Handling MWEs in Walenty, a new valence dictionary for Polish [WG2]

Agnieszka Patejuk

aep@ipipan.waw.pl

INSTYTUT PODSTAW INFORMATYKI
POLSKIEJ AKADEMII NAUK
ul. Jana Kazimierza 5, 01-248 Warszawa

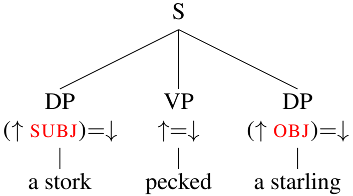
PARSEME meeting, Athens, March 2014

What is this presentation about?

- modelling Polish MWEs together with their syntactic structure
- framework: Lexical-Functional Grammar (LFG)
- platform: Xerox Linguistic Environment (XLE)
- Walenty, a valence dictionary of Polish:
  - open source, available from: `zil.ipipan.waw.pl/Walenty`
  - developed since 2012, spans 3 projects
  - contains 38874 schemata for 8644 verbs
  - created on the basis of attested data
  - can be used by various formalisms (currently: LFG)
  - accounts for coordination (syntactic positions as sets)
  - accounts for MWEs:
    - internal structure (NP/PP, fixed phrase)
    - interactions with syntax (case assignment for NPs)
    - displayed modification pattern

LFG formalism

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



S

DP (↑ SUBJ)=↓ a stork

VP ↑=↓ pecked

DP (↑ OBJ)=↓ a starling

$$\begin{bmatrix} \text{PRED} & \text{`PECK}\langle\boxed{1},\boxed{2}\rangle\text{'} \\ \text{SUBJ}\ \boxed{1} & \begin{bmatrix}\text{PRED} & \text{`STORK'}\end{bmatrix} \\ \text{OBJ}\ \boxed{2} & \begin{bmatrix}\text{PRED} & \text{`STARLING'}\end{bmatrix} \\ \text{TENSE} & \text{PAST}\end{bmatrix}$$

- analyses of diverse languages (English, Warlpiri, Russian, Urdu. . .)
- LFG grammars may be implemented in XLE
- attempts at commercial use (Bing search engine)

POLFIE

- an LFG grammar of Polish implemented in XLE
- based on previous grammars (DCG, HPSG)
- morphological information from analyser, treebank or corpus
- valence information from converted dictionary
- coverage: parses 32% of sentences from 1M sample of the National Corpus of Polish (NKJP; `nkjp.pl`)
- structure bank is being created
- plans for the (near) future: adding semantics
- open source, available from: `zil.ipipan.waw.pl/LFG`

About

- valence dictionary developed since 2012, spans 3 projects
- contains 38874 schemata for 8644 verbs (as of 5/03/2014)
- created on the basis of attested data (from NKJP, from the web)
- open source, available from: `zil.ipipan.waw.pl/Walenty`

Formalism

- syntactic positions (separated by "+") are sets (enclosed in "{}")
- realisations of the position are members of the relevant set (separated by ";")
- realisations belong to the same set if they may be coordinated
  `subj{np(str)} + obj{np(str)} + {np(inst)}`
  `+ {prepnp(o,loc); prepncp(o,loc,że)}`
- some positions are explicitly assigned a grammatical function

More features

- non-canonical realisations of arguments, unlike category coordination
  `subj{np(str); cp(int); ncp(str,int); ncp(str,że)}`
  `+ {np(str)}`
- structural case marked explicitly
- control relations (for infinitival and predicative complements)
- adverbial complements classified according to semantic type

MWE types

- fixed expressions:
  - cannot be modified in any way, the exact string is given
  - `fixed(string)`
- lexicalised phrases:
  - nominal: `lexnp(case,number,lemma,mod)`
  - prepositional: `preplexnp(preposition,case,number,lemma,mod)`
  - typical information: case, preposition form
  - extra information: number, lemma, modification pattern

Modification patterns

- natr: modification not allowed
- atr: modification allowed (though not necessary)
- ratr: modification required (often possessive, NP or adjective)
- batr: specific modification required (possessive: SWÓJ or WŁASNY, 'own')

Examples

Zbił ich na (*bardzo) kwaśne jabłko/*jabłka.
beat then for very sour apple.SG/PL
‘He beat them to a pulp.’
(literally: ‘He beat them into a sour apple.’)

constraints:

- modification not allowed → natr

```
subj{np(str)} + obj{np(str)} + {fixed('na kwaśne
jabłko')}
```

Examples 2

(Gorąca)	krew/*krwie	płynie/*płyną	w	*(jej/Marysi/tych)	żyłach/*żyle.
hot	blood.SG/PL	flow.SG/PL	in	her/Mary's/those	vein.PL/SG

'(Hot) blood flows in her/Mary's/those veins.'

constraints:

- modification allowed (though not necessary) → atr
- modification required (often possessive, NP or adjective) → ratr

```
subj{lexnp(str,sg,'krew',atr)} +
{preplexnp(w,loc,pl,'żyła',ratr)}
```

Examples

Daję	(*swoją/mądrą)	głowę/*głowy,	że	przyjdą.
give	own/wise.SG	head.SG/PL	that	come.FUT

‘I’m sure that they will come.’

(literally: ‘I give (my) head that they will come.’)

constraints:

- modification not allowed → natr

```
subj{np(str)} + {cp(że)} +
{lexnp(str,sg,'głowa',natr)}
```

Examples

Doręczyli to jej do rąk *(własnych).
delivered it her to hands own
‘They delivered it to her as hand delivery.’
(literally: ‘They delivered it to her to (her) own hands.’)

constraints:

- specific modification required (possessive: SWÓJ or WŁASNY, ‘own’) → batr

```
subj{np(str)} + obj{np(str)} + {np(dat)} +
{preplexnp(do,gen,pl,’ręka’,batr)}
```

Conversion process

- python script (around 1K lines)
- takes entries from Walenty, returns XLE lexical entries
- grammatical function (GF) chosen on the basis of contents of the set corresponding to the relevant position (roughly: on the basis of morphosyntax)

Converting MWEs into LFG constraints

- number: $(\uparrow \text{GF NUMBER}) =_c \text{NUM}$
- lemma: $(\uparrow \text{GF PRED FN}) =_c \text{LEMMA}$
- modification:
  - `fixed`: same modification constraints as `natr`
  - `natr`: $\neg(\uparrow \text{GF ADJUNCT})\ \neg(\uparrow \text{GF POSS})$
  - `atr`: no constraint needed (modification allowed but not required)
  - `ratr`: $\{\ (\uparrow \text{GF ADJUNCT}) \mid (\uparrow \text{GF POSS})\ \}$
  - `batr`: $(\uparrow \text{GF ADJUNCT}\ \$\ \text{PRED FN}) \in_c \{\text{SWÓJ WŁASNY}\}$

Issues

- not all modification constraints can be expressed in Walenty
- no information about category corresponding to `fixed`
- only 3 lexicalised categories: NP, PP, `fixed`
- semantics: compositional vs non-compositional

Conclusion

- Walenty: a new valence dictionary for Polish
- large and still growing
- can be used by various grammar formalisms (so far: LFG)
- MWEs (apart from `fixed`) have internal syntactic structure
- constraints can be imposed on MWEs:
  - lemma
  - number
  - modification pattern

Walenty

`zil.ipipan.waw.pl/Walenty`

Questions?

Thank you for your attention

Walenty

`zil.ipipan.waw.pl/Walenty`

Conclusion

- Walenty: a new valence dictionary for Polish
- large and still growing
- can be used by various grammar formalisms (so far: LFG)
- MWEs (apart from `fixed`) have internal syntactic structure
- constraints can be imposed on MWEs:
  - lemma
  - number
  - modification pattern

Walenty

`zil.ipipan.waw.pl/Walenty`