

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

- 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?

Pracownicy Zespołu Inżynierii Lingwistycznej:

- **dr Anna Andrzejczuk** —
słownik walencyjny Walenty
- **mgr Tomasz Bartosiak** —
Słowa! — 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

Pracownicy Zespołu Inżynierii Lingwistycznej:

- **dr hab. Agnieszka Mykowiecka** —
ekstrakcja terminologii, przetwarzanie tekstów dziedzinowych
- **mgr inż. Bartłomiej Nitoń** —
Słowa! — 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

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 (Świga)
- **dr Alina Wróblewska** (obecnie urlop macierzyński) —
parsowanie zależnościowe
- **mgr Bartosz Zaborowski** —
implementacja analizatora powierzchniowskładniowego Spejd, wyszukiwarka korpusowa Poliqarp 2
- **współpracownicy projektowi** (do kilkudziesięciu).

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 Ogródniczuk
- **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 wyrażenia opinii w języku polskim**, NCN PRELUDIUM, Alex Wawer

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



NARODOWY KORPUS JĘZYKA POLSKIEGO

CONSORTIUM



IJP

Poliqarp search engine for NKJP data

[QUERY](#)
[SETTINGS](#)
[FILE A BUG](#)
[HELP](#)

Query:

Corpus:



Results

Found 1000 results

Displaying results 7—8

[Previous 2](#)

[Next 2](#)

[Z](#) zaczęło odpadać mięso. Ręka oderwała się od **korpusu** [korpus.subst.sg.gen.m3] i upadła na ziemię. Gdzie jest

8. . Poddawało się moim rękom jak wypchany trocinami **korpus** [korpus.subst.sg.nom.m3] lalki, ale twarz wciąż pozostawała władcza

[Previous 2](#)

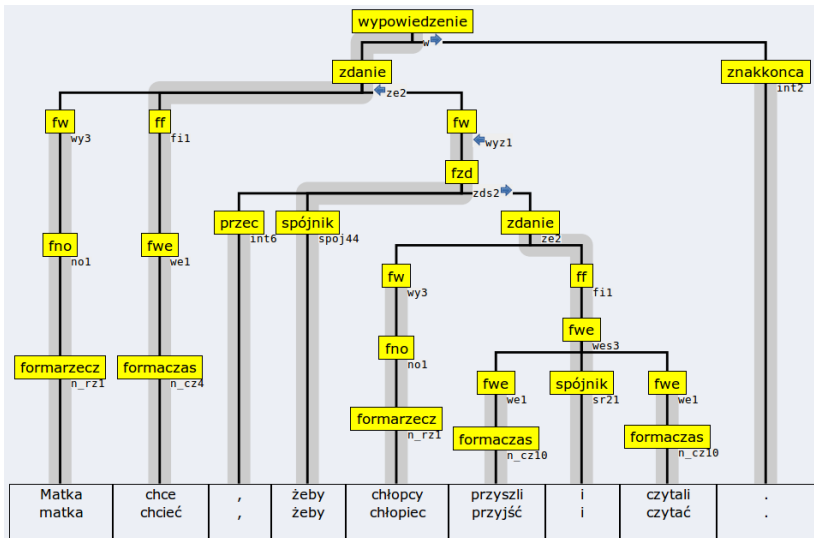
[Next 2](#)

Context

Na moich oczach twarz starzała się coraz bardziej. Skóra wiotczała niczym balon, z którego uchodzi powietrze, aż wreszcie zaczęła odpadać od kości. Wstrząsając się z obrzydzenia, uwolniłem się od ściskającej moją dłoń ręki, z której zaczęło odpadać mięso. Ręka oderwała się od **korpusu** i upadła na ziemię. Gdzie jest Joanna?! - krzyknąłem do kobiety w czerni. Czy to nie wszystko jedno? - odparła. - Twoja Joanna została na szosie. Zastąpię ci ją. Czyż nie jestem do niej podobna? Masz siwe włosy - powiedział

Metadata

- 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/>



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 przeszukiwanie 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



Poliqarp2 strona projektu na sourceforge Zaloguj Polski / Polish ▾

Znalazłem 7 wyników

Kolory przypisane zmiennym: [\\$X,t_o](#)

doc named p s seg sub syng syng syng

Podniosły był, i rozmówili się ["po leśnemu"](#)

```
graph TD
    doc[doc] --- p[p]
    doc --- s[s]
    s --- syng1[syng:NG]
    s --- syng2[syng]
    s --- syng3[syng]
    s --- syng4[syng]
    s --- syng5[syng]
    s --- syng6[syng:AdvG]
    s --- syng7[syng]
    s --- syng8[syng]
    syng1 --- syng1_s1[syng]
    syng1 --- syng1_s2[syng]
    syng1_s1 --- byl[był]
    syng1_s2 --- byl2[był]
    syng2 --- comma[,]
    syng3 --- i[i]
    syng4 --- rozmowili[rozmówili]
    syng4 --- sie[się]
    syng6 --- po[po]
    syng6 --- leśnemu[leśnemu]
    syng7 --- asterisk1[*]
    syng8 --- asterisk2[*]
```

W tym roku [po rosyjsku i angielsku](#) wydany zostanie "Castorpi", który w zeszłym sezonie zebrał w Niemczech wiele recenzji.

- 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.

STATYSTYKI

Klasa trudności tekstu:

5 / 7



Tekst trudniejszy, zrozumiały dla ludzi wykształconych

[rozwiń »](#)

LEGENDA

- Aa** Fragment wyraźnie trudniejszy od reszty tekstu
- Aa** Fragment trudniejszy od reszty tekstu
- Aa** Bardzo długie zdanie
- Aa** Trudne słowo wymagające zmiany

DANE

Tekst po analizie

Metamorfoza – w literaturze (oraz kulturze) gruntowna przemiana kogoś, jego sposobu myślenia, postępowania i niekiedy również wyglądu.

Przemiana bohatera jest jednym z najpierwotniejszych tematów występujących w literaturze. Już w starożytności obserwujemy fascynację nad zmiennością. Motyw metamorfozy często spotykany jest w mitologiach wielu narodów. Występuje również w ludowej baśni, leżąc u podstaw **metamorfozy bohatera** stanowiący główny chwyt konstrukcyjny **Metamorfozy Apulejusza (Metafrazysy osioła)**.

Motyw, m. pojawia się także we współczesnej literaturze. Są to przemiany bohatera, którego w doborze i odwrotnie, przemiany całkowite obejmujące również wygląd, bądź też takie metamorfozy, które polegają na ewoluowaniu bohatera, gdzie jego postawa i poglądy stają się bardziej dojrzałe. Przemiana bohatera jest ilustracją dynamicznej natury człowieka.

FOG-Base: 8,0

Metamorf

Proponowane zastępcze
słowa: zmiany, przemiany,
zamiany.

Multiserwis to serwis webowy udostępniający narzędzia analizy lingwistycznej, umożliwiające 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/>

Raw text Segmentation Morphosyntax Named entities Words Groups Mentions Coreference Dependency parse

1 Maria od zawsze kochała Jana. Gdy poprosił ją o rękę, była szczęśliwa.

Raw text Segmentation Morphosyntax Named entities Words Groups Mentions Coreference Dependency parse

1 Maria od zawsze kochała Jana . Gdy poprosił ja o rękę , była szczęśliwa .

Raw text Segmentation Morphosyntax Named entities Words Groups Mentions Coreference Dependency parse

1 Maria od zawsze kochała Jana. Gdy poprosił ją o rękę, była szczęśliwa.

Raw text Segmentation Morphosyntax Named entities Words Groups Mentions Coreference Dependency parse

1 Maria od zawsze kochała Jana. Gdy poprosił ją o rękę, była szczęśliwa.

Raw text Segmentation Morphosyntax Named entities Words Groups Mentions Coreference Dependency parse

1 Maria od zawsze kochała Jana. Gdy poprosił ją o rękę, była szczęśliwa.

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)

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/>)