POLFIE: an LFG grammar of Polish accompanied by a structure bank

Agnieszka Patejuk and Adam Przepiórkowski

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. Jana Kazimierza 5, 01-248 Warszawa

Wrocław, 17 October 2015
POLFIE grammar (Patejuk and Przepiórkowski 2012):

- **Lexical-Functional Grammar** approach (LFG),
- implemented in **XLE** (platform dedicated to LFG),
- created by maximising the use of existing Polish resources:
  - previous implemented grammars,
  - morphosyntactic information from **Morfeusz**, 
  - valency information from **Walenty**, 
- used for the construction of **structure bank**, 
- uses **OT marks** for disambiguation, 
- available on an open source license: **GPL3**.
LFG formalism in a (tiny) nutshell

- **constraint-based**, highly **lexicalised**,
- parallel levels of **representation**:

![Diagram of LFG formalism]

- analyses of many **typologically diverse** languages (English, Warlpiri, Russian, Urdu...),
- LFG grammars may be **implemented** (in XLE),
- attempts at commercial use (Bing search engine).
Maximising the use of existing resources

Roughly, LFG grammars consist of:

- rules,
- lexicon.

Saving resources... by building on existing resources for Polish:

- **rules** created on the basis of previous grammars: DCG, HPSG;
- **lexicon** built using:
  - morphosyntactic information from Morfeusz,
  - valence information from Walenty,
  - lexicalised information from grammars.
Maximising the use of existing resources

Roughly, LFG grammars consist of:

- rules,
- lexicon.

Saving resources... by building on existing resources for Polish:

- **rules** created on the basis of previous grammars: DCG, HPSG;
- **lexicon** built using:
  - morphosyntactic information from Morfeusz,
  - valence information from Walenty,
  - lexicalised information from grammars.
Grammars of Polish

- **GFJP2** (Woliński 2004):
  - DCG,
  - based on Świdziński 1992,
  - reasonable coverage,
  - limited linguistic description (due to the representation);

- **FOJP** (Przepiórkowski et al. 2002):
  - HPSG,
  - based on numerous HPSG works of authors,
  - very limited coverage,
  - sound linguistic description.
About Morfeusz

Morfeusz (Woliński 2006, 2014):

- **state-of-the-art** morphological analyser for Polish,
- extra information: **names**, **register**,
- provides **segmentation**,
- **fast, reliable**, easy to use,
- **very good coverage**,
- its dictionary can be **modified**.
General information (see the poster after lunch!):

- valency dictionary of Polish,
- human- and machine-readable,
- framework-independent, uses its own formalism,
- can be converted to various formalisms,
- schemata illustrated with attested examples,

Linguistic features (Przepiórkowski et al. 2014):

- 2 levels: morphosyntax, semantics,
- structural case, passivisation, control relations,
- explicit account of coordination,
- some arguments defined by semantics rather than category (e.g. manner, location, duration, path),
- lexicalised arguments: rich component.
About Walenty (CLARIN-PL resource)

**General** information (see the poster after lunch!):
- valency dictionary of **Polish**,
- **human- and machine-readable**,
- **framework-independent**, uses its own formalism,
- **can be converted** to various formalisms,
- schemata illustrated with **attested examples**,
- **open source**, available from:
  http://walenty.ipipan.waw.pl/.

**Linguistic** features (Przepiórkowski *et al.* 2014):
- 2 levels: **morphosyntax**, **semantics**,
- **structural case**, **passivisation**, **control** relations,
- explicit account of **coordination**,
- some arguments defined by **semantics** rather than category (e.g. manner, location, duration, path),
- **lexicalised** arguments: rich component.
Methods of testing

Results for the grammar version in the structure bank:

- **quantitative:**
  - reparsing **Składnica** (77.62% out of 9011):
    - DCG treebank,
    - sentences selected from NKJP1M,
    - relatively short sentences were chosen,
    - not very big: currently around 10K good sentences,
    - ensures compatibility with the DCG grammar;
  - parsing **NKJP1M** (33.25% out of 85663):
    - manually annotated subcorpus of NKJP,
    - 1.2M segments, over 85.5K sentences,
    - more variety, longer sentences, sentences without finite verb,
    - checks coverage and robustness;

- **qualitative:** manually disambiguated structure bank.
parsebank: LFG structures produced by XLE,
disambiguated using INESS system (Rosén et al. 2012),
6 500 disambiguated sentences,
58 000 segments,
solutions (before disambiguation):
  average: 458.5,
  median: 12;
each sentence disambiguated independently by 2 people,
still under development and growing...
Availability

- **grammar:**
  - license: GPL3,
  - package for local installation (requires XLE),
  - web service: XLE-Web in INESS (no sign in);

- **structure bank:**
  - under construction,
  - release in December,
  - license: TBA, probably GPL3.
Via INESS (CLARINO infrastructure)

A system perfectly suited for work with LFG:

- visualisation,
- disambiguation,
- structure bank development,
- online parsing,
- search.

Resources shared using INESS (Bergen or Warsaw installation):

- grammar via XLE-Web:
  - web service,
  - open: no sign in,
  - grammars to choose from: stable, with OT marks, dev,
  - parse a sentence and disambiguate it;

- structure bank:
  - browsing,
  - search.
Via INESS (CLARINO infrastructure)

A system perfectly suited for work with LFG:
- visualisation,
- disambiguation,
- structure bank development,
- online parsing,
- search.

Resources shared using INESS (Bergen or Warsaw installation):
- grammar via XLE-Web:
  - web service,
  - open: no sign in,
  - grammars to choose from: stable, with OT marks, dev,
  - parse a sentence and disambiguate it;

- structure bank:
  - browsing,
  - search.
Sample structures from XLE-Web (grammar with OT marks)
Applications

- **browsing** and **searching** the structure bank (INESS),
- **parsing** sentences in XLE-Web (INESS),
- creating your **own grammar** on the basis of POLFIE,
- **training** parsers on the structure bank,
- further processing of structures: **semantics**.
User experience (comments following a workshop)

User experience:
- (very) useful: 90%,
- ready for use: 90%,
- easy to use: 80%,
- clear interface: 70%.

Comments:
- “very interesting”, “fascinating”,
- “solid theoretical background”,
- “useful after getting to know LFG”,
- “make access easier and more open”,
- “not automatic (requires disambiguation)”.

Summary

- maximising the use of existing resources,
- open source,
- mature and powerful formalism,
- focus on high quality of representation,
- quantitative and qualitative evaluation,
- INESS: very convenient search and visualisation.

POLFIE

http://zil.ipipan.waw.pl/LFG/

XLE-Web (local Warsaw installation)

http://iness.mozart.ipipan.waw.pl/iness/xle-web
"Dziękujemy za uwagę."

Thank you for your attention.
Summary

- maximising the use of **existing resources**, 
- **open source**, 
- mature and powerful **formalism**, 
- focus on high quality of **representation**, 
- quantitative and qualitative **evaluation**, 
- **INESS**: very convenient search and visualisation.

**POLFIE**

http://zil.ipipan.waw.pl/LFG/

**XLE-Web (local Warsaw installation)**

http://iness.mozart.ipipan.waw.pl/iness/xle-web