Multiword Verbs in WordNets

Veronika Vincze[1], **Attila Almási**[2] and **János Csirik**[1]
[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
Tisza Lajos krt. 103., 6720 Szeged, Hungary
{vinczev, csirik}@inf.u-szeged.hu
[2]University of Szeged, Department of Informatics
Árpád tér 2., 6720 Szeged, Hungary
vizipal@gmail.com

Abstract

In this paper, we describe how wordnets treat multiword verbs. We pay special attention to the English and Hungarian wordnets and we argue that from a multilingual perspective it is recommended to store idioms and light verb constructions as a whole rather than listing their parts separately. In order to enhance their applicability in multilingual applications, a unified treatment should be applied for subtypes of multiword verbs.

1 Introduction

In natural language processing, one of the most challenging tasks is the proper treatment of multiword expressions (MWEs). Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002). To put it differently, they are lexical items that contain space or "idiosyncratic interpretations that cross word boundaries". Multiword expressions are frequent in language use and they usually exhibit unique and idiosyncratic behavior, thus, they often pose a problem to NLP systems.

In this paper, we describe how wordnets treat multiword expressions and we pay special attention to multiword verbs. Multiword verbs comprise phrasal verbs, light verb constructions and idioms[1], however, we focus on idioms and light verb constructions in our investigations as they represent two different levels of compositionality: while idioms are totally non-compositional, light verb constructions are semi-compositional (i.e. the meaning of the noun plays and important role in computing the meaning of the whole structure). We concentrate on English and Hungarian and we argue that from a multilingual perspective, it is more advisable to store multiword expressions as a whole rather than listing their parts separately.

The structure of the paper is as follows. First, the decomposability of multiword expressions is discussed, then it is shown how idioms and light verb constructions should be treated in wordnets. The paper concludes with a comparison of methods offered for these two types of multiword verbs.

2 The decomposability of multiword expressions

Multiword expressions can be classified according to their semantic decomposability (Sag et al., 2002; Nunberg et al., 1994). If the parts of the MWE can be interpreted as having a special sense unique to this construction, that is, there can be a word-to-word mapping between the lexical and the semantic level, it is called a decomposable MWE. An English and a Hungarian example are offered here:

to spill the beans
'to reveal a secret'
beans = 'secret'
spill = 'reveal'

[1]Idioms usually consist of a verb phrase and they are semantic predicates, thus, their grammatical function is similarly to that of verbs. This is why we consider them as a subtype of multiword verbs.

veszi *a* *lapot*
take-3SGOBJ the card-ACC
'to understand the message'

vesz = 'understand'
lap = 'message'

It should be noted that in the English example, the definite article in the idiom corresponds to an indefinite one on the semantic level, however, all words in the idiom can be mapped to another one on the semantic level. If no such correspondence can be found, the MWE is considered to be non-decomposable. An example is *to bite the dust* 'to die' or its Hungarian equivalent *fűbe harap* (grass-ILL bites) which meaning cannot be decomposed in a way to match the single words within the expression.

The above distinction may have interesting implications for wordnet building. If the parts of a MWE can be attributed a special distinct meaning, the question arises whether this meaning should be added to the sense inventory of the given words or not, in other words, to decompose its meaning or not. From another perspective, should decomposable MWEs be stored as one unit in wordnets (i.e. as one synset) or should their parts be separately listed in synsets with the corresponding senses? In order to answer this question from a multilingual aspect, we first examine how the Princeton WordNet (PWN) (Miller et al., 1990) and the Hungarian WordNet (Miháltz et al., 2008) treat multiword verbs.

3 Idioms in the English and the Hungarian wordnets

We can find the following synset in PWN:

{gutter:2, sewer:3, toilet:3}

These literals are parts of idioms, which are not listed as a whole in PWN. The PWN synset means "misfortune resulting in lost effort or money", however, it is not obvious from the representation that this sense is valid only within the idiom, i.e. in combination with *go* or *be* and a preposition.

The Hungarian equivalent of the above synset is a non-lexicalized one[2]:

(WC, ablak, csatorna; kidobhatod az ablakon) 'toilet, window, gutter; you can throw it out the window'

Thus, it seems that the above PWN synset has no lexicalized Hungarian counterpart although there are Hungarian idioms that express the same meaning, e.g. *kidobhatja az ablakon* (out.throw-MOD-3SGOBJ the window-SUP) or *lehúzhatja a WC-n* (down.flush-MOD-3SGOBJ the toilet-SUP). Thus, it would have been feasible to create a Hungarian synset with the nominal parts of the idioms such as:

(WC, ablak, csatorna) 'toilet, window, gutter'

As Osherson and Fellbaum (2010) propose, the connection between the parts of idioms can be signaled by idiom-specific relations between synsets. However, the major problem with this approach is that not all members of the synset can be paired with the same verb: for instance, in Hungarian, there are no phrases like **lehúzhatja az ablakot* 'to flush the window' or **kidobhatja a WC-n* 'to throw it out the toilet'. Thus, it would be complicated to signal which literal can be paired with which verb if the nominal parts of the idioms with similar meanings are included in the very same synset.

From a multilingual perspective, it is interesting to note that most multiword expressions have an equivalent in other languages, however, it may well be the case that the linguistic structure of the MWE with the same meaning in two languages do not coincide or one of them is decomposable (the parts of the MWE can be interpreted as having a special sense unique to this construction, that is, there can be a word-to-word mapping between the lexical and the semantic level) and the other one is not as in:

to be on cloud number 9

örül, *mint* *majom* *a*
be.glad as monkey the
farkának
tail-3SGPOSS-DAT
'to be extremely happy'

[2]Creating the HuWN database practically meant rendering the PWN synsets into Hungarian, that is, Hungarian equivalents had to be found for PWN synsets. Whenever this was not possible, e.g. due to differences in culture, language use or grammar, the synset was marked as *non_lexicalized* in Hungarian and an approximate definition was given for the English concept.

Here the English idiom is decomposable – *cloud number 9* corresponds to *happiness* – while in Hungarian, the verb *örül* 'be glad' corresponds to the "happy" component in the meaning of the idiom, however, *majom* 'monkey' and *farkának* 'to his tail' cannot be matched to any meaning component. On the other hand, in English, *cloud nine* is listed in a synset denoting happiness (bliss:1, blissfulness:1, cloud nine:1, seventh heaven:1, walking on air:1) in accordance with the proposal found in Osherson and Fellbaum (2010) but in Hungarian, no mention of the idiomatic usages is made in the corresponding synset (elragadtatás:2, mennyei boldogság:1, üdvösség:1), although *hetedik mennyország* 'seventh heaven' could have been listed here as there is an idiom with a similar meaning (*a hetedik mennyországban érzi magát* (the seventh heaven-INE feel-3SGOBJ himself-ACC) 'to be in seventh heaven'). However, none of the components of the idiom *örül, mint majom a farkának* could be included in this synset in HuWN since there is no noun in the idiom corresponding to *cloud nine* that can be included in the nominal hierarchy. This also highlights that the treatment of idioms is somewhat problematic in HuWN: sometimes, synsets corresponding to idiom parts in PWN are marked as *non_lexicalized* in HuWN, or no idiom parts are mentioned within the synset. In order to solve this problem, we propose to include the whole idiom as a lexical unit in the verbal parts of wordnets, which can be easily matched to another idiomatic synset in the other language, without being forced to find a nominal component in both languages that have the same meaning within the MWE. Thus, the following synsets can be proposed:

> {be in the gutter, go down the sewer, be in the toilet}
>
> {lehúzhatja a WC-n (down.flush-MOD-3SGOBJ the toilet-SUP), kidobhatja az ablakon (out.throw-MOD-3SGOBJ the window-SUP)}

Although Osherson and Fellbaum (2010) suggest that parts of decomposable idioms should be consequently included in wordnets on the basis of the fact that there are idioms with the same or similar meanings, thus, their components may form a single synset (compare *seventh heaven* and *cloud nine*), they also admit that prepositions and other function word elements of idioms cannot be given in this way since PWN only includes nouns, verbs, adjectives and adverbs. In Hungarian, the situation is somewhat more complicated since nouns get suffixes in sentences, which cannot be signaled in any way by listing only parts (lemmas) of idioms. We argue, however, that including the whole idiom as one lexical unit is more beneficial from the aspect of multilinguality for it is easier to find the other language equivalent of idioms than the equivalent of idiom parts, and, on the other hand, the whole idiom is listed and not only its nominal, adjectival and verbal parts. Nouns in idioms also occur in the right grammatical form (i.e. with the correct suffix). In this way, non-lexicalized synsets related to idiom parts can also be eliminated. On the other hand, decomposable and non-decomposable idioms are treated in the same way: they are both listed as a whole. With this solution, idioms that share the same meaning should be treated similarly to single synonymous word, that is, they can be included within one synset.

4 Multiword verbs in the Hungarian WordNet

In the following, it is shown how multiword verbs are included in the conceptual hierarchy of the Hungarian Wordnet.

Among the 3607 verbal synsets of the Hungarian WordNet, 84 contain at least one multiword verb (106 altogether). Among them, 10 phrases consist of an adjective in the translative case and the verb *tesz* 'make', e.g. *jobbá tesz* (better-TRANSL makes) 'to ameliorate'. The English equivalents of these synsets are typically single verbs one meaning component of which is 'make' as it is shown in their definition, for instance:

> ID: ENG20-00498510-v
> Synonyms: {disable:1, disenable:1, incapacitate:1}
> Definition: make unable to perform a certain action

In Hungarian, the meaning component 'make' is explicitly expressed by the verb *tesz* 'make'.

Although there are some idiomatic expressions such as *dűlőre jut* (brink-SUB gets) 'to come to an

agreement' among multiword verbs in HuWn, most of them belong to the category of light verb constructions. Light verb constructions consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent. The Hungarian WordNet treats them as separate lexical units, that is, they behave as normal literals.

When constructing the Hungarian Wordnet, wordnet builders were given special instructions to include the most frequent light verb constructions in synsets (frequency data were estimated on the basis of the Hungarian National Corpus (Váradi, 2002)). They can be found in synsets together with their verbal counterparts as in:[3]

> ID: ENG20-00777368-v
> Synonyms: {engedélyez:1, **engedélyt ad:1**}
> Definition: Hatóság vagy hivatal engedélyt megad.

> ID: ENG20-00777368-v
> Synonyms: {authorize:1, authorise:2, pass:24, clear:4}
> Definition: Grant authorization or clearance for.

Sometimes, there are more than one light verb constructions within one synset, which entails that they are synonyms:

> ID: ENG20-00862885-v
> Synonyms: {**hálát ad:1**, **köszönetet mond:1**, **köszönetet nyilvánít:1**, megköszön:1, köszön:1}
> Definition: Köszönetét fejezi ki valakinek valamiért.

> ID: ENG20-00862885-v
> Synonyms: {thank:1, **give thanks:1**}
> Definition: Express gratitude or show appreciation to.

In certain cases, the synset contains only one light verb construction, that is, it is regarded as a separate lexical unit (having one entry in the dictionary or rather forming one synset in the wordnet):[4]

> ID: ENG20-00992244-v
> Synonyms: {**szóba hoz:1**}
> Definition: Szól róla, megemlíti.

> ID: ENG20-00992244-v
> Synonyms: {raise:19, bring up:6}
> Definition: Put forward for consideration or discussion.

Based on these examples, HuWN can be considered as a database in which light verb constructions are treated as separate lexical entries.[5] However, as wordnets contain several lexical relations among synsets, it would prove useful to link the synset of the nominal component to that of the light verb construction, e.g. the relation *derivative*[6] might connect them to each other, thus signaling their morphological and semantic interrelatedness (e.g. *engedély* 'permission' is paired with {engedélyez:1, engedélyt ad:1} 'authorize'). This extension of relations between synsets would be fruitful in the sense that the synsets of the construction and its components would be directly connected hence they could inherently be matched without any further steps. From a multilingual perspective, although *make a decision* and *döntést hoz* are translational equivalents, this cannot be deduced without analyzing the definition. In order to enrich the applicability of wordnets in the automatic translation of multiword expressions, we also suggest that light verb constructions be included in wordnets in a more systematic way, i.e. they should be literals within the synset and not only parts of the definition.

[3] The corresponding English synsets are imported from the Princeton WordNet, thus, they do not always contain a light verb construction and translation is not always word-by-word.

[4] However, the definition itself contains single word equivalents of the concept.

[5] As for PWN, light verb constructions are sometimes treated as lexical units (e.g. *give thanks*) but in other cases, it is the definition that contains the light verb construction equivalent of the literals (e.g. {decide:1, make up one's mind:1, determine:5} is defined as "reach, make, or come to a decision about something").

[6] In HuWN, no derivative relations have been included so far.

5 Discussion

We argue that multiword verbs such as idioms and light verb constructions should be listed as one unit in wordnets. In this way, there is no difference in treating decomposable and non-decomposable idioms and other language equivalents of the expressions are easier to find. From a theoretical point of view, this means that multiword expressions are regarded as a separate lexical unit, reflecting the semantic unity of the construction. This is in line with construction grammars (see e.g. Goldberg (1995)), where contents and forms are paired to form a construction with typically unpredictable meaning.

However, there is a difference between idioms and light verb constructions as regards their linking to the synsets of their members. Since the nominal component of light verb constructions preserves its original meaning to some degree, we proposed to connect the nominal component to the synset that contains the light verb construction. On the other hand, in the case of idioms no detectable connection between meanings of the parts of the idiom and the whole phrase can be established that is why we suggest not connecting them.

6 Conclusions

In this paper we have suggested that it is advisable to treat multiword verbs such as idioms and light verb constructions as one unit in wordnets. First, their semantic unity is reflected in this way, second, it is easier to match other language equivalents of the same units. We have also argued that a unified treatment should be applied for types of multiword verbs in order to enhance their applicability in multilingual applications. The revision of idioms and light verb constructions in wordnets is, however, left for future work.

Acknowledgments

This work was supported by the Project "TÁMOP-4.2.1/B-09/1/KONV-2010-0005 – Creating the Center of Excellence at the University of Szeged", supported by the European Union and co-financed by the European Regional Development Fund.

References

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1934–1940, Las Palmas.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Su Nam Kim. 2008. *Statistical Modeling of Multiword Expressions*. Ph.D. thesis, University of Melbourne, Melbourne.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 311–320, Szeged. University of Szeged.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70:491–538.

Anne Osherson and Christiane Fellbaum. 2010. The Representation Of Idioms In WordNet. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fifth Global WordNet Conference (GWC 2010)*, Mumbai, India. Narosa Publishing House.

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

Tamás Váradi. 2002. The Hungarian National Corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 385–389, Las Palmas de Gran Canaria. European Language Resources Association.