Ludzie, projekty, narzędzia analizy lingwistycznej

w Zespole Inżynierii Lingwistycznej IPI PAN

Maciej Ogrodniczuk

INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. Jana Kazimierza 5, 01-248 Warszawa

Rozszerzone seminarium „Przetwarzanie języka naturalnego”
22 lutego 2016

Plan prezentacji

1. Przegląd tematów badawczych i osób zajmujących się nimi w ZIL.
2. Przegląd aktualnie prowadzonych projektów.
3. Przegląd narzędzi i zasobów ZIL (oprócz tych, które będą szerzej prezentowane w dalszej części seminarium).
4. Gdzie szukać dalszych informacji o narzędziach do komputerowego przetwarzania polszczyzny?

ZIL — pracownicy, tematy badawcze i narzędzia

Pracownicy **Zespołu Inżynierii Lingwistycznej**:

- **dr Anna Andrzejczuk** —
  słownik walencyjny Walenty
- **mgr Tomasz Bartosiak** —
  Slowal — system do przeglądania i tworzenia słownika walencyjnego Walenty (warstwa semantyczna)
- **dr hab. Elżbieta Hajnicz** —
  słownik walencyjny Walenty
- **dr inż. Łukasz Kobyliński** —
  multitager PoliTa
- **mgr Katarzyna Krasnowska-Kieraś** — współpraca analizatora Morfeusz 2 z parserem POLFIE (LFG)
- **dr Małgorzata Marciniak** —
  ekstrakcja terminologii, przetwarzanie tekstów dziedzinowych

ZIL — pracownicy, tematy badawcze i narzędzia

Pracownicy **Zespołu Inżynierii Lingwistycznej**:

- **dr hab. Agnieszka Mykowiecka** —
  ekstrakcja terminologii, przetwarzanie tekstów dziedzinowych
- **mgr inż. Bartłomiej Nitoń** —
  Slowal — system do przeglądania i tworzenia słownika walencyjnego Walenty (warstwa semantyczna),
  Jasnopis do badania zrozumiałości tekstu
- **dr Maciej Ogrodniczuk** —
  referencja, dyskurs
- **dr Agnieszka Patejuk** —
  analiza składniowa (LFG), Walenty (warstwa składniowa)
- **dr hab. Adam Przepiórkowski** —
  analiza składniowa i semantyczna, NKJP

ZIL — pracownicy, tematy badawcze i narzędzia

Pracownicy **Zespołu Inżynierii Lingwistycznej**:

- **dr Piotr Rychlik** —
  ekstraktor terminologii TermoPL
- **dr inż. Aleksander Wawer** —
  analiza sentymentu
- **dr Marcin Woliński** —
  analiza morfologiczna (Morfeusz 2, SGJP), analiza składniowa (Świgra)
- **dr Alina Wróblewska** (obecnie urlop macierzyński) —
  parsowanie zależnościowe
- **mgr Bartosz Zaborowski** —
  implementacja analizatora powierzchniowoskładniowego Spejd, wyszukiwarka korpusowa Poliqarp 2
- **współpracownicy projektowi** (do kilkudziesięciu).

Projekty realizowane w ZIL

Projekty kierowane przez członków ZIL:

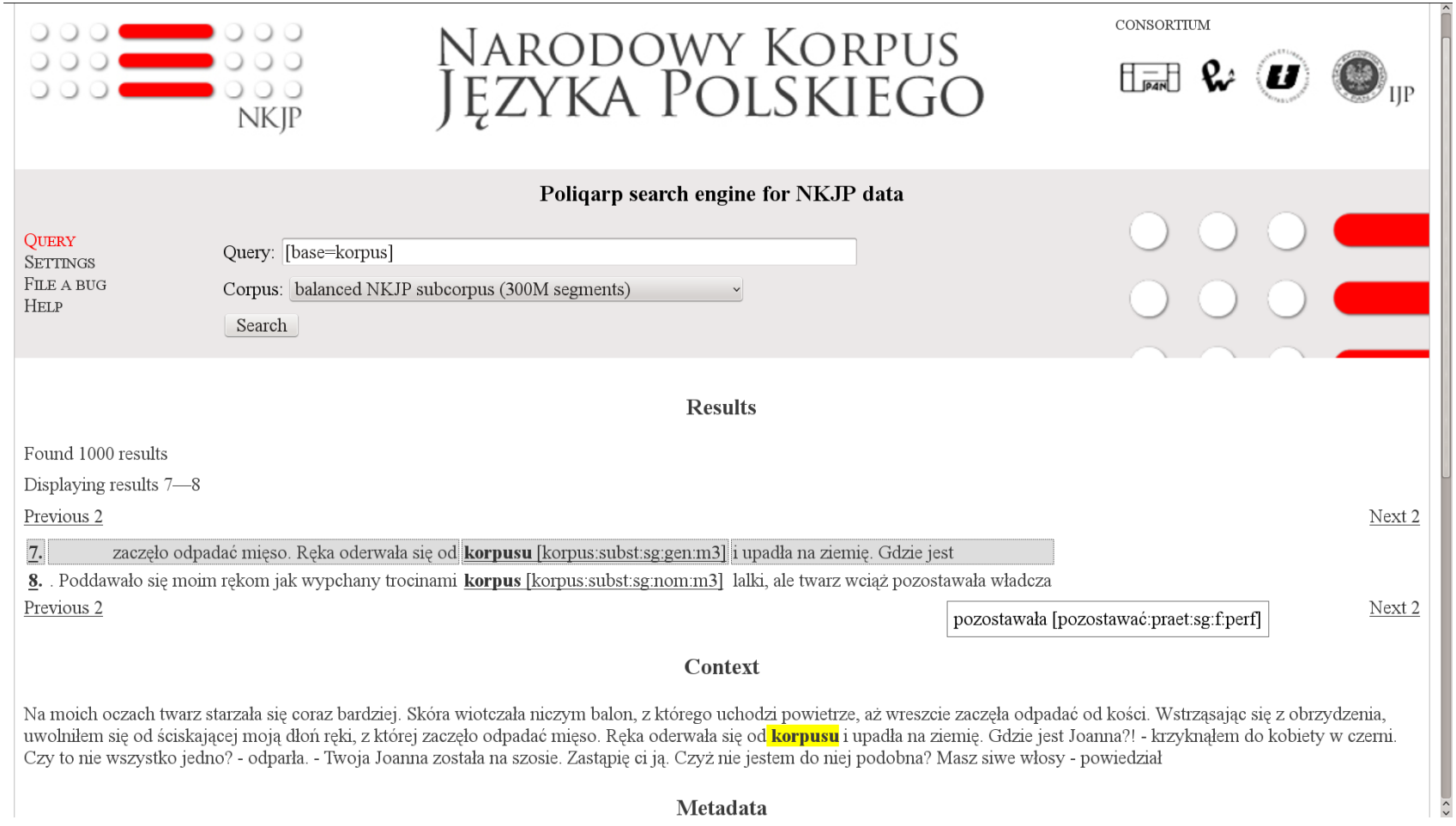
- **CLARIN-PL** (oraz dodatkowy projekt wspierający), MNiSW – budowa infrastruktury, Adam Przepiórkowski
- **PARSEME**, Akcja COST (Komisja Europejska), IPI PAN: Grant Holder, Adam Przepiórkowski: vice chair
- **Wykorzystanie metod kompozycyjnej semantyki dystrybucyjnej do identyfikacji i rozróżniania znaczeń w języku polskim**, NCN OPUS, Agnieszka Mykowiecka
- **COTHEC – Ujednolicona teoria koreferencji w języku polskim i jej korpusowa weryfikacja**, NCN OPUS, Maciej Ogrodniczuk
- **Chronofleks – Model formalny diachronicznego opisu fleksji polskiej i jego komputerowa implementacja**, NCN OPUS, Marcin Woliński
- **Scwad – Kompozycyjno-dystrybucyjne modelowanie semantyki języka polskiego**, NCN SONATA, Alina Wróblewska
- **OPTA – Automatyczne metody rozpoznawania przedmiotów i wyrażeń opinii w języku polskim**, NCN PRELUDIUM, Alex Wawer

Projekty realizowane w ZIL

Projekty z udziałem członków ZIL:

- **DARIAH-PL** – Infrastruktura badawcza dla humanistyki cyfrowej, Maciej Ogrodniczuk
- **Jasnopis** – Badanie zrozumiałości polskich tekstów użytkowych, NCN OPUS; Bartłomiej Nitoń
- **KORBA** – Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do roku 1772), projekt NPRH; Marcin Woliński, Maciej Ogrodniczuk
- **SYNAMET** – Mikrokorpus metafor synestezyjnych, NCN OPUS; Maciej Ogrodniczuk
- **TextLink**, Akcja COST, Maciej Ogrodniczuk

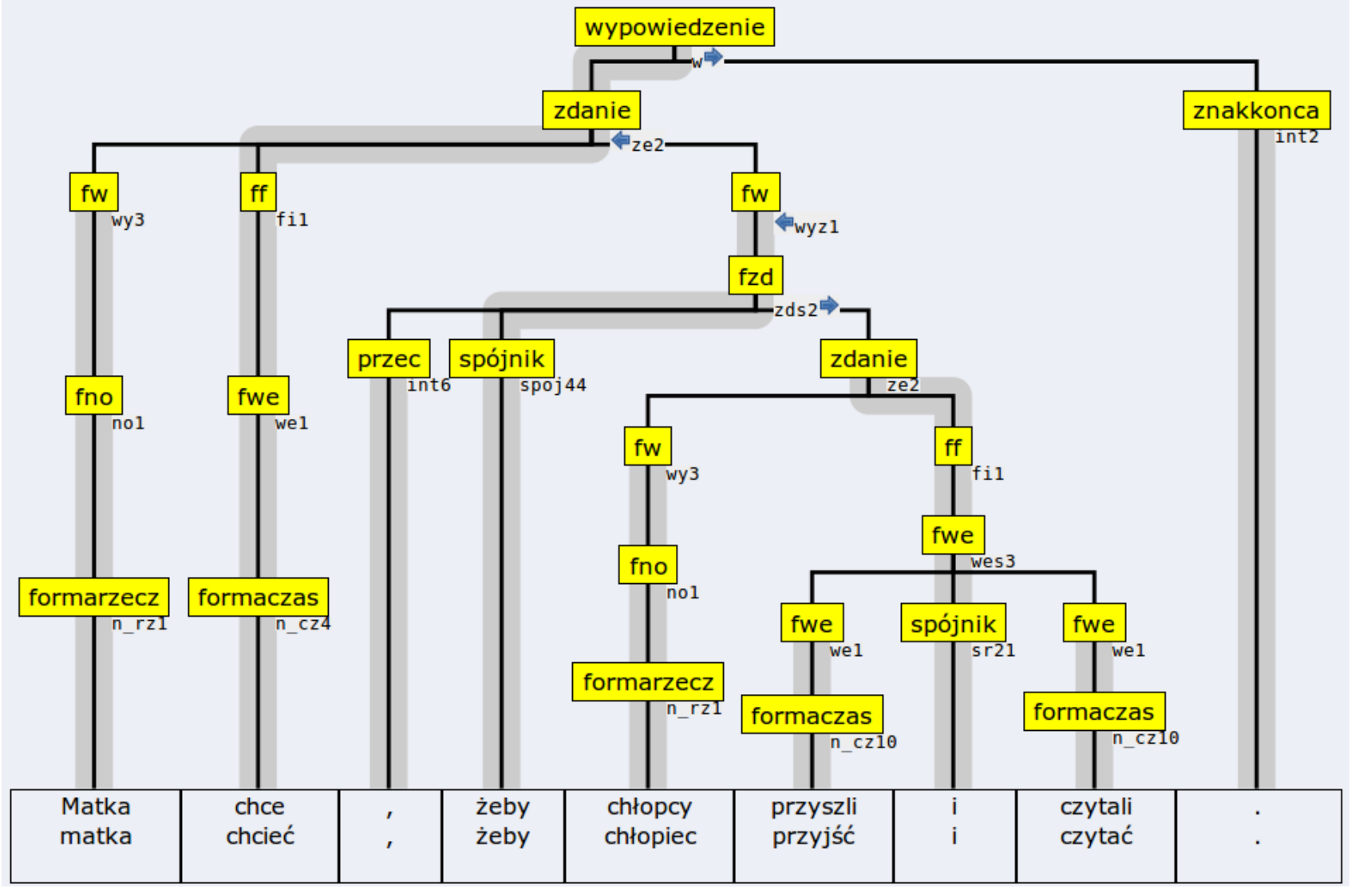
Zasoby ZIL: Narodowy Korpus Języka Polskiego



Narzędzia ZIL: Świgra 2

- analizator składniowy języka polskiego w formalizmie DCG,
- generuje drzewa składnikowe,
- gramatyka wywodząca się z *Gramatyki formalnej języka polskiego* Marka Świdzińskiego (1992),
- podstawa treebanku Składnica,
- demo: `http://swigra.nlp.ipipan.waw.pl/`

Narzędzia ZIL: Świgra 2



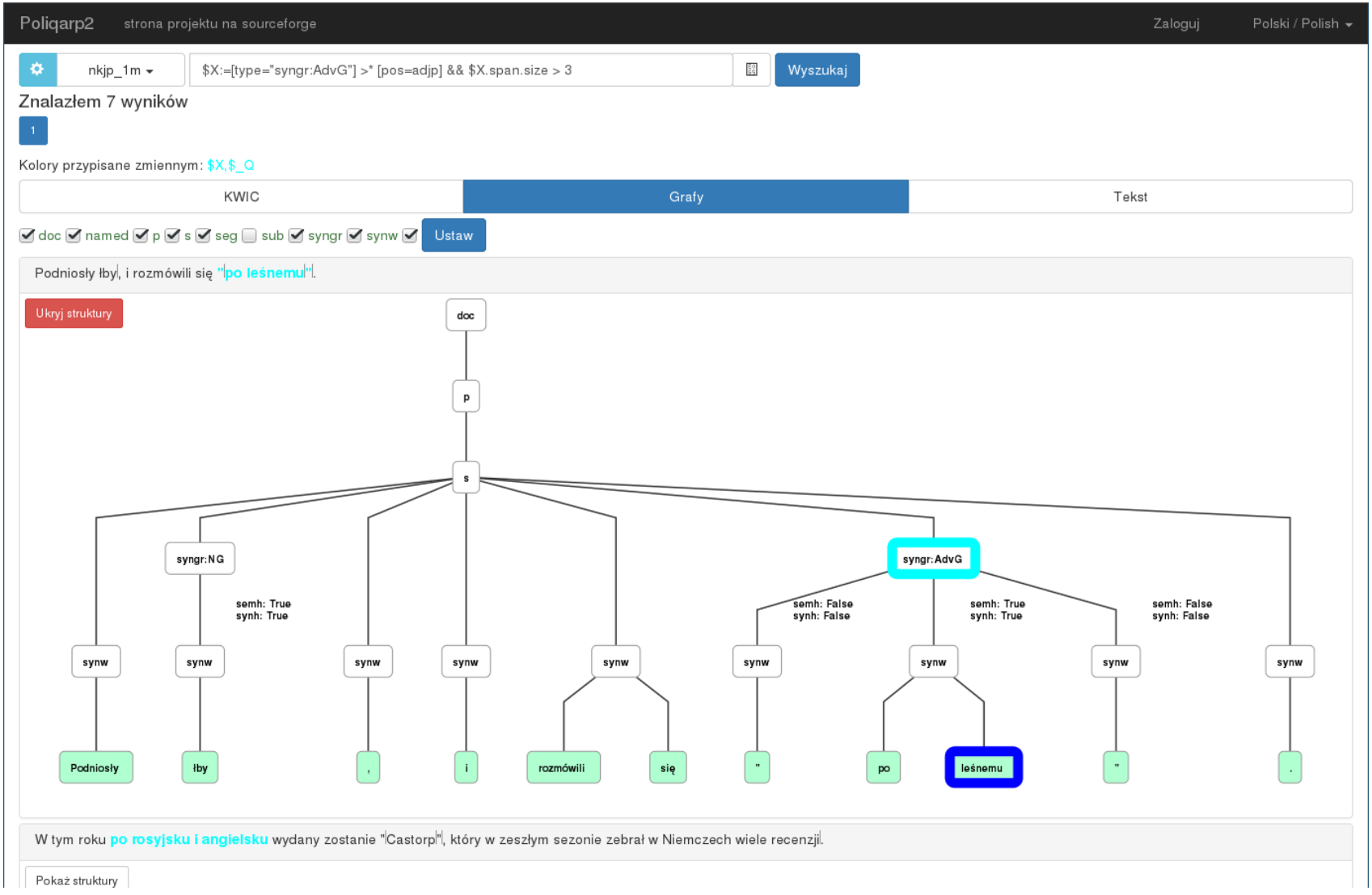
Narzędzia ZIL: Poliqarp 2

Poliqarp 2 to wydajna wyszukiwarka korpusowa tworzona z myślą o przeszukiwaniu dużych korpusów anotowanych na wielu różnych poziomach, np. NKJP. Unikalne cechy narzędzia:

- elastyczny model danych: struktura dokumentu jako graf, anotacje powiązane z węzłami i krawędziami – łatwo modelowalna segmentacja, morfoskładnia, drzewa, struktury LFG i podobne,
- duża siła wyrazu języka zapytań (wyrażenia regularne na węzłach grafu, także wzdłuż tekstu; kwantyfikowane zmienne),
- postprocessing podobny do SQL,
- najważniejsze: efektywne przszukiwanie dużych korpusów bez utraty powyższych możliwości.

Wyszukiwarka jest w tej chwili w ostatniej fazie budowy i w najbliższych miesiącach zostanie udostępniona publicznie.

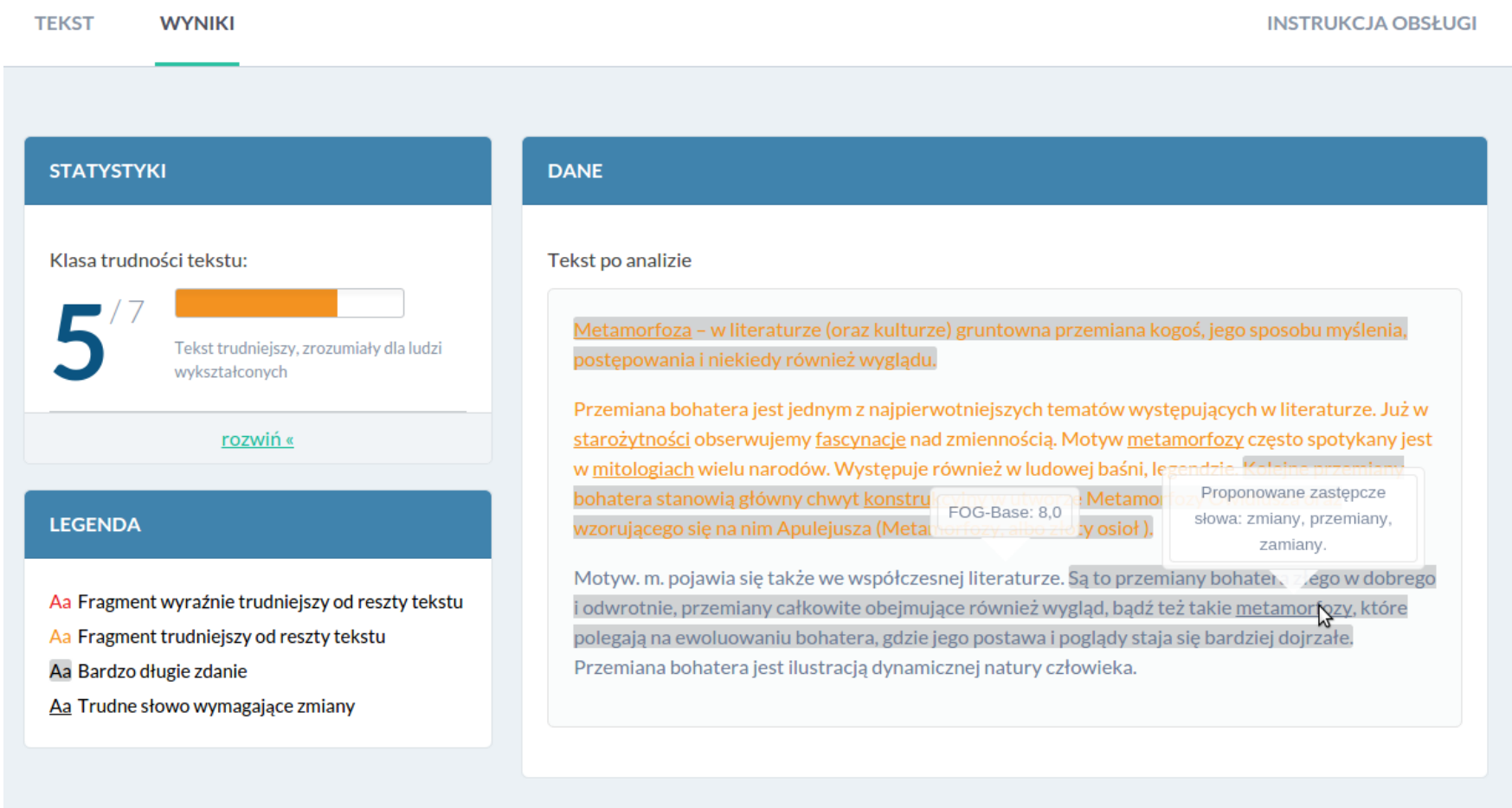
Narzędzia ZIL: Poliqarp 2



Narzędzia ZIL: Jasnopis

- Jasnopis jest programem komputerowym, z którego można bezpłatnie korzystać online pod adresem: `www.jasnopis.pl/aplikacja`.
- Za pomocą Jasnopisu można mierzyć trudność (zrozumiałość) tekstu polskiego; trudność jest wyrażona liczbowo w skali 1–7.
- Jasnopis podaje oceny wyliczane innymi, wcześniej znanymi metodami (indeks FOG, wzór Pisarka).
- Jasnopis podaje także „statystyki tekstu", które mogą się przydać bardziej zaawansowanym użytkownikom.
- Jasnopis zaznacza niektóre trudniejsze fragmenty tekstu (akapity, zdania, wyrazy) i proponuje dla nich poprawki.

Narzędzia ZIL: Jasnopis



Narzędzia ZIL: Multiserwis

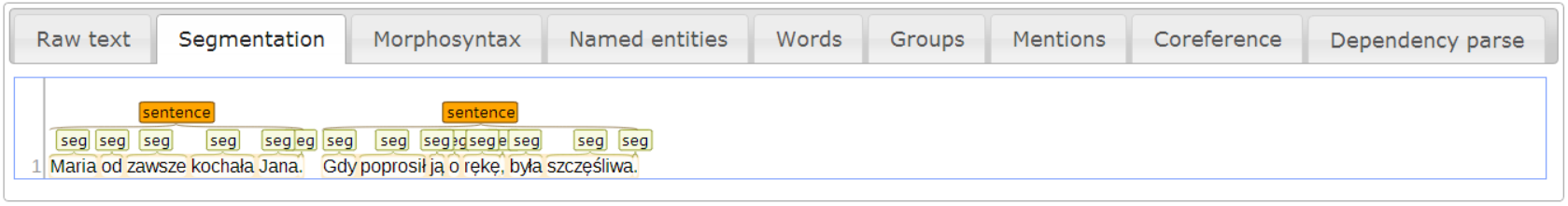
Multiserwis to serwis webowy udostępniający narzędzia analizy lingwistycznej, umożliwiający łączenie ich w łańcuchy i wizualizację wyników.

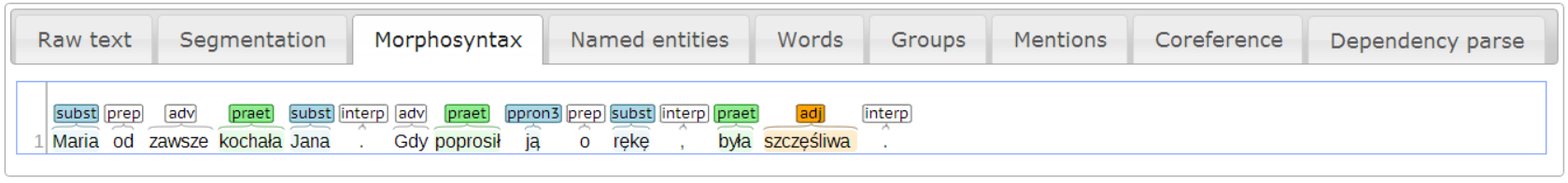
Aktualnie dostępne narzędzia:

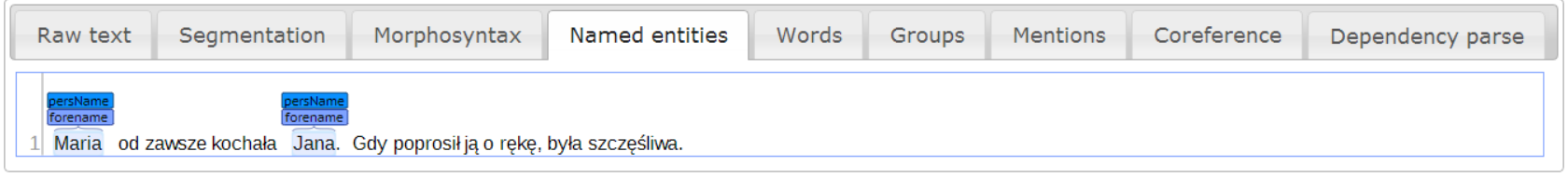
- tagery: Pantera, Concraft, WCRFT, WMBT, PoliTa
- parsery: Spejd, parser zależnościowy
- wykrywacz nazw własnych Nerf
- narzędzie do analizy sentymentu Sentipejd
- narzędzia koreferencyjne: MentionDetector, Ruler, Bartek
- narzędzia do streszczania tekstów: OpenTextSummarizer, Lakon, streszczarka Joanny Świetlickiej

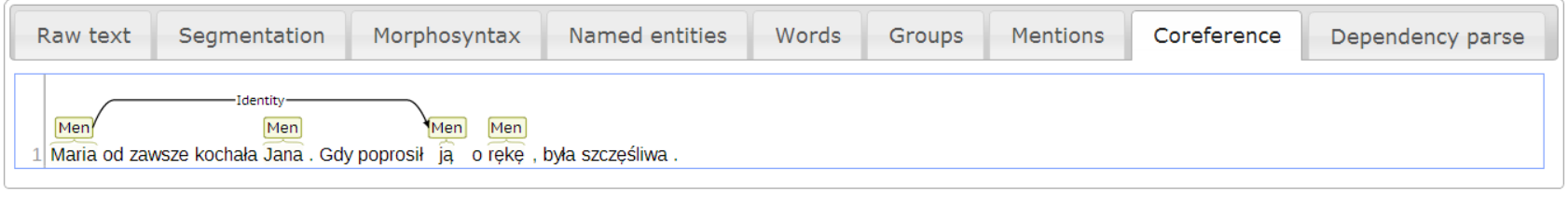
`http://multiservice.nlp.ipipan.waw.pl/`

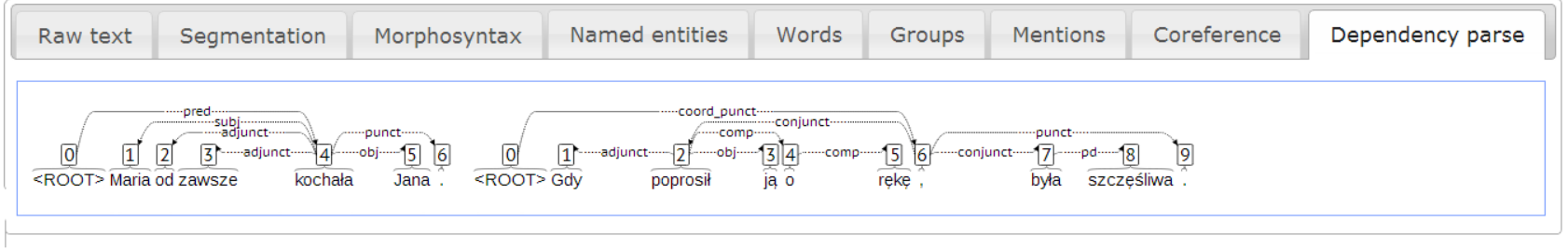
Narzędzia ZIL: Multiserwis











Współpraca

Współpraca międzynarodowa:

- **liczna** w ramach projektów PARSEME i TextLink
- **liczna** w ramach nieformalnego projektu PARGRAM
- Uniwersytet Karola w Pradze, Czechy
- Université François Rabelais Tours, Francja
- Institut für Deutsche Sprache, Mannheim, Niemcy
- Uniwersytet w Bergen, Norwegia

Najbliższe konferencje organizowane przez ZIL:

- ELRC Polska: warsztat European Language Resource Coordination: `http://lr-coordination.eu/pl/poland` (9 marca)
- HeadLex16: Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar: `http://headlex16.ipipan.waw.pl/` (24–29 lipca)

Inne

Seminarium ZIL (`http://zil.ipipan.waw.pl/seminarium`):

- 16 spotkań w 2015 r.
- średnio 25 uczestników

Journal of Language Modelling (`http://jlm.ipipan.waw.pl/`)

Strona CLIP (`http://clip.ipipan.waw.pl/`)