Constraint Based Description of Polish Multi-word Expressions

Roman Kurc, Maciej Piasecki, Bartosz Broda

Institute of Informatics, Wrocław University of Technology, Poland
roman.kurc,maciej.piasecki,bartosz.broda@pwr.wroc.pl

Abstract

We present an approach to the description of Polish Multi-word Expressions (MWEs) which is based on expressions in the WCCL language of morpho-syntactic constraints instead of grammar rules or transducers. For each MWE its basic morphological form and the base forms of its constituents are specified but also each MWE is assigned to a class on the basis of its syntactic structure. For each class a WCCL constraint is defined which is parametrised by string variables referring to MWE constituent base forms or inflected forms. The constraint specifies a minimal set of conditions that must be fulfilled in order to recognise an occurrence of the given MWE in text with high accuracy. Our formalism is focused on the efficient description of large MWE lexicons for the needs of utilisation in text processing. The formalism allows for the relatively easy representation of flexible word order and discontinuous constructions. Moreover, there is no necessity for the full specification of the MWE grammatical structure. Only some aspects of the particular MWE structure can be selected in way facilitating the target accuracy of recognition. On the basis of a set of simple heuristics, WCCL-based representation of MWEs can be automatically generated from a list of MWE base forms. The proposed representation was applied on a practical scale for the description of a large set of Polish MWEs included in plWordNet.

Keywords: multi-word expression representation, multi-word expression recognition, morphosyntactic constraints, WCCL, plWordNet, Polish

1. Motivations

Contemporary state-of-the-art morphological analysers and taggers provide precise morphological description for almost every single words in many different languages. At the same time research on large lexical resources shows that multi word expressions (MWE) constitute a substantial part of the lexicon, and thus their effective large scale description and recognition in text is important for Language Technology applications, e.g. Information Extraction and Question Answering. In facat (Sag et al., 2002) claims that MWEs are at least as numerous as single words and (Orliac and Dillinger, 2003) emphasised that MWE are "the key to producing more acceptable output" in Machine Translation. MWE often express syntagmatic relations between word forms, e.g. in frozen expressions, collocations, compound nouns, phrasal verbs, idioms etc. There is no strict definition of MWE. However MWE are described as sharing the following characteristics (Savary, 2008): "they are compose of two or more graphical words; they show some degree of morphological, syntactic, distributional or semantic non-compositionality; they have unique and constant references".

Two major problems emerge in the context of MWE: *discovery* and *recognition*. The task of discovering MWE has been already well studied. In general it is possible to either use dictionaries (but they usually give a very limited coverage of MWEs) or discover MWEs in corpora with the help of statistical methods. The latter approach seems to be more appropriate in the real life applications because of its suitability for a large scale processing of text. However it comes at a cost of lower quality of MWEs. Whatever is the result of discovery phase one has to be able to recognize multi-word expression. We will address this problem in this paper.

MWE recognition is a procedure of marking a sequence of tokens in text as representing an already known MWE described in a kind of MWE dictionary. In our work, we aim at a lexicon-based recognition method which is application oriented, i.e. is focused on efficiency and flexibility. Both when text is processed and also when MWE descriptions are prepared. Due to the existence of large coverage morphological analysers (e.g. Morfeusz SGJP for Polish[1]) we decided to provide only structural features in the MWE description without keeping the whole morphological analysis. There are many approaches to this task, see Sec. 2., and while following the main lines identified in the literature, we would like to propose a constraint-based description of MWE lexico-syntactic structure. The proposed format is based on the WCCL (Wrocław Corpus Constraint Language) language (Radziszewski et al., 2011) of lexico-morpho-syntactic constraints. It has originated from a language of tagging rules but was finally extended for the needs of shallow parsing, shallow semantic analysis and Information Extraction, in general.

2. Formats for describing Multi-word Expressions

Description formats proposed for MWE can be divided into several classes. They are either based on lexicon, unification grammar or finite state automata (FSA). In lexicon based approaches, e.g. developed at (Laboratoire d'informatique documentaire et linguistique, University of Paris 7, France), the DELA (Dictionnaires èlectroniques du LADL), lists of inflected MWEs are generated. Inflection is controlled by detailed formal description prepared by linguists. The description specifies morphological characteristics of constituents, together with operators that can be applied to transform MWE constituent base forms into inflected forms. A typical FSA based description consists of simple regular expressions that model single words. The

[1] http://sgjp.pl/morfeusz/

regular expressions are then combined to form more elaborate descriptions of MWEs. Examples of FSA-based format are: Lexc[2] or IDAREX [3] that was build over the Lexc. Unification grammar enables MWE description in terms of lexico-syntactic structures enriched with variables mediating different forms of structural depedencies, e.g. agreement. Examples of the this class are: LinGo (Sag et al., 2002), (Copestake et al., 2002), (Villavicencio et al., 2004) or FASTR (Jacquemin, 2001).

3. Phenomena in MWE inflection

(Savary, 2008) pointed to a number of language phenomena that are relevant for MWE recognition.

3.1. Morphosyntactic Compositionality

When MWE is compositional, its inflectional properties can be deduced from MWE constituents. In addition, one of the MWE constituents determines morphological properties in context of a sentence for the whole MWE. The constituent is called a head word and can substitute the whole MWE in a sentence. In Polish language compositionality is typical for compounds. For example *żywy trup* 'living dead'. The word *trup* 'dead' is a head in this MWE and *żywy* 'living' is agreed with the head i.e. they share the same case and number. When MWE is non-compositional one has to consider the MWE as a whole i.e. inflection must be considered on the lexical level.

3.2. Morphosyntactic non-compositionality

Most MWE are at least a bit non-compositional. That means there is a need for an appropriate handling of the variations at the morphosyntactic level. The simpliest case, exocentric MWE fe. *Piotr i Paweł* 'Peter and John', does not have a head word. As far as both constituents have the same gender, case and number, neither of them can be considered a main element of the MWE. Next there may occur irregularities in the agreement inside MWE, e.g. noun constituents that are in apposition are expected to aggree but they only have the same case. F.e in *majster klepka* 'handyman' headword *majster* has different gender than the other constituent. Majster is masculine and klepka is feminin noun. Last type of morphosyntactic non-compositionality is caused by a defective inflection paradigm of the whole MWE. Such MWEs my not have some of the inflected forms or when the form is used it may not bear the same meaning, e.g. a *zimne ognie* 'fireworks'. *zimny ogień* does not really make any sense.

3.3. Inflection and variation

Savary states that MWEs undergo a more general phenomenon i.e. terminological variation. It may be understood as a partial indepedence of constituents that can form the MWE. The phenomenon is described on the lexical level as:

- insertion i.e. an additional constituent in MWE can be used but will not change the basic meaning of the MWE;
- omission i.e. a constituent can be skipped from the basic form of MWE without a change in the meaning – *nauczyciel języka angielskiego*'teacher of the English language' and *nauczyciel angielskiego* 'English teacher'. Whearas the first MWE is a complete title of profession, the other is still correct and is easily recognized;
- order change f.e. *areszt tymczasowy* 'executive detention' is usually the same as *tymczasowy areszt*;
- derivational transformation;
- semantically motivated replacements and abbreviations i.e. using acronisms and initializms.

Different variations can occur together.

3.4. Inflectional paradigm and base form

An inflectional paradigm for a highly inflected language can contain many word forms for a MWE. For example the inflectional paradigm of Polish adjectives has got about hundred forms bacuse a form of an adjective differs depending on gender, number, case. It is similar for other gramatical classes. As MWE can have many constituents, therefore we come to the conclusion that a MWE description must be as compact as possible. It means that we need to avoid preparing exhaustive lists of all word forms for an MWE. It seams that having contemporary state-of-the-art morphological analysers and taggers at hand one can describe MWE using only lemmas and morphological constraints imposed on the MWE constituents.

3.5. Discontinuous MWEs

Last phenomenon that is associated with MWEs, mostly those containing verbs, are 'gaps' i.e. tokens that do not belong to a given MWE may be mingled with occurrences of the constituents. For example *wolna (..) wola* 'free will' can be easily separated by other adjectives f.e. *wolna i niczym nie skrepowana wola*. Such continuous MWEs are difficult to recognize because they can span over other tokens or phrases.

4. WCCL

WCCL originated from a language of morpho-syntactic tagging rules and works primarely on the level of word tokens and word-to-word relations. However, token sequences can be also marked as chunks and later referred to in WCCL expressions. Both boolean constraints and rules for tagging (syntactic and semantic), as well as for tag elimination can be expressed in WCCL. Constraints written in WCCL can be used as a source of knowledge in Machine Leaning, e.g. in chunking or Named Entity recognition.
A detailed description of WCCL would take a lot of space, the most important WCCL properties were mentioned below and a more detailed description can be found in (Radziszewski et al., 2011). WCCL enables:

[2] http://www.cis.upenn.edu/ cis639/docs/lexc.html

[3] The formalism and Finite State Compiler have been developed at Rank Xerox Research Centre by L. Karttunen, P. Tapanainen and G. Valetto

- accessing values of morpho-syntatic features of individual tokens such as case or gender (set of values in the case of non-disambiguated tokens),
- testing different forms of morpho-syntactic agreement supported by built-in operators,
- iterate across token sequence and write constraints sensitive to properties of token sequencies,
- express complex constraints based on combining simpler constraints and with the help of variables of different types,
- provide access to variable values on the outside,
- apply filtering based on frequency lists,
- transform word forms and lemmas on the basis of user-supplied dictionaries.

WCCL can work with any positional tagset (tagset attributes automatically become valid functions). A MWE can be described by complex constraints referring to both: lemmas and/or word forms of constituents, as well as to relations between tokens corresponding to MWE constituents. WCCL-based description allows for discontinues MWE occurrences and linear order variants. Examples of WCCL expressions will be presented and explained in the next section.

5. CB-MWE format - CCL based description format

Three aspects of MWE description must be taken into account when our goal is to recognise MWE occurrences in text:

1. word forms (inflectional) of MWE constituents,
2. linear order of constituents,
3. and MWE sequence continuity.

Concerning the MWE word forms, we assume that complete morphological descriptions of MWE occurrence constituents can be read from the results returned by the morphological analyser. What is left is to check whether the given sequence of word forms represent really a given MWE. Thus we must verify presence of the certain lemmas and those morphological features that are important for the MWE lexico-syntactic structure. Each MWE can be expressed by a complex WCCL constraint. As our analysis showed, MWE lexico-syntactic structures for Polish can be grouped into a limited number of classes (at least MWEs described in plWordNet and proper names in a huge gazetteer), their structural properties can be expressed by a limited number of templates of WCCL complex constraints parametrised by MWE lemmas and word forms from the basic form of MWE.

The second and the third aspects are also encompassed by the complex WCCL constraint template. However, application of WCCL constraints to text is time consuming, and the recognition can be more efficient, when the constraints are applied only in selected areas of text. On the basis of MWE lemma sequence and knowledge whether the sequence is *fix* or *flexible* – the second aspect, as well as, whether the given MWE allows for 'gaps' inside its occurrences – the third aspect, an efficient pre-recognition of potential MWE occurrences can be performed. Only next, WCCL constraints are run for pre-selected text areas.

A MWE description based on a formal grammar could also be used for MWE structural description, but constraint based approach gives more flexibility in simplifying the description (e.g. in terms of workload) and encompassing by it only those MWE properties that are crucial for its recognition in text.

The WCCL based format for MWE (CB-MWE) is encoded in XML. Description schema for a MWE is as following:

```
<mwegroup type='...' name='...' class='...'>
   <condition>
     ...
   </condition>
   <instances>
       <MWE/>
   </instances>
</mwegroup>
```

where `mwegroup` groups the whole set of MWE described with the same WCCL constraint template with the same inflection pattern. The `type` attribute of `mwegropup` defines the MWE linear order (values: *fix* and *flex*). Fixed MWEs have strict order and must be continuous. Flexible MWEs may include 'gaps' filled with other tokens. The `name` attribute represents the name of the whole group. The `class` represents grammatical class (generalised Part of Speech) associated with the whole MWE. The WCCL constraint template is expressed in `condition`. Constraint occurring in the condition section is used to determine the expected behaviour and the dependencies between those tokens in the sentence that correspond to the MWE constituents, e.g. morphological features, word order, morphological arrangement between tokens or across token sequences. The `instances` section contains a list of MWEs described by the constraint in `condition` tag.

For instance, MWE *chleb powszedni* 'daily bread' was included into the class:

```
 <mwegroup type='flex' name='SubstAdjPlFlex'
           class='subst'>
```

This MWE is defined as `flex` as both: *chleb powszedni* and *powszedni chleb* are acceptable. The `SubstAdjPlFlex` constraint is presented below.

Each MWE description is parametrised by its base form, WCCL expression for head word recognition and a list variables for the WCCL constraint template. The description for *chleb powszedni* is presented below:

```
<MWE base='chleb powszedni'>
  <head>in(class[0],{subst,ger,depr})</head>
  <var name='S'>chleb</var>
  <var name='A'>powszedni</var>
</MWE>
```

The `base` attribute is a MWE lemma (a morphological base form) as used in lexicons, e.g. plWordNet. A head

word is defined as the leftmost element of MWE that satisfies the WCCL expression in `head`. Each variable on the list is identified by a name. The name is then used in the condition section. A variable can be of string, boolean or numeral type.
The above MWE is structurally described by the constraint presented below with instantiated by the string variables: `$s:A` and `$s:S`

```
<condition>
or(
 and(
  inter(base[1],$s:A),
  inter(class[1],adj),
  inter(base[0],$s:S),
  inter(class[0],subst,ger,depr),
  agrpp(0,1,nmb,gnd,cas),
  setvar($Pos1, 0),
  setvar($Pos2, 1)  ),//and
 and(
  inter(base[0],$s:A),
  inter(class[0],adj),
  inter(base[1],$s:S),
  inter(class[1],subst,ger,depr),
  agrpp(0,1,nmb,gnd,cas),
  setvar($Pos1, 1),
  setvar($Pos2, 0)   )
)//or
</condition>
```

In the above constraint the `or` operator was applied to describe two possible linear orders of a MWE. In both variants, lemmas of the first two words are compared with the string variables (constraint parameters) and the grammatical classes with the values expected. Because the applied morpho-syntactic tagger can leave more than one tag per word, the intersection (`inter`) between the expected values and values assigned to a word is checked. The `agrpp` operator tests morpho-syntactic agreement between two tokens at the specified positions with respect to: number (`nmb`), gender (`gnd`) and case (`cas`). Finally, the successfully identified positions are assigned to the output variables to make them externally readable.
MWE *kobieta . . . życia* lit. ‘woman of life’ – ‘a woman of (sombody’s) life’ – is an example of an MWE with a gap inside that can be filled with expressions of the limited types.

```
<mwegroup type=’fix’ name=’SubstSubstGenGapFix’
         class=’subst’>
<condition>
  and(
   inter(base[0],$s:S),
   inter(class[0],subst),
   rlook(1,end,$G, and(
         inter(base[$G],$s:SG)
         inter(class[$G],subst),
         inter(cas[$G],gen),
 )),
   only(1,$G-1,$N, or(
      inter(class[$N],adv,qub),
and(
  or( inter(class[$N],adj,ppas,pact),
      inter(base[$N],"mój") ),
  agrpp($G,$N,nmb,gnd,cas)
))),
  setvar($Pos1, 0),
  setvar($Pos2, $G)
 )
</condition>
<instances>
 <MWE base=’kobieta życia’>
   <head>in(class[0],subst,ger,depr)</head>
   <var name=’S’>kobieta</var>
   <var name=’SG’>życie</var>
 </MWE>
</instances>
</mwegroup>
```

In the above constraint , the sequence is expected to be started by *kobieta*, but next we are searching along the rest of the sentence for the second constituent. The `rlook` iterates across words till the end of the sentence (`end`) until the internal condition (built by `and`) is not fulfilled, i.e. the appropriate noun in the genitive case has not been found. Next, a potential gap is checked with the help of `only` operator. Here, only adverbs (`adv` and `qub`) are expected or pronoun *mój* ‘mine’ and adjectives that agree with the second constituent.
MWE, especially verbal, can impose conditions on the nearest syntactic context, e.g. *dobre wyjście na . . .* ‘coming off well out of something’ introduces an open position in the prepositional phrase that must be filled with a noun phrase of the appropriate case. Due to the WCCL our MWE representation provides means to specify such mutual dependencies. For instance:

```
<mwegroup type=’fix’ name=’AdjSubstPrepFix’
          class=’subst’>
<condition>
  and(
    inter(base[0],$s:A),
    inter(class[0],adj),
    inter(base[1],$s:A),
    inter(class[1],subst),
    agrpp(0,1,nmb,gnd,cas)
    equal(base[2],$s:P),
    equal(class[2],prep),
    rlook(3,end,$N,
        in(base[$N],subst,ger,
               depr,ppron3,fin,praet,imps)
 )
    in(base[$N],subst,ger,depr,ppron3),
    inter(cas[$N],loc),
    setvar($Pos1, 0),
    setvar($Pos2, 1),
    setvar($Pos3, 2)
  )
</condition>
<instances>
  <MWE base=’dobre wyjście na’>
     <head>in(class[0],subst,ger,depr)</head>
     <var name=’A’>dobry</var>
     <var name=’S’>wyjście</var>
     <var name=’P’>na</var>
  </MWE>
</instances>
</mwegroup>
```

Table 1: comparision with Multiflex fotmat

	Multiflex (2005)	CB-MWE (2011)
Exocentric MWEs	√	√
Irregular agreement	√	√
Defective paradigms	√	√
Insertions and omissions	√	√
Order change	√	√
Duplications	√	
Derivational variants		√
Semantic variants		√
Abbreviations		√
Unification	√	√
Non-abstract lemmas	√	
Non-contiguous MWEs		√
Non-redundancy	√	
Infl. analysis	√	√
Infl. generation	√	
Automated MWE lexicon creation	√	√
Sense computation		
Formal tool	graphs, FSTs	constraint language
Number of MWEs described	2 822	6 954
Language	Serbian	Polish

6. Comparison with other formats

In order to assess the expressive power of CB-MWE, we compared it with the Multiflex format. The comparison was done for the set of phenomena identified in (Savary, 2008).

Multiflex provides detailed morphological description of MWE and its constituents that is expanded to the full list of MWE word forms. The list is next used as a basis for MWE recognition in text. In our approach only lexico-morpho-syntactic constraints are stored that are next applied to the morphological analyses of tokens in text produced by a morphological analyser compatible with the assumed tagset. There is no need to generate all MWE forms in advance.

Due to the constraint-based representation of MWEs in our format, *exocentric*, irregular agreement and defective paradigm MWE can be all described by complex lexico-morphosyntactic constraints. A similar result is obtained in Multiflex by providing detailed description of inflection rules connected with each MWE.

Insertions, *omissions* and *order changes* in MWEs are possible to be modelled in CB-MWE due to the presence of 'and' and 'or' operator in WCCL. The operators allow us to describe many permutations of input words in one condition. So, e.g., change of order is described by writing a number of constraints for every valid sequence of words of the given MWE. In Multiflex these phenomena are easier to be modelled because morphological descriptions of single units are connected in graph like fashion. Therefore insertion and omission is simply encoded as adding a link between non-consecutive single units.

However, there are some phenomena that cannot be modelled by Multiflex, but are easy to be expressed in our format – namely: *abbreviations*, and also *semantic and derivational variants*. They may be expressed by' forming a set of candidate words for a position out of variables feed to a complex WCCL constraint (`condition`). Moreover, special WCCL iteration operators: *llook* and *rlook* facilitates description of non-continuous MWEs, as they enable skipping tokens not belonging to the given MWE and long distant search for all constituents (see the last example in 5.)
.

Duplication can be expressed using the mechanism of regular expression embedded in WCCL constraints.

We gave a deep thoughts to redundancy and it seems that we could express conditions in modular manner. This way we could avoid redundancy which is now apparent as we describe many times similar conditions varying only on one or two positions.

We always use non-abstract lemmas i.e. base forms from the dictionary.

7. Performance

Precision and recall of our hand written MWE descriptions were evaluated on the IPI PAN Corpus (Przepiórkowski, 2004) [4]. We prepared baseline using MWEs' constituents i.e. their base and word forms. From the corpus we kept sentences in which we found all constituents of at least one MWE (about 10mln sentences).

Next, a sample of 400 (10% level of confidence) sentences was randomly selected out of the non-discarded sentences. This step was performed in order to check whether the baseline contains errors i.e. sentences that actually did not contain any MWE. We found out that such kind of filtering was very lenient and results contained only 20% of sentences with MWEs.

Then we applied MWE descriptions to the remaining sentences. For estimating precision we used all the sentences where instances of MWEs were found. We selected a random sample of 400 (10% level of confidence) sentences and checked if marked tokens formed a complete MWE. Recall was calculated as a ratio between all sentences thought to have correctly marked (based on precision) MWE and the baseline. The precision was at the level of 99%, and recall (regarding the the number of correct sentences containing MWEs in baseline) reached the level of 61%.

8. Conclusions

In this article we presented a novel, constraint-based format, called CB-MWE, for the description of Multi-word Expressions (MWE) applied to Polish MWEs. Its comparison with Multiflex was discussed. We showed advantages and disadvantages of our format. CB-MWE does not require generation of the whole dictionary of all MWE inflected forms and utilises morphological data from the tagged texts. In addition it enables expressing discontinuous MWE. The format has been already integrated with our

[4] ≈250 mln. tokens

tools for corpus analysis i.e. SuperMatrix, NER and Chunker. Next, we evaluated MWE descriptions in CB-MWE prepared by linguists in terms of the recognition precision and recall. The experiments showed that our approach is sufficient for real life applications. We plan to work on automation of the description process. Preliminary results seems to be promising. However we have still to solve problems related to recognition of the order and continuity of MWEs. We would also like to be able to merge syntactic and semantic variants as well as abbreviations.

Acknowledgments

Financed by the Polish National Centre for Research and Development, project SyNaT.

9. References

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bondþ, Timothy Baldwiný, Ivan A. Sagý, and Dan Flickingerý. 2002. Multiword expressions: Linguistic precision and reusability. In *In Proc.*

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*, volume 10. The MIT Press, New York, NY, USA, June.

Brigitte Orliac and Mike Dillinger. 2003. Collocation extraction for machine translation. In *Machine Translation Summit IX*.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science PAS.

Adam Radziszewski, Adam Wardynski, and Tomasz Sniatowski. 2011. Wccl: A morpho-syntactic feature toolkit. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 434–441. Springer.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *CICLing*, pages 1–15.

Agata Savary. 2008. Computational inflection of multiword units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 0(0).

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of mwes. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, pages 80–87, Stroudsburg, PA, USA. Association for Computational Linguistics.