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

Agnieszka Patejuk, Institute of Computer Science, Polish Academy of Sciences

The aim of this paper is to present how multiword expressions (MWEs) are handled in Walenty (Przepiórkowski *et al.* 2014), a new valence dictionary of Polish. The information stored in Walenty was converted automatically to yield a lexicon used by POLFIE (Patejuk and Przepiórkowski 2012), an LFG grammar of Polish implemented in XLE (`http://www2.parc.com/isl/groups/nltt/xle/`).

Apart from the fact that Walenty takes phenomena such as control, raising, passivisation and structural case assignment into account, the distinctive feature of Walenty is that it pays a considerable amount of attention to the issue of coordination, including unlike category coordination. Walenty takes a new approach to modelling valence through the use of sets: each syntactic position (argument) is a set. The members of this set are possible categorial realisations of the given argument: if two (or more) realisations can be coordinated, it is assumed that they fill the same syntactic position. By contrast, if more than one realisation of the relevant position is possible but these categorial realisations cannot be coordinated, they belong to different schemata in Walenty. If a given argument can only be realised as one syntactic category, a singleton set is used. A sample valence schema (syntactic positions are separated by `+`, set elements are separated using `;`):

(1) `subj{np(str)} + obj{np(str)} + {np(inst)}`
`+ {prepnp(o,loc); prepncp(o,loc,że)}`

(2) `subj{np(str); cp(int); ncp(str,int); ncp(str,że)} + {np(str)}`

Walenty distinguishes two syntactic positions which are explicitly assigned a grammatical function (it precedes the set to which it corresponds): subject (`subj`) and object (`obj`); the latter applies to passivisable objects, regardless of case marking. In (1) both subject and object correspond to singleton sets which contain `np(str)`, an NP marked for structural case (case assignment is handled by the grammar, it takes information about syntactic environment into consideration). The remaining positions have no corresponding grammatical functions marked explicitly in Walenty – instead they are assigned in the process of conversion to dictionaries used by particular grammars (if need be: not all grammars use or need the notion of grammatical function). The third position of the schema is a singleton set containing an NP marked for instrumental case, `np(inst)`. The last position is a set which contains two elements: `prepnp(o,loc)` and `prepncp(o,loc,że)`. The first element is a PP which requires a certain preposition form (`o`) and a nominal marked for locative case. The second element is a PP which requires a locative correlative NP, a nominal taking a clausal complement (containing the complementiser `że`).

A schema featuring coordination of unlikes is provided in (2): the subject can be a structural NP, an interrogative clause (`cp(int)`) and two correlative NPs: taking an interrogative clause (`ncp(str,int)`) or a clause with `że` (`ncp(str,że)`) as its complement.

While plain categories defined in Walenty have no lexical restrictions on how they can be realised, there are three categories which are subject to such constraints: `fixed`, `lexnp` and `preplexnp`. Some examples featuring these categories are provided below:

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

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

In (3) there is a category `fixed`: it can only be realised as the string given as its argument – it must be *na kwaśne jabłko* (lit. "[beat sb] into a sour apple"), it does not accept modifiers (*na (*bardzo) kwaśne jabłko*, lit. "into a very sour apple"), it cannot appear in a different form (case, number, etc.). (4) contains two other categories mentioned above, `lexnp` and

`preplexnp`, where the nominal used in these phrases is restricted lexically – the required lemma is provided as the penultimate argument of the relevant category. Unlike in their unrestricted counterparts (`np` and `prepnp`, respectively), it is possible to constrain the number of the nominal (`sg`, `pl` or `_` – any number) as well as its requirements with regard to modification (the last argument). There are four possible modification constraints: `natr` means that modification is not possible, `atr` allows modification, `ratr` requires a modifier (often possessive: an NP or an adjective), `batr` requires specific possessive modifiers, namely forms of adjectives SWÓJ ("self") or WŁASNY ("own"). If modification is possible, agreement between the nominal and its modifier is handled by the grammar using standard agreement mechanisms: it takes into consideration the constraints stated in the valence dictionary (such as case and number).

Walenty was designed so as to store valence information in a way that can be used by various grammars – information from Walenty can be converted into an appropriate format used by a given grammar. Let us briefly discuss how entries related to MWEs are converted to the format of POLFIE, an LFG grammar of Polish.

(5) (↑ OBL PRED)$=_c$ 'NA KWAŚNE JABŁKO' ∧ ¬(↑ OBL ADJUNCT) ∧ ¬(↑ OBL POSS)

(6) (↑ SUBJ PRED)$=_c$ 'KREW' ∧ (↑ SUBJ CASE)$=_c$ NOM

(7) (↑ OBL PFORM)$=_c$ W ∧ (↑ OBL PRED)$=_c$ 'ŻYŁA' ∧ (↑ OBL CASE)$=_c$ LOC ∧ (↑ OBL NUM)$=_c$ PL ∧ [(↑ OBL ADJUNCT) ∨ (↑ OBL POSS)]

The constraints provided in (5) correspond to the syntactic position realised by `fixed` category in (3): (↑ OBL PRED)$=_c$ 'NA KWAŚNE JABŁKO' requires that the semantic form (PRED) of the given grammatical function (OBL) is the string NA KWAŚNE JABŁKO ("into a sour apple"). The remaining constraints ensure that no modifiers are attached to this element: no adjectival modifiers (¬(↑ OBL ADJUNCT)) and no possessive modifiers (¬(↑ OBL POSS)). In (4) there are two MWE arguments, they are used in the construction which translates literally as "(some) blood flows in sb's veins". Constraints related to the nominal subject are provided in (6): the first constraint restricts the form of the subject to KREW ("blood"), the second one requires nominative case marking from the subject (this is how structural case is realised in this context). There are no constraints related to modifiers in (6) because, according to the information provided in (4), this argument allows modification (`atr`), which is the default situation – no special constraints are needed. The constraints in (7) correspond to the prepositional argument which is the only possible realisation of the relevant position: (↑ OBL PFORM)$=_c$ W ensures that the appropriate preposition form is used (W), (↑ OBL PRED)$=_c$ 'ŻYŁA' checks that the semantic form of the nominal is ŻYŁA ("vein"), (↑ OBL CASE)$=_c$ LOC handles appropriate case marking of the nominal (locative), (↑ OBL NUM)$=_c$ PL makes sure that the plural form of the nominal is used. Finally, the last element, [(↑ OBL ADJUNCT) ∨ (↑ OBL POSS)], is a disjunction of two existential constraints which ensure appropriate modification: an adjunct or a possessive modifier is required (it corresponds to the `ratr` in (4)).

Bibliography

• Patejuk, A. and Przepiórkowski, A. (2012). Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, Istanbul, Turkey. ELRA. • Przepiórkowski, A., Skwarski, F., Hajnicz, E., Patejuk, A., Świdziński, M., and Woliński, M. (2014). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*. To appear.