Potencjał technologii języka i sztucznej inteligencji: w jakim punkcie jesteśmy, dokąd powinniśmy zmierzać?

Maciej Piasecki

CLARIN-PL

& Politechnika Wrocławska

(Wrocław University of Science and Technology)

CLARIN ERIC

CLARIN

CLARIN-PL
Common Language Resources and Technology Infrastructure

Wrocław University of Science and Technology

Plan

- Język naturalny w sztucznej inteligencji.
- Technologia językowa i jej zastosowania.
- Szanse i ograniczenia współczesnej głębokiej rewolucji.
- Wyzwanie: rozumienie języka naturalnego w przetwarzaniu języka naturalnego.
- Budowa infrastruktury technologii językowej w Polsce
  - naukowej i badawczo-rozwojowej.
- Kierunki rozwoju, wyzwania i zaproszenie do współpracy.

Język naturalny w sztucznej inteligencji

- Źródła danych i wiedzy.
- Środek komunikacji.
- Narzędzie oddziaływania.
- Sztuczna inteligencja ukierunkowana na język – „*language-centric AI*” (Projekt ELE, Raport 1.2, 2021):
  - elastyczne wyszukiwanie informacji:
    - również wielomedialne,
    - wyszukiwanie w dużych kolekcjach tekstów,
  - rozpoznawanie pisma,
  - systemy rekomendacji,
  - tłumaczenie maszynowe,
  - automatyczne rozpoznawanie mowy,
  - systemy czatbotowe (ale jeszcze nie dialogowe).

Technologia językowa

- Przetwarzanie języka naturalnego — *Natural Language Processing* (NLP):
  - dziedzina sztucznej inteligencji,
  - konstrukcja programów przetwarzających informację zapisaną w języku naturalnym,
  - podejście inżynierskie, niekoniecznie uwzględniające wiedzę lingwistyczną o języku.
- Inżynieria języka naturalnego – *Natural Language Engineering* (NLE):
  - konstrukcja praktycznych, wielkoskalowych systemów przetwarzających język naturalny, nie tylko ‚eksperymentalnych'
- Technologia językowa – *Language Technology* (LT) / *Human Language Technology*
  - budowa zasobów i narzędzi językowych
  - architektura komponentowa
- Lingwistyka informatyczna — *Computational Linguistics* (CL) (lingwistyka komputerowa):
  - nauka o języku naturalnym: struktura, znaczenia i jego użycie w procesie komunikacji,
  - stosowanie metod formalnych do opisu języka: formułowanie teorii, hipotez, weryfikacja, …
  - konstrukcja teorii łatwych do implementacji.

Zastosowania technologii językowej

- Użytkownicy indywidualni (sondaż ELE, raport D.17, rozdz. 4.3, 2022)
  - najbardziej rozpoznawalne pojęcia: „machine translation”, „chatbot”, „smart personal assistant”,
  - najczęściej używane narzędzia: „machine translation, search tools and proofing tools”,
  - przyszłość: „personal assistant tools are the most desired tool for the future in many European languages”.

Zastosowania technologii językowej

- Naukowcy z dziedziny nauk humanistycznych i społecznych (wg CLARIN-PL), np.:
  - budowa korpusów tekstów i dostęp do danych językowych
  - analiza leksykalno-statystyczna tekstów:
    - przeszukiwanie korpusów tekstu i mowy,
    - również na poziomie znaczeń leksykalnych,
  - wykrywanie i klasyfikacja wystąpień nazw własnych,
  - analiza tematyczna, potencjalnie rozszerzona o nazwy własne i terminy, czas i metadane źródeł,
  - analiza wydźwięku, a także emocji,
  - automatyczna transkrypcja pisma i mowy,
  - automatyczne etykietowanie tekstu wg kategorii zadanych przez przykłady
    - (permanentny problem niedostatecznych danych treningowych),
  - stylometria – w oparciu o różne cechy i w różnej perspektywie,
  - rzutowanie tekstu na kolekcje (sieci pojęć) (ang. *entity linking*),
  - wyszukiwanie semantyczne w tym podobnych tekstów.

Zastosowania technologii językowej

- Instytucje publiczne (wg CLARIN-PL), np.:
  - analiza opinii publicznej:
    - analiza statystyczno-leksykalna tekstów i porównywanie zbiorów,
    - rozpoznawanie wydźwięku i jego ukierunkowania,
    - analiza tematyczna,
  - transkrypcja mowy,
  - wyszukiwanie semantyczne
    - wydobywanie informacji w pewnym zakresie,
  - wykrywanie relacji pomiędzy fragmentami tekstów – wnioskowanie w języku naturalnym.

Zastosowania technologii językowej

- Podmioty gospodarcze (wg CLARIN-PL-Biz), np.:
  - modele językowe: ogólne, dziedzinowe, dostosowane pod zadanie,
  - rozpoznawanie wydźwięku i emocji, w tym aspekt, dziedzinowość,
  - rozpoznawanie i klasyfikacja nazw własnych, wyrażeń temporalnych,
  - wydobywanie słów kluczowych, indeksowanie semantyczne,
  - wyszukiwanie semantyczne z elementami odpowiadania na pytania,
  - klasyfikacja dziedzinowa, filtrowanie semantyczne,
  - systemy dialogowe:
    - automatyzacja konstrukcji zasobów, segmentacja, inteligentna generacja odpowiedzi, rozszerzone śledzenie stanu dialogu, połączenie z systemami odpowiadania na pytania i wyszukiwania.

Bariery dla rozwoju technologii językowej

- Czy możemy efektywnie posługiwać się językiem naturalnym w komunikacji z maszyną, szerzej w ramach systemów informatycznych przetwarzających język naturalny?
- Czy programy komputerowe rozumieją język naturalny?
- Projekt ELE (Raport 3.4, 2022)

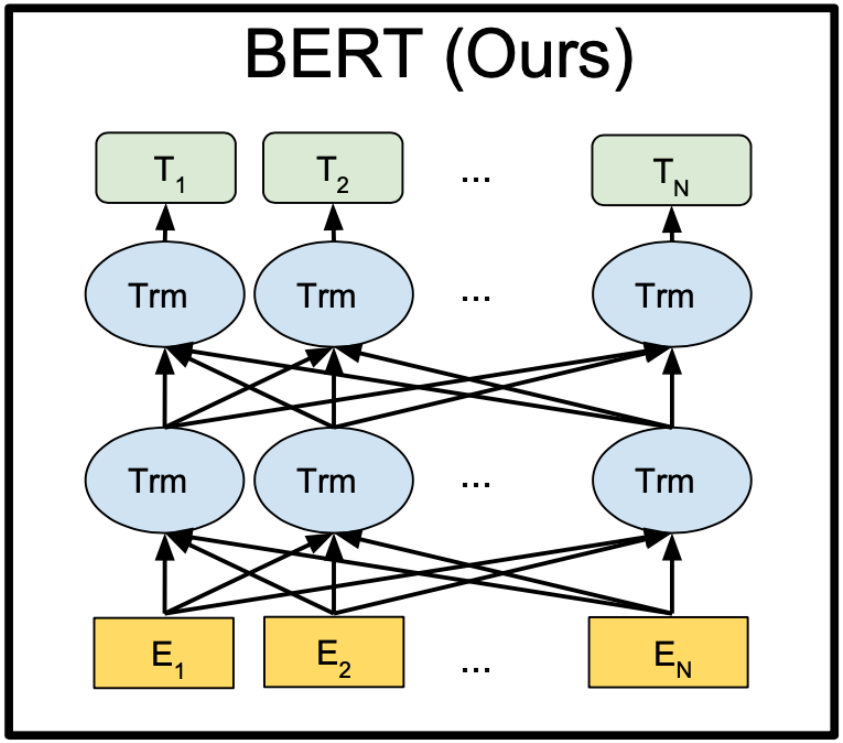
  "But can we fully use our own language in our digital interactions? Is our language adequately supported and ready to keep pace with the technological advancements of the AI era?"
- Dwa aspekty:
  - czynniki przyczyniające się do dużego postępu w PJN w ostatnich latach,
  - poziom TJ dla różnych języków.

Ery w rozwoju technologii językowej

- Era symboliczna (lata 50. – wczesne 90.):
  - gramatyki formalne, modele sformalizowane, bazy danych.
- Era statystyczna (późne 90. do dziś):
  - paradygmat statystyczny :
    - duże zbiory tekstów,
    - statystyczne modele językowe, np. prawdopodobieństwo wystąpienia słowa w kontekście,
    - semantyka dystrybucyjna
      - słowa i frazy o podobnym użyciu reprezentowane przy pomocy podobnych wektorów,
  - paradygmat głębokich modeli neuronalnych:
    - bardzo duża pojemność informacyjna,
    - kondensacja informacji w gęstych wektorach,
    - wrodzona zdolność do grupowania reprezentacji wyrażeń językowych.

Głębokie modele neuronalne

- Przykład – BERT (Devlin i in., 2018)
  - wielowarstwowy dwukierunkowy transformer:
    - architektura kodująco-dekodująca,
    - **340 mln. parametrów**



- Ogromna liczba parametrów – **setki miliardów**
  → potrzebne bardzo duże zbiory danych do wytrenowania
  - efekt ogromnej pamięci asocjacyjnej,
  - ryzyko przeuczenia przy mniejszych danych,
  - niepewność odnośnie działania.

Przykład – ograniczenia modelu GPT-3

- GPT-3
  - generatywny model językowy oparty na głębokich sieciach neuronowych,
  - dane treningowe: ok. 480 tokenów (części wyrazów),
  - liczba parametrów: 175 miliardów → 750 GB pamięci (sic!),
  - wiele spektakularnych zastosowań, nie tylko generowanie tekstu.
- Przykłady nieoczekiwanego wnioskowania (Marcus i Davis, MIT Technology Review, 2020)
  - *You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So* **you drink it. You are now dead.**
  - *You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to* **remove the door. You have a table saw, so you cut the door in half and remove the top half.**

Ograniczenia głębokich modeli neuronalnych

- Dostępność bardzo dużych wolumenów danych:
  - ograniczenia licencyjne,
  - największe światowe zbiory w rękach kilku firm,
  - bardzo niezrównoważona reprezentacja języków naturalnych
    - głównie język angielski i chiński (ELE, Raport 1.2, 2021).
- Dostępność danych dla różnych dziedzin i typów zadań.
- Ograniczone zbiory treningowo-testowe:
  - opisane metadanymi (zwykle przez ludzi) → trenowanie głębokich sieci do zadania,
  - te same standardowe zbiory wzorcowe (ang. *benchmark data sets*) – wybiórcze i odległe od danych z praktyki.
- Wymagana bardzo duża moc obliczeniowa:
  - zarówno dla trenowania jak i zastosowania,
  - wysoki ślad węglowy.

Wyzwania

- **Bardziej realistyczne zbiory danych** reprezentujące praktyczne zadania.
- **Głębokie rozumienie języka naturalnego** („Deep Natural Language Understanding”) – raport o strategii rozwoju (Projekt ELE - European Language Equality, Raport 3.4, 2022)
  - „highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale.”
  - „ to accurately and seamlessly integrate modalities, situational and linguistic context, general knowledge,”
  - „Deep understanding is understood in the sense that the resulting application using LT is able to explain itself: why did it make the decision it made given the linguistic context, the situa- tional context (across modalities), linguistic knowledge, world knowledge etc.”

Wyzwania

- **Bardziej strukturalna i dogłębna analiza** wypowiedzi językowych:
  - uwzględnienie struktury logicznej wypowiedzi,
  - wnioskowanie w oparciu o informację i wiedzę z tekstu,
  - w powiązaniu z zasobami symbolicznej wiedzy (np. sieci semantyczne).
- **Małe, często przypadkowe i zaszumione zbiory danych językowych**:
  - szczególnie ręcznie opisane metadanymi,
  - licencje (!) – naukowcy nie posiadają praw do większości zbiorów danych.
- **Reprodukowalność badań i eksperymentów**
  - bardzo trudno znaleźć system, który działa w praktyce zgodnie z obietnicami w opisującej go pracy.

Wyzwania

- **Wielomodalne przetwarzanie**
  - łączne przetwarzanie danych pochodzących z różnych mediów I sygnałów.
- **Nieustannie rosnący koszt obliczeń** na potrzeby co raz bardziej skomplikowanych neuronalnych modeli językowych:
  - coraz trudniejsza sytuacja instytucji naukowych,
  - (które dodatkowo mają problem z zatrzymaniem naukowców).
- **Wyjaśnialne systemy PJN (NLP)**, szczególnie systemy oparte na uczeniu głębokim (ang. Deep Learning)
  - głębokie sieci neuronowe działające w oparciu o wektory z głębokich neuronalnych modeli językowych (sic!).
- **Adaptacja dziedzinowa I personalizacja:**
  - zależność od dziedziny vs aplikacje wielodziedzinowe,
  - adaptacja do grup użytkowników, a nawet pojedynczych użytkowników.
- Konstrukcja w oparciu o ograniczone dane - działanie dla pojedynczych przykładów
  - nie do końca tożsame z *a few shot learning* i *zero shot learning*
- Podejścia neurosymboliczne.

Rozwój technologii językowej dla języka polskiego

- Infrastruktura technologii językowych:
  - CLARIN-PL – finansowana przez MNiE oraz EU (POIG, POIR 4.2),
  - wsparcie CLARIN-PL dla DARIAH.PL i projektu DARIAH.Lab (POIR 4.2).
- Projekty naukowe:
  - kilka ośrodków naukowych,
  - szersze grono zespołów i badaczy.
- Otwarta technologia językowa:
  - wynikająca z CLARIN-PL i wielu projektów naukowych,
  - seria konkursów PolEval zorganizowanych przez IPI PAN.
- Rozwiązania komercyjne:
  - duża liczba małych i średnich firm (POIR 1.1.1, „szybka ścieżka”),
  - kilka dużych firm – szczególnie w dziedzinie budowy dużych neuronalnych modeli językowych.

CLARIN-PL – infrastruktura technologii językowej dla języka polskiego

- CLARIN ERIC (clarin.eu):
  - europejskie konsorcjum państw typu ERIC,
  - część europejskiej mapy drogowej infrastruktury naukowej,
  - obszar działania: nauki humanistyczne i społeczne.
- CLARIN-PL (clarin-pl.eu):
  - polskie konsorcjum naukowe,
  - realizuje polski wkład w budowę i utrzymanie CLARIN ERIC,
  - polski węzeł infrastruktury – Centrum Technologii Językowych,
  - **infrastruktura naukowa dla nauk humanistycznych i społecznych**
- CLARIN-PL-Biz (clarin.biz)
  - projekt infrastrukturalny: „POIR, Priorytet IV: Zwiększenie potencjału naukowo-badawczego, Działanie 4.2: Rozwój nowoczesnej infrastruktury ”
  - **infrastruktura badawczo-rozwojowa i wsparcie dla SI (AI)**
  - obszar: **wszystkie dziedziny nauk oraz zastosowania w biznesie**

CLARIN ERIC – Europa

22 członków:

Austria
Bułgaria
Chorwacja
Cypr
Czechy
Dania
Estonia
Finlandia
Grecja
Holandia
Islandia
Litwa
Łotwa
Niemcy
Norwegia
Polska
Portugalia
Słowenia
Szwecja
Węgry
Włochy

2 obserwatorów:

Francja
Wielka Brytania
RPA

1 instytucja:

Carnegie Mellon University

CLARIN

- CLARIN =
  - Common Language Resources and Technology Infrastructure,
  - wspólne zasoby językowe i infrastruktura technologiczna.
- Rozproszona infrastruktura badawcza technologii językowych dla nauk humanistycznych i społecznych:
  - 24 centra technologiczne i 25 wiedzy w 27 krajach,
  - działających jako jeden wspólny system.
- Zakres działania:
  - zbiory danych i bazy danych opisujące język naturalny oraz jego użycie,
  - programy komputerowe do analizy tekstu i mowy na różnych poziomach analizy języka naturalnego,
  - aplikacje badawcze wspierające badania w obszarach nauk humanistycznych i społecznych.

Oferta CLARIN

- Łatwy dostęp do zasobów językowych:
  - federacja repozytoriów – Virtual Language Observatory
  - federacyjne przeszukiwanie korpusów – Federated Content Search
  - rodziny popularnych zasobów – Resource Families
- Automatyczna analiza tekstu i mowy:
  - sugestie narzędzi – Language Resource Switchboard
  - paleta gotowych do użycia narzędzi językowych:
    - usługi sieciowe (ang. *Web Services*) i aplikacje narzędziowe,
    - dostęp poprzez repozytoria, stronę usług CLARIN-PL
- Aplikacje badawcze
  - konkretne potrzeby, często budowane we współpracy z użytkownikami, np. LEM, WebSty

Oferta CLARIN

- Infrastruktura Wiedzy (*Knowledge Infrastructure*):
  - centra CLARIN typu K: fizyczne i wirtualne,
  - dostęp do wiedzy eksperckiej i wsparcie,
  - poradniki i punkty informacyjne, warsztaty, itd.
- Wymogi dla centrum typu K (wybrane):
  - strona WWW z jasno określonym zakresem usług,
    - np. pomoc, wsparcie techniczne lub technologiczne, kursy, …
  - zapewniają reakcję na kontakt,
    - np. odpowiedzi na zapytania użytkowników w ciągu 2 dni roboczych,
  - aktywne, np. seminaria, warsztaty, szkolenia,
  - dysponują odpowiednią kadrą naukową.

CLARIN Polska: CLARIN-PL

CENTRUM TECHNOLOGII JĘZYKOWYCH
CLARIN-PL

Centrum Technologii Językowej CLARIN-PL (http://clarin-pl.eu)

repozytorium danych językowych

CLARIN B CENTRE

CORE TRUST SEAL

CLARIN Cloud – prywatna chmura danych dla naukowców (https://nextcloud.clarin-pl.eu/)

zasoby językowe dla języka polskiego i innych

usługi i aplikacje do analizy tekstów i mowy (https://ws.clarin-pl.eu)

PolLinguaTec - Centrum Wiedzy Technologii Językowych dla Języka Polskiego

http://kcentre.clarin-pl.eu/

CLARIN CENTRE K

CLARIN-PL – wsparcie dla naukowców

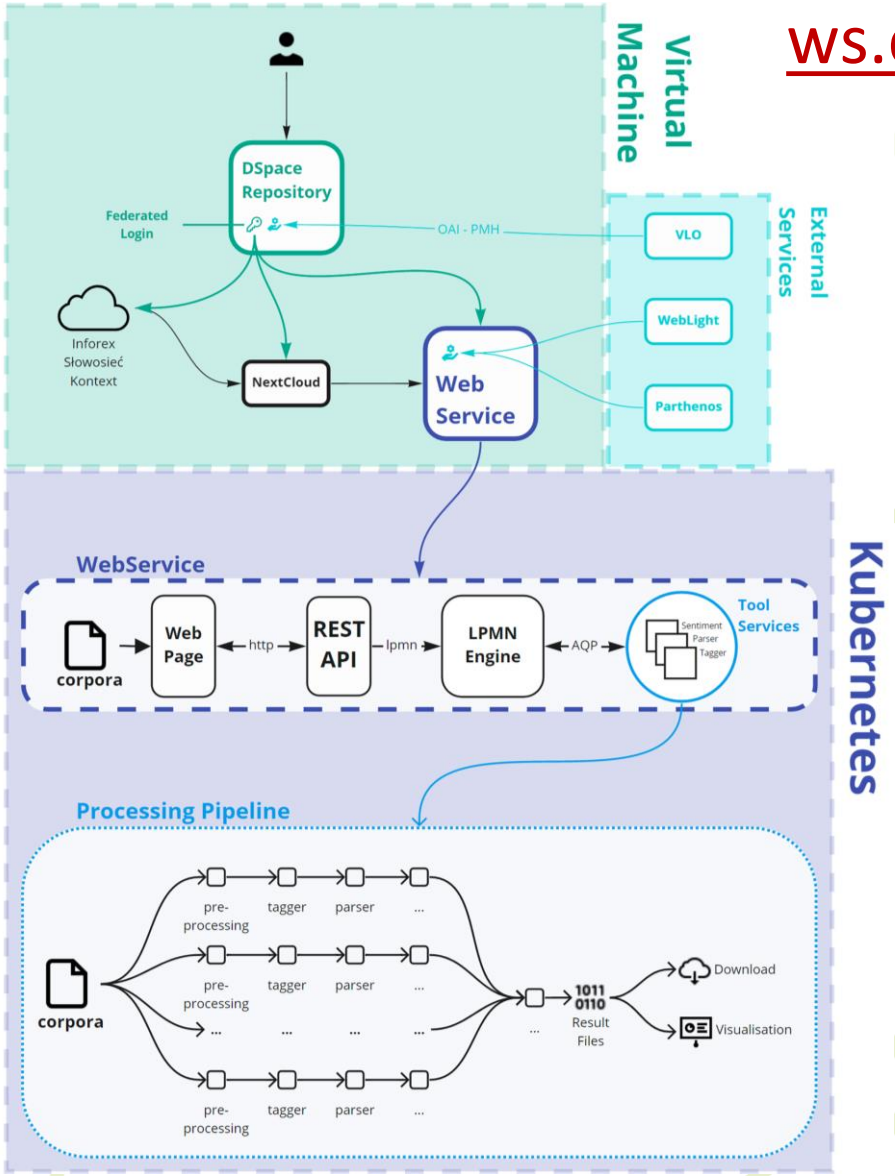
- Twórcy (konsorcjum):
  - Katedra Sztucznej Inteligencji, Politechnika Wrocławska (lider),
  - Zespół Inżynierii Lingwistycznej, Instytut Podstaw Informatyki PAN,
  - Instytut Slawistyki, PAN,
  - Polsko-Japońska Akademia Technik Komputerowych
  - Instytut Anglistyki, Uniwersytet Łódzki,
  - Uniwersytet Wrocławski.
- Beneficjenci:
  - **Wszystkie** jednostki naukowe i **Naukowcy** w Polsce, szczególnie z obszaru Nauk Humanistycznych i Społecznych,
  - również **sztuczna inteligencja** w szerokim rozumieniu (CLARIN-PL-Biz)

CLARIN-PL – wsparcie dla naukowców

- **Otwarta nauka** (FAIR):
  - zasoby językowe i dane, oprogramowanie i aplikacje,
  - wiedza i bezpośrednie wsparcie,
  - **bez opłat** – sponsor **Ministerstwo Edukacji i Nauki** (dziękujemy!).
- Zasoby językowe:
  - korpusy: języka polskiego i wielojęzyczne, bogatoanotowane,
  - leksykalne bazy danych: morfologiczne, gramatyczne, semantyczne – jedne z największych na świecie.
- Podstawowe narzędzia językowe –
  - analiza morfologiczna, gramatyczna, wydobywanie informacji.
- Aplikacje badawcze –
  - budowa i analiza korpusów, analiza statyczna, stylometryczna, wydźwięku, analiza semantyczna.
- Bezpośrednie wsparcie dla naukowców:
  - zespoły, projekty i indywidualni naukowcy oraz studenci,
  - od pomysłu, poprzez problem, zadania, wniosek po realizację badań.

Centrum Technologii Językowych CLARIN-PL

ws.clarin-pl.eu



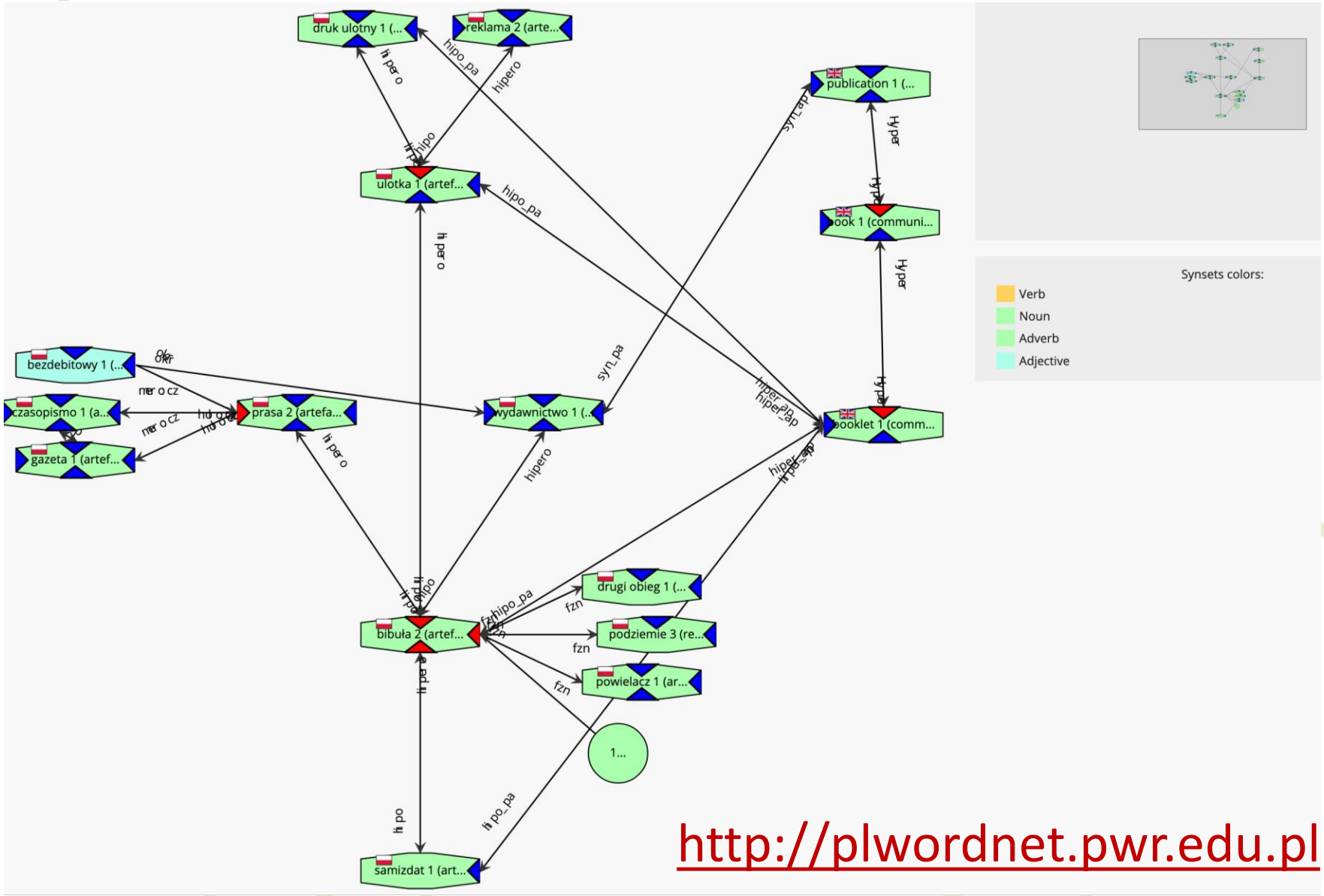
Oferta CLARIN-PL: korpusy

- Korpusy języka polskiego.
- Wielojęzyczne korpusy i systemy korpusowe ukierunkowane na język polski
  - dwujęzyczne i wielojęzyczne korpusy równoległe wyposażone w specjalistyczne systemy do edycji i przeszukiwania.
- Systemy do edycji i analizy jednojęzycznych korpusów:
  - **Inforex** – webowy system do edycji i anotacji korpusów,
  - **Spokes** – wyszukiwarka dla korpusów mowy,
  - **Paralela** – wyszukiwarka dla korpusów równoległych,
  - **ChronoPress** – przeszukiwanie i analiza statystyczna korpusów historycznych,
  - **Korpusomat** – system do budowy i przeszukiwania własnych korpusów użytkowników.

Oferta CLARIN-PL: zasoby językowe

1. System bardzo dużych zasobów leksykalnych:
   - **Słowosieć (plWordNet)** – wielki relacyjny słownik semantyczny języka polskiego wraz z dwujęzycznym (ręcznym) rzutowaniem na Princeton WordNet
     - jak również na wiele zasobów z *Linked Open Data,*
   - **MWELexicon** – bardzo duży leksykon wyrażeń wielowyrazowych,
   - **Walenty** – bardzo duży słownik walencyjny języka polskiego
   - **HASK** – słownik kolokacji polskich i angielskich.
2. Systemy do edycji i publikowania zasobów leksykalnych
   - **WordnetLoom**
     - system do edycji i przeglądania leksykalnych sieci semantycznych, np. wordnetów i słowników relacyjnych,
   - **Lexical Platform (Platforma Leksykalna)**
     - system do przeszukiwania dynamicznie łączonych kolekcji polskich słowników.

plWordNet (Słowosieć)



http://plwordnet.pwr.edu.pl

Oferta CLARIN-PL: narzędzia językowe dla języka polskiego

- System podstawowych narzędzi językowych:
  https://ws.clarin-pl.eu
  - analiza morfosyntaktyczna:
    - lematy i gramatyczne własności słów,
  - narzędzia do wydobywania informacji (ang. *Information Extraction*, *Text Mining*)
    - nazwy, odniesienia do bytów, momentów i okresów czasu, przestrzeni, wybranych relacji semantycznych,
  - ujednoznacznianie znaczeń słów (ang. *Word Sense Disambiguation*) w tekście
    - automatyczne rzutowanie słów na znaczenia w Słowosieci (plWordNet),
    - również na elementy *Linked Open Data*,
  - wydajna i efektywna analiza składniowa:
    - struktury zależnościowe,
    - relacje zależnościowe pomiędzy słowami w tekście.

CLARIN-PL – idea aplikacji badawczych

- **Czytanie z bliska** – ang. *close reading*,
  - inaczej *czytanie bliskie*,
  - uważne czytanie i analiza tekstu – danych językowych – przez badacza.
- **Czytanie na odległość** – ang. *distant reading*,
  - inaczej *czytanie dalekie*, *czytanie odległe*, *czytanie na dystans*,
  - analiza zbiorów tekstów,
  - rozpoznawanie statystycznych cech,
  - szukanie grup i prawidłowości,
  - wydobywanie ukrytej struktury zbiorów danych,
  - objęcie analizą szerszego kontekstu.
- Komplementarne zmiany optyki –
  - czytanie na odległość → czytanie z bliska i *vice versa*.

Nie tylko humanistyka

- Nauki społeczne:
  - badanie ludzi i społeczności,
  - ankiety, badania eksperymentalne, ale również **wywiady pogłębione**,
  - badanie ludzi poprzez ich wytwory, w tym zapisy komunikacji językowej.
- Językowy obraz świata – świat poprzez językowy pryzmat.
- Ludzki umysł poprzez językowy przyrząd projekcyjny.
- Sukcesja wiedzy:
  - wiedza i informacje zapisane w języku,
  - czytanie na odległość źródeł.

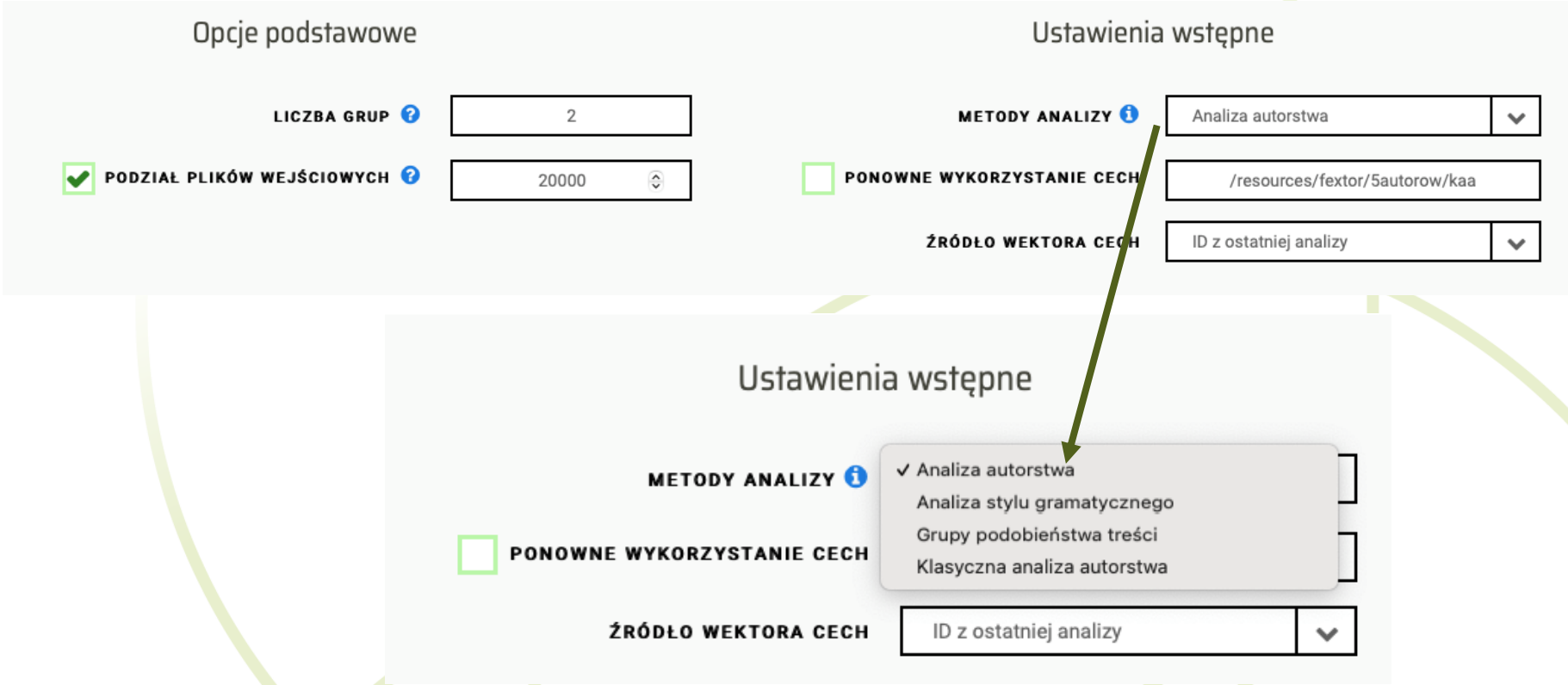
CLARIN-PL: aplikacje badawcze

- Aplikacje webowe nie wymagające instalowania ani specjalistycznej wiedzy
  - warunek trudny do zrealizowania w zakresie lingwistyki i technologii językowych,
  - kompleksowe rozwiązania wraz z wizualizacją,
  - oparte na potokach i architekturze CTJ CLARIN-PL,
  - dostępne również jako usługi sieciowe
    - łatwa integracja z zewnętrznymi systemami.
- Realizacja potrzeb użytkowników
  - użytkownicy kluczowi z różnych obszarów NHiS,
  - potrzeby badawcze vs możliwości i ograniczenia technologii językowej,
  - twórcza współpraca z CLARIN-PL (**bezpłatna**).

WebSty: przetworzenie korpusu

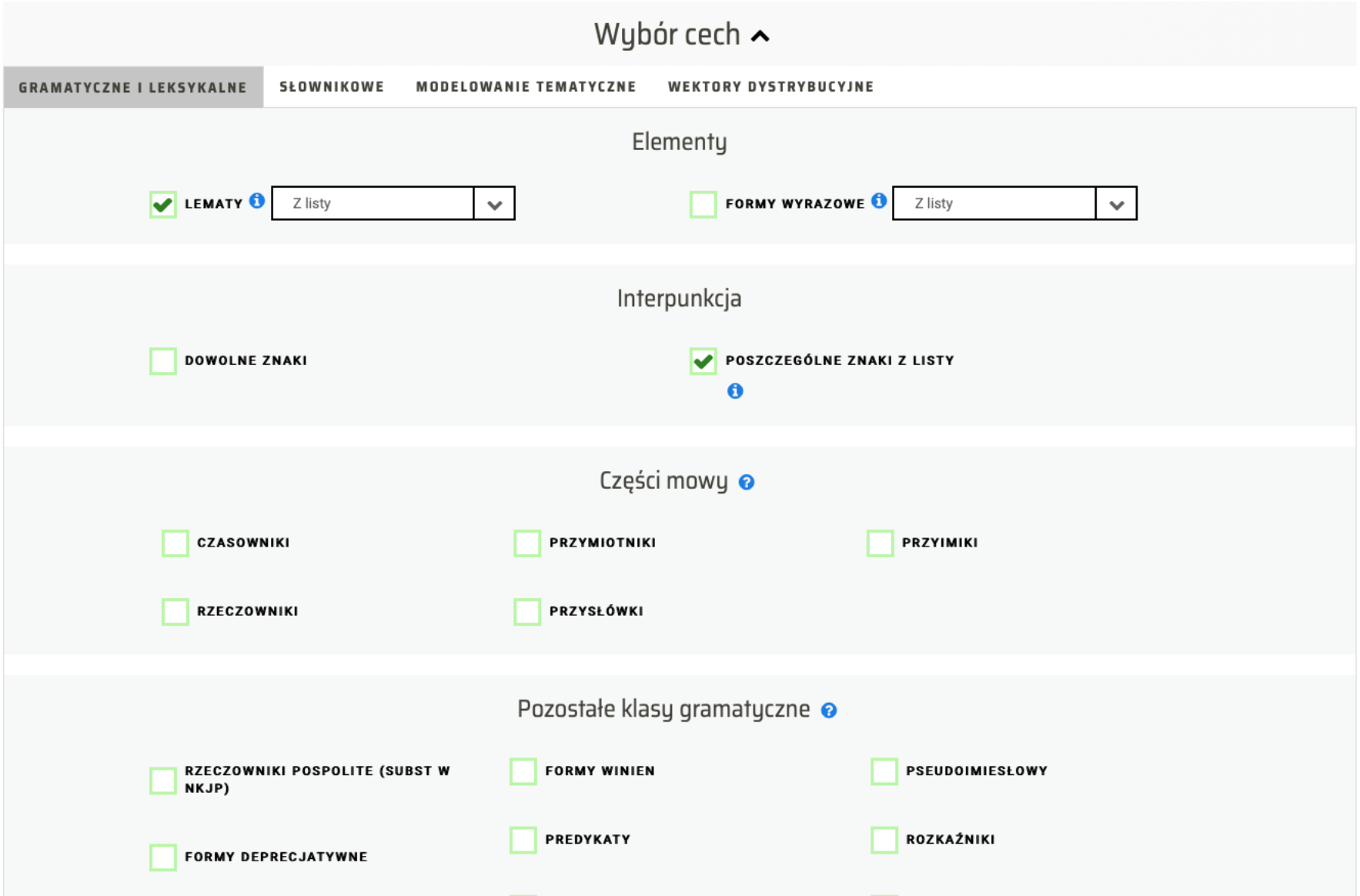
http://ws.clarin-pl.eu/websty.shtml

- Gotowe ustawienia dla typowych zadań



WebSty: przetworzenie korpusu

http://ws.clarin-pl.eu/websty.shtml



Prezentacja wyników

INTERAKTYWNY DENDROGRAM

MAPA CIEPŁA

SKALOWANIE WIELOWYMIAROWE

KALOWANIE WIELOWYMIAROWE Z WIZULIZACJĄ 3D

WYKRES RADAROWY

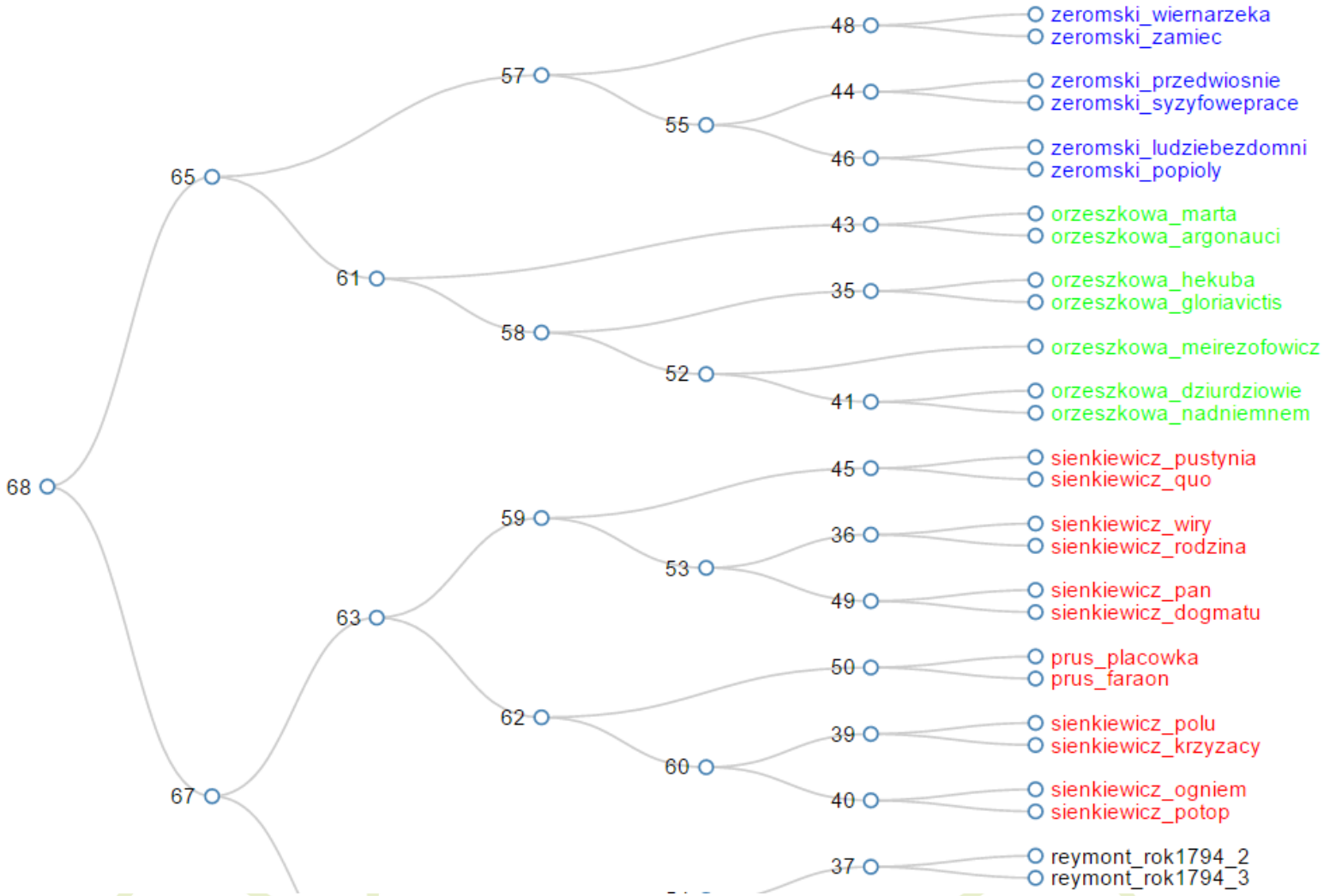
WYKRES KOŁOWY

PLIK XSLX

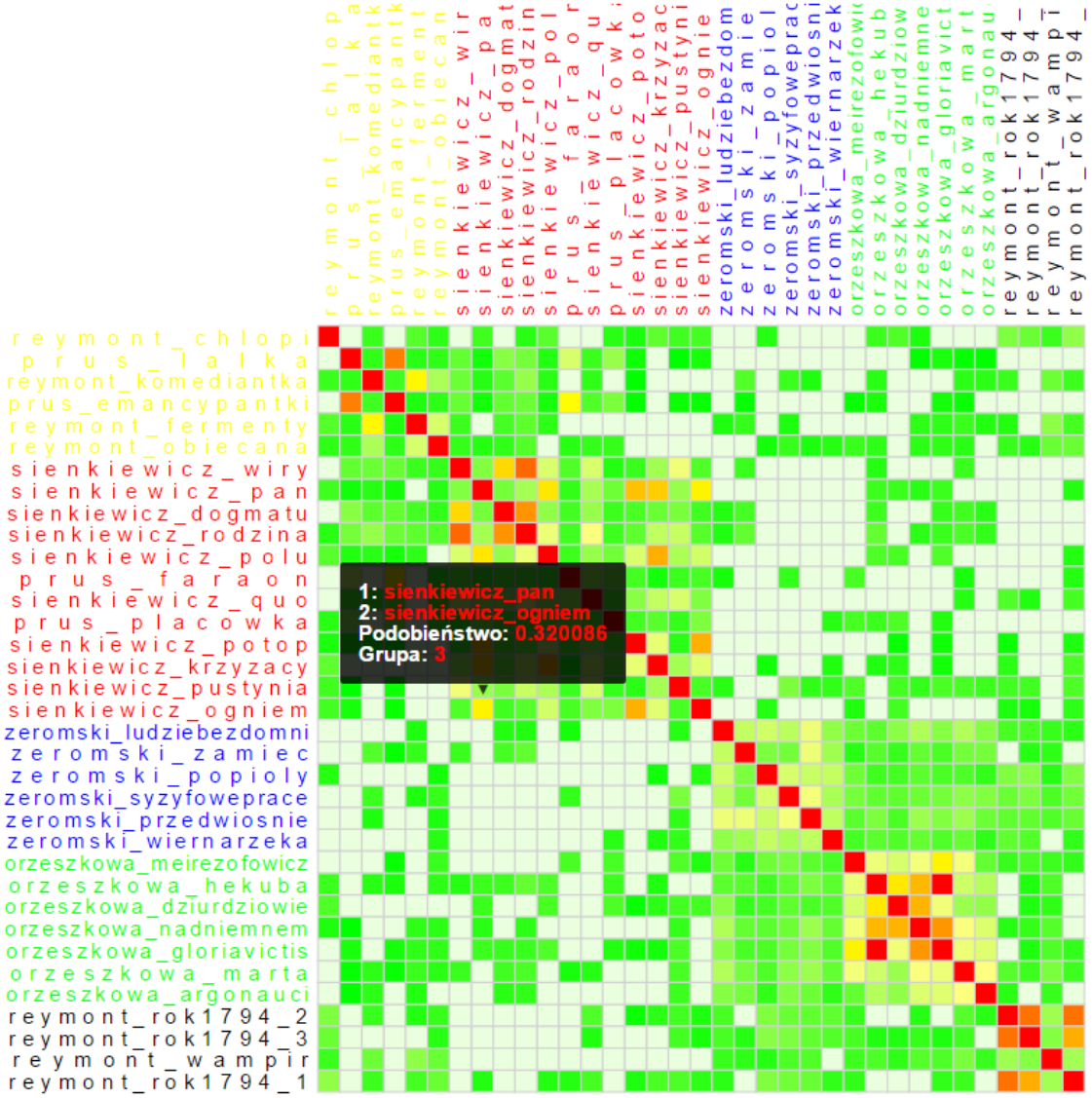
ISTOTNOŚĆ CECH

REZULTATY

Prezentacja wyników – drzewo interaktywne



Prezentacja wyników – mapa ciepła

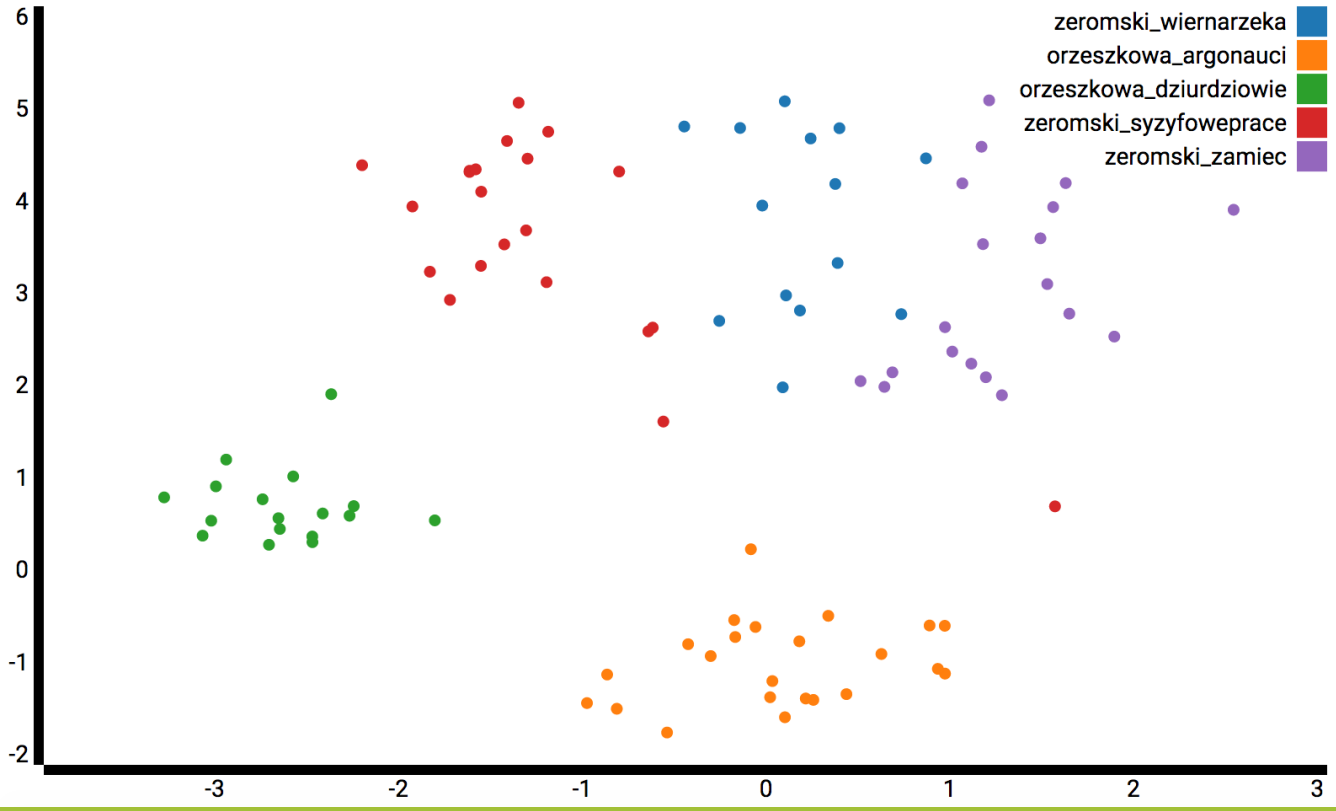


Skalowanie wielowymiarowe

METODA SKALOWANIA TSNE

PERPLEXITY 40

Przelicz



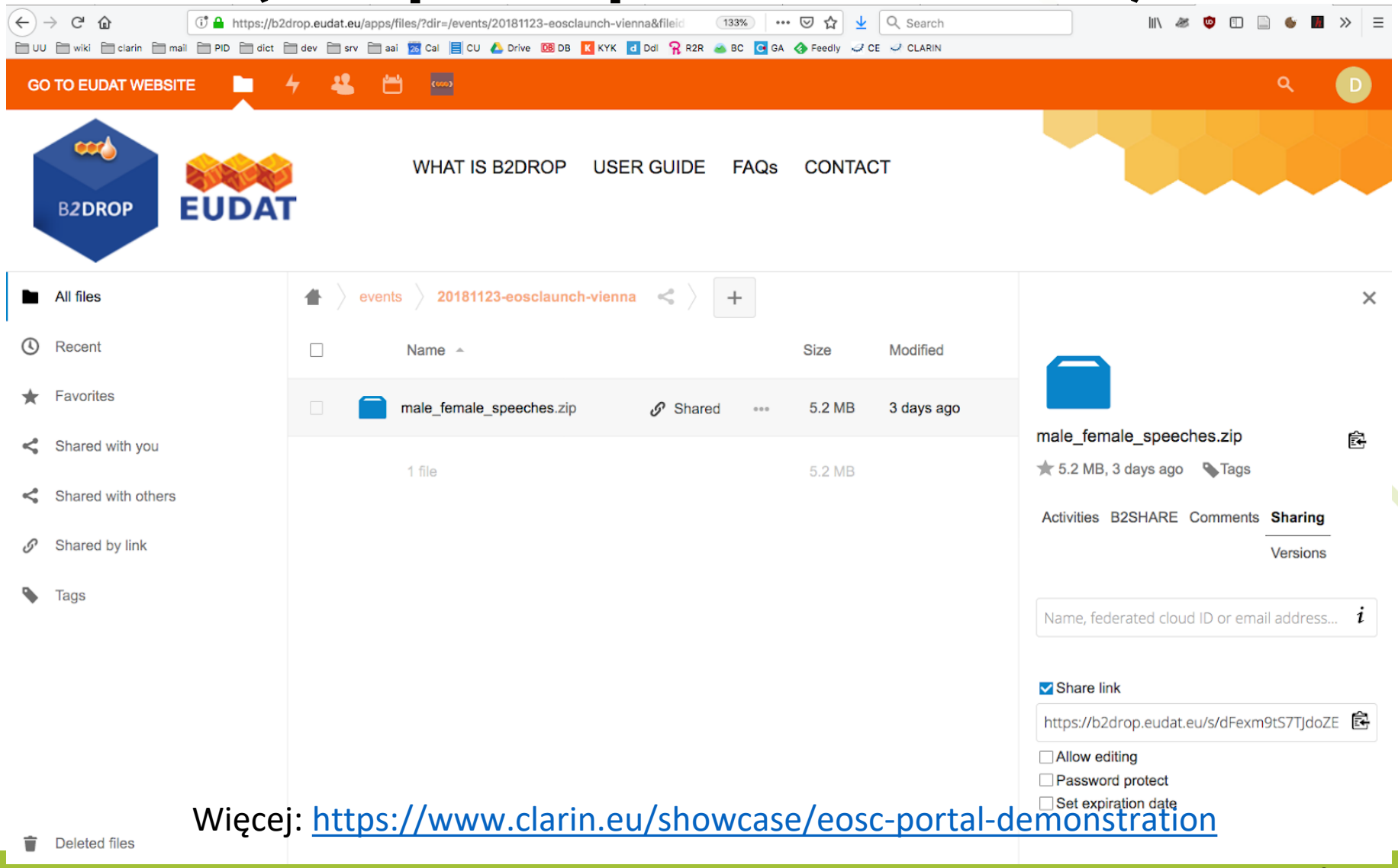
ETYKIETY

PODZIAŁ NA GRUPY

ostatni poziom

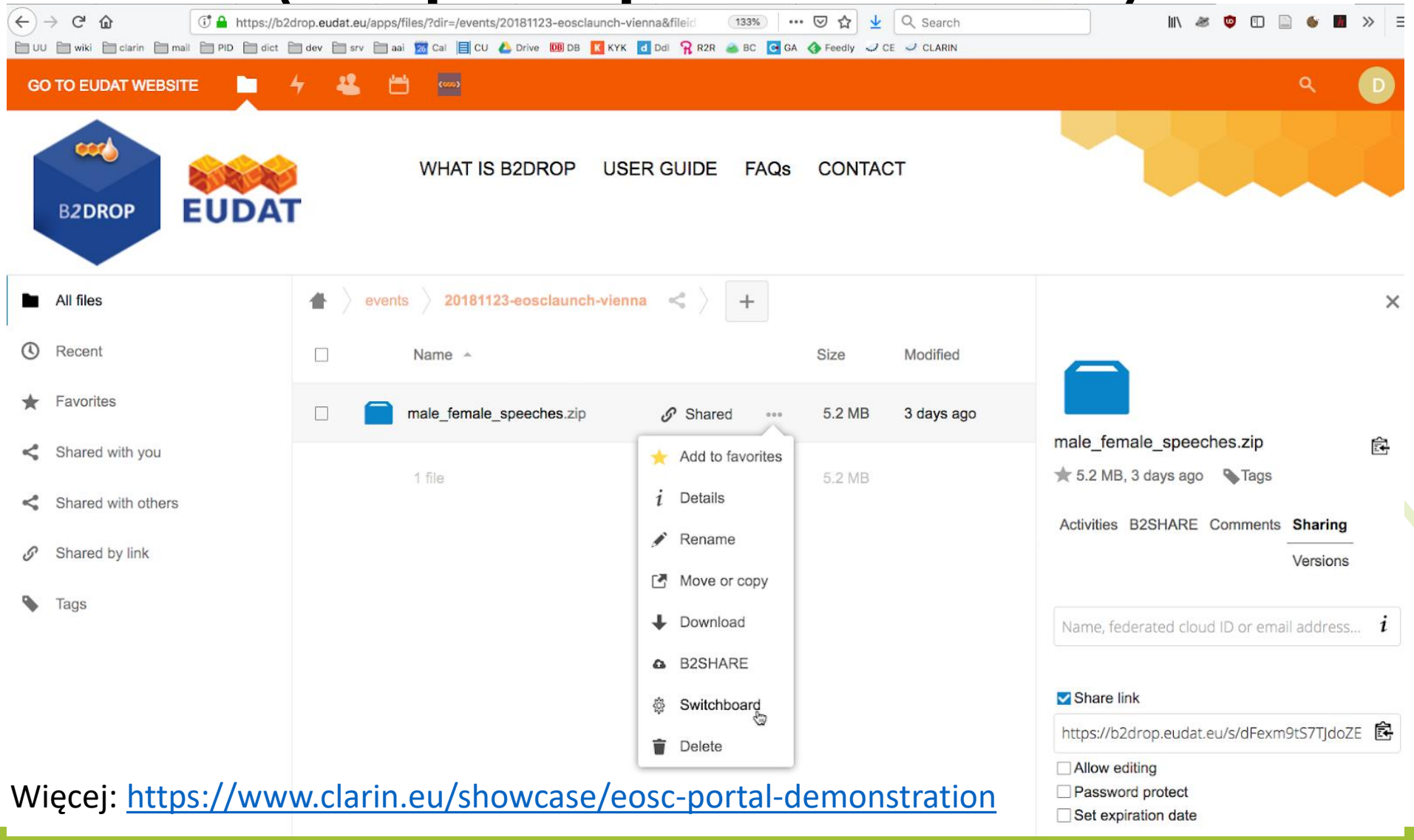
ETYKIETY GRUPY

CLARIN-PL w SSHOC (SSH Open Cloud) - EOSC (European Open Science Cloud)



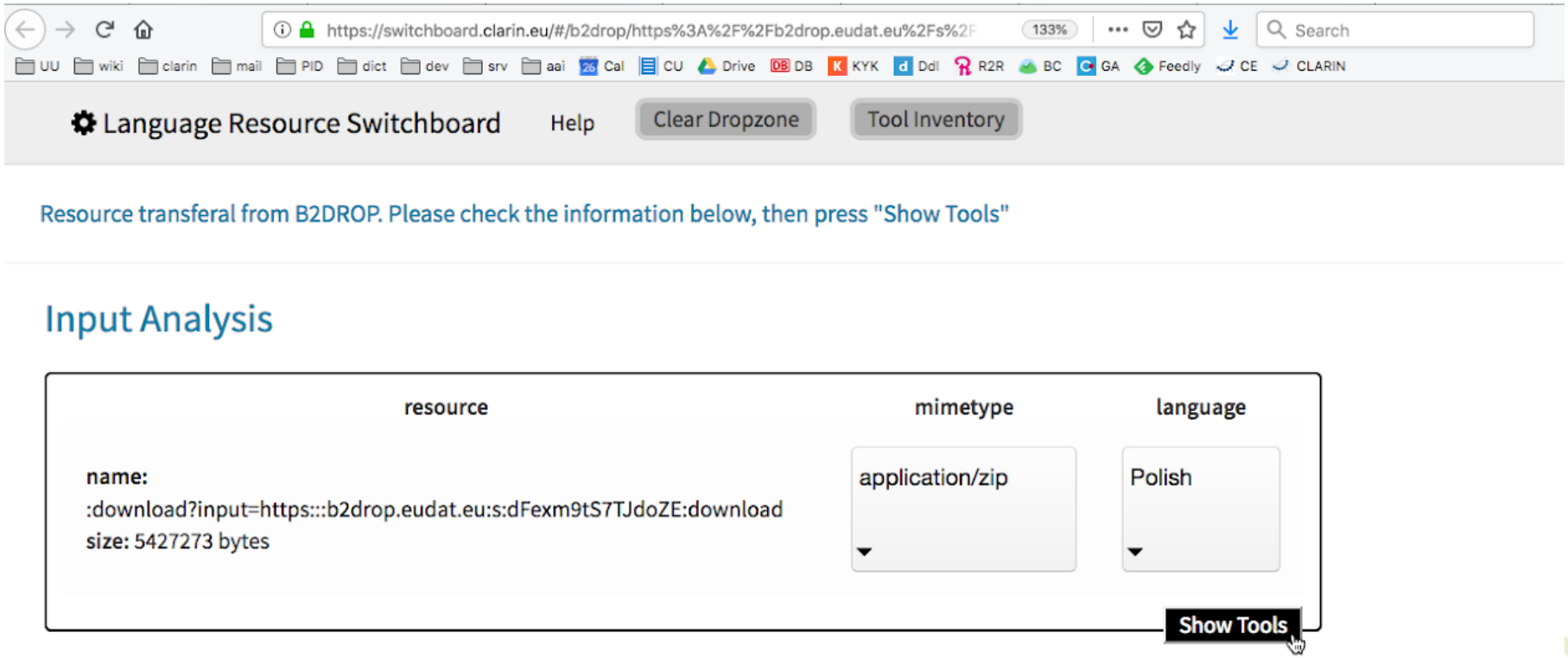
Więcej: https://www.clarin.eu/showcase/eosc-portal-demonstration

CLARIN-PL w SSHOC (SSH Open Cloud) - EOSC (European Open Science Cloud)



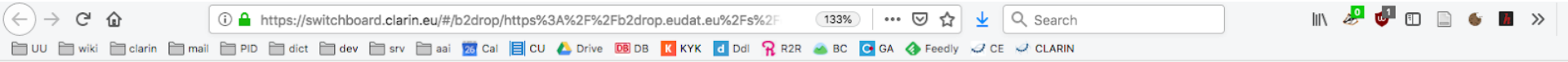
Więcej: https://www.clarin.eu/showcase/eosc-portal-demonstration

CLARIN Switchboard i EOSC



https://switchboard.clarin.eu/

EOSC → Switchboard → WebSty (CLARIN-PL)



Tools

Only Tools | Both Tools & Web Services | Only Web Services | Sort by Task | Order Alphabetically

Stylometry

WEBSTY

Similarity and clustering of texts in Polish. The tools used include: Morfeusz 2 with SGJP dictionary (for morphological analysis), wcrft2 (for tagging), Liner2 (for named entities recognition), Fextor (for extraction of feaures from texts); Cluto (for clustering), result visualisation: D3.js, D3-tip.

http://ws.clarin-pl.eu

no

application/octet-stream

Click to start tool

Wrocław, Poland

tomasz.walkowiak@pwr.edu.pl

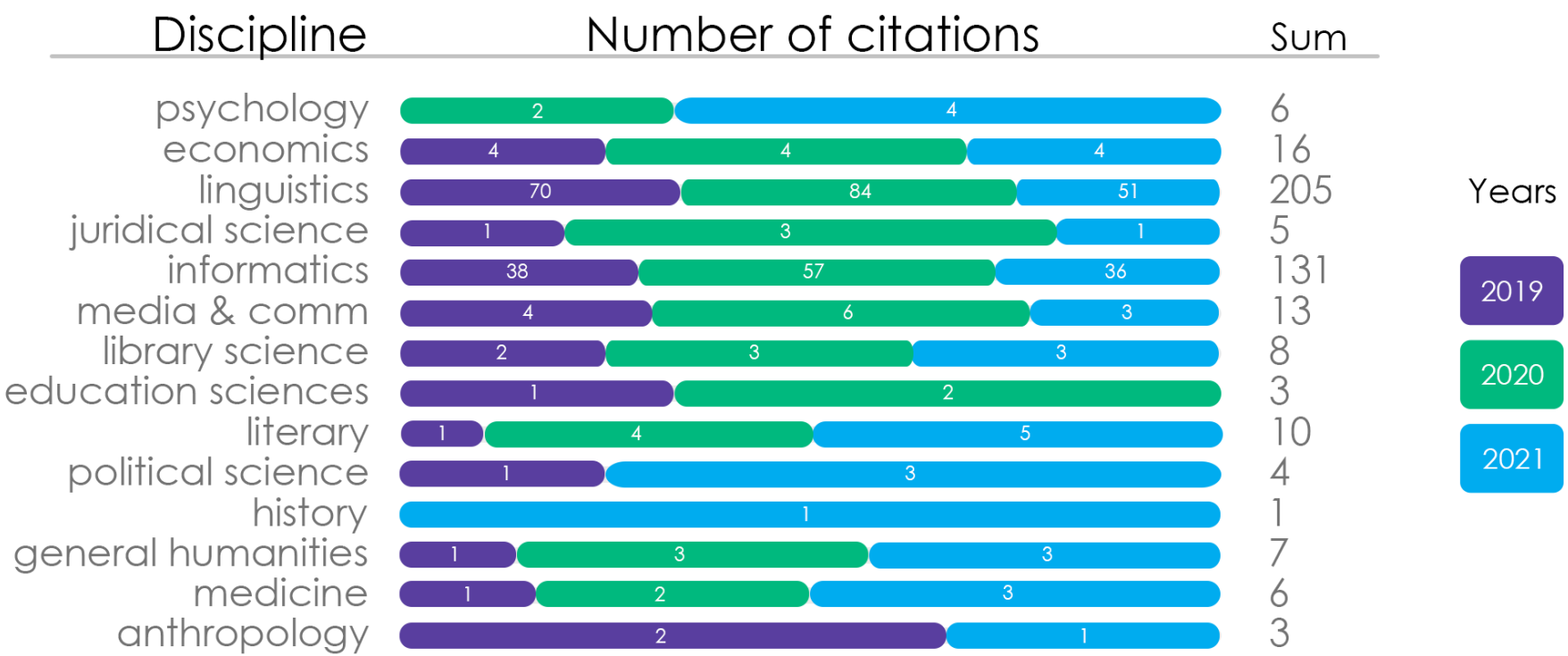
https://switchboard.clarin.eu/

Użytkownicy CLARIN-PL (VI 2018 – VI 2021)

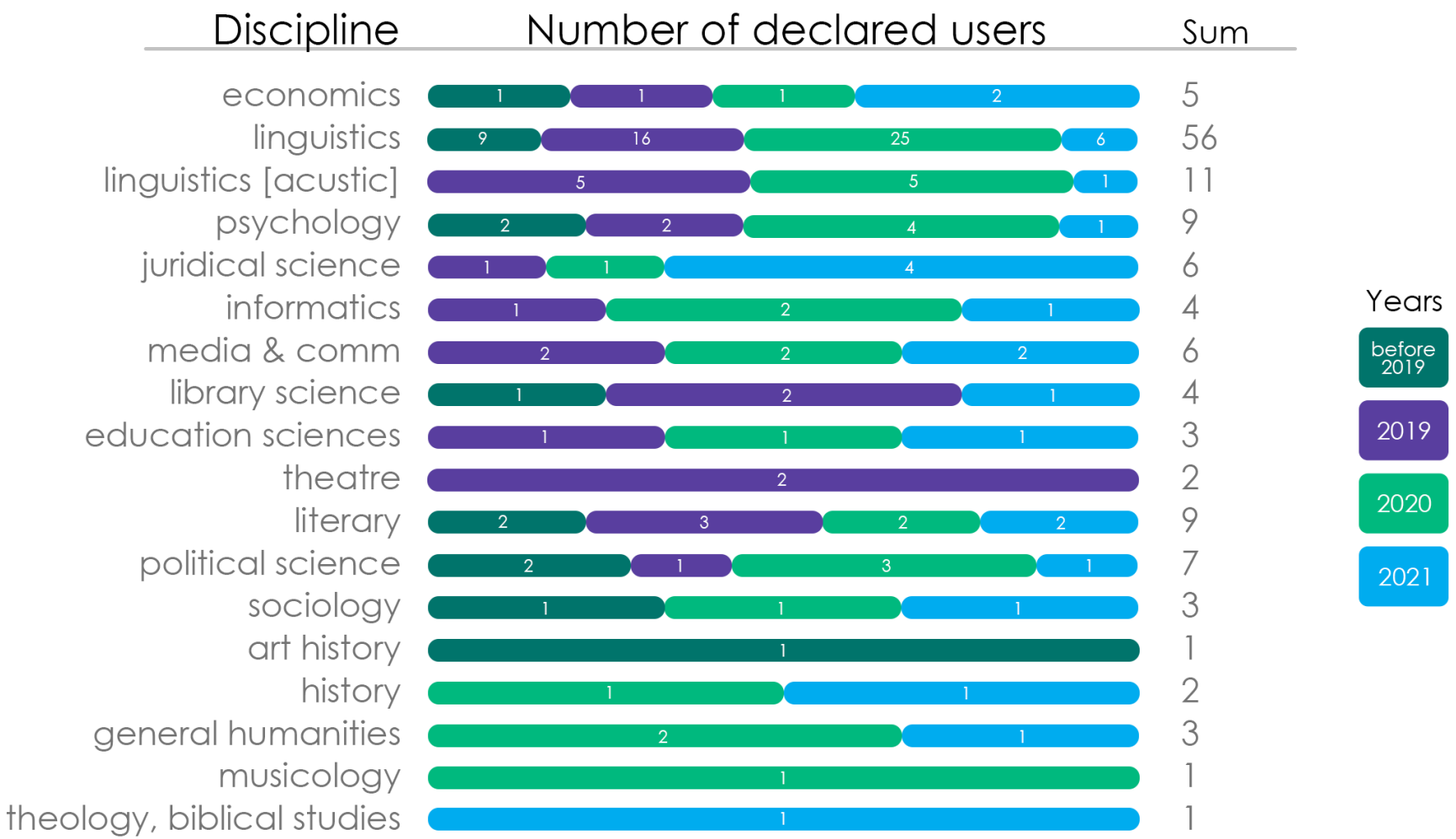
- Znani użytkownicy CLARIN-PL (zasoby, narzędzia i aplikacje):
  - **łącznie 199 zespołów i indywidualnych badaczy**,
  - działania:
    - od używania zasobów i narzędzi, aż do zaangażowania się zespołu CLARIN-PL (**bezpłatnie**) w prace nad rozszerzaniem zasobów, adaptowaniem narzędzi i budowaniem nowych wersji aplikacji dla użytkowników.
- Spontaniczni użytkownicy: 442
  - rozpoznani na podstawie cytowań i wzmianek w sieci.
- Użytkownicy z dziedziny edukacji: 94
- Statystyki użycia:
  - 6 871 828 zadań przetwarzania na ws.clarin-pl.eu
  - 25 195 294 przetworzonych dokumentów,
  - o **łącznej objętości: 405 736 [QV = QuoVadises]**

1 [QV] = ilość tekstu w „Quo Vadis” Henryka Sienkiewicza

Użytkownicy CLARIN-PL – cytowania



Użytkownicy CLARIN-PL



CLARIN-PL-Biz: CLARIN-PL dla sztucznej inteligencji oraz prac B+R

- „CLARIN-PL-Biz (2020–2023), POIR 4.2, Priorytet IV, Działanie 4.2: Rozwój nowoczesnej infrastruktury ”
- Budżet: **136,1 mln. PLN**, w tym dofinansowanie 105 mln. PLN i **19,8 mln. PLN wkładu od Partnerów Gospodarczych**
- Konsorcjum:
  - Naukowi Partnerzy: PWr. (lider), IS PAN, IPI PAN, UWr., UŁ
  - Gospodarczy Partnerzy: Netguru S.A., SentiOne Sp. Z o.o., SmartNet Research & Solutions Sp. z o.o., KODA z o.o., Polska Agencja Prasowa S.A., Equilibrium Solutions Sp. z o.o., Q&Q Sp. z o.o., VOICELAB.AI sp. z o.o., Literacka Sp. z o.o., INNECT FEE COMPASS sp. z o.o., W3A Sp. z o.o., Techmo sp. z o.o., Damovo, Sages Sp. z o. o., Findwise Sp. z o.o., MakoLab SA, Pragmatists Sp. z o.o. Sp.k., Silver Bullet Solutions Sp. z o. o., Technicenter Sp. z o.o., Press-Service Monitoring Mediów Sp. z o.o., STERMEDIA Sp. z o.o., Funmedia sp. z o.o.
- Wspierający Partnerzy Gospodarczy:
  - EIP S.A. , Event Registry, Slovenska Tiskovna Agencija, Intel Technology Poland Sp. z o.o., IQS Sp. z o.o., Carrot Search s.c., Lingventa Sp. z o.o., Polska Press Sp. z o.o.

CLARIN-PL-Biz

- Hybrydowy model działania (wymóg POIR 4.2)
  - [60%] otwarta warstwa podstawowej i naukowej infrastruktury,
  - [40%] płatne, wyspecjalizowane usługi rozwijane wg wymogów przemysłu.
- **Centrum technologiczne** – duże repozytorium i klaster obliczeniowy dla PJN (NLP) i SI (AI):
  - tekst, mowa, zasoby multimodalne i połączonych danych numeryczno-symbolicznych,
  - obliczenia ukierunkowane na uczenie maszynowe, w tym głębokie (ang. Deep Learning).
- **Zbiory danych na potrzeby uczenia maszynowego:**
  - korpusy mowy i dialogowe,
  - anotowane korpusy pytań-odpowiedzi (QA), dla wydobywania informacji (*Information Extraction*) i analizy emotywnej.
- **Narzędzia językowe** zaadaptowane do potrzeb SI i biznesu
  - rozszerzone i poprawione,
  - różne formy wdrażania i udostępniania.
- **Aplikacje i specjalistyczne usługi przetwarzania danych językowych**

Centrum technologiczne CLARIN-PL-Biz

- **Zasób obliczeniowy** – min. 3PFLOPS:
  - węzły obliczeniowe z CPU min. 96 rdzeni i 6GB na rdzeń,
  - węzły obliczeniowe – 1 TB RAM i 4xGPU Nvidia H100,
  - węzły obliczeniowe – 1TB RAM i 4xGPU/IPU o min. 80GB RAM,
  - węzły z kartami do przetwarzania (inferencji),
  - sieć międzywęzłowa Infinibad HDR 200GB/s.
- **Przestrzeń dyskowa**:
  - na potrzeby przetwarzania danych – min. 2PB oparte na NVMe,
  - repozytorium danych w dostępie obiektowym i plikowym – min. 5PB.
- Usługi podstawowe:
  - IaaS, PaaS, CaaS, składowania danych, dostęp do klastra obliczeniowego.
- **Rozszerzona architektura konfigurowalnych potoków** przetwarzania z CLARIN-PL:
  - wydajność, autoryzacja, bezpieczeństwo,
  - narzędzia i usługi wyspecjalizowane pod potrzeby biznesu.

CLARIN-PL-Biz – zbiory danych

- Rozszerzone zbiory danych CLARIN-PL
  - w tym KPWr (PWr) - anotacja korpusu zgodna z najnowszymi specyfikacjami (m.in. ujednoznacznienie morfosyntaktyczne, całostki składniowe, jednostki identyfikacyjne, wyrażenia przestrzenne i temporalne, sytuacje, słowa kluczowe), Korpus Dyskursu Parlamentarnego (IPI PAN), Chronopress (UWr)
- Korpus Wieszczów (*wspierany przez CLARIN-PL-Biz*):
  - korpus tekstów czterech autorów (A. Mickiewicz, J. Słowacki, C.K. Norwid, Z. Krasiński),
  - przetworzonych łącznie około 25 tysięcy stron,
  - około 9 tysięcy różnego rodzaju tekstów (od dedykacji i notatek, przez listy, po poematy i prozę artystyczną).
- Korpusy dialogów i struktur dyskursywnych:
  - korpusy mowy (UŁ, PWr i PJATK) – min. 1000 godzin nagrań danych medialnych, wywiadów i rozmów formalnych i swobodnych,
  - DiaBiz (UŁ, PJATK) – ~400 godzin transkrybowanych i nagrań telefonicznych rozmów klient-agent przeprowadzonych na podstawie scenariuszy biznesowych opracowanych dla 9 dziedzin,
  - DiaBiz.Kom (PWr) - podkorpus 1100 dialogów ręcznie anotowany w ramach struktury funkcjonalnej opartej na aktach dialogowych,
  - *Korpus metatekstowy - korpus ręcznie anotowany relacjami dyskursywnymi (retorycznymi, metatekstowymi) zbudowany na bazie Polskiego Korpusu Koreferencyjnego.*

CLARIN-PL-Biz – zbiory danych

- Korpus idiolektów (UŁ)
  - 100 idiolektów, różne kanały komunikacji, tematyka.
- Korpus emocji i wydźwięku emocjonalnego:
  - MultiEmo (PWr) – wydźwięk opinii, 8 000 tekstów, 57 000 zdań, 3 anotatorów, 6 wymiarów, 10 języków,
  - CLARINEmo (PWr) – emocje w opiniach, 1 100 tekstów i 8 600 zdań, 6 anotatorów, 11 wymiarów emocjonalnych
  - MultiAspectEmo (PWr) – polaryzacja aspektów opinii, 1 800 tekstów, 26000 anotacji, 3 anotatorów, 5 języków
  - CLARINAffect (PWr) – subiektywne reakcje afektywne, 3 500 komentarzy, 6 anotatorów, 28 wymiarów.
- Korpusy pytań i odpowiedzi (IPI PAN i PWr)
  - PoQuAD – standardowy zbiór pytań oparty na artykułach z Wikipedii,
  - Complex Questions PL – zbiór pytań złożonych, które wymagają wielokrokowego wnioskowania.
- Korpus wnioskowania tekstowego
  - anotowany relacjami tekstowego wynikania (ang. *Natural Language Inference*) wzorowany na Stanford NLI, z dodatkową warstwą CST np. zawieranie, parafraza, krzyżowanie, sprzeczność itd.

CLARIN-PL-Biz – zbiory danych

- Zasoby leksykalne i powiązane (PWr)
  - Słowosieć (plWordNet) – (wielki) wordnet języka polskiego powiązany z angielskim Princeton WordNet
  - korpusy ujednoznacznione na poziomie leksykalnym (WSD)
    - KPWr 100 – 100 zaanotowanych dokumentów Korpusu Politechniki Wrocławskiej
    - GLEX – korpus glos Słowosieci
    - SPEC – "Pstrokata opaska" A. C. Doyle'a (z cyklu "Przygód Sherlocka Holmesa")
  - VeSNet – sieć semantyczna w formacie SKOS-RDF łącząca Słowosieć i Open Multilingual WordNet z dziedzinowymi tezaurusami otwartych danych połączonych (ang. Linked Open Data) – ponad 30 różnorodnych zasobów słownikowych
- Lepiszcze (http://lepiszcze.ml/)
  - *benchmark* i *leaderboard* dla modeli uczenia maszynowego języka polskiego,
  - biblioteka do testowania modeli (CLARIN-PL/embeddings: Embeddings: State-of-the-art Text Representations for Natural Language Processing tasks, an initial version of library focus on the Polish Language (github.com)),
  - zbiór modeli językowych zbudowanych w ramach CLARIN-PL.

CLARIN-PL-Biz: narzędzia językowe

- Poziomu wyrazowego
  - Punctuator (PWr) – przywrócenie interpunkcji,
  - Konfigurowalna biblioteka do czyszczenia korpusów (PWr)
  - Anonimizator (PWr) – inteligentne narzędzie do anonimizacji tekstów
- Analiza morfosyntaktyczna i składniowa:
  - CMCTagger (PWr) – tager morfosyntaktyczny dla tekstów z mediów społecznościowych,
  - Całostkownik Neuronalny (PWr) – (ang. *chunker*) narzędzie do płytkiej analizy składniowej,
  - Combo (IPI PAN) – złożone narzędzie do tokenizacji, tagowania i parsowania zależnościowego,
  - Lematyzator bezkontekstowy (PWr) – sprowadzanie wyrazów do formy podstawowej, bez potrzeby reprezentacji w kontekście.
- Analiza semantyczna poziomu leksykalnego:
  - WSD (PWr) – ujednoznacznianie słów w tekście względem Słowosieci,
  - Elinker (PWr) – narzędzie do rzutowania tekstu na sieć semantyczną.

CLARIN-PL-Biz: narzędzia językowe

- Wydobywanie informacji
  - Easymatcher (PWr) – elastyczne rozpoznawanie wystąpień haseł słownikowych w tekście
  - Liner2 i PolDeepNer2 (PWr) – programy do rozpoznawania i klasyfikacji wystąpień nazw własnych,
  - Narzędzia do rozpoznawania i klasyfikacji wyrażeń przestrzennych (PWr),
  - narzędzia do rozpoznawania, klasyfikacji oraz normalizacji wyrażeń temporalnych (IPI PAN),
  - adaptowalne narzędzie do rozpoznawania relacji semantycznych (PWr),
  - *Narzędzie do rozpoznawania opisów sytuacji (PWr),*
  - *Wykrywanie koreferencji (IPI PAN).*
- Analiza wydźwięku i emocji
  - MultiEmo – rozpoznawanie wydźwięku tekstów, zdań lub akapitów dla ponad 100 języków (PWr),
  - HateSpeech – spersonalizowane rozpoznawanie agresji dla ponad 100 języków (PWr),
  - MultiAspectEmo* – rozpoznawanie polaryzacji wydźwięku aspektów opinii dla 6 języków (PWr+IPI),
  - CLARINAffect* – spersonalizowana predykcja reakcji afektywnych wzbudzanych przez tekst (PWr).

CLARIN-PL-Biz: narzędzia językowe

- Analiza dialogów:
  - *adaptowalny, dziedzinowy system rozpoznawania mowy*,
  - adaptowalne narzędzie do wykrywania obiektów, atrybutów i wartości w oparciu o minimalne anotowane dane,
  - narzędzie do rozpoznawania aktów, funkcji i struktur dialogowych,
  - *rozpoznawanie ram semantycznych w danych dialogowych (w oparciu o FrameNet),*
  - rozpoznawanie i klasyfikacja pytań w ramach dialogów.
- Narzędzia do wykrywania relacji pomiędzy fragmentami tekstów:
  - wnioskowanie w języku naturalnym,
  - rozpoznawanie relacji tekstowego wynikania (NLI) oraz międzytekstowych relacji dyskursu (CST).

CLARIN-PL-Biz: aplikacje i usługi analizy

- Konfigurowalny, współbieżny, wydajny potok przetwarzania (PWr)
- Aplikacje do budowy i przeszukiwania korpusów:
  - Inforex (PWr) – anotacja korpusów, zarządzanie zespołami anotatorów,
  - Korpusomat (IPI PAN) – budowa i przeszukiwanie korpusów,
  - Chronocorpus – przeszukiwarka (PWr i UWr)
  - Spokes, Paralella (UŁ) – budowa i przeszukiwanie korpusów: mowy i równoległych
  - Platforma Mowy (PJATK) – aplikacja do zbierania korpusów mowy,
  - rozszerzenie Doccano (PWr) m.in. o wsparcie procesów opartych na uczeniu aktywnym (ang. Active Learning):
    - konstrukcja spersonalizowanych systemów do rozpoznawania emocji i stanów afektywnych w reakcji na teksty,
    - budowa zasobów do konstrukcji systemów dialogowych (aktywne tworzenie zbiorów uczących).
- Wydobywanie terminów i wielowyrazowych jednostek leksykalnych: TermoPL (IPI PAN), MeWeX (PWr).
- Analiza statystyczna i tematyczna, porównywanie korpusów tekstów:
  - LEM (PWr), Topic (PWr), CompCorp (PWr).

CLARIN-PL-Biz: aplikacje i usługi analizy

- Analizy stylometrycznej:
  - WebSty (PWr), WebStyML (PWr) – wielojęzyczny,
  - analiza krótkich tekstów na poziomie idiolektu.
- Klasyfikacja semantyczna i grupowanie:
  - zarówno tekstów dłuższych, jak i krótkich,
  - adaptowalne dziedzinowo.
- System do wyszukiwania semantycznego i odpowiadania na pytania:
  - metody nienadzorowane dostrajające do dziedziny,
  - metody nadzorowane, jak i korzystanie z wiedzy eksperckiej i ontologii,
  - generowanie odpowiedzi, jak i ekstrakcja odpowiedzi z tekstu.

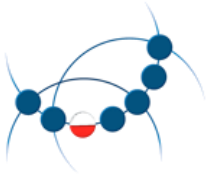
Do zrobienia we współpracy z CLARIN-PL

- Zadania z praktyki (również pod względem skali) i dane z praktycznych problemów i zastosowań.
- Odpowiednia automatyzacja budowy zasobów językowych
  - anotowane (etykietowane) zbiory danych językowych.
- Poprawa naukowej ewaluacji rozwiązań
  - problem ‚standardowych' zbiorów wzorcowych.
- Poprawa reprodukowalności.
- Reprezentacja wypowiedzi językowych:
  - łącząca aspekty dystrybucyjne i strukturalne,
  - w kompleksowy sposób uwzględniająca kontekst interpretacji.

Dziękuję za uwagę!

www.clarin.eu
clarin-pl.eu
or
clarin@clarin.eu
clarin-pl@pwr.edu.pl
maciej.piasecki@pwr.edu.pl

CLARIN-PL
Common Language Resources and Technology Infrastructure



http://clarin-pl.eu/
http://clarin.biz

http://clarin.eu/