

# Handling MWEs in Walenty, a new valence dictionary for Polish [WG1]



Agnieszka Patejuk

Institute of Computer Science, Polish Academy of Sciences

## INTRO

### Aim

Modelling Polish MWEs together with their internal syntactic structure

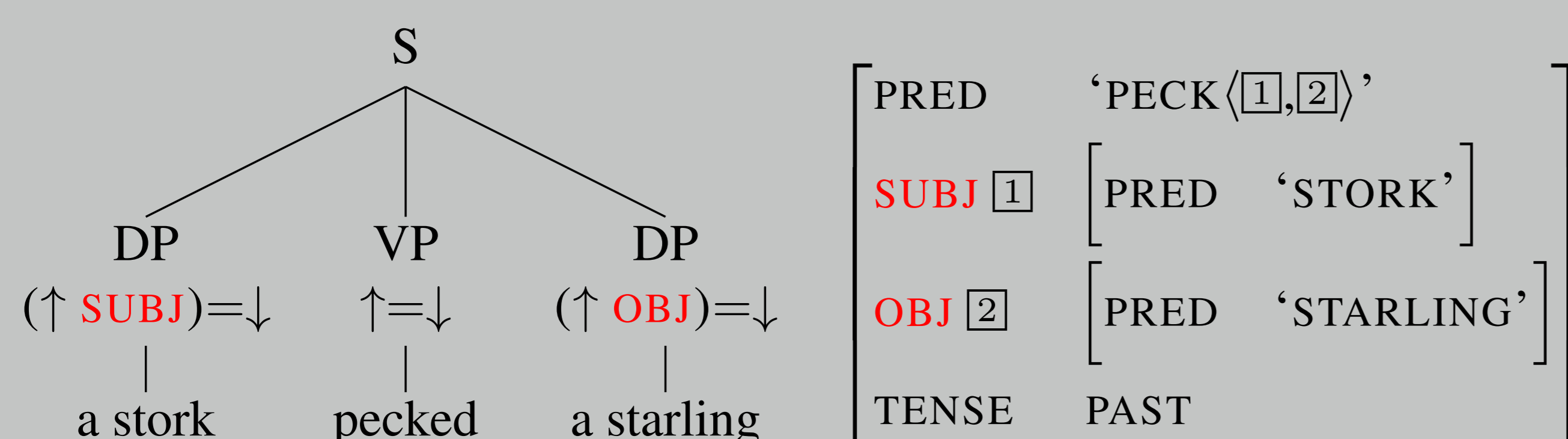
### Means

- framework: Lexical-Functional Grammar (LFG)
- platform: Xerox Linguistic Environment (XLE)
- valence dictionary: Walenty

## 1. LFG

### Formalism

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



- 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](http://nkjp.pl))
- structure bank is being created
- plans for the (near) future: adding semantics
- open source, available from: [zil.ipipan.waw.pl/LFG](http://zil.ipipan.waw.pl/LFG)

## 2. WALENTY

### 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](http://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  
 $\text{subj}\{\text{np}(\text{str})\} + \text{obj}\{\text{np}(\text{str})\} + \{\text{np}(\text{inst})\}$   
 $+ \{\text{prepn}(\text{o}, \text{loc}); \text{prepn}(\text{o}, \text{loc}, \text{ze})\}$
- some positions are explicitly assigned a grammatical function

### More features

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

## 3. MWES IN WALENTY

### MWE types

- fixed expressions:
  - cannot be modified in any way, the exact string is given
  - fixed(string)
- lexicalised phrases:
  - nominal:  $\text{lexnp}(\text{case}, \text{number}, \text{lemma}, \text{mod})$
  - prepositional:  $\text{preplexnp}(\text{preposition}, \text{case}, \text{number}, \text{lemma}, \text{mod})$
  - 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

- $\text{subj}\{\text{np}(\text{str})\} + \text{obj}\{\text{np}(\text{str})\} + \{\text{fixed}(\text{'na kwaśne jabłko'})\}$   
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.')
- $\text{subj}\{\text{lexnp}(\text{str}, \text{sg}, \text{'krew'}, \text{atr})\} + \{\text{preplexnp}(\text{w}, \text{loc}, \text{pl}, \text{'żyła'}, \text{ratr})\}$   
(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.'
- $\text{subj}\{\text{np}(\text{str})\} + \{\text{cp}(\text{ze})\} + \{\text{lexnp}(\text{str}, \text{sg}, \text{'głowa'}, \text{natr})\}$   
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.')
- $\text{subj}\{\text{np}(\text{str})\} + \text{obj}\{\text{np}(\text{str})\} + \{\text{np}(\text{dat})\} + \{\text{preplexnp}(\text{do}, \text{gen}, \text{pl}, \text{'ręka'}, \text{batr})\}$   
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.')

## 4. CONVERTING WALENTY TO LFG

### 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:  $(\uparrow \text{GF ADJUNCT } \$ \text{ PRED FN}) \in_c \{\text{SWÓJ WŁASNY}\}$

## 5. ISSUES

- not all modification constraints can be expressed in Walenty
- no information about category corresponding to fixed
- semantics: compositional vs non-compositional