Whalley	428
11	raz	644	26	woda	402
12	coś	633	27	pokład	392
13	nic	592	28	Jim	390
14	ręka	579	29	okręt	385
15	wszystko	575	30	parowiec	384
			31	ludzie	372

Przykłady analiz — Krzysztof Varga

Korpus

- dwie powieści (Masakra, Trociny) i jednak książka eseistyczna (Langosz w jurcie),
- 337 tys. segmentów.

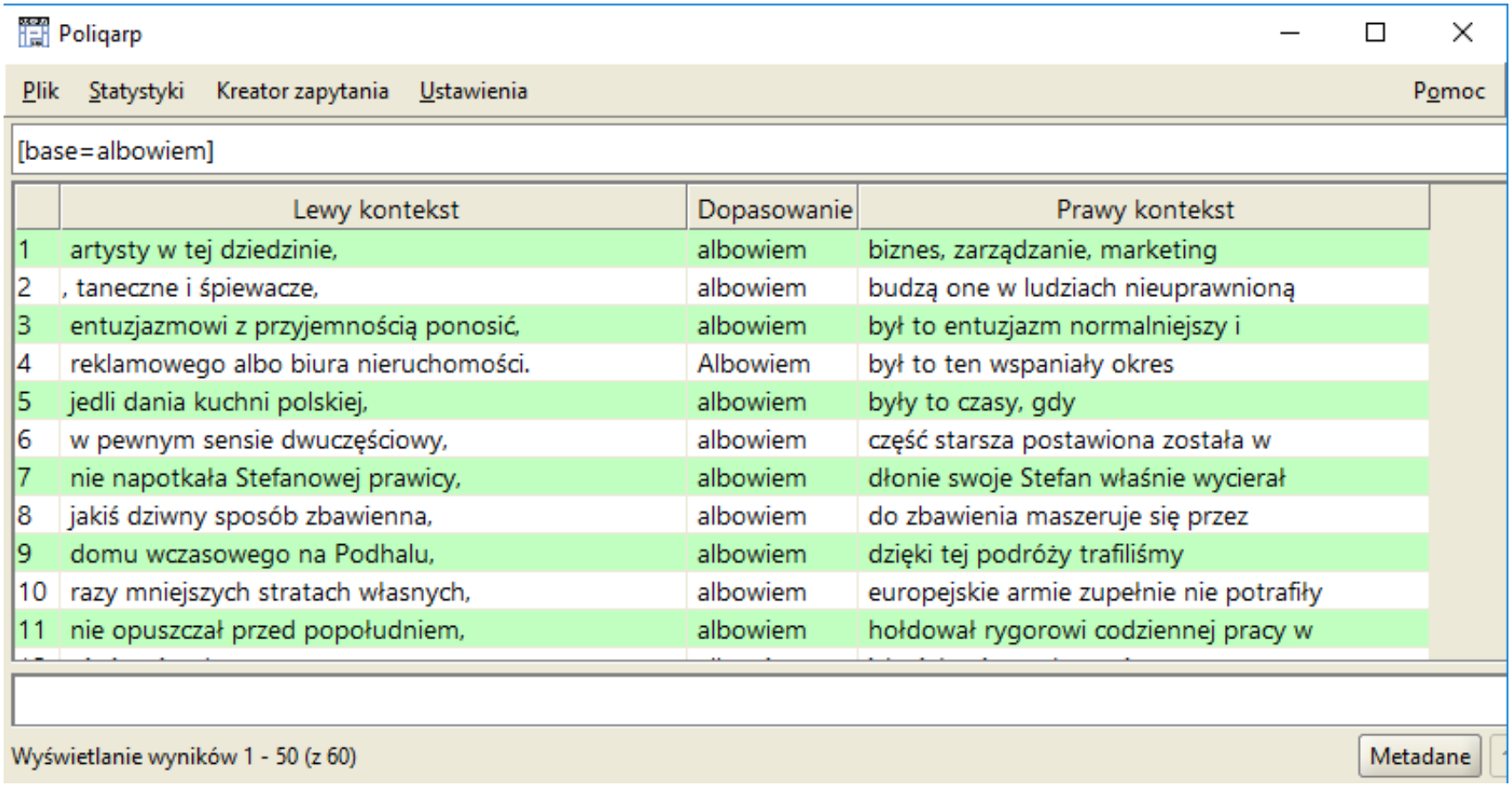
Cel

Sprawdzić, czy Varga faktycznie nadużywa spójnika ALBOWIEM.

Rezultat

W korpusie tekstów Vargi: 60 wystąpień, a liście rangowej spójników podrzędnych 18. miejsce.
Dla porównania w NKJP1M (prawie 4 razy większy korpus): tylko 11 wystąpień, 30. miejsce na liście rangowej.
Ergo: Varga używa ALBOWIEM wyraźnie częściej. Widać też, że w przypadku innych spójników podrzędnych nie ma aż takich różnic.

Przykłady analiz — Krzysztof Varga



	Lewy kontekst	Dopasowanie	Prawy kontekst
1	artysty w tej dziedzinie,	albowiem	biznes, zarządzanie, marketing
2	, taneczne i śpiewacze,	albowiem	budzą one w ludziach nieuprawnioną
3	entuzjazmowi z przyjemnością ponosić,	albowiem	był to entuzjazm normalniejszy i
4	reklamowego albo biura nieruchomości.	Albowiem	był to ten wspaniały okres
5	jedli dania kuchni polskiej,	albowiem	były to czasy, gdy
6	w pewnym sensie dwuczęściowy,	albowiem	część starsza postawiona została w
7	nie napotkała Stefanowej prawicy,	albowiem	dłonie swoje Stefan właśnie wycierał
8	jakiś dziwny sposób zbawienna,	albowiem	do zbawienia maszeruje się przez
9	domu wczasowego na Podhalu,	albowiem	dzięki tej podróży trafiliśmy
10	razy mniejszych stratach własnych,	albowiem	europejskie armie zupełnie nie potrafiły
11	nie opuszczał przed popołudniem,	albowiem	hołdował rygorowi codziennej pracy w

Przykłady analiz — Newsweek

Korpus

- Newsweek, rocznik 2016, wszystkie 52 numery,
- ok 2,4 mln segmentów.

Zapytanie

```
[base=polityka & case=$1 & number=$2 & gender=$3][pos=adj &
case=$1 & number=$2 & gender=$3] group by 1.base;2.base sort
by scp min 5 count all
```

Rezultat

Kolokacje, uszeregowane od najbardziej prawdopodobnych:

- zagraniczny, historyczny, gospodarczy,
- społeczny, wewnętrzny.

Przykłady analiz — Newsweek

=polityka & case=$1 & number=$2 & gender=$3][pos=adj & case=$1 & nu

	1.base	2.base	c(1.base)	c(2.base)	c(1.base;2.base)	scp
1	polityka	zagraniczny	323	62	62	0,192
2	polityka	historyczny	323	55	55	0,170
3	polityka	gospodarczy	323	20	20	0,062
4	polityka	społeczny	323	14	14	0,043
5	polityka	wewnętrzny	323	13	13	0,040
6	polityka	pieniężny	323	12	12	0,037
7	polityka	imigracyjny	323	9	9	0,028
8	polityka	europejski	323	9	9	0,028
9	polityka	kadrowy	323	8	8	0,025
10	polityka	klimatyczny	323	6	6	0,019
11	polityka	rodzinny	323	6	6	0,019
12	polityka	migracyjny	323	6	6	0,019
13	polityka	fiskalny	323	5	5	0,015
14	polityka	kulturalny	323	5	5	0,015
15	polityka	obronny	323	5	5	0,015

Przykłady analiz — Newsweek (2)

Zapytanie

```
[pos=subst][pos=conj][pos=subst] group by 1.base; -1.base
sort by scp min 10 count all
```

Rezultat

Wiele ciekawych przykładów, np. „Schetyna i Petru” (ale nie „Petru i Schetyna”!), „komunista i złodziej” (z wyrażenia „cała Polska z was się śmieje, komuniści i złodzieje”). Ale też wiele konwersów:

- ręka + noga,
- mężczyzna + kobieta,
- imię + nazwisko,
- brat + siostra,
- ojciec + syn,
- śmierć + życie.

Przykłady analiz — Newsweek (2)

[pos=subst][pos=conj][pos=subst] group by 1.base; -1.base sort by scp min 10 count all

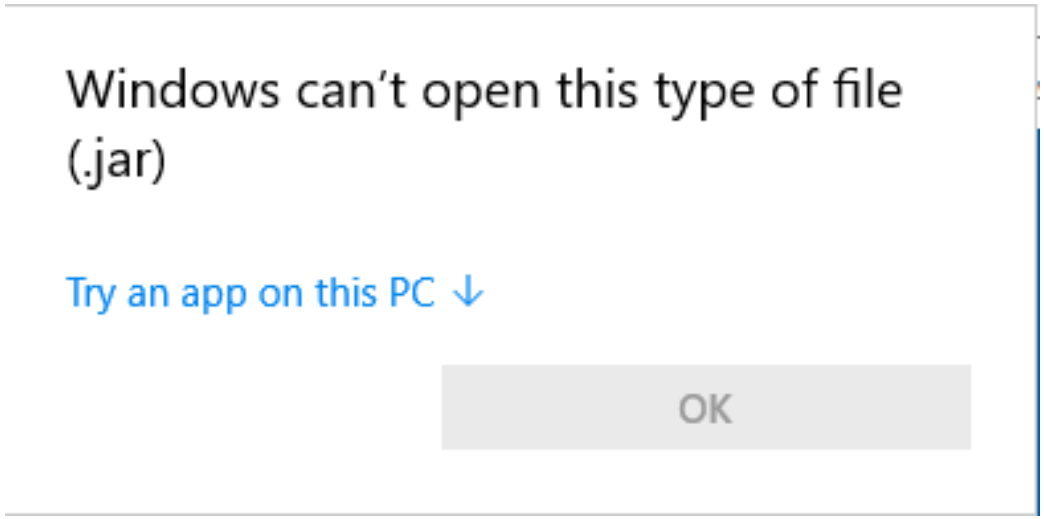
	1.base	-1.base	c(1.base)	c(-1.base)	c(1.base; -1.base)	scp
1	popiół	diament	25	26	24	0,886
2	ręka	noga	18	14	14	0,778
3	reżyser	producent	64	65	54	0,701
4	Aleksandra	Jacek	12	19	12	0,632
5	prawo	sprawiedliwość	155	99	96	0,601
6	Schetyna	Petru	19	15	13	0,593
7	kompozytor	dziennikarz	55	83	52	0,592
8	imię	nazwisko	16	23	14	0,533
9	mężczyzna	kobieta	21	27	17	0,510
10	krew	kość	33	21	18	0,468
11	radio	telewizja	21	30	17	0,459
12	kobieta	mężczyzna	63	30	29	0,445
13	brat	siostra	13	20	10	0,385
14	komunista	złodziej	20	19	12	0,379
15	ojciec	syn	30	30	14	0,218
16	prezydent	premier	36	26	12	0,154
17	to	owo	69	10	10	0,145

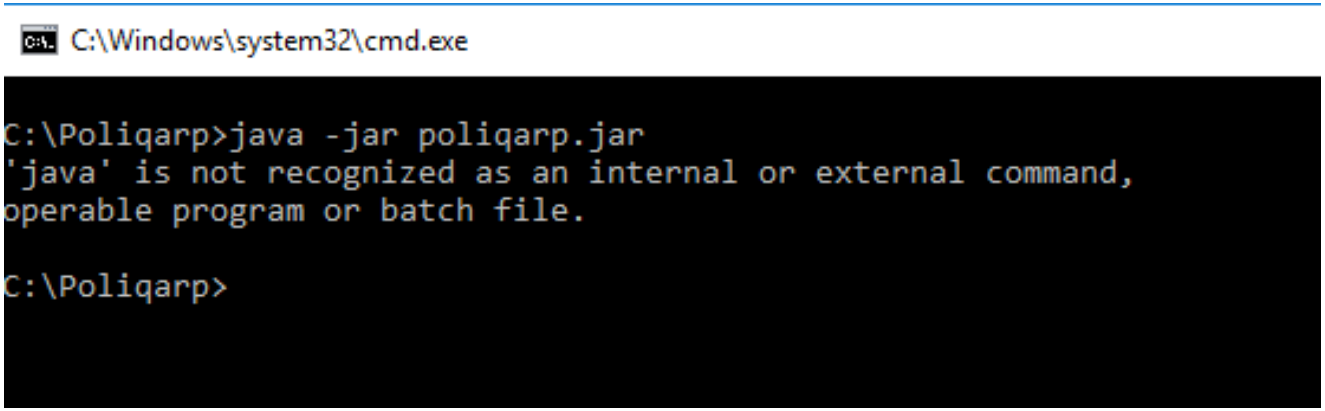
Wnioski po warsztatach DARIAH-PL/CLARIN-PL

Pytania i problemy

- czy możliwe będzie analizowanie innych języków (czeski, rosyjski, niemiecki)?
- czy dane przesyłane do Korpusomatu nie będą dalej udostępniane?
- ...,
- dlaczego Poliqarp nie działa? (szczególnie Windows 10),
- **jak uruchomić Poliqarpa?!**

Problem z aplikacjami desktopowymi (1)





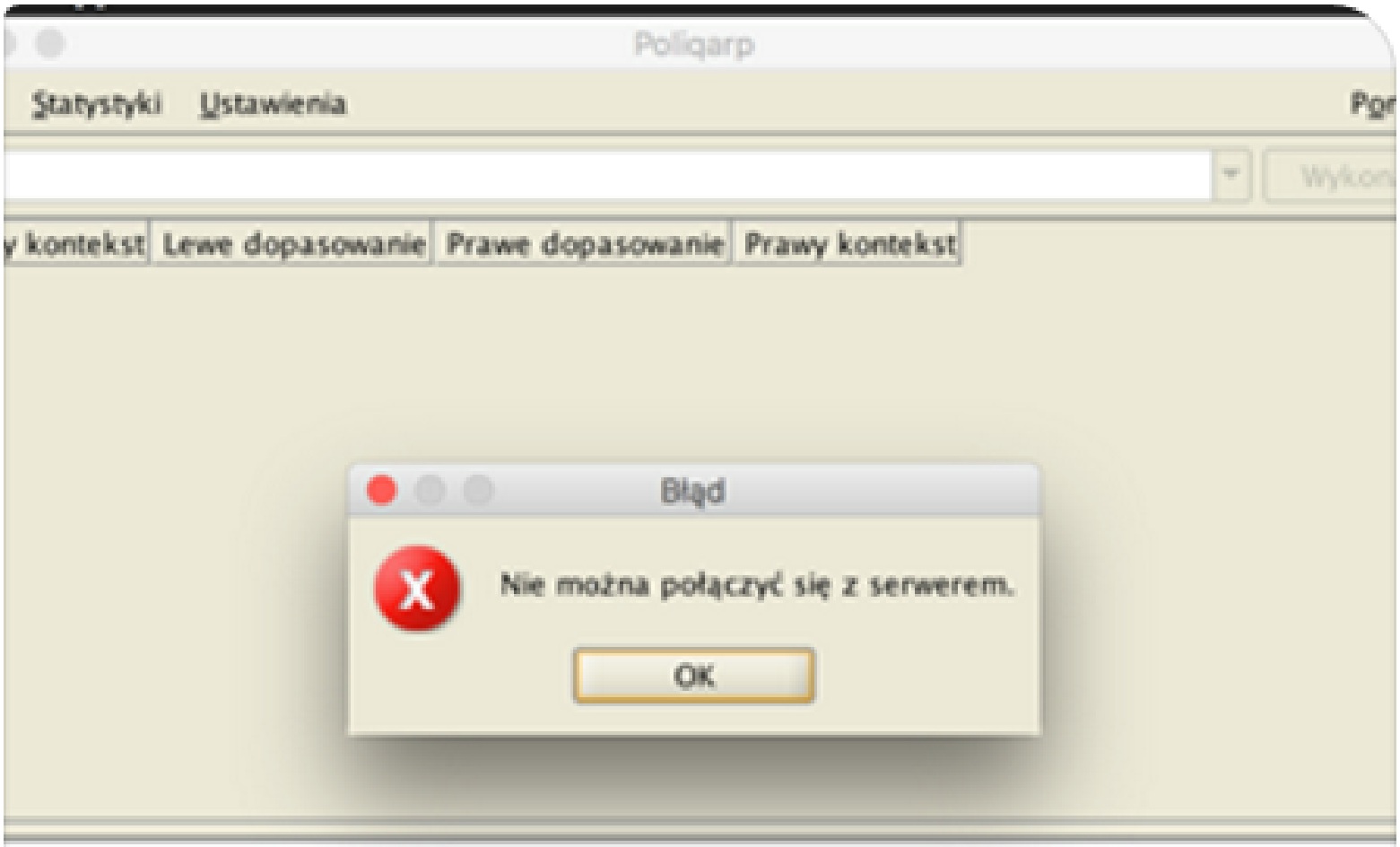
Problem z aplikacjami desktopowymi (2)

Szanowny Panie

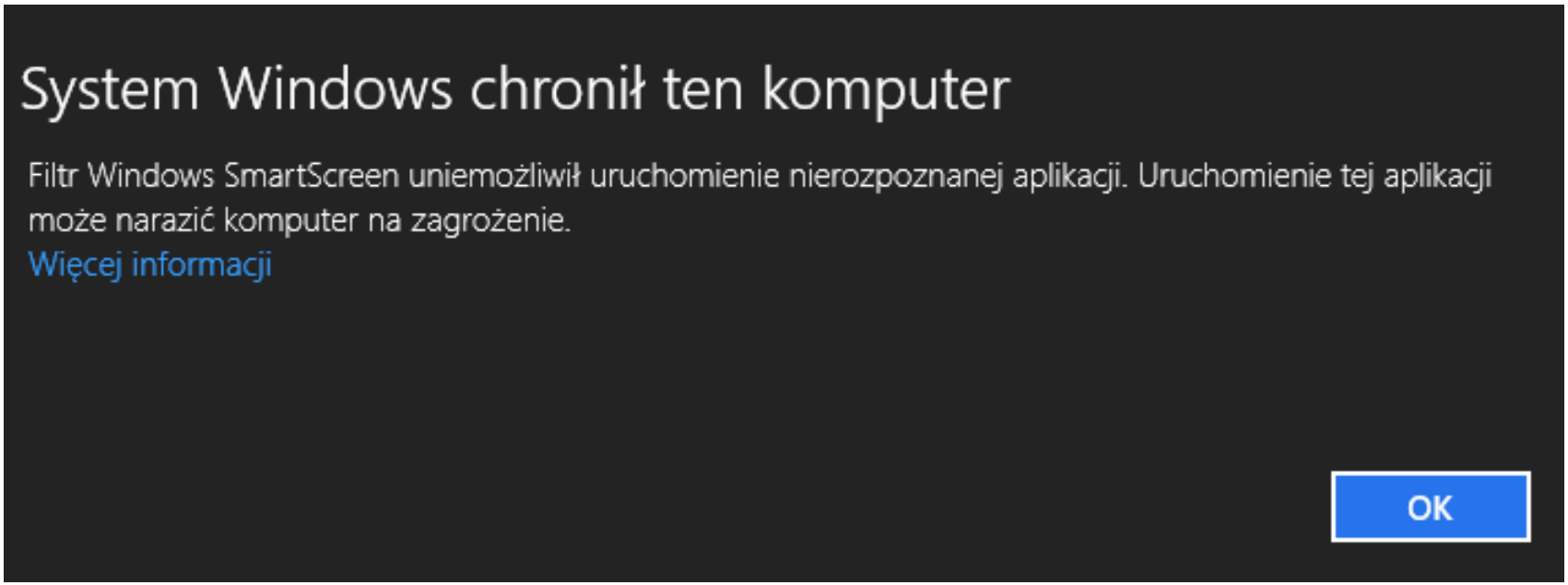
chciałem skorzystać z korpusomatu i poliqarpa na macosx więc zrobiłem próbny korpus w korpusomacie, zainstalowałem poliqarpa i utknałem na komunikacie

nie można połączyć z serwerem

Problem z aplikacjami desktopowymi (2)



Problem z aplikacjami desktopowymi (3)



Dalsze plany

Pomysły na dalsze plany rozwoju Korpusomatu

- interfejs webowy do Poliqarpa / Poliqarpa 2,
- pobieranie źródłowej wersji korpusu (XML).

Sugestie mile widziane!

Dziękujemy!

Dziękujemy za uwagę.