DEPENDENCY PARSING IN POLISH

Alina Wróblewska – Institute of Computer Science, Polish Academy of Sciences, Warsaw
alina@ipipan.waw.pl

Dependency Parsing

Dependency parsing is a process of automatic annotating (syntactic analysis) an input sentence with dependency relations.

Dependency Structure

Dependency structure is annotated as a graph:

- **nodes**: tokens in a sentence,
- **arcs**: dependency relations between two tokens (dependent and its governor).

Dependency Graph (DG)

- **spanning property**: DG spans over all tokens in a sentence,
- **connectedness property**: DG forms a connected syntactic structure with all tokens linked by syntactic dependencies,
- **directness property**: DG contains only directed edges (arcs),
- **single-head property**: any node depends on just one unique governor-node in DG,
- **acyclicity property**: no cycles in DG,
- **root property**: DG is a rooted tree with one node root not governed by any other node.
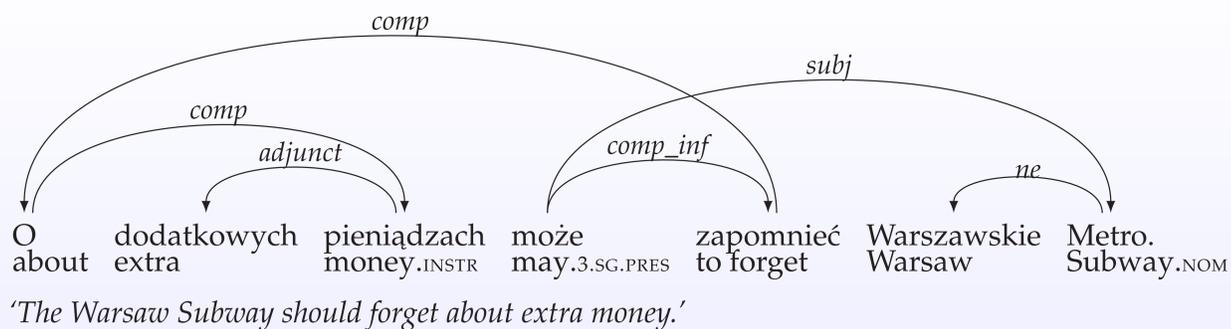
Grammatical Functions

Arcs linking lexical nodes are labelled with grammatical functions.

Polish grammatical functions:

- **arguments**: *comp* (complement)
  comp_fin (clausal complement)
  comp_inf (infinitical complement)
  obj (object)
  obj_th (dative object)
  pd (predicative complement)
  subj (subject)

- **syntactically motivated non-arguments**:
  abbrev_punct (abbreviation marker)
  adjunct
  complm (complementizer)
  imp (imperative marker)
  pred (predicate)
  punct (punctuation mark)
  refl (reflexive marker)

- **morphologically motivated non-arguments**:
  aglt (mobile inflection)
  aux (auxiliary)
  cond (conditional marker)
  neg (negation marker)

- **semantically motivated non-aruments**:
  app (apposition)
  mwe (multiword expression)
  ne (named entity)

- **coordination**: *conjunct*
  coord (coordinating conjunction)
  coord_punct (punctuation conjunction)
  pre_coord (pre-conjunction)

Description of Polish dependency relations:
zil.ipipan.waw.pl/FunkcjeZaleznosciowe.

Example: Polish Dependency Structure

'The Warsaw Subway should forget about extra money.'

Polish Dependency Bank (PDB)

Polish Dependency Bank (*Składnica Zależnościowa*):

- 7500 syntactically annotated sentences,
- dependency structures meet dependency graph properties and are labelled with grammatical functions defined for Polish,
- dependency structures are automatically converted from constituent trees (*Składnica Frazowa*).

Constituency-to-dependency conversion:

- all language-specific syntactic phenomena encoded in constituent trees are covered and annotated with correctly chosen dependencies,
- even if conversion is a relatively straightforward process, a certain amount of reorganising is necessary, in order to make dependency structures meet annotation principles.

constituent tree   conversion   dependency structure
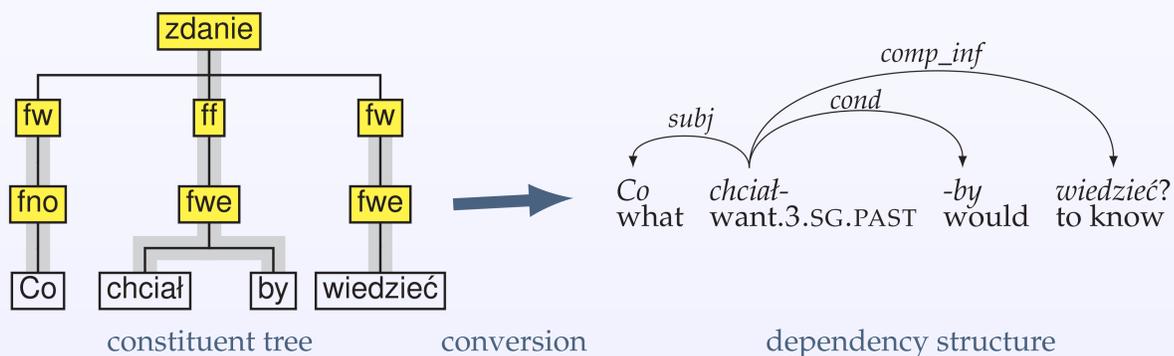
Polish Dependency Bank in CoNLL format (Składnica-zależnościowa-0.5.conll.gz) is available on http://zil.ipipan.waw.pl/Składnica.

Polish Dependency Parser

Polish dependency parser is trained on the Polish Dependency Bank with the publicly available parsing system **MaltParser** (Nivre et al., 2006).

Parameters of Polish MaltParser:

- **parsing algorithm**: built-in `stackeager` algorithm (Nivre, 2009) designed for non-projective dependency structures,
- **feature model** defined in terms of token attributes, i.e., word form (FORM), part-of-speech tag (POS), morphological features (FEATS), and lemma (LEMMA) available in input data, or dependency types (DEPREL) extracted from partially built dependency graphs,
- **classifier** trained with the LIBLINEAR library.

Parsing performance:

Polish MaltParser trained on 6832 sentences is evaluated against 759 sentences (9.8 tokens/sentence) taken from PDB and a set of 50 manually annotated sentences (17.8 toks/sentence).

Parsing model	PDB test sets (759 sentences)		Additional test set (50 sentences)	
	LAS	UAS	LAS	UAS
Polish MaltParser	84.7	90.5	68.5	72.2

The dependency parsing model for Polish is available on http://zil.ipipan.waw.pl/PolishDependencyParser.

Acknowledgement