A Practical Classification of Multiword Expressions

Radosław Moszczyński
Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warszawa, Poland
rm@ipipan.waw.pl

Abstract

The paper proposes a methodology for dealing with multiword expressions in natural language processing applications. It provides a practically justified taxonomy of such units, and suggests the ways in which the individual classes can be processed computationally. While the study is currently limited to Polish and English, we believe our findings can be successfully employed in the processing of other languages, with emphasis on inflectional ones.

1 Introduction

radosław moszczyńskiIt is generally acknowledged that multiword expressions constitute a serious difficulty in all kinds of natural language processing applications (Sag et al., 2002). It has also been shown that proper handling of such expressions can result in significantly better results in parsing (Zhang et al., 2006).

The difficulties in processing multiword expressions result from their lexical variability, and the fact that many of them can undergo syntactic transformations. Another problem is that the label "multiword expressions" covers many linguistic units that often have little in common. We believe that the past approaches to formalize the phenomenon, such as IDAREX (Segond and Breidt, 1995) and Phrase Manager (Pedrazzini, 1994), suffered from trying to cover all multiword expressions as a whole. Such an approach, as is shown below, cannot efficiently cover all the phenomena related to multiword expressions.

Therefore, in the present paper we formulate a proposal of a taxonomy for multiword expressions, useful for the purposes of natural language processing. The taxonomy is based on the stages in the NLP workflow in which the individual classes of units can be processed successfully. We also suggest the tools that can be used for processing the units in each of the classes.

2 An NLP Taxonomy of Multiword Expressions

At this stage of work, our taxonomy is composed of two groups of multiword expressions. The first one consists of units that should be processed before syntactic analysis, and the other one includes expressions whose recognition should be combined with the syntactic analysis process. The next sections describe both groups in more detail.

2.1 Morphosyntactically Idiosyncratic Expressions

The first group consists of morphosyntactically idiosyncratic units. They follow unusual morphological and syntactic patterns, which causes difficulties for automatic analyzers.

By morphological idiosyncrasies we mean two types of units. First of all, there are bound words that do not inflect and cannot be used independently outside of the given multiword expression. In Polish, there are many such units, which are typically prepositional phrases functioning as complex adverbials, e.g.:[1]

[1]The asterisk in this and the following examples indicates an untranslatable bound word.

Proceedings of the ACL 2007 Student Research Workshop, pages 19–24,
Prague, June 2007.

(1)

na	*wskroś*
on	*

'thoroughly'

Secondly, there are unusual forms of otherwise ordinary words that only appear in strictly defined multiword expressions. An example is the following unit, in which the genitive form of the noun 'daddy' is different than the one used outside this particular construction:

(2)

nie	*rób*	*z*	*tata*
Neg	do-Imperative	of	*daddy-Gen
wariata			
fool			

'stop making a fool of me'

Morphological idiosyncrasies can be referred to as "objective" in the sense that it can be proved by doing corpus research that particular words only appear in a strictly limited set of constructions. Since outside such constructions the words do not have any meaning of their own, it is pointless to put them in the lexicon of a morphological analyzer. From the processing point of view, they are parts of complex multiword lexemes which should be considered as indivisible wholes.

Syntactically idiosyncratic phrases are those whose structure or behavior is incorrect from the point of view of a given grammar. In this sense, they are "subjective", because they depend on the rules underlying a particular parser.

A typical parser of Polish is expected to accept full sentences, i.e. phrases that contain a finite verb phrase, but possibly not many phraseologisms that are extremely common in texts and speech, and do not constitute proper sentences from the point of view of the grammar. This qualifies such phrases to be included and formalized among the first group we have distinguished. In Polish, such phrases include, e.g.:

(3)

Precz	*z*	*łapami!*
off	with	hands-Inst

'Get your hands off!'

Another group of multiword expressions that should be processed before parsing consists of complex adverbials that do not include any bound words, but that could be interpreted wrongly by the syntactic analyzer. Consider the following multiword expression:

(4)

na	*kolanach*
on	knees-Loc

'on one's knees' ('groveling')

This expression can be used in constructions of the following type:

(5)

Na	*kolanach*	*Kowalskiego*
on	knees-Loc	Kowalski-Gen

będą	*błagać.*
be-Future;Pl;3rd	beg-Infinitive

'They will beg Kowalski on their knees.'

In the above example *na kolanach* is an adjunct that is not subcategorized for by any of the remaining constituents. However, since *Kowalskiego* is genitive, the parser would be fooled to believe that one of the possible interpretations is 'They will beg on Kowalski's knees', which is not correct and semantically odd. Such complex adverbials are very common in Polish, which is why we believe that formalizing them as wholes would allow us to achieve better parsing results.

The last type of units that it is necessary to formalize for syntactic analysis are multiword text cohesion devices and interjections, whose syntactic structure is hard to establish, as their constituents belong to weakly defined classes. They can also directly violate the grammar rules, as the coordination in the English example does:

(6)

bądź	*co*	*bądź*
be-Imperative;Sg	what	be-Imperative;Sg

'after all'

(7) *by and large*

Since the recognition and tagging of all the above units will be performed before syntactic analysis, it seems natural to combine this process with a generalized mechanism of named entity recognition. We intend to build a preprocessor for syntactic analysis, along the lines of the ideas presented by Sagot and Boullier (2005). However, in addition to the set of named entities presented by the authors, we also intend to formalize multiword expressions of the types presented above, possibly with the use of `lxtransduce`.[2] This will allow us to prepare the input to the parser in such a way as to eliminate all the unparsable elements. This in turn should result in significantly better parsing coverage.

2.2 Semantically Idiosyncratic Expressions

The other group in our classification consists of multiword expressions that are idiosyncratic from the point of view of semantics. It includes such units as:

(8) *NP-Nom* *wziąć* *nogi* *za*
NP-Nom to take legs-Acc under
pas
belt-Acc
'to run away'

From the syntactic analysis point of view, such units are not problematic, as they follow regular grammatical patterns. They create difficulties in other types of NLP-based applications, as their meaning is not compositional, and cannot be predicted from the meaning of their constituents. Examples of such applications include electronic dictionaries, which should be able to recognize idioms and provide an appropriate, non-literal translation (Prószéky and Földes, 2005).

Such expressions can be extremely complex due to the lexical and word order variations they can undergo, which is especially the case in such languages as Polish. The set of syntactic variations that are possible in unit (8) is very large. First of all, there is the subject (NP-Nom). English multiword expressions are usually encoded disregarding the subject, as it can never break the continuity of the other constituents. In Polish it is different — the subject can be absent altogether, it can appear at the very beginning of the multiword expression without breaking its continuity, but it can also appear after the verb, between the core constituents. The subject can be of arbitrary length and needs to agree in morphosyntactic features (number, gender, and person) with the verb.

The verb can be modified with adverbial phrases, both on the left hand side and the right hand side. However, if the subject is postponed to a position after the verb, all the potential right hand side adverbials need to be attached after the subject, and not directly after the verb. Thus, taking all the variation possibilities into account, it is not unlikely to encounter such phrases in Polish:

(9) *Wziął* *pan*
take-1sg;Masc;Past you-1sg;Masc;Nom
przed *wszystkimi* *nogi* *za*
before everyone legs-Acc under
pas!
belt-Acc
'You ran away before everyone else!'

Some of the English multiword expressions also display properties that make them difficult to process automatically. Although the word order is more rigid, it is still necessary to handle, e.g., passivization and nominalization. This concerns the canonical example of *spill the beans*, and many others.

It follows that the units in the second group should not, and probably cannot, be reliably encoded with the same means as the simpler units from Section 2.1, which can be accounted for properly with simple methods based on regular grammars and surface processing.

One possible solution is to encode the complex units with the rules of a formal grammar of the given language. Another solution could be constructing an appropriate valence dictionary for verbs in such expressions. Both possibilities imply that the recognition process should be performed simultaneously with syntactic analysis.

3 Rationale

The above classification was formulated during an examination of the available formalisms for encoding multiword expressions, which was a part of the present work.

The attempts to formalize multiword expressions for natural language processing can be roughly divided into two groups. There are approaches that aim at encoding such units with the rules of an existing formal grammar, such as the approach described by Debusmann (2004). On the other hand, specialized, limited formalisms have been created,

[2] `http://www.cogsci.ed.ac.uk/~richard/ltxml2/lxtransduce.html`

whose purpose is to encode only multiword expressions. Such formalisms include the already mentioned IDAREX (Segond and Breidt, 1995) and Phrase Manager (Pedrazzini, 1994).

The first approach has two drawbacks. One of them is that using the rules of a given grammar to encode multiword expressions seems to have sense only if the rest of the language is formalized in the same way. Thus, such an approach makes the lexicon of multiword expressions heavily dependant on a particular grammar, which might make its reuse difficult or impossible.

The other disadvantage concerns complexity. While full-blown grammars do have the means to handle the most complex multiword expressions and their transformational potential, they create too much overhead in the case of simple units, such as idiomatic prepositional phrases that function as adverbials, which have been presented above.

Thus, we decided to encode Polish multiword expressions with an existing, specialized formalism. However, after an evaluation of such formalisms none of the ones we were able to find proved to be adequate for Polish. This is mostly due to the properties of the language — Polish is highly inflectional and has a relatively free word order. Both of these properties also apply to multiword expressions, which implies that in order to capture all their possible variations in Polish, it is necessary to use a powerful formalism (cf. the example in (9)).

Our analysis revealed that IDAREX, which is a simple formalism based on regular grammars, is not appropriate for handling expressions that have a very variable word order and allow many modifications. In IDAREX, each multiword unit is encoded with a regular expression, whose symbols are words or POS-markers. The words are described in terms of two-level morphology, and can appear either on the lexical level (which permits inflection) or the surface level (which restricts the word to the form present in the regular expression). An example is provided below:

(10) `kick:  :the :bucket;`

Encoding the multiword expression in (8) with IDAREX in such a way as to include all the possible variations leads to a description that suffers from overgeneration. Also, IDAREX does not include any unification mechanisms. This makes it unsuitable for any generation purposes (and reliable recognition purposes, too), as Polish requires a means to enforce agreement between constituents.

Phrase Manager makes encoding multiword expressions difficult for other reasons. The methodology employed in the formalism requires each expression to be assigned to a predefined syntactic class which determines the unit's constituents, as well as the modifications and transformations that it can undergo:[3]

(11)
```
SYNTAX-TREE
(VP V (NP Art Adj N AdvP))
MODIFICATIONS
V >
TRANSFORMATIONS
Passive, N-Adj-inversion
```

Since it is sometimes the case that multiword expressions belonging to the same class differ in respect of the syntactic operations they can undergo, the classes are arranged into a tree-like structure in which a class might be subdivided further on into a subclass that allows passivization, another one that allows nominalization and subject-verb inversion, etc.

The problem with this approach is that it leads to a proliferation of classes. At least in Polish, multiword expressions that follow the same general syntactic pattern often differ in the transformations they allow. Besides, the formalism creates too much overhead in the case of simple multiword expressions. Consider the following example in Polish:

(12) *No nie!*
oh no
'Oh, come on!'

In Phrase Manager it would be necessary to define a syntactic class for this unit, which seems to be both superfluous and problematic, as it is hard to establish what parts of speech are the constituents without taking purely arbitrary decisions.

To complicate matters further, the expression in the example has a variant in which both constituents

[3] The transformations need to be defined with separate rules elsewhere. The whole description is abbreviated.

switch their positions (with the meaning preserved). In the case of such a simple expression, it is impossible to "name" this transformation and assign any syntactic or semantic prominence to it — it can safely be treated as a simple permutation. However, Phrase Manager requires each operation to be named and precisely defined in syntactic terms, which in this case is more than it is worth.

In our opinion both those formalisms are inadequate for encoding all the phenomena labeled as "multiword expressions", especially in inflectional languages. Such approaches might be successful to a large extent in the case of fixed order languages, such as English — both IDAREX and Phrase Manager are reported to have been successfully employed for such purposes (Breidt and Feldweg, 1997; Tschichold, 2000). However, they fail with languages that have richer inflection and permit more word order variations. When used for Polish, the surface processing oriented IDAREX reaches the limits of its expressiveness; Phrase Manager is inadequate for different reasons — the assumptions it is based on would require something not far from writing a complete grammar of Polish, a task to which it is not suitable due to its limitations. And on the other hand, it is much too complicated for simple multiword expressions, such as (12).

4 Previous Classifications

There are numerous classifications available in linguistic literature, and we considered three of them in turn. From the practical point of view, none of them proved to be adequate for our needs. More precisely, none of them partitioned the field of multiword expressions into manageable classes that could be handled individually by uniform mechanisms.

The classification presented by Brundage et al. (1992) approaches the whole problem from an angle similar to what is required in Phrase Manager. It is based on a study of ca. 300 English and German multiword expressions, which were divided into classes based on their syntactic constituency and the transformations they are able to undergo.

Such an approach seems to be a dead end for exactly the same reasons that Phrase Manager has been criticized above. The study was limited to 300 units, which made the whole undertaking manageable. We believe that a really extensive study would lead to an unpredictable proliferation of very similar classes, which would make the whole classification too fine-grained and unpractical for any processing purposes.

The categorization that has been examined next is the one presented by Sag et al. (2002). It consists of three categories: fixed expressions (absolutely immutable), semi-fixed expressions (strictly fixed word order, but some lexical variation is allowed), syntactically-flexible expressions (mainly decomposable idioms — cf. (8)), and institutionalized phrases (statistical idiosyncrasies). Unfortunately, such a categorization is hard to use in the case of some Polish multiword expressions. Consider this example:

(13)	*Niech*	*to*	*szlag*	*trafi!*
	let	it-Acc	*	hit-Future
	'Damn it!'			

It is hard to establish which of the above categories does it belong to. The only lexically variable element is *it*, which can be substituted with another noun. This would qualify the expression to be included in the second category. However, it has a very free word order (*Niech to trafi szlag!*, *Szlag niech to trafi!*, and *Niech trafi to szlag!* are all acceptable). This in turn qualifies it to the third category, but it is not a decomposable idiom, and the word order variations are not semantically justified transformations, but rather permutations, as in (12). To make matters worse, the main element — *szlag* — is a word with a very limited distribution. This intuitively makes the unit fit more into the first category of unproductive expressions. This is even more obvious considering the fact that the word order variations do not change the meaning.

Another classification was presented by Guenthner and Blanco (2004). Their categories are very numerous, and the whole undertaking suffers from the fact that they are not formally defined. It also lacks a coherent purpose – it is neither a linguistic, nor a natural language processing classification, as it tries to put very different phenomena into one bag.

The categories are sometimes more lexicographically, and sometimes more syntactically oriented. For example, on the one hand the authors distinguish compound expressions (nouns, adverbs, etc.), and on the other hand collocations. In our opinion the categories should not be considered as parts of the same classification, as members of the former category belong to the lexicon, and the latter are a purely distributional phenomenon. Therefore, in the present form, the classification has no practical use.

5 Conclusions and Further Work

We have shown that trying to provide a formal description of *all* phenomena labeled as multiword expressions as a *whole* is not possible, which becomes obvious if one goes beyond English and tries to describe multiword expressions in heavily inflectional and relatively free word order languages, such as Polish. We have also shown the inadequacy of the available classifications of multiword expressions for computational processing of such languages.

In our opinion, a successful computational description of multiword expressions requires distinguishing two groups of units: idiosyncratic from the point of view of morphosyntax and idiosyncratic from the point of view of semantics. Such a division allows for efficient use of existing tools without the need of creating a cumbersome formalism.

We believe that the practically oriented classification presented above will allow us to build robust tools for handling both types of multiword expressions, which is the aim of our further research. The immediate task is to build the syntactic preprocessor. We also plan to extend the classification to make it slightly more fine-grained, which hopefully will make even more efficient processing possible.

References

Elisabeth Breidt and Helmut Feldweg. 1997. Accessing foreign languages with COMPASS. *Machine Translation*, 12(1/2):153–174.

Jennifer Brundage, Maren Kresse, Ulrike Schwall, and Angelika Storrer. 1992. Multiword lexemes: A monolingual and contrastive typology for NLP and MT. Technical Report IWBS 232, IBM Deutschland GmbH, Institut für Wissenbasierte Systeme, Heidelberg.

Ralph Debusmann. 2004. Multiword expressions as dependency subgraphs. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Frantz Guenthner and Xavier Blanco. 2004. Multilexemic expressions: an overview. In Christian Lèclere; Éric Laporte; Mireille Piot; Max Silberztein, editor, *Syntax, Lexis, and Lexicon-Grammar*, volume 24 of *Linguisticæ Investigationes Supplementa*, pages 239–252. John Benjamins.

Sandro Pedrazzini. 1994. *Phrase Manager: A System for Phrasal and Idiomatic Dictionaries*. Georg Olms Verlag, Hildeseim, Zürich, New York.

Gábor Prószéky and András Földes. 2005. An intelligent context-sensitive dictionary: A Polish-English comprehension tool. In *Human Language Technologies as a Challenge for Computer Science and Linguistics. 2nd Language & Technology Conference April 21–23, 2005,*, pages 386–389, Poznań, Poland.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Benoît Sagot and Pierre Boullier. 2005. From raw corpus to word lattices: robust pre-parsing processing. *Archives of Control Sciences, special issue of selected papers from LTC'05*, 15(4):653–662.

Frédérique Segond and Elisabeth Breidt. 1995. IDAREX: Formal description of German and French multi-word expressions with finite state technology. Technical Report MLTT-022, Rank Xerox Research Centre, Grenoble.

Cornelia Tschichold. 2000. *Multi-word units in natural language processing*. Georg Olms Verlag, Hildeseim, Zürich, New York.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia. Association for Computational Linguistics.