

Phraseological information in an XLE/LFG grammar of Polish

Agnieszka Patejuk

aep@ipipan.waw.pl

Introduction

There is hardly any systematic work on the representation of multi-word expressions (MWEs) in theoretical LFG or in implemented LFG grammars,¹ and there are very few mentions of the LFG treatment of specific MWEs in particular languages.² The aim of this poster is to provide an – admittedly technical – presentation of a large scale treatment of phraseology in an XLE (Crouch *et al.* 2011) implementation of an LFG grammar of Polish.

The source of this MWE information is Walenty, a comprehensive valence dictionary of Polish (Przepiórkowski *et al.* 2014b; <http://zil.ipipan.waw.pl/Walenty>), which contains a rich phraseological component describing lexicalised arguments (Przepiórkowski *et al.* 2014a). So, the main difficulty is to appropriately translate this detailed information into XLE constraints.

Lexicalised arguments in Walenty

A rather complex example of a valence schema from Walenty, for the verb WITAĆ ‘welcome’, is given below:

```
(1) subj{np(str)} + obj{np(str)} +  
    {lex(preppnp(z,inst),pl,XOR('ramię','ręka'),  
      ratrl({lex(adjp(agr),agr,agr,'otwarty',  
        atrl({lex(advp(mod),'szeroko',natr)}))}))}
```

There are three arguments in (1): they are enclosed in curly brackets and separated by +. Only two uncontroversial grammatical functions are marked in Walenty explicitly: the subject and the object. Here, the schema includes both: the structurally-cased (usually nominative) NP subject (`subj{np(str)}`) and the structurally-cased (normally accusative, see Patejuk and Przepiórkowski 2014) object (`obj{np(str)}`). The third argument is lexicalised (`lex`).

Lexicalised arguments are formalised in Walenty in a uniform way: a metacategory, `lex`, takes any base category used in Walenty (e.g., `np` for a nominal phrase or `preppnp` for a prepositional phrase) as the first parameter, followed by parameters imposing constraints appropriate for the relevant base category (the number of such parameters depends on the base category) and finally the displayed modification pattern. Here, the lexical argument is a prepositional phrase (`preppnp`) headed by the preposition `Z` ‘with’ which takes an instrumental (`inst`) NP in the plural (`pl`). This NP must be headed by either a form (specifically, the instrumental plural form) of `RAMIĘ` ‘arm’ or a form (again, instrumental plural) of `REKA` ‘hand’. The final parameter of this `lex` specification expresses the complex information that this head noun *must* be modified (`ratrl`) by an adjectival phrase agreeing with the noun in case (`adjp(agr)`), as well as in number and gender (the other two occurrences of `agr`), and headed by an appropriate form of the adjective `OTWARTY` ‘open’. This adjective *may* in turn be modified (`atrl`) by an adverbial phrase headed by `SZEROKO` ‘widely’, which may not be modified any further (`natr`).

In summary, the valence schema in (1) describes the Polish version of the phraseological expression *somebody welcomes somebody with arms wide open* in a rather precise way and illustrates the three main modification types defined in Walenty which indicate the internal structure of such lexicalised arguments:

(2) no modification: `natr`

(3) optional modification: `atrl(1)`

(4) obligatory modification: `ratrl(1)`

(3)–(4) have two variants, allowing for or requiring, respectively: any number of modifiers, `(r)atrl`, or exactly one, `(r)atrl`. In (1) the lexicalised prepositional argument (`preppnp`) requires exactly one modifier (`ratrl`), i.e., an adjectival phrase (`adjp`) headed by `OTWARTY` ‘open’, which in turn may be modified by exactly one (`atrl`) adverbial phrase (`advp`) headed by `SZEROKO` ‘widely’, which cannot be further modified (`natr`).

Conversion: selecting the grammatical function

In order to use valence information from Walenty in an XLE/LFG grammar, all arguments must be assigned a grammatical function so as to impose relevant constraints. The same applies to lexicalised arguments whose modification pattern is formalised as in (1). However, as mentioned above, Walenty explicitly marks only two grammatical functions: the subject and the object.

Lexicalised arguments such as `preppnp`, `adjp` and `advp` in (1) are dependents of their respective heads. While all dependents of a verb are assumed to be its arguments (the lexicalised `preppnp`), this does not hold for embedded lexicalised arguments – `adjp` and `advp` in (1) – which are dependents of relevant predicates, but no information is provided concerning their argument/adjunct status.

The grammatical function linking the head with its dependent can be determined on the basis of their respective categories, in a similar way to the strategy which is used for arguments of verbs. In (1) the adjectival phrase (`adjp`) is the adjunct of the nominal head of the prepositional phrase (`preppnp`), while the adverbial phrase (`advp`) is an adjunct of the adjectival phrase (`adjp`).

¹Some preliminary work on MWEs in an Arabic implemented grammar is reported in Attia 2006.

²But see, e.g., Asudeh *et al.* 2013.

It may happen, however, that more than one grammatical function is possible – for instance, prepositional dependents of nominals may be an oblique argument or an adjunct. In such situations functional uncertainty may be used in the constraints – a disjunction of potentially relevant grammatical functions (see (6) below). When parsing, such underspecification should be narrowed down to one possibility by the lexical entry of the nominal head, which either takes an oblique or not.

Conversion: imposing constraints

The `natr` modification pattern (disallowing modification) is handled using negation coupled with functional uncertainty:

(5) `~(PATH GF)`

(6) `GF = {SUBJ|OBJ|OBL(-?*)|(X)COMP|...|ADJUNCT}`.

(6) defines `GF` as a disjunction of all possible grammatical functions. The constraint in (5) makes sure that the `f`-structure in `PATH` contains no grammatical function defined in (6), making it impossible to attach dependents. In (1) the `PATH` for `advp` is `^ OBL ADJUNCT $ ADJUNCT $` – an adjunct of an adjunct of the oblique argument.

The constraint provided in (7), where `DEP` is the grammatical function assigned to the modifier, corresponds to `ratr` (which requires modification):

(7) `(PATH DEP ATTR)=c val`

(8), which corresponds to `ratr1`, contains an additional constraint ensuring that there are no other dependents: the fragment `GF-DEP` removes `DEP` from the list of grammatical functions in (6), disallowing all except `DEP`.

(8) `(PATH DEP ATTR)=c val`
`~(PATH GF-DEP)`

However, when the grammatical function corresponding to the modifier is `ADJUNCT`, the constraint in (9) must be used instead of (8):

(9) `(PATH ADJUNCT $)=%DEP`
`(%DEP ATTR)=c val`
`~[(PATH ADJUNCT $) <h %DEP]`
`~[%DEP <h (PATH ADJUNCT $)]`

(9) assigns a local variable (`%DEP`) to an element of the adjunct set (first line), requiring that it satisfies relevant constraints (second line). The remaining lines ensure that there are no other elements of the adjunct set apart from `%DEP` using head precedence operator (`<h`) – the third line ensures there is no set element to the left of `%DEP`, the fourth line checks there is none to its right.

The constraints corresponding to optional modification patterns, `atr(1)`, are analogous to the ones provided in (7)–(9) for `ratr(1)` – the only difference is that the relevant constraints must be placed in curly brackets so as to ensure optionality, as shown for `atr` in (10) (compare with `ratr` in (7)):

(10) `{(PATH DEP ATTR)=c val}`

Conclusion

For reasons of space, only the very basics of the procedure of converting the lexicalised information provided in Walenty into XLE constraints could be presented here. The procedure is currently being applied to the almost 6000 valence schemata of Walenty which contain lexicalised arguments (i.e., almost 10% of all Walenty schemata), to complement an earlier effort to convert non-lexicalised Walenty information to the XLE/LFG grammar of Polish. This way the XLE/LFG grammar of Polish is becoming one of the largest such grammars, comparable to the grammars of English, German and Norwegian (of which, to the best of our knowledge, only the last one is still actively developed), and probably the only one that contains such detailed and extensive phraseological information.

References

- ◆ Asudeh, A., Dalrymple, M., and Toivonen, I. (2013). Constructions with Lexical Integrity. *Journal of Language Modelling*, 1(1), 1–54.
- ◆ Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, eds., *Advances in Natural Language Processing: Proceedings of the 5th International Conference on Natural Language Processing, FinTAL 2006, Turku, Finland*, pp. 87–98, Berlin Heidelberg. Springer-Verlag.
- ◆ Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE documentation. http://www2.parc.com/isl/groups/nlft/xle/doc/xle_toc.html.
- ◆ Patejuk, A. and Przepiórkowski, A. (2014). Structural case assignment to objects in Polish. In M. Butt and T. H. King, eds., *The Proceedings of the LFG'14 Conference*, pp. 429–447, Stanford, CA. CSLI Publications.
- ◆ Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014a). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pp. 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.
- ◆ Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014b). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pp. 2785–2792, Reykjavík, Iceland. ELRA.