Polish LFG grammar

Agnieszka Patejuk and Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences

INTRO

Aim

Creating a parasitic, wide-coverage LFG grammar of Polish.

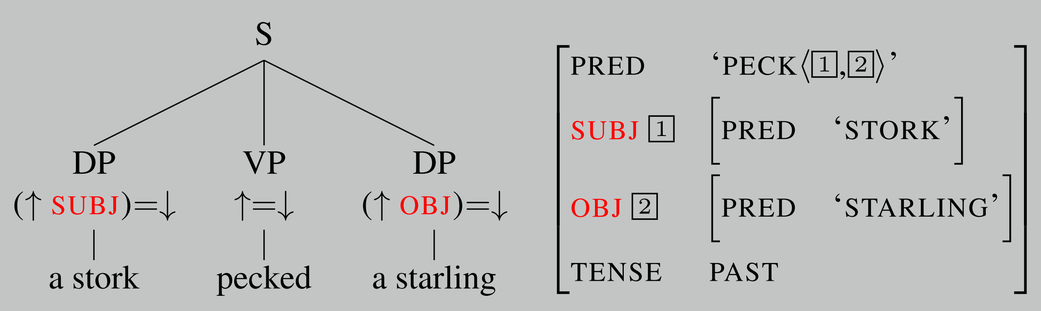
Means

- framework: Lexical-Functional Grammar (LFG),
- platform: Xerox Linguistic Environment (XLE),
- resources: grammars, valence dictionary, testsuites, treebank, corpus.

1. LFG BASICS

Formalism

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



- analyses of diverse languages (English, Warlpiri, Russian, Urdu...),
- attempts at commercial use (Bing search engine).

Possible extensions

- additional level of structure: semantics,
- ranking mechanisms:
  - Optimality Theory,
  - probability scores.

2. RESOURCES

Previous grammars

- GFJP (DCG):
  - c-structure, valence dictionary,
  - reasonable coverage,
  - limited linguistic description;
- FOJP (HPSG):
  - f-structure,
  - very limited coverage,
  - sound linguistic description.

Morphological analyser, corpus

- Morfeusz: fast, reliable, wide coverage;
- National Corpus of Polish (NCP):
  - largest currently available corpus of Polish,
  - 1-million manually annotated subcorpus,
  - uses NCP tagset (similar to Morfeusz),
  - disambiguated annotation.

Składnica treebank

- sentences from NCP (manually annotated subcorpus),
- almost 20000 sentences, 8227 have a good parse,
- one parse per sentence, selected by human annotators,
- OOV verbs parsed using default frames.

3. PUTTING THINGS TOGETHER

Morphosyntactic info

Kobiety dostrzegają.
women.NOM/ACC notice
‘Women notice (someone).’ or ‘(They) notice women.’

- Morfeusz output (form,lemma,pos(:tags)):
  kobiety,kobieta,subst:sg:gen:f|subst:pl:nom.acc.voc:f
  dostrzegają,dostrzegać,fin:pl:ter:imperf
- converted:

```
kobiety      SUBST * @(SUBST KOBIETA PL NOM F);
                   @(SUBST KOBIETA PL ACC F).
dostrzegają  FIN   * @(TRANS DOSTRZEGAĆ +)
                     @(FIN PL 3 IMPERF).
```

- alternative sources: Składnica, NCP.

Valence info

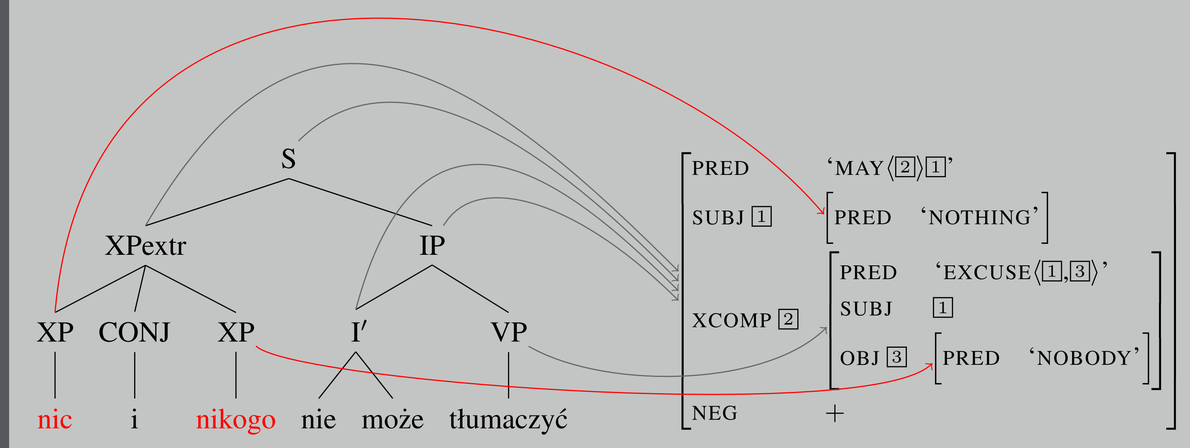
- Original entry (phrasal categories): dostrzegać V np(bier)
- converted (grammatical functions): (TRANS DOSTRZEGAĆ +)

```
TRANS(P O) = "passivisable two argument verb"
@(PASS (^ PRED)='P<(^ SUBJ)(^ OBJ)>')
@(STRLEX OBJ O).
```

- extra information:
  - arguments: passivisable/predicative,
  - case: structural/lexical,
  - control and raising;
- alternative source: Składnica (frames chosen implicitly),
- a better dictionary, Walenty (also presented in this session), is on its way.

Rules rewritten with extensions

- hierarchical c-structure, coordination,
- case assignment, reflexive haplology,
- agreement, extraction.



4. TESTING

Constructed testsuites

- backwards compatibility,
- isolated phenomena, complex interactions, negative examples.

Reparsing Składnica

Treebank testing:

- robustness,
- performance issues,
- real-life text coverage,
- unexpected interactions.

Out of 8227 sentences:

parsed	7138	86.8%
failed	150	1.8%
timed out	931	11.3%
out of memory	7	0.0%

5. OUTLOOK

- stable version release: January 2013,
- further extensions: treebank-based development,
- LFG structure bank (c- and f-), ParGram compliance,
- ranking parses, adding semantic layer.

agnieszka.patejuk@gmail.com, adamp@ipipan.waw.pl