



PSI-TOOLKIT - zestaw narzędzi dla lingwistów oraz inżynierów języka naturalnego

autorzy: Krzysztof Jassem, Filip Graliński



PSI – Pracownia Systemów Informatycznych

- Projekt MNiSW „Narzędzia do automatycznego przetwarzania języka polskiego udostępnione publicznie”
- Terminy: 21.04.2011 – 20.04.2013
- Wykonawcy:
Krzysztof Jassem, Filip Graliński, Marcin Junczys-Dowmunt, Roman Grundkiewicz, Paweł Skórzewski, Marcin Walas





Teza



Informatyk



Lingwista

- Większość lingwistów komputerowych wywodzi się ze środowiska informatycznego, a nie lingwistycznego
- Przyczyna?
Narzędzia NLP nie są zachęcające dla lingwistów
- **PSI – Toolkit**
 - Dostarcza narzędzi dla inżynierów języka (L4LT)
 - Pomaga lingwistom (LT4L)



Agenda



- Przykłady NLP-toolkitów z punktu widzenia lingwistów
- PSI-Toolkit
 - Sesja
 - Cechy charakterystyczne
 - Wnioski



CZĘŚĆ 1.

Przegląd NLP-toolkitów





Projekt UIMA Project



- UIMA – Unstructured Information Management applications
- Przetwarza tekst i wyszukuje:
 - osób, miejsc, organizacji
 - relacji, np. kto pracuje gdzie
- Można wykorzystać aplikacje UIMA celu:
 - identyfikacji języka
 - podziału na tokeny i zdania
 - wyszukiwania jednostek nazwanych



UIMA Project

- "Smartest Machine on Earth"
 - Quiz Jeopardy! (Va Bank)
Człowiek kontra komputer
14 – 16 April 2011
 - Watson pokonał Brada Ruttera, najefektywniejszego gracza i Kena Jenningsa, rekordzisty kolejnych zwycięstw (74).¹
 - Watson stosował Apache UIMA
- Aplikacje napisane są w Javie
 - Muszą być kompilowane w Javie
 - Wymagają przynajmniej podstawowych umiejętności programowania w Javie





NLTK

- "Natural Language Toolkit to biblioteka narzędzi NLTK napisana w Pythonie
- Narzędzia uruchamia się z GUI Pythona (z punktu widzenia lingwisty GUI Pythona to praktycznie wiersz poleceń)
- Sposób użycia
 - Pobierz i zainstaluj Pythona
 - Załaduj korpusy tekstów opracowane przez autorów NLTK lub...
 - Przetwarzaj swoje teksty, ale przedtem musisz je skonwertować do formatu NLTK





Sesja NLTK (1)



```
>>> import nltk
>>> text3
<Text: The Book of Genesis>
>>> text7
<Text: Wall Street Journal>
>>> len(text3)
44764
>>> text3.count("conference")
0
>>> text7.count("conference")
12
>>> text7.concordance("conference")
```

Displaying 14 of 14 matches:

Maxwell , said 0 *T*-1 at a news conference . He said 0 more than 90 % of the
sal * to build a \$ 19 million *U* conference center for a joint meeting of the
fore Japan The Senate-House conference committee is used *-1 when a bill



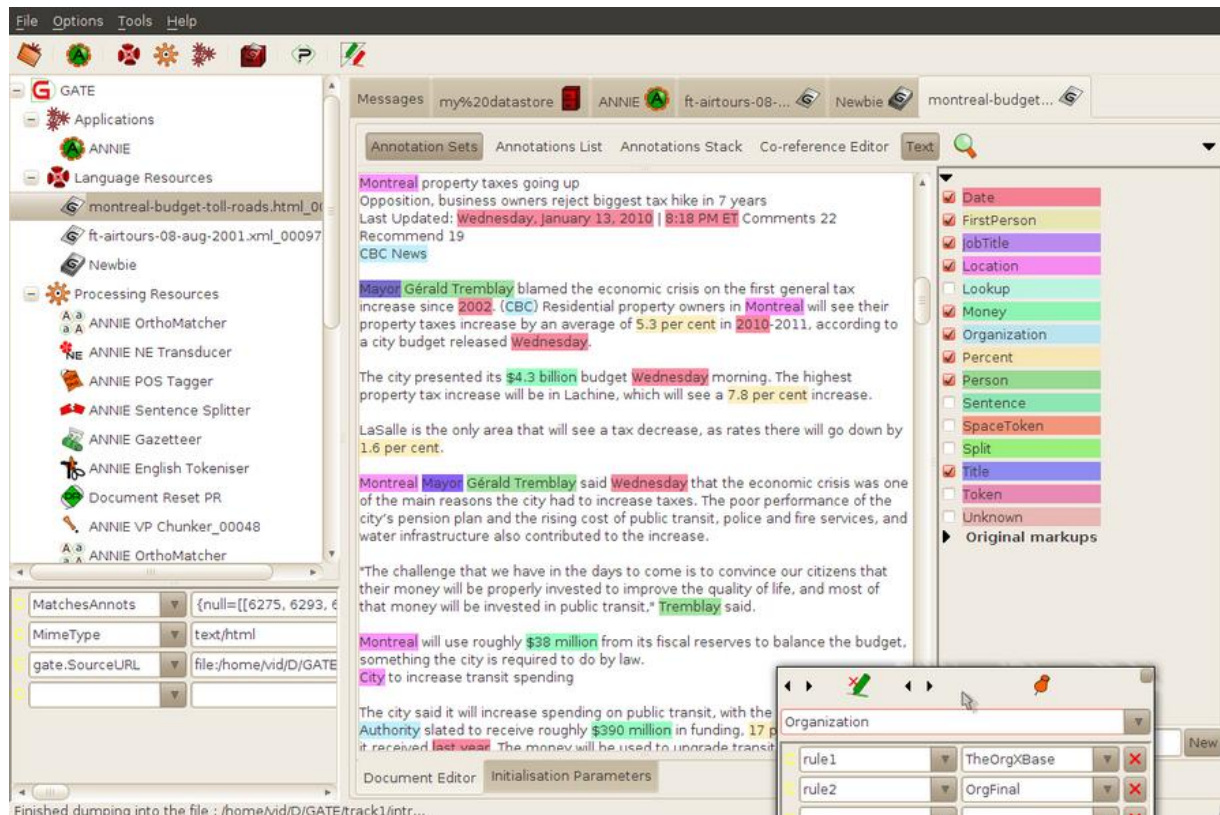
Sesja NLTK (2)



```
>>> text = "I've been to many countries, e.g. Germany, France, Canada. I enjoy travelling."
>>> tokens = nltk.word_tokenize(text)
>>> tokens
['I', "'ve", 'been', 'to', 'many', 'countries', ',', 'e.g.', 'Germany', ',', 'France', ',', 'Canada.', 'I', 'enjoy',
 'travelling', '.']
>>> nltk.pos_tag(tokens)
Traceback (most recent call last):
...
  import numpy
ImportError: No module named numpy
>>> nltk.pos_tag(tokens)
[('I', 'PRP'), ("'ve", 'VBP'), ('been', 'VBN'), ('to', 'TO'), ('many', 'JJ'), ('countries', 'NNS'), (',', ','),
 ('e.g.', 'NNP'), ('Germany', 'NNP'), (',', ','), ('France', 'NNP'), (',', ','), ('Canada.', 'NNP'), ('I',
 'PRP'), ('enjoy', 'VBP'), ('travelling', 'VBG'), ('.', '.')]
>>>
```



„GATE is open source software capable of solving almost any text processing problem”





- GATE został zaprojektowany dla:
 - Inżynierów języka (mogą tworzyć programy autorskie)
 - Lingwistów:
 - Mogą korzystać z gotowych narzędzi
 - Mogą tworzyć własne słowniki lub gramatyki
- Aby otrzymać anotację, lingwista musi:
 - Załadować specjalny plugin (CREOLE pugin)
 - Zdefiniować nowe „zasoby przetwarzające” (anotatory)
 - Zdefiniować aplikację jako pipeline anotatorów
 - A wtedy...
 - Aplikacja zapewne się nie uruchomi, gdyż:
 - Każdy proces aplikacji musi być przypisany do dokumentu
 - Nie każdy „pipeline” jest poprawny



Stanford Core NLP

- Stanford CoreNLP

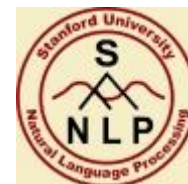
Zintegrowany zestaw narzędzi NLP dla j. angielskiego napisany w Javie; zawiera:

- tokenizację
- POS tagging
- NER
- parsing,
- rozwiązywanie koreferencji

- Licencja: GPL



Stanford online



Stanford CoreNLP

Output format: Visualise

Please enter your text here:

I've been to many countries, e.g. France,
Germany, Canada.

Wyślij zapytanie

Clear

Part-of-Speech:

1 I've been to many countries, e.g. France, Germany, Canada.

PRP VBP VBN TO JJ NNS , FW NNP , NNP , NNP ,

Named Entity Recognition:

1 I've been to many countries, e.g. France, Germany, Canada.

Loc Location Loc



Stanford Core NLP offline



„Before using Stanford CoreNLP, it is usual to create a configuration file (a Java Properties file)”

annotators = tokenize, ssplit, pos, lemma, ner, parse, dcoref

„In particular, to process the included sample file input.txt you can use this command in the distribution directory Stanford CoreNLP Command”

```
java -cp stanford-corenlp-2012-01-08.jar:stanford-corenlp-2011-12-27-models.jar:xom.jar:joda-time.jar-Xmx3g  
edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators  
tokenize,ssplit,pos,lemma,ner,parse,dcoref -file input.txt
```



CZĘŚĆ 2.

PSI-Toolkit



PSI-Toolkit

<http://psi-toolkit.wmi.amu.edu.pl>

- Zestaw narzędzi do przetwarzania języka:
 - polskiego, angielskiego, francuskiego, tureckiego,...
- (Lesser) General Public License
- Dla:
 - Inżynierów języka
 - Biblioteki: Java, Perl, Python
 - Paczki dla dystrybucji linuksowych
 - językoznawcy
 - Dostęp przez serwis WWW



Sesja PSI-Toolkit (1)

Who is it for?

PSI-Toolkit is designed for linguists, language engineers or any users who need to process natural language texts and get the result now!

PSI-Toolkit provides an easy-to-use web-service as well as a set of advanced tools for users with some computer skills.

Binding to programming languages

Using PSI-Toolkit within popular programming languages is as simple as possible. Let's have a look at this Perl code:

```
my $psi = PSIToolkit::Simple::PipeRunner->new
  ("tokenize --lang en ! perl-simple-writer");
$runner->run_for_perl('PSI-Toolkit is Skynet.');
```

It is simple, isn't?

Ubuntu packages!

No more wasting time trying cumbersome toolkits for processing! PSI-Toolkit is available in Linux distributions:

Download

Write command:

[random example](#)

Write input text:

[...or click here to select a txt, doc or pdf file!](#)

Result:

```
I've been to many countries, e.g. Germany, France, Canada.
I enjoy travelling.
```



Sesja PSI-Toolkit (2)

psi
toolkit

AboutHelpDown

Write command:

[random example](#)

Write input text:

[...or click here to select a txt, doc or pdf file!](#)

Result:

```
beauty+subst|beautiful+adj|excellent+adj  
game+subst|play+verb  
last+verb|stay+verb|abide+verb|permanent+adj|durable+adj|lasting+adj  
long+adv
```



PSI-Toolkit Session (3)

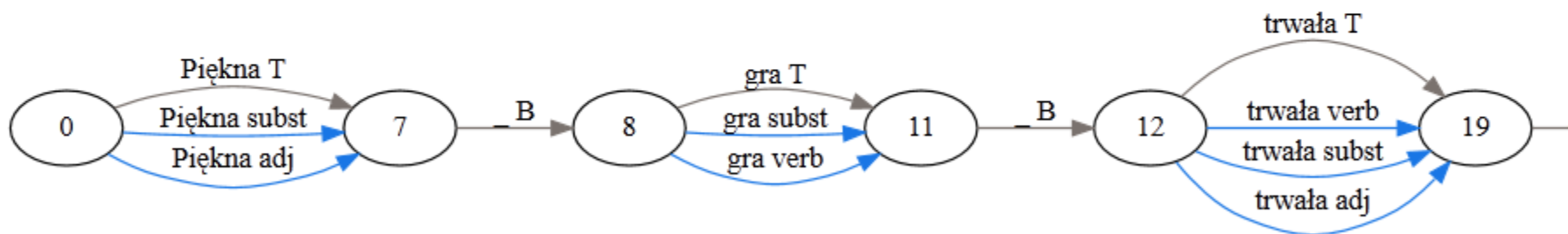
psi toolkit [About](#) [Help](#) [Download](#)

Write command: [random example](#)

Write input text: [...or click here to select a txt, doc or pdf file!](#)

Result:

Change output format type:



Or download the output file:





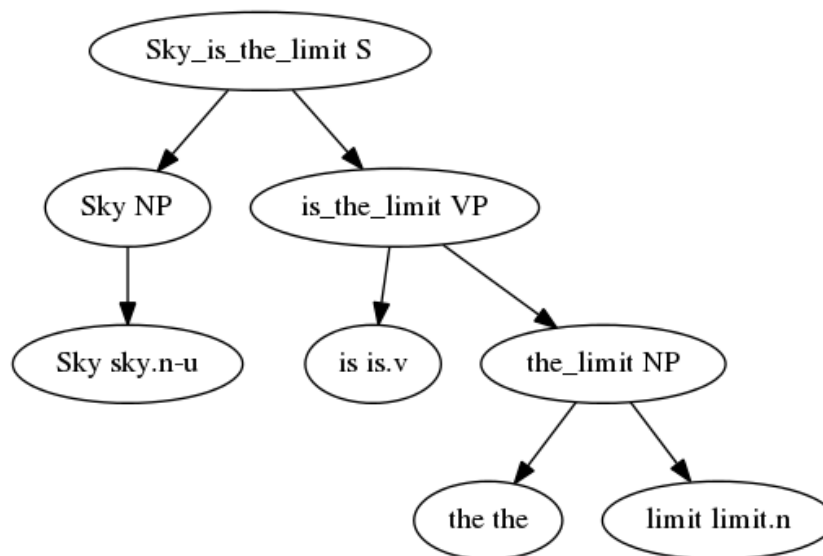
Sesja PSI-Toolkit (4)

psi toolkit [About](#) [Help](#) [Download](#)

Write command: [random example](#)

Write input text: [...or click here to select a txt, doc or pdf file!](#)

Result:





PSI-Toolkit - procesory

■ Readery

- Czytają dane w różnych formatach (włącznie z .pdf, .djvu, .doc)
- Wczytywanie danych w wewnętrznym formacie (PSI-Lattice)

■ Annotatory

- Tokenizatory
- Segmentatory
- Lematyzatory
- Parsery
- Słowniki dwujęzyczne

■ Writery

- Graficzne
- Tekstowe



Zewnętrzne zasoby

- Możliwość włączenia zewnętrznych narzędzi
 - Poprzez użycie wiersza poleceń
- Jako procesory PSI-Toolkit:
 - *Morfologik* (polski)
 - *Link-parser* (angielski)
 - SFST/trmorph (turecki)
 - aspell
- Konwertery tagsetów



PSI-Toolkit w Linuksie

- PSI-PIPE

- Paczki dostępne dla kilku dystrybucji
- PSI-Pipeline (potoki)

```
>echo 'I've been to many countries e.g. France.' |  
psi_pipe ! tokenize --lang en ! psi-writer | sort
```

- PSI-SERVER

- Dostęp przez *Web service*

Download PSI-Toolkit:

Ubuntu packages

- Ubuntu 12.04 LTS Precise Pangolin - amd64
- Ubuntu 12.04 LTS Precise Pangolin - i386
- Ubuntu 11.10 Oneiric Ocelot - amd64
- Ubuntu 11.10 Oneiric Ocelot - i386
- Ubuntu 11.04 Natty Narwhal - amd64
- Ubuntu 10.04 LTS Lucid Lynx - amd64

Debian packages

- Debian Unstable (sid) - amd64

Mint packages

Corresponding Ubuntu packages can be used:

- Mint 13 LTS Maya - 64-bit
- Mint 13 LTS Maya - 32-bit
- Mint 12 Lisa - 64-bit
- Mint 12 Lisa - 32-bit
- Mint 11 Katya - 64-bit
- Mint 9 Isadora - 64-bit

Arch Linux packages

- Arch Linux x86_64
- Arch Linux i686





Biblioteki

- Dowiązania do wybranych języków programowania
 - Java
 - Python
 - Perl
 - Ruby

```
my $psi = PSIToolkit::Simple::PipeRunner->new  
  ("tokenize --lang en ! perl-simple-writer");  
$runner->run_for_perl('PSI-Toolkit is Skynet.');
```

```
#returns: ['PSI-Toolkit', 'is', 'Skynet', '.'];
```



Podsumowanie: PSI-Toolkit

- Różne języki
- Włączanie zewnętrznych narzędzi
- Dostęp on-line
- Uruchamianie przez wiersz poleceń
- Dowiązania do ważniejszych języków programowania
- Dostęp przez *Web service* (JSON)