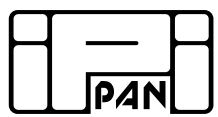
ICS PAS Multiservice

A linguistic Web service for Polish

Maciej.Ogrodniczuk@ipipan.waw.pl



INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. J. K. Ordona 21, 01-237 Warszawa

CLARIN/D-SPIN workshop

16-17 November 2010

Lingustic Engineering Group ICS PAS

LEG (`http://nlp.ipipan.waw.pl/`) @ ICS PAS in Warsaw:

- group headed by Adam Przepiórkowski,
- recent research and development topics:
  - morphosyntactic analysis,
  - syntactic parsing (deep and shallow),
  - named entity recognition (NER),
  - word sense disambiguation (WSD),
  - information extraction (IE),
  - discovery of linguistic knowledge,
  - semantic analysis of spoken data,
  - corpus linguistics,
  - LRT standards.
- currently 6 externally funded projects (various projects recently finished, other being reviewed).

What's available (standalone tools)

Currently available LRTs:

- the IPI PAN Corpus of Polish,
- morphosyntactic analysers,
- two taggers,
- DCG-like efficient parser and Polish DCG grammar,
- shallow parser and shallow Polish grammar,
- syntactic valence dictionary of Polish,
- various smaller resources and tools.

What's coming soon

Within half a year:

- the National Corpus of Polish,
- Polish treebank.

Within the next 1–2 years:

- wide coverage version of the DCG-like grammar,
- LFG extensions of this grammar,
- more efficient shallow parser and larger shallow grammar of Polish,
- syntactico-semantic valence dictionary of Polish,
- WSD systems for Polish,
- NER systems for Polish,
- various IE and QA systems for Polish.

ICS PAS Multiservice

ICS PAS Multiservice provides a common interface and coherent linguistic annotation of Polish texts basing on individual offline tools.

Following annotation layers are supported:

- paragraph-level segmentation,
- sentence-level segmentation,
- tokenization,
- lemmatization,
- morphological analysis,
- tagging,
- shallow parsing,
- deep parsing.

Overview

Currently:

- input: plain text, UTF-8-encoded,
- output: packaged TEI P5-based and National Corpus of Polish-based format,
- linguistic features preserved using TEI-embedded feature structure formalism.

Integrated tools:

- Morfeusz – sentence splitting, tokenization, lemmatization and morphological analysis,
- TaKIPI – tagging,
- Spejd – shallow parsing,
- Świgra – deep parsing.

Segmentation

Simple paragraph splitter; unstructured texts represented as single paragraph:

```
<text xml:id="p-text1">
  <body>
    <p xml:id="p-p1">Ważny list? Tu go miałem.</p>
  </body>
</text>

<text xml:id="s-text1">
  <body>
    <p xml:id="s-p1" corresp="p-p1">
      <!-- Ważny list? -->
      <seg xml:id="s-seg1"
         corresp="#string-range(p-p1,0,11)"/>
    </p>
  </body>
</text>
```

Tokenization

Tokens are string fragments from the paragraph layer;
variants are identified.

```
<!-- list -->
<seg xml:id="t-seg2" corresp="#string-range(p-p1,6,4)"/>
...
<!-- miałem -->
<seg xml:id="t-seg6" corresp="#string-range(p-p1,18,6)"/>
<!-- miał -->
<seg xml:id="t-seg7" corresp="#string-range(p-p1,18,4)"/>
<!-- em -->
<seg xml:id="t-seg8" corresp="#string-range(p-p1,22,2)"/>
```

Lemmatization

Lemmata are features of tokens.

```
<!-- list -->
<seg xml:id="l-seg2" corresp="t-seg2">
  <fs type="lex">
    <f name="base">
      <vAlt>
        <string xml:id="l-string2">
          list</string>
        <string xml:id="l-string3">
          lista</string>
      </vAlt>
    </f>
  </fs>
</seg>
```

Morphological analysis

Morphological analysis of a token is provided for all identified lemmata; POS information is separated from MSD.

```
<!-- list -->
<seg xml:id="m-seg2" corresp="t-seg2">
  <fs type="morph">
    <f name="interps">
      <vAlt>
        <fs type="lex">
          <f name="base"
             fval="l-string2"/>
          <f name="ctag">
            <symbol value="subst"/>
          </f>
          <f name="msd">
            <symbol xml:id="m-symbol5"
             value="sg:nom:m3"/>
            <symbol xml:id="m-symbol6"
             value="sg:acc:m3"/>
          </f>
        </fs>
        <fs type="lex">
          <f name="base"
             fval="l-string3"/>
          <f name="ctag">
            <symbol
             value="subst"/>
          </f>
          <f name="msd">
            <symbol
             xml:id="m-symbol7"
             value="pl:gen:f"/>
          </f>
        </fs>
```

Tagging

Tagging selects single morphological analysis (which corresponds to a definite lemma).

```
<!-- list -->
<seg xml:id="g-seg2"
     corresp="t-seg2">
  <fs type="morph">
    <f name="disamb"
       fVal="m-symbol5"/>
  </fs>
</seg>
```

Shallow parsing

Syntactic groups represented in the annotation model contain pointers to immediate constituents of the group — tokens or other syntactic groups.

```
<!-- ważny list -->
<seg xml:id="c-seg1">
  <ptr type="head"
       target="t-seg1"/>
  <ptr type="nonhead"
       target="t-seg2"/>
</seg>
```

Deep parsing

Deep parsing produces a shared parse forest, i.e. a collection of parse subtrees stored in a packed graph format, with each unique subtree stored only once.

```
<seg xml:id="d-seg2" corresp="s-seg2" type="forest">
  <seg xml:id="d-seg2.4" type="nt" subtype="formaczas">
    <fs>
      <f name="czas">    <symbol value="ter"/>  </f>
      <f name="liczba">  <symbol value="poj"/>  </f>
      <f name="osoba">   <symbol value="1"/>    </f>
    </fs>
    <seg type="children" subtype="n_cz4">
      <ptr type="head" target="d-seg2.5"/>
    </seg>
  </seg>
  <seg xml:id="d-seg2.5" type="terminal">
    <fs>
      <f name="msd-ref" fVal="morph_2.1.1.1-msd"/>
    </fs>
  </seg>
</seg>
```

Conclusions

Plans:

- involvement of pre-processing tools (e.g. converting input from other formats and encodings to plain UTF-8 text),
- integration of further annotation layers (e.g. coreference, named entities etc.)
- integration of variants of tools (e.g. new Brill tagger – Pantera, other morphological analysers – Morfologik, UAM Text Tools etc.)

Fresh ideas:

- conversion to (and from?) new formats, such as TCF,
- integration with other frameworks (WebLicht?)