Mapping a MWE lexicon on a treebank

Agata Savary and Jakub Waszczuk

Université François-Rabelais Tours, LI, France

Rationale

MWE-annotated treebanks (Rosén *et al.*, 2015):

- ◇ are essential for linguistic study of MWEs,
- ◇ are prerequisites for joint parsing and MWE recogn.,
- ◇ cover selected MWE types only (mostly NEs),
- ◇ rarely cover verbal MWEs.

MWE lexicons (Losnegaard *et al.*, 2016):

- ◇ develop more rapidly than MWE-annotated treebanks,
- ◇ exist for a large number of languages,
- ◇ are often distributed under open licenses.

Objective: use MWE lexicons to complete existing treebanks with annotation layers dedicated to MWEs.

Walenty

- ◇ Polish large-scale valence dictionary,
- ◇ expressive and theory-neutral formalism (Przepiórkowski *et al.*, 2016),
- ◇ compact syntax, both lexicographer-friendly and NLP-applicable (Patejuk, 2015),
- ◇ 50,000, 3,700 3,000, and 1,000 subcat frames for verbs, nouns, adjectives, and adverbs resp.,
- ◇ phraseological component with 8,000 verbal MWEs (Przepiórkowski *et al.*, 2014).

(1) Nie umiem w tych sprawach **trzymać języka za zębami**.
Not know in these affairs hold.INF tongue.SG.GEN behind teeth.
(lit.) I cannot hold my tongue behind my teeth in such cases.
'I cannot hold my tongue in such cases'

```
trzymać: subj{np(str)}+
   obj{lex(np(str),sg,'język',natr)}+
   {lex(prepnp(za,inst),pl,'ząb',natr)}
```

- ◇ subject: NP in structural case (normally nominative, genitive with nominalized head verb),
- ◇ direct object: NP in structural case (normally accusative, genitive under negation scope), head lemma *język* 'tongue' in singular (`sg`), never modified (`natr`),
- ◇ complement: Prep NP headed by the preposition *za* 'behind', governing the instrumental case (*inst*), and a lexicalized non-modifiable (`natr`) noun with the lemma *ząb* 'tooth' in plural (`pl`).

Składnica

- ◇ Polish constituency treebank, 9,000 sentences, manually disambiguated (Świdziński and Woliński, 2010),
- ◇ no MWE annotations,
- ◇ Walenty-compatible tagset (but Polish mnemonics),
- ◇ dependents of the verbs are explicitly marked: arguments (`fw`) or adjuncts (`fl`).

Walenty → Składnica mapping

Challenges:

- ◇ realizations of **structural cases** depend on complex syntactic constraints,
- ◇ (required) **arguments** (e.g. subject) can regularly be **omitted** in Polish sentences,
- ◇ constraints can concern deeply **embedded** nodes.

Checking Walenty constraints in a Składnica subtree:

- ◇ straightforward tagset mapping: `np`→`fno` 'nominal phrase', `mian`→`nom` 'nominative',
- ◇ conditional statements over combinations of FSs to validate structural (`str`) and agreeing (`agr`) case,
- ◇ arguments with lexicalized components - compulsory,
- ◇ non-lexicalized arguments - optional.

Results:

- ◇ 499 occurrences candidate verbal MWEs,
- ◇ 390 true positives, 27 compositional occurrences, 82 false positives,
- ◇ idiomaticity rate **0.93** (El Maarouf and Oakes, 2015),
- ◇ sources of errors: relieving too many constraints.

Extensions:

- ◇ mapping other MWE resources (recently done for NEs and compounds),
- ◇ allowing for more fine-grained constraints,
- ◇ tuning the degree of flexibility in constraint validation for optimal precision and recall.