Computational Lexicography of Multi-Word Units: How Efficient Can It Be?

Filip Graliński
Adam Mickiewicz University
filipg@amu.edu.pl

Agata Savary
Université François Rabelais, Institute of Computer Science Polish Academy of Sciences
agata.savary@univ-tours.fr

Monika Czerepowicka
University of Warmia and Mazury
czerepowicka@gmail.com

Filip Makowiecki
University of Warsaw
f.makowiecki@student.uw.edu.pl

Abstract

The morphosyntactic treatment of multi-word units is particularly challenging in morphologically rich languages. We present a comparative study of two formalisms meant for lexicalized description of MWUs in Polish. We show their expressive power and describe encoding experiments, involving novice and expert lexicographers, and allowing to evaluate the accuracy and efficiency of both implementations.

1 Introduction

Multi-word units (MWU) are linguistic objects placed between morphology and syntax: their general syntactic behavior makes them similar to free phrases, while some of their idiosyncratic (notably from the morphological point of view) properties call for a lexicalized approach in which they are treated as units of description. Moreover, MWUs, which encompass such classes as compounds, complex terms, multi-word named entities, etc., often have unique and constant references, thus they are seen as semantically rich objects in Natural Language Processing (NLP) applications such as information retrieval. One of the main problems here is the conflation of different surface realizations of the same underlying concept by the proper treatment of orthographic (*head word* vs. *headword*), morphological (*man servant* vs. *men servants*), syntactic (*birth date* vs. *birth of date*), semantic (*hereditary disease* vs. *genetic disease*) and pragmatic (*Prime minister* vs. *he*) variants (Jacquemin, 2001).

In this paper we are mainly interested in orthographic, morphological, and partially syntactic variants of contiguous MWUs (i.e. not admitting insertions of external elements). Describing them properly is particularly challenging in morphologically rich languages, such as Slavic ones.

We believe that the proper treatment of MWUs in this context calls for a computational approach which must be, at least partially, lexicalized, i.e. based on electronic lexicons, in which MWUs are explicitly described. Corpus-based machine learning approaches bring interesting complementary robustness-oriented solutions. However taken alone, they can hardly cope with the following important phenomenon: while MWUs represent a high percentage of items in natural language texts, most of them, taken separately, appear very rarely in corpora. For instance, (Baldwin and Villavicencio, 2002) experimented with a random sample of two hundred English verb-particle constructions and showed that as many as two thirds of them appear at most three times in the Wall Street Journal corpus. The variability of MWUs is another challenge to knowledge-poor methods, since basic techniques such as lemmatisation or stemming of all corpus words, result in overgeneralizations (e.g. *customs office* vs. **custom office*) or in overlooking of exceptions (e.g. *passersby*). Moreover, machine learning methods cannot reliably be used alone for less resourced languages. In such cases an efficient annotation of a large corpus needed for machine learning usually requires the pre-existence of e-lexicons (Savary and Piskorski, 2010).

Despite these drawbacks machine learning allows robustness and a rapid development, while

Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 2–10,
Beijing, August 2010

knowledge-based methods in general have the reputation of being very labor intensive. In this paper we try to show how effective tools of the the latter class can be. We present two formalisms and tools designed in view of lexicalized MWU variant description: *Multiflex* and *POLENG*. We discuss their expressivity, mainly with respect to Polish. We also show their applications and perform their qualitative and quantitative comparative analysis.

2 Linguistic Properties and Lexical Encoding of MWUs

Compounds show complex linguistic properties including: (i) heterogeneous status of separators in the definition of a MWU's component, (ii) morphological agreement between selected components, (iii) morphosyntactic non-compositionality (exocentricity, irregular agreement, defective paradigms, variability, etc.), (iv) large sizes of inflection paradigms (e.g. dozens of forms in Polish). A larger class of verbal multiword expressions additionally may show huge variability in word order and insertion of external elements.

For instance in the Polish examples below: (1) requires case-gender-number agreement between the two first components only, in (2) the components agree in case and number but not in gender, (3) admits a variable word order, (4) shows a depreciative paradigm (no plural), (5) includes a foreign lexeme inflected in Polish manner, (6) is characterized by a shift in gender (masculine animate noun is the head of a masculine human compound[1]), and (7) is a foreign compound with unstable Polish gender (masculine, neuter or non-masculine plural).

(1) *Polska Akademia Nauk* 'Polish Academy of Sciences'

(2) *samochód pułapka* 'car bomb'

(3) *subsydia zielone, zielone subsydia* 'green subsidies'

(4) *areszt domowy* 'house arrest'

(5) *fast food, fast foodzie*

(6) *ranny ptaszek* 'early bird'

(7) *(ten/to/te) public relations*

Due to this complex behavior, as well as to a rich semantic content, MWUs have been a hot topic in international research for quite a number of years (Rayson et al., 2010) in the context of information retrieval and extraction, named entity recognition, text alignment, machine translation, text categorization, corpus annotation, etc. In this study we are interested in lexical approaches to MWUs, i.e. those in which MWUs are explicitly described on the entry-per-entry basis, in particular with respect to their morpho-syntax. Earlier examples of such approaches include *lexc* (Karttunen et al., 1992), *FASTR* (Jacquemin, 2001), *HABIL* (Alegria et al., 2004), and *Multiflex* discussed below. They mainly concentrate on contiguous nominal and adjectival MWUs, sometimes considering limited insertions of external elements. More recent approaches, such as (Villavicencio et al., 2004), (Seretan, 2009) and (Grégoire, 2010), increasingly address verbal and other non contiguous multi-word expressions (MWEs). These studies are complemented by recent advances in parsing: robust and reliable syntactic analysis now available can be coupled with MWEs identification, and possibly also translation. The *POLENG* formalism discussed below belongs to some extent to this class of tools. While the processing of non contiguous MWEs is an important step forward, the morphological phenomena in MWUs should still be addressed with precision, in particular in inflectionally rich languages. Therefore we present below a comparative study of *Multiflex* and *POLENG* based on an experiment with encoding nominal and adjectival MWUs in Polish.

3 Multiflex

Multiflex (Savary, 2009) (Savary et al., 2009) is a graph-based cross-language morpho-syntactic generator of MWUs relying on a 'two-tier approach'. First, an underlying morphological module for simple words allows us to tokenize the MWU lemma, to annotate its components, and to generate inflected forms of simple words on demand. Then, each inflected MWU form is seen as

[1]There are three subgenders of the masculine in Polish.

a particular combination of the inflected forms of its components. All inflected forms of an MWU and their variants are described within one graph. Compounds having the same morpho-syntactic behavior are assigned to the same graph. A unification mechanism accounts for compact representation of agreement within constituents. For instance, Fig. 1 presents the inflection graph for compounds inflecting like example (3). Its first path combines the first component $1 (here: *subsydia*) inflected in any case with the unchanged second component $2 (here: space) and a case-inflected third component $3 (here: *zielone*). The common unification variable $c imposes case agreement between components $1 and $3. The second path describes the inverted variant of this term, in any of the cases. The description between the paths says that each resulting compound form agrees in case with components $1 and $3, and inherits its gender (*Gen*) and number (*Nb*) from component $1 as it appears in the MWU lemma (here: neutral-2 plural).

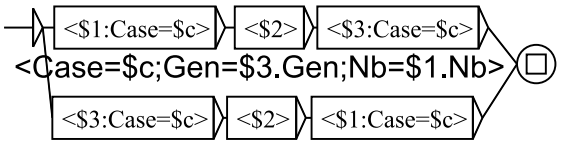

Figure 1: *Multiflex* inflection graph for compounds inflecting like *subsydia zielone*.

The main drawbacks of the formalism include: (i) the difficulty of conflating variants of MWUs containing numerical expressions (*ulica XI Poprzeczna, ulica Jedenasta Poprzeczna* '11th Cross Street'), (ii) impossibility of expressing relations existing between an MWU and external elements (e.g. in German *die Vereinten Nationen, Vereinte Nationen* 'United Nations'). Last but not least, *Multiflex* is meant for describing only contiguous compounds, i.e. those that admit no insertions of external elements (*He made up his bloody mind.*).

For the current study we are using a MWU encoding environment *Toposław* (Woliński et al., 2009), which integrates *Multiflex* along with the morphological analyser and generator for Polish *Morfeusz* (Savary et al., 2009), and the graph editor from *Unitex* (Paumier, 2008). *Toposław* speeds up the automated controlled encoding of MWUs by automatic look-up of constituents, filtering of MWUs entries, as well as automatic graph creation, debugging and filtering.

4 POLENG Formalism

By the "POLENG formalism" we mean the formalism used in the *POLENG* rule-based machine translation system (Jassem, 1996; Jassem, 2004) for the purposes of morphosyntactic description of MWUs in bilingual lexicons.

The *POLENG* formalism was designed with simplicity, conciseness and practical applicability for the MWU recognition and generation in mind, rather than care for nuances and theoretical coherence or elegance. As in Multiflex, a two-tier approach was used; however all inflected forms of a MWU are described by means of a compact, linear string rather than a graph. (One of the advantages of using such an approach is that MWU descriptions can be edited within a regular text input control and can be easily stored in a single database field.) For instance the term *subsydia zielone* from example (3) has the following description:

(8) `N:5p[subsydium_N! zielony_A]`

where:

- `N` is a part-of-speech tag (`N` = *noun*, i.e. it is a nominal phrase),
- additional morphosyntactic flags are given after the colon – `5` stands for the fifth (neuter) gender, `p` – stands for *plural* (i.e. the phrase is used only in plural),
- the description of individual components is given in square brackets, namely the first component of *subsydia zielone* is the lexeme identified with `subsydium_N` (i.e. the noun *subsydium* 'subsidy') and the second[2] one – the lexeme identified with `zielony_A` (i.e. the adjective *zielony* 'green'); the main (head) component is marked with `!`.

[2]The space is not considered a MWU component.

Note that case, number and gender agreement between the MWU components is not imposed explicitly. It is rather assumed implicitly (by default, all inflected components of a nominal MWU must agree in case, number and gender). Such assumptions have to be hard-coded into MWU recognition/generation modules for particular languages – this is the price one pays for the simplicity of the formalism.

The order of the components of a MWU is assumed to be fixed (except for verbal MWUs, more on this later), e.g. *zielone subsydia* is not covered by (8), i.e. a separate entry *zielone subsydia* described as `N:5p[zielony_A subsydium_N!]` must be entered.[3]

The identifier of a lexeme is usually its base form followed by an underscore and its part-of-speech tag (e.g. `subsydium_N`). In case of homonyms of the same part of speech, consecutive numbers are appended. For instance, the Polish verb *upaść* 'fall down' is denoted with `upaść_V` and its homonym *upaść* 'fatten up' is denoted with `upaść_V2`.[4] Homonym identifiers are assigned roughly in order of frequency. In *POLENG*, lexeme identifiers can be abbreviated to the POS tag (followed by a number, if necessary) on condition that its base form is the same as the form that is used in the base form of the MWU. For instance, in Example (a) in Table 1[5] `N:3[N! A]` is an abbreviation for `N:3[system_N! operacyjny_A]`.

A component of a MWU which is not inflected (in that particular MWU) is referred to simply as `0`, see Example (b) in Table 1.

A lexeme identifier may be followed by a hyphen and a so-called *sublexeme* tag if a subset of inflected forms can be used in a given MWU, see Example (c) in Table 1 (`PA` denotes active participle forms and `GR` – gerundial forms). Also additional flags may be specified, for instance in Example (d) the flag `u` is used (it means that the upper case of the first letter is required).

Polish verbal MWUs are treated in a different manner than other types of MWUs. Namely the fixed order of components is not assumed, for instance, in Example (e) in Table 1 each of the six permutations of the main verb *chodzić* 'walk', the adverb *boso* 'barefoot' and the prepositional phrase *po rosie* 'through the dew' is acceptable (the flag `I` denotes the imperfective aspect). The only restriction is the fixed order of the components of the PP. This restriction is specified using round brackets. What's more, a verbal phrase does not need to be contiguous in a given sentence to be recognized by the *POLENG* system. For example, the verbal MWU *chodzić boso po rosie*, described as in Example (e), will be detected in the following sentence:

(9) Po rosie Anna chodziła dziś boso.
Through dew Anna walked today barefoot.
'Anna walked barefoot through the dew today.'

POLENG allows for describing required (but not fixed) constituents, using so-called *slots*, see Example (f) in Table 1, where `$L$` is a slot for a noun phrase in locative (note that slots are given in the "base form" of a MWU, not in its description, where a slot is simply marked with `0`).

It is also possible to describe some relations between MWUs and external elements (e.g. between a German MWU and an article, cf. *die Vereinten Nationen, Vereinte Nationen* 'United Nations') within the *POLENG* formalism. However, this is achieved by rather ad hoc methods.

The descriptions of MWUs does not have to be entered manually. The *POLENG* machine translation system is equipped with a special "translation" direction in which a phrase can be "translated" automatically into its description as a MWU. New MWUs are usually described in this automatic manner and are corrected manually if necessary (e.g. while entering equivalents in other languages). There are also tools for the automatic detection of anomalies in MWU descriptions (e.g., cases when a Polish MWU was described as a nominal phrase and its English equivalent as a verbal phrase).

[3] Note that the position of the adjective may affect the meaning of a Polish MWU, e.g. *twardy dysk* is a disk that happens to be hard, whereas *dysk twardy* is a term (*hard disk*, *HDD*).

[4] Both verbs have the same base form but different valence and inflected forms.

[5] All the examples in Table 1 are real entries from the lexicon of the *POLENG* Polish-English machine translation system.

	MWU	English equivalent	description
a.	*system operacyjny*	*operating system*	`N:3[N! A]`
b.	*jądro systemu operacyjnego*	*kernel of an operating system*	`N:5[N! 0 0]`
c.	*lekceważące mrugnięcie*	*deprecating wink*	`N:5[lekceważyć_V-PA mrugnąć_V-GR!]`
d.	*Rzeczpospolita Polska*	*Republic of Poland*	`N:4[rzeczpospolita_N:u! polski_A:u]`
e.	*chodzić boso po rosie*	*walk barefoot through the dew*	`V:I[V! 0 (0 0)]`
f.	*być orłem w L*	*be a wizard at something*	`V:I[V! 0 (0 0)]`

Table 1: Examples of MWUs annotated within the *POLENG* formalism.

5 Comparative Evaluation

5.1 Existing Data

Both *POLENG* and *Multiflex* have proved adequate for the large-scale lexicalized description of MWUs in several languages and in different applications. Table 2 lists the lexical resources created within both formalisms.

The *Multiflex* formalism has been used for the construction of language resources of compounds in various applications (Savary, 2009): (i) general-purpose morphological analysis, (ii) term extraction for translation aid, (iii) named entity recognition, (iv) corpus annotation. The *Multiflex* implementation has been integrated into several NLP tools for corpus analysis and resource managment: *Unitex* (Paumier, 2008), *WS2LR* (Krstev et al., 2006), *Prolexbase* (Maurel, 2008), and *Toposław* (Woliński et al., 2009).

	Language	Type of data	# entries
POLENG	Polish		286,000
	English		356,000
	Russian		26,000
	German		59,000
Multiflex	English	general language	60,000
		computing terms	57,000
	Polish	general language	1,000
		urban proper names	8,870
		economic terms	1,000
	Serbian	general language	2,200
	French	proper names	3,000
	Persian	general language	277

Figure 2: Existing MWU resources described with *POLENG* and *Multiflex*.

The *POLENG* formalism has been used mainly for the description of MWU entries in Polish-English, Polish-Russian and Polish-German bilingual lexicons. Another application of the *POLENG* formalism was the description of multi-token abbreviations[6] for the purposes of text normalization in a Polish text-to-speech system (Graliński et al., 2006). The MWUs described in this manner can be taken into account in the stand-alone, monolingual (Polish, English, German or Russian) *POLENG* parser as well. Descriptions compatible with the *POLENG* formalism are also dynamically generated by the *NERT* (named entity recognition and translation) module of the *POLENG* machine translation system, e.g. for named entities denoting persons (Graliński et al., 2009).

5.2 Describing New Data

In order to perform a qualitative and quantitative comparative analysis of *POLENG* and *Multiflex* we have performed an experiment with encoding new linguistic data. By "encoding" we mean assigning a *Multiflex* inflection graph or a *POLENG* MWU description to each MWU. Four distinct initial lists of about 500 compounds each have been prepared: (i) two lists with compounds of general Polish, (ii) two lists with economical and financial terms. About 80% of the entries consisted of 2 words. One or two novice lexicographers were to encode one list of (i) and one of (ii).[7] The two remaining lists were to be dealt with by an expert lexicographer. Almost all entries were compound common nouns although some contained proper name components (*reguła Ramseya* 'Ramsey rule') and some were compound adjectives (*biały jak śmierć* 'as white as death').

Table 2 shows the time spent on each part of the experiment. The training phase of each system consisted in watching its demo, reading the user's documentation, making sample descriptions, and discussing major functionalities with experts. The

[6] Such Polish expressions as, for example, *prof. dr hab.*, *sp. z o.o.*, *nr wersji*.

[7] The data was encoded by two novice lexicographers (one list each) in case of *Multiflex* and by one novice lexicographer in case of *POLENG*.

	POLENG			Multiflex		
	novice		expert	novice		expert
	training	encoding	encoding	training	encoding	encoding
General language (about 500 entries)	5.5 h	6 h	4 h	3 h	23 h	7.5 h
Terminology (about 500 entries)	4 h	5 h	3 h	3 h	20 h	12 h

Table 2: Encoding time for two categories of lexicographers and two types of data.

further encoding phase was performed by each lexicographer on his own with rare interaction with experts.

Describing general language data proves slightly more time consuming for novice lexicographers due to exceptional behavior of some units, such as depreciativity, gender variation, etc. With *Multiflex*, the average speed of a novice lexicographer is of 21 and 27 entries per hour for the general and terminological language, respectively. In the case of an expert, these figures are of 36 and 67 entries per hour. Thus, the encoding by an expert is about 1.6 and 2.5 times faster than by a novice for terminological and general language, respectively. The big difference in expert encoding time between both data categories can be justified by the fact that terminological data require domain-specific knowledge, and contain more components per entry and more embedded terms. Nevertheless, the general language compounds present more grammatical idiosyncrasies such as depreciativeness, gender change, etc. The two novice lexicographers reported that it took them about 6 to 7.5 hours of personal efforts (training excluded) in order to gain confidence and efficiency with the formalism and the tools, as well as with the rather rich Polish tagset. The *Multiflex* expert spent about 50% of her time on creating graphs from scratch and assigning them to MWUs. As these graphs can be reused for further data, the future encoding time should drop even more. Both novice and expert lexicographers heavily used the block working mode and filtering options.

With *POLENG*, the lexicographers were given the MWU descriptions generated automatically by the *POLENG* system (see Section 4). As most of these descriptions (90%) were correct, the lexicographers' work was almost reduced to revision and approval. Most errors in the descriptions generated automatically involved non-trivial homonyms and rare words, not included in the *POLENG* lexicons (e.g. names of exotic currencies).

Table 3 shows the quantitative analysis of MWU inflection paradigms created by the expert lexicographer.[8] Unsurprisingly, the 5 most frequent paradigms cover up to 77% of all units. They correspond to 3 major syntactic structures (in *Multiflex*, possibly embedded): *Noun Adj* (*agencja towarzyska* 'escort agency'), *Noun* $Noun_{genitive}$ (*dawca organów* 'organ donor'), and *Adj Noun* (*biały sport* 'winter sport'), with or without number inflection (*adwokat/adwokaci diabła* 'devil's advocate/advocates' vs *dzieła wszystkie* 'collected works'), and some of them allowing for inversion of components (*brat cioteczny, cioteczny brat* 'cousin'). Conversely, 33% through 57% of all *Multiflex* paradigms (about 50% for *POLENG*) concern a single MWU each. In *Multiflex* delimiting embedded compounds allows to keep the number of paradigms reasonably low, here 23 and 3 embedded MWU were identified for terminological and general language, respectively (embedded MWUs are not allowed in *POLENG*).

With *Multiflex* some data remain erroneously or only partially described after the experiment. Table 4 shows the typology and quantities of problems encountered by novice lexicographers:

- For general language, the high percentage of errors in inflection paradigms is due to one repeated error: lack of the number value. As the full list of all inflection categories relevant to a class is explicitly known, this kind of errors may be avoided if the encoding tool automatically checks the completeness of morphological descriptions.

[8] For the purposes of this analysis, POLENG lexeme identifiers were reduced to POS-tags and some redundant morphosyntactic flags (gender and aspect flags) were erased.

	POLENG			Multiflex		
	# inflection paradigms	coverage of 5 most frequent paradigms	# single-entry paradigms	# inflection paradigms	coverage of 5 most frequent paradigms	# single-entry paradigms
General language	58	72%	30	36	77%	12
Terminology	46	77%	23	52	67%	30

Table 3: Distribution of inflection paradigms defined in the experiment by the expert lexicographer.

	POLENG			Multiflex				
	Entries			Inflection paradigms		Entries		
	incomplete	errors	non-MWUs in *POLENG*	errors	redundancies	incomplete	errors	non-optimal description
General language	2%	1.6%	0.4%	41%	22%	5%	1%	3%
Terminology	3%	2.3%	0%	0%	23%	14%	0.7%	5%

Table 4: Errors and imprecisions committed by novice lexicographers.

- Redundancies in graphs are mainly due to identical or isomorphic graphs created several times. A tool allowing to automatically detect such cases would be helpful.
- The incompletely described entries are mainly due to unknown single components. Despite its very high coverage, the morphological analyzer and generator *Morfeusz* lacks some single items[9]: general language lexemes (*radarowiec* 'radar-operating policeman'), rare currency units (*cedi*), foreign person names (inflected in Polish, e.g. *Beveridge'owi*), and borrowed terms (*forwardowy* 'forward-bound'). Some rare words are homonyms of common words but they differ in inflection (*lek* 'Albanian currency unit'). It is thus necessary to incorporate an encoding tool for new general language or application-dependent single units.
- We consider the description of an entry non optimal if the data helpful for determining the inflection graph are not correctly indicated. The effective graphs are however correct here, and so are the resulting inflected forms.
- The rate of actual errors, i.e. inflection errors resulting from inattention or badly understood formalism, is very low ($\leq 1\%$)

Some further problems stem from the limits of either *Multiflex* or *Morfeusz* design. Firstly, unlike *POLENG*, *Multiflex* does not allow to describe compounds having a lexically free but grammatically constrained element ('slots', cf sec. 4). Secondly, inflection variants of single words, such as *transformacyj* 'transformation$_{gen.pl.}$' are not distinguished in *Morfeusz* by grammatical features, thus it is impossible to forbid them in compounds via feature constraints (*transformacji wolnorynkowych* but not **transformacyj wolnorynkowych* 'free market transformations'). Thirdly, since depreciativity is modeled in *Morfeusz* as inflectional class rather than category it is not easy to obtain depreciative forms of nouns from their base forms (*chłopi/chłopy na schwał* 'lusty fellows').

The following problems were encountered during the descriptions of MWUs with the *POLENG* formalism:

- As was the case with *Multiflex*, some single components (mainly of economical and financial compounds) were absent in the *POLENG* Polish lexicon. Nonetheless, inflected forms of an unknown component can be recognized/generated provided that they end in frequent and regular suffixes (e.g. in suffixes typical of adjectives such as *-owy*, *-cyjny*) – i.e. "virtual" lexemes are created if needed. Otherwise, an unknown component

[9] Some problems with unknown words could be solved by introducing a token boundary inside a word, thus obtaining a non inflected prefix and a known inflected core word, e.g. *pół|hurtowy* 'half-wholesale'.

makes the recognition/generation of a given MWU impossible. However, the description can be entered anyway, and as soon as a missing lexeme is entered into the *POLENG* lexicon, the MWU will be correctly recognized/generated.

- What is a multi-word unit is defined by the POLENG tokenizer. Some of the terms described in the experiment, such as *by-pass*, *quasi-pieniądz* (*quasi-money*), are tokenized as single terms by the *POLENG* tokenizer and, consequently cannot be covered by the *POLENG* MWU formalism.

- As it was mentioned in Section 4, it is not possible to cover variability in word order with one description in the *POLENG* formalism (unlike in *Multiflex*), the only exception being totally free order of verbal phrases. The same limitation applies to MWUs with alternative or optional components. In such cases, multiple MWUs have to be entered and described separately. However, in order to avoid redundancy in bilingual lexicons, it is possible to link variant MWUs with so-called *references* (i.e. an equivalent in the target language has to be specified for just one of them).

- The rate of actual errors is higher than in *Multiflex*. Most of them involve non-trivial homonyms and words absent from the *POLENG* lexicon. If MWUs with such words were marked in some way for a lexicographer, the error rate would probably be much lower.

6 Conclusions

MWUs show a complex linguistic behavior, particularly in inflectionally rich languages, such as Slavic ones. They call for descriptive formalisms that allow to account for their numerous morphological, syntactic and semantic variants. We have presented two formalisms used for the description of MWUs in Polish, and we have performed a comparative analysis of the two formalisms. *Multiflex* aims at a precise and explicit description, as well as at adaptivity to different languages and morphological models. It allows to conflate many types of MWUs variants such as acronyms, inversions etc. However its use is relatively slow, and non contiguous units, or units containing semantically free elements ('slots'), cannot be described. See also (Savary, 2008) for a detailed contrastive analysis of *Multiflex* with respect to 10 other systems for a lexical description of MWUs in different languages such as (Karttunen et al., 1992), (Jacquemin, 2001), and (Alegria et al., 2004).

POLENG offers a complementary approach: it includes a faster semi-controlled encoding process, allows for the treatment of non contiguous units or 'slots', and was applied to more massive data in professional machine translation. Its formalism is however more implicit, thus less interoperable, and variant conflation can be done to a limited degree only.

Encoding experiments involving both novice and expert lexicographers showed that both tools can be efficiently used for creating morphological resources of MWUs. They also allowed to put forward further improvements of our tools such as verifying the completeness of morphological description, checking paradigm identity, and encoding new single-word entries. Both tools are used for the morphological description of MWUs in different languages, notably Slavic ones, which show a rich inflection system. They have been used in various NLP applications: computational lexicography, machine translation, term extraction, named entity identification, and text normalization.

References

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. 2004. Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the ACL'04 Workshop on Multiword Expressions*, pages 48–55.

Baldwin, Timothy and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104.

Graliński, Filip, Krzysztof Jassem, Agnieszka Wagner, and Mikołaj Wypych. 2006. Text normalization as

a special case of machine translation. In *Proceedings of International Multiconference on Computer Science and Information Technology (IMCSIT'06)*, pages 51–56, Katowice. Polskie Towarzystwo Informatyczne.

Graliński, Filip, Krzysztof Jassem, and Michał Marcińczuk. 2009. An environment for named entity recognition and translation. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT'09)*, pages 88–96, Barcelona.

Grégoire, Nicole. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2).

Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Jassem, Krzysztof. 1996. Elektroniczny słownik dwujęzyczny w automatycznym tłumaczeniu tekstu. PhD thesis. Uniwersytet Adama Mickiewicza. Poznań.

Jassem, Krzysztof. 2004. Applying Oxford-PWN English-Polish dictionary to Machine Translation. In *Proceedings of 9th European Association for Machine Translation Workshop, "Broadening horizons of machine translation and its applications", Malta, 26-27 April 2004*, pages 98–105.

Karttunen, Lauri, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-Level Morphology with Composition. In *Proceedings of the 14 th International Conference on Computational Linguistics (COLING'92), Nantes*, pages 141–148.

Krstev, Cvetana, Ranka Stanković, Duško Vitas, and Ivan Obradović. 2006. WS4LR: A Workstation for Lexical Resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, pages 1692–1697.

Maurel, Denis. 2008. Prolexbase. A multilingual relational lexical database of proper names. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Marocco*, pages 334–338.

Paumier, Sébastien. 2008. Unitex 2.1 User Manual.

Rayson, Paul, Scott Piao, Serge Aharoff, Stefan Evert, and Bego na Villada Moirón, editors. 2010. *Multiword expression: hard going or plain sailing*, volume 44 of *Language Resources and Evaluation*. Springer.

Savary, Agata and Jakub Piskorski. 2010. Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish. In *Intelligent Information Systems, Siedlce, Poland*, pages 141–154.

Savary, Agata, Joanna Rabiega-Wiśniewska, and Marcin Woliński. 2009. Inflection of Polish Multi-Word Proper Names with Morfeusz and Multiflex. *Lecture Notes in Computer Science*, 5070:111–141.

Savary, Agata. 2008. Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2):1–53.

Savary, Agata. 2009. Multiflex: a Multilingual Finite-State Tool for Multi-Word Units. *Lecture Notes in Computer Science*, 5642:237–240.

Seretan, Violeta. 2009. An integrated environment for extracting and translating collocations. In *Proceedings of the 5th Corpus Linguistics Conference, Liverpool, U.K.*

Villavicencio, Aline, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical Encoding of MWEs. In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, pages 80–87.

Woliński, Marcin, Agata Savary, Piotr Sikora, and Małgorzata Marciniak. 2009. Usability improvements in the lexicographic framework Toposław. In *Proceedings of Language and Technology Conference (LTC'09), Poznań, Poland*, pages 321–325. Wydawnictwo Poznańskie.