(WG 1 and WG 2)

Separating the regular from the idiosyncratic: An object-oriented lexical encoding of MWEs using XMG

Timm Lichte[1], **Yannick Parmentier**[2], **Simon Petitjean**[1], **Agata Savary**[3] **& Jakub Waszczuk**[3]
[1]CRC 991, University of Düsseldorf, Germany
[2]Université d'Orléans, France
[3]Université François Rabelais Tours, France

Abstract We present a general object-oriented approach to the lexical encoding of multi-word expressions (MWEs) that is couched into the framework of eXtensible MetaGrammar (XMG). We think that XMG provides the flexibility and power needed to account for both regular and idiosyncratic aspects of MWEs, which enables the lexicographer to encode MWEs in a transparent and yet factorized way. We compare XMG with two other existing formats for lexical encoding of MWEs, DuELME and Walenty, which have been coupled with real-size grammars and provide mechanisms to avoid description redundancy. We claim that XMG offers additional facilities that reinforce the virtues of its competitors. In this work we confine ourselves to syntax and morphology.

DuELME DuELME (Dutch Electronic Lexicon of Multiword Expressions, [4]) is an electronic lexicon comprising roughly 5000 Dutch multiword expressions. An example entry for *zijn kansen waarnemen* ('to seize the opportunity') is shown in Figure 1. DuELME distinguishes two sorts of descriptions, pattern descriptions and MWE descriptions, which are composed of non-intersecting sets of predefined fields. Pattern descriptions contain regular templates of syntactic structure (see `PATTERN` in line 4), which can be referred to in the MWE descriptions (see the field `PATTERN_NAME` in line 10). However, there is no such notion of reference, or reuse, among the 141 pattern descriptions that DuELME comprises [3]. Hence this distinction between patterns and MWE descriptions introduces only some limited degree of factorization, i.e., the inheritance hierarchy is bound to depth two. Moreover, neither the full set of syntactic constraints (e.g. linearization and diathesis) nor any semantic content can be expressed.[1] Another shortcoming gets evident in the MWE description in Figure 1: One would like to express that the subject and the possessive determiner of the object agree in person, number and gender. This cannot be expressed by enforcing the equality of "parameters" (i.e. the features enclosed by square brackets in line 9) by, e.g., the use of variables. Yet there is a special feature available in DuELME to hold the "binding type" of a pronoun [2, Table 5].

Walenty Walenty is a Polish large-scale valence dictionary offering a rather expressive formalism [5] including notably an elaborate phraseological component [6]. Figure 4 shows a sample MWE entry of (1), which exhibits several interesting constraints and idiosyncrasies.

(1) dobrze [KOMUŚ] z oczu patrzy
well someone.DAT from eyes.GEN looks
'Someone looks like a good person.'

Firstly, the syntactic subject is prohibited here although the head verb *patrzeć* 'look' does take a subject as a stand-alone verb. This fact is expressed in Walenty simply by omitting the `subj` argument in the valence frame. Secondly, the adverb *dobrze* ('well', encoded in Figure 4 by a more generic, non lexicalized, `advp(misc)` requirement of a "true" adverbial clause) should usually precede the prepositional complement and the verb. However, linearization constraints cannot presently be expressed in Walenty, even though a conservative extension of the formalism to include them was proposed by [5]. Thirdly, while the indirect object can typically be skipped, it is compulsory in this MWE. It seems that this fact is covered by simply including the `np(dat)` argument in the entry. Fourthly, several morphological constraints arise. The verb *patrzeć* ('look') is always in the 3rd person singular (any tense or mood), although it has a complete inflection paradigm as a stand-alone verb. Such paradig-

[1]In DuELME, syntactic constraints can be expressed implicitly by assigning special patterns whose implicit meaning is somehow known to the NLP system.

matic constraints imposed on the head verbs cannot currently be expressed in Walenty. Since, however, impersonal finite verbs typically occur in the 3rd person singular in Polish, the expression of this fact is probably left to the grammar. Finally, within the lexicalized prepositional group (`lex(prepnp(...))`, which does not admit modification (`natr`), the preposition *z* ('from') requires its nominal complement *oczy* ('eyes') to be in genitive plural (`((z,gen),pl,'oko')`.

This brief case study shows that the Walenty format seems to offer sufficient means to encode many properties of MWEs, even challenging ones. Still, Walenty does not allow for the encoding of word order constraints, and it leaves the borderline between regular and idiosyncratic properties rather implicit.

eXtensible MetaGrammar The framework of eXtensible MetaGrammar (XMG, [1]) provides description languages and dedicated compilers for generating a wide range of linguistic resources.[2] Descriptions are organized into CLASSES, alluding to the class concept in object-oriented programming. Similarly, classes have encapsulated name spaces and inheritance relations may hold between them. The crucial elements of a class are DIMENSIONS. They can be equipped with specific description languages and are compiled independently, thereby enabling the grammar writer to treat the levels of linguistic information separately. In the following we will be using the standard dimension `<syn>` for the syntax, skipping over other available dimensions for descriptions of semantic representations or morphological structure. Note that `<syn>` contains tree descriptions where nodes may carry untyped feature structures.

Figure 2 shows part of a tentative XMG encoding of the Dutch MWE *zijn kansen waarnemen*. First thing to notice when comparing it to the DuELME counterpart in Figure 1: there is no principled distinction between "patterns" and "MWE descriptions". Rather they are equally represented as classes, yet of varying specificity. Crucially, the classes stand in inheritance relations, here marked with the import statement. For example, the most basic class shown in Figure 2, `intransitive[]`, imports two other classes, `subject[]` and `verb[]` (see line 6). On the other hand, `intransitive[]` is further handed down to `transitive[]`, just adding `object[]`. Finally, `transitive[]` gets imported into `zijn_kansen_waarnemen[]`, which is the class of the MWE. Hence, `transitive[]` contains the regular properties of the MWE, and `zijn_kansen_waarnemen[]` the idiosyncratic ones. The corresponding inheritance hierarchy of the classes is shown in Figure 3. In general, classes that correspond to irregular properties of lexical entries appear as leaves, whereas regular aspects are assigned to dominating classes.[3] Hence, "patterns" can be arbitrarily factorized, which is in sharp contrast to the DuELME encoding format. Another difference is the general availability of variables in XMG, which are commonly prefixed with a question mark. This is exploited in `zijn_kansen_waarnemen[]` when expressing agreement between the subject and the possessive determiner using the variables `?NUM`, `?PERS`, and `?GEND` (see line 31 and 33). Note that features and variables can be freely added to XMG, for example features to indicate constraints on modification (`modifiable`) or passivization.

The preliminary XMG encoding of the Polish MWE *dobrze* [*KOMUS*] *z oczu patrzy* is presented in Figure 5. Again, the class that corresponds to the MWE, `dobrze_z_oczu_patrzy[]`, inherits from more abstract (and "regular") classes, which can be also seen from the inheritance hierarchy in Figure 6. Here, the `impers_intransitive[]` class encodes the fact that the subject is absent (as only the verb phrase and its subordinate verb are listed), and that the (impersonal) verb must occur in the third person singular. The `impers_intransitive_IndObj_PP[]` class expresses the requirement of a prepositional complement and of a direct object dominated by the verb phrase. Finally, the `dobrze_z_oczu_patrzy[]` class reuses the previous class and adds the compulsory adverb. Moreover, certain nodes, identified by shared variables, are further specified for lemmas (specified between double quotes) and all idiosyncratic morphological constraints are listed. Notably, the noun governed by the preposition *z* 'from' is restricted to the lemma `oku` 'eye' and to plural, and its modification is prohibited. Note that the genitive case of *oko* is not specified in this class, as it is imposed by agreement rules inherited from the `prep_compl[]` class. Finally, lineariza-

[2] `https://sourcesup.cru.fr/xmg/`

[3] This is reminiscent of type hierarchies in HPSG. However, the lexical entries proposed there seem far from being theory-neutral. It remains to be seen whether and how HPSG could be used as a general encoding format.

tion constraints on the adverb appear in lines 29–30, with >>+ being the transitive, non-reflexive precedence operator (recall that neither the encoding format of DuELME nor the one of Walenty includes precedence operators). Thus, all the necessary constraints imposed on this MWE can be covered at various abstraction levels, while factorizing information in such a way that the dobrze_z_oczu_patrzy[] class only contains the constraints which are specific to the MWE.

Note that XMG comes with a solver for these classes, and a viewer. Hence the solutions can be inspected independently of a specific application belonging to some specific framework.

Prospects In future work we want to extend the coverage of the XMG descriptions in order to see the benefit of factorization more clearly, and also address the semantics of MWEs using the semantic dimensions that are already available in XMG.

References

[1] Crabbé, B., D. Duchier, C. Gardent, J. Le Roux & Y. Parmentier. 2013. XMG: eXtensible MetaGrammar. *Computational Linguistics* 39(3). 1–66.

[2] Grégoire, N. 2007. *MWE lexicon for Dutch: Encoding protocol*.

[3] Grégoire, N. 2007. *MWE lexicon for Dutch: Overview of pattern descriptions*.

[4] Grégoire, N. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1–2). 23–39.

[5] Przepiórkowski, A., J. Hajič, E. Hajnicz & Z. Urešová. To appear. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography* .

[6] Przepiórkowski, A., E. Hajnicz, A. Patejuk & M. Woliński. 2014. 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)*, 83–91. Dublin, Ireland.


```
1  % Pattern description
2  PATTERN_NAME ec1
3  POS d n v
4  PATTERN  [.VP [.obj1:NP [.det:D (1) ]
5    [.hd:N (2) ]] [.hd:V (3) ]]
6
7  % MWE description
8  EXPRESSION zijn kansen waarnemen
9  CL zijn kans[pl] waar_nemen[part]
10 PATTERN_NAME ec1
```

Figure 1: DuELME pattern description ec1 (from [3]) and MWE description of *zijn kansen waarnemen* ('to seize the opportunity', from [4])

```
1  %%%%%%%%%%%
2  % PATTERNS %
3  %%%%%%%%%%%
4
5  class intransitive
6  import subject[] verb[]
7  { <syn> {
8      ?Subj >>+ ?V
9  } }
10
11
12 class transitive
13 import intransitive[] object[]
14 { <syn> {
15     ?Subj >>+ ?Obj;
16     ?Obj >>+ ?V
17 } }
18
19 %%%%%%%
20 % MWE %
21 %%%%%%%
22
23 class zijn_kansen_waarnemen
24 import transitive[]
25 declare ?NUM ?PERS ?GEND
26 { <syn> {
27     ?Subj[num=?NUM,pers=?PERS,gend=?GEND];
28     ?Obj [] {
29       [cat=d,num=pl,possnum=?NUM,pers=?PERS,
                gend=?GEND] "zijn"
30       [cat=n,modifiable=-,num=pl] "kans"};
31     ?V[] "waar_nehmen"
32 } }
```

Figure 2: XMG encoding of *zijn kansen waarnemen* ('to seize the opportunity')

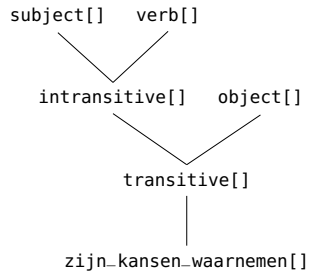

Figure 3: Inheritance hierarchy of XMG classes according to the code in Figure 2

```
patrzeć:  np(dat)+advp(misc)+lex(prepnp(z,gen),pl,'oko',natr)
```

Figure 4: Description of *dobrze* [KOMUŚ] *z oczu patrzy* ('someone looks like a good person') in Walenty

```
%%%%%%%%%%%
% PATTERNS %
%%%%%%%%%%%
class impers_intransitive
export ?VP ?V
declare ?VP ?V
{ <syn>{
    ?VP [cat=vp] { ?V [cat=v,pers=3,num=pl] }
} }

class impers_intransitive_IndObj_PP
import impers_intransitive[] indir_object[]
    prep_compl[]
{ <syn> {
    ?VP -> ?PP;
    ?VP -> ?IndObj
} }

%%%%%%%
% MWE %
%%%%%%%
class dobrze_z_oczu_patrzy
import impers_intransitive_IndObj_PP[]
    adverb[]
{ <syn> {
    ?AP [] { ?A [] "dobrze"};
    ?PP [] {
      [cat=p,case=gen] "z"
      [cat=np] { [cat=n,num=pl,modifiable=-]
          "oko" }};
    ?V "patrzeć";
    ?AP >>+ ?PP;
    ?AP >>+ ?V
} }
```

Figure 5: XMG encoding of *dobrze* [KOMUŚ] *z oczu patrzy* ('someone looks like a good person')

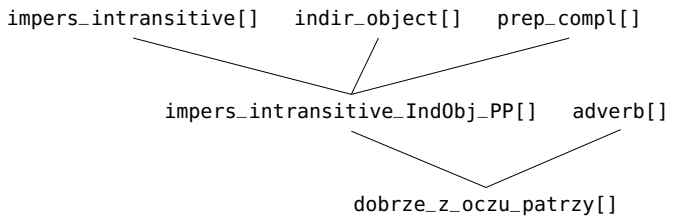

Figure 6: Inheritance hierarchy of the XMG classes in Figure 5