Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus

Veronika Vincze

Hungarian Academy of Sciences
Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

Abstract

In this paper, we describe the first English–Hungarian parallel corpus annotated for light verb constructions, which contains 14,261 sentence alignment units. Annotation principles and statistical data on the corpus are also provided, and English and Hungarian data are contrasted. On the basis of corpus data, a database containing pairs of English–Hungarian light verb constructions has been created as well. The corpus and the database can contribute to the automatic detection of light verb constructions and they can enhance performance in several fields of NLP (e.g. parsing, information extraction/retrieval and machine translation).

Keywords: light verb constructions, English–Hungarian parallel corpus, multilinguality

1. Introduction

In natural language processing (NLP), one of the most challenging tasks is the proper treatment of multiword expressions (MWEs). MWEs are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Calzolari et al., 2002). Light verb constructions form a subtype of multiword expressions. They consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb loses its original sense to some extent, e.g. *to give advice*, *to take into account*, *the problem lies (in)*. They are frequent in language use and because of their idiosyncratic behavior, they often pose a problem to NLP systems.

In this paper, we describe SzegedParalellFX, the first English–Hungarian parallel corpus annotated for light verb constructions. We believe that the corpus can contribute to the research on multiword expressions and more specifically, to the development of algorithms aiming at detecting light verb constructions.

The structure of the paper is as follows. First, related corpora and related work on the NLP treatment of multiword expressions are presented. Then the corpus is described together with annotation principles. Some statistical data on corpus data are also provided, which is followed by a qualitative analysis and a comparison of English and Hungarian data. The paper concludes with illustrating how the corpus and the database can be exploited in several fields of NLP.

2. Related work

Lately, multiword expressions have been received special interest in the NLP research community (Rayson et al., 2010). This also holds for multiword verbs since they constitute a subtype of multiword expressions, e.g. Sag et al. (2002) classify them as a subtype of lexicalized phrases and flexible expressions. The automatic identification of multiword verbs has been studied in several languages. Cook et al. (2007) differentiate between literal and idiomatic usages of verb and noun constructions in English. Their basic hypothesis is that the canonical form of each construction occurs mostly in idioms since they show syntactic variation to a lesser degree than constructions in literal usage. Hence, they make use of syntactic fixedness of idioms when developing their unsupervised method.

Van de Cruys and Moirón (2007) describe a semantic-based method for identifying verb-preposition-noun combinations in Dutch. Their method relies on selectional preferences for both the noun and the verb and they also make use of automatic noun clustering when considering the selection of semantic classes of nouns for each verb.

Bannard (2007) seeks to identify verb and noun constructions in English on the basis of syntactic fixedness. He examines whether the noun can have a determiner or not, whether the noun can be modified and whether the construction can have a passive form, which features are exploited in the identification of the constructions.

Gurrutxaga and Alegria (2011) extract idioms and light verb constructions from Basque texts by employing statistical methods. Since Basque is a free word-order language, they hypothesized that a wider window would yield more significant cooccurrence statistics, however, their initial experiments did not confirm this.

Tu and Roth (2011) classify verb + noun object pairs as being light verb constructions or not. They operate with both contextual and statistical features and conclude that on ambiguous examples, local contextual features perform better.

Vincze et al. (2011a) exploit shallow morphological features in identifying English light verb constructions and the domain specificity of the problem is emphasized in Nagy T. et al. (2011).

Parallel corpora are of high importance in the automatic identification of multiword expressions: it is usually one-to-many correspondence that is exploited when designing methods for detecting multiword expressions. On the other hand, aligned parallel corpora can also enhance the identification of multiword expressions in different languages: if an algorithm is implemented for one language, data from the other language can also be gathered with the help of aligned units.

For instance, Caseli et al. (2010) developed an alignment-based method for extracting multiword expressions from parallel corpora. The first step is to align the corpus on the sentence level, which is followed by POS-tagging. After this, sentence alignment units are word-aligned. Candidates for multiword expressions are produced by the word aligner and the POS-tagger as well, then they are filtered according to some empirically defined patterns or frequency data.

Zarrieß and Kuhn (2009) argue that multiword expressions can be reliably detected in parallel corpora by using dependency-parsed, word-aligned sentences. For one-to-many translation pairs, they apply a generate-and-filter strategy: first, aligned syntactic configurations are generated, which are then filtered and post-edited.

Sinha (2009) detects Hindi complex predicates (i.e. a combination of a light verb and a noun, a verb or an adjective) in a Hindi–English parallel corpus by identifying a mismatch of the Hindi light verb meaning in the aligned English sentence. Although the method requires the generation of all possible light verbs, it seems to be applicable to languages of the Indo Aryan family.

Many-to-one correspondence is also exploited in Attia et al. (2010) when identifying Arabic multiword expressions relying on asymmetries between entry titles of Wikipedia.

Tsvetkov and Wintner (2010) identify Hebrew multiword expressions by searching for misalignments in an English–Hebrew parallel corpus. MWE candidates are then ranked and filtered based on monolingual frequency data.

With regard to their NLP treatment, a database of light verb constructions and an annotated corpus might be of great help in the automatic recognition of light verb constructions. They can serve as a training database when implementing an algorithm for identifying those constructions, and they can also have an essential role in evaluating the methods developed.

There already exist some monolingual corpora annotated for light verb constructions. For instance, Grégoire (2010) presents a lexicon of Dutch multiword expressions (DuELME). Kaalep and Muischnek (2008) describe an Estonian database and a corpus of multiword verbs and Krenn (2008) developed a database of German PP-verb combinations. The Prague Dependency Treebank is also annotated for light verb constructions (Cinková and Kolářová, 2005). NomBank (Meyers et al., 2004) contains the argument structure of common nouns, including those occurring in support verb constructions as well. Literal and idiomatic usages of English verb + noun combinations are annotated in the VNC-Tokens dataset (Cook et al., 2008). An example of corpus-based identification of light verb constructions in English is described in Tan et al. (2006). An annotated corpus and a database containing Hungarian light verb constructions has been recently developed (Vincze and Csirik, 2010).

To the best of our knowledge, no parallel corpora have been manually annotated for light verb constructions. With this motivation in mind, we developed an English–Hungarian parallel corpus in which light verb constructions are annotated. Our corpus can prove useful in the automatic evaluation of methods that are able to identify English and/or Hungarian light verb constructions in texts.

3. The corpus

In this section, the English-Hungarian parallel corpus annotated for light verb constructions will be presented. Annotation principles and statistical data will also be provided. Finally, some qualitative analysis of data and interlingual differences will be discussed.

Texts to be annotated were selected on the basis of their topics from the SzegedParalell English–Hungarian parallel corpus (Tóth et al., 2008), which contains 99,745 manually aligned sentence alignment units (SAUs). Since it is primarily texts on economics, law and the like written in an official style that are expected to contain a number of light verb constructions as earlier results on monolingual data indicated (Vincze and Csirik, 2010), texts belonging to these domains were all annotated for light verb constructions together with some novels and language book sentences. With this selection of texts for annotation, it can be examined whether there are any differences between texts

- from different domains (e.g. between economic-legal texts and literature);
- from different source languages;
- from different periods (Jonathan Swift's *Gulliver's Travels* was published in 1726, Mark Twain's *The Man That Corrupted Hadleyburg* in 1900 and Frigyes Karinthy's *Tanár úr kérem* (*Please, Sir!*) in 1916,[1] thus, differences between earlier and contemporary language use (found in magazine texts or language books) might also be revealed).

Data on annotated texts can be seen in Table 1.

Subcorpus	# of texts	# of SAUs
EU	30	1518
Bilingual magazines	151	5320
Language book sent.	7	3496
Literature	3	3232
Miscellaneous	5	695
Total	196	14,261

Table 1: Texts and sentence alignment units in SzegedParalellFX.

3.1. Types of light verb constructions

Light verb constructions may occur in various forms due to their syntactic flexibility. Besides the prototypical noun + verb combination in Hungarian and the verb + noun combination in English (VERB), light verb constructions may be present in different syntactic structures, that is, in participles (PART, e.g. *photos taken*) and they may also undergo nominalization, yielding a nominal compound (NOM, e.g. *service provider*).[2] Split light verb constructions (SPLIT, e.g. ***a decision*** *has been recently* ***made***),

[1] Their translations were published in 1906, 1955 and 1968, respectively.

[2] It should be mentioned that nominal components occurring without the verb (e.g. *decision on the future*) are sometimes considered as a type of light verb constructions, e.g. in Laporte et al. (2008). However, we restricted ourselves to annotate cases where both the nominal component and the verb are present.

where the noun and the verb are not adjacent, are also annotated and tagged. It must be mentioned that split light verb constructions are a subclass of the VERB category, however, they were marked distinctively and the nominal and the verbal component are also marked within the construction because in this way, their identification becomes possible and the database can be used for training an algorithm that automatically recognizes (split) light verb constructions.

These types are all annotated in the corpus texts since they also occur relatively frequently (see Table 2). Furthermore, it is also important to identify all these types since applications like machine translation should also treat them in a specific way (i.e. it is not only verbal occurrences that should be translated as lexical units but participles and nominalized forms as well).

3.2. Annotation principles

Three native speakers of Hungarian who could speak English at an advanced level carried out the annotation. Corpus texts contain single annotation, i.e. one annotator worked on each text. The same annotator worked on both the source and the target language versions of each text. Texts contain stand-off annotation, that is, original texts and the annotation are stored in different files.

In order to decide whether a noun + verb combination is a light verb construction or not, annotators were suggested to make use of a test battery including questions such as *Can a verb (derived from the same root as the nominal component) substitute the construction?*, *When omitting the verb (e.g. in a possessive construction), can the original action be reconstructed?*, *Can the construction itself be nominalized?*, *Can the construction be passivized?* etc. Although there exist some extraction-based methods developed for collecting MWEs from natural language texts (Ramisch et al., 2010), we argue that it is important to annotate each occurrence of MWEs in text. The reason behind this is that in certain cases a given text span functions as an MWE while in other contexts, it does not. For instance, *make decisions* is definitely a light verb construction in *The government will make decisions on foreign policy issues* whereas in *They will make decisions on the issues publicly available* it is the causative verb *make* that precedes the noun *decisions* and thus they do not form a light verb construction. If it was only known that the sequence *make decisions* is a light verb construction (e.g. based on MWE lists), the latter occurrence would also be annotated as such. Another example is *give a ring*: when *ring* means "calling", then it is a light verb construction, on the other hand, when *ring* is a piece of jewellery, it is just a verb-object pair. By annotating the whole corpus for light verb constructions (that is, deciding whether the candidate text span is a light verb construction in the given context), such pseudo-light verb constructions can be discarded, and the frequency of such cases can also be estimated.

Besides the prototypical occurrences of light verb constructions (i.e. a bare common noun + verb[3]), other instances were also annotated in the corpus. For instance, the noun might be accompanied by an article or a modifier (recall that phrase boundaries were considered during annotation) or – for word order requirements – the noun follows the verb as in:

Ő	*hozta*	*a*	*jó*	*döntést.*
he	bring-PAST-3SG-OBJ	the	good	decision-ACC

'It was him who made the good decision.'

For the reasons mentioned in Section 3.1., a single light verb construction manifests in several different forms in the corpus. However, each occurrence was manually paired with its prototypical (i.e. bare noun + verb) form. The lists of prototypical forms for both languages are available at the corpus website (`http://www.inf.u-szeged.hu/rgai/mwe`).

3.3. Statistics on corpus data

The total number and the number of the subtypes of light verb constructions are presented in Table 2. In each cell, the first number refers to the English data and the second to the Hungarian data.

The number of English and Hungarian light verb constructions is approximately the same, thus, approximately the same percentage of sentence alignment units contains a light verb construction (see Table 2). However, it does not entail that each light verb construction in the corpus has an equivalent in the other language – in other words, the translational equivalents of certain constructions are single verbs rather than constructions (e.g. *break into smile – elmosolyodik*).

In the Hungarian part of the corpus, there are 1377 occurrences of 703 light verb constructions, thus, a specific construction occurs 1.96 times in the corpus on average. Concerning English, 727 light verb constructions occur altogether 1371 times (1.89 times each on average). In Hungarian, 9.66% of SAUs contain a light verb construction on average whereas in English this percentage is 9.61%. These numbers are comparable to the ratio of light verb constructions in the Hungarian Szeged Treebank and in the Wiki50 corpus: 8.2%, and 8.46%, respectively (Vincze and Csirik, 2010; Vincze et al., 2011b). This suggests that in different types of texts, the average ratio of light verb constructions per sentence is about 8-9%, which is true for Hungarian and English as well.

As for the types of light verb constructions, it is revealed that the number of verbal and nominal occurrences is (basically) the same in the two languages, on the other hand, there is a considerable difference between the number of participles and split constructions. This may be the result of grammatical differences between English and Hungarian. For example, most instances in the category SPLIT form a passive construction in English, where the nominal component of the construction functions as the subject hence it is not adjacent to the verb, while passive constructions are hardly used in present-day Hungarian and split

[3] As opposed to other languages where prototypical light verb constructions consist of a verb + a noun in accusative or a verb + a prepositional phrase – see e.g. Krenn (2008) –, in Hungarian, postpositional phrases rarely occur within a light verb construction. However, annotators were told to annotate such cases as well.

Subcorpus	VERB	PART	NOM	SPLIT	Total	LVC/sentence%
EU	132 / 158	30 / 76	24 / 32	41 / 29	227 / 295	14.95 / 19.43
Bilingual magazines	356 / 387	55 / 120	31 / 42	83 / 53	525 / 602	9.87 / 11.32
Language book sent.	158 / 79	5 / 21	14 / 4	22 / 15	199 / 119	5.69 / 3.4
Literature	270 / 261	15 / 24	6 / 5	119 / 57	410 / 347	12.69 / 10.74
Miscellaneous	7 / 12	1 / 1	1 / 0	1 / 1	10 / 14	1.44 / 2.01
Total	923 / 897	106 / 242	76 / 83	266 / 155	1371 / 1377	9.61 / 9.66

Table 2: English/Hungarian light verb constructions in SzegedParalellFX.

light verb constructions are typically due to changes in the information structure of the sentence. Concerning the category PART, a premodifier before the nominal component requires the presence of the participle form of the verbal component in Hungarian, however, in English, its equivalent is mostly a postmodifier, which may or may not be accompanied with a participle, as in

az emberi jogokba ***vetett*** *hit*
the human right-PLUR-INE cast-PAST-PART belief
"a belief in human rights"

The data gained from the parallel corpus were manually converted into a database of pairs of English–Hungarian light verb constructions, that is, English light verb constructions were paired with their Hungarian equivalents. Moreover, the verbal counterparts of the light verb constructions are also included in the list (wherever applicable), which contains 344 pairs of light verb constructions and is available at the corpus website (`http://www.inf.u-szeged.hu/rgai/mwe`).

3.4. Inter-annotator agreement

In order to compare the difficulty of annotating light verb constructions in both English and Hungarian, 928 sentence alignment units were annotated by all the annotators and later differences were resolved, yielding the gold standard annotation.

Agreement rate was calculated at two levels: first, it was only considered whether the given light verb construction was marked (i.e. no type was taken into account). At this level, the average agreement rate among the annotations was 78.15% on the English data and 74.23% on the Hungarian data (agreement rates are given in F-measure). Second, the type of the light verb construction was also taken into consideration, that is, if the construction was marked but with a different label (e.g. PART instead of NOM), it also counted as an error. At this stricter level of measurement, the average agreement rates were 64.79% and 71.18% on the English and Hungarian data, respectively. At Level 2, the metrics Jaccard index and κ-measure were also calculated: on the English data, the agreement rates are 0.5049 and 0.5934 whereas on the Hungarian data, 0.5754 and 0.6575, which can be regarded as fairly good agreement rates.

The above data shed light on the fact that on average, annotation was somewhat easier for Hungarian than for English. According to the κ-measure metrics, moderate agreement can be reached on English data while substantial agreement on Hungarian. This may be traced back to the fact that the annotators were native speakers of Hungarian who could speak English at an advanced level, however, the latter was not their mother tongue. Still, it is interesting to see that at Level 1, better results can be achieved in English than in Hungarian (78.15% vs. 74.23%). It might be the case that reading in the mother tongue and reading in a foreign language requires different concentration skills and techniques and probably more effort, thus, while reading in Hungarian they were more prone to overlook certain constructions.

However, differentiating between types of light verb constructions (i.e. annotating at Level 2) usually led to considerable decline of performance, which is especially true for the English data, where Annotator 1 often labeled English gerunds as PART while the others considered them NOMs. Since the grammatical forms of gerunds and present participles coincide in English (i.e. they both have the ending *-ing*), this might – at least partially – serve as an explanation for this huge difference between the two levels in English (13.36% vs. 3.05% in Hungarian). In Hungarian, there is no such ambiguity of wordforms in the corpora, thus, the difference between the two levels is not substantial.

Interesting differences can be also revealed if the performance of annotators are contrasted. Annotator 1 achieved much better results on the Hungarian data than on the English data. Her moderate performance on the the English data may be explained by the errors related to the NOM and PART categories (see above). However, in Hungarian she could achieve substantial agreement with the gold standard annotation. Annotator 2 achieved moderate results in both languages, however, his performance on the English data was better than on the Hungarian data. Annotator 3 had the most experience in annotating linguistic corpora, which manifested in perfect precision. Thus, in her case, annotation errors were related only to recall. In other words, she failed to recognize some instances of light verb constructions in text, but the text spans she marked were indeed light verb constructions.

3.5. Comparing English and Hungarian data

The comparison of the English and Hungarian verbal components reveals that there is not much difference between the two languages: the translational equivalents *ad – give*, *vesz – take*, *tesz – make/do/put*, *tart – hold/keep* and *hoz – bring* all occur among the most frequent verbal components. There is one notable exception: *have* does not have a direct equivalent in Hungarian since there is no separate verb of possession in Hungarian. However, in the English data this verb is the fourth most frequent one.

As for the domains of the texts, it is revealed that economic

and legal texts (i.e. texts on the European Union) contain the most light verb constructions (on average). However, in miscellaneous texts and language book sentences there are hardly such constructions, which might suggest that it is mostly grammatical aspects that were considered when creating the sentences instead of aspects of vocabulary acquisition.

A cross-linguistic difference is that English literary texts contain light verb constructions in a much bigger rate than their Hungarian counterparts – this is especially true for *Gulliver's Travels* (20% of sentence alignment units contain a light verb construction, which is the highest rate in the English subcorpus). However, it must also be admitted that *Gulliver's Travels* was published in 1726, thus it reflects the early 18th century language use, which might be a reason for the difference between English and Hungarian literary texts. Nevertheless, it might be too hasty to conclude that the English language of that period contained more light verb constructions than contemporary English – more substantial research in historical linguistics is needed to investigate this issue.

When examining the matching of light verb constructions across languages, it can be found that in texts on the EU, constructions can be paired with their target language equivalents (in other words, if one language applies a construction, it is highly probable that the other language also employs a construction). Nevertheless, this tendency does not hold for literary texts. First, English texts contain much more light verb constructions than Hungarian ones (see above) except for the Mark Twain novel, where their number is almost the same, second, it is very common that the equivalent of the light verb construction is not a construction (or not even a verb), for instance:

*The emperor **gave orders** to have a bed prepared for me.*

A	*császár*	***parancsára,***	*fekvőhelyet*	
the	emperor	order-3SGPOSS-SUB	bed-ACC	
készítettek	*nekem.*			
make-PAST-3PL	for.me			

In this example, the English light verb construction corresponds to a noun in Hungarian.

From the above it can be concluded that literary texts are less likely candidates to be used as training or test databases for algorithms that aim to automatically align light verb constructions from different languages – as opposed to e.g. legal or economic texts or newspaper articles.

In certain cases, one language applies a construction while the other a verb – typically derived from the same root as the nominal component of the construction:

*It **decided** to welcome 10 more countries to join the EU on 1 May 2004.*

A	*Tanács*	***meghozta***						***a***
the	Council	PREVERB-bring-PAST-3SG-OBJ						the
döntést	*arról,*	*hogy*	*2004.*	*május*				
decision-ACC	that-DEL	that	2004	May				
1-jén	*10*	*új*	*államot*		*vesznek*	*fel*		
1-3SGPOSS-SUP	10	new	state-PL-ACC		take-3PL	up		
az	*Unió*	*tagállamai*						
the	Union	member.state-3SGPOSS-PL						
sorába.								
line-3SGPOSS-ILL								

The Hungarian construction contains the nominal component *döntés* "decision", which is derived morphologically from the verb *dönt* "decide", however, in English, it is the verb *decide* that appears in the sentence.

A final interesting fact is that English passive constructions are frequently paired with light verb constructions including the verb *kerül* "get":

The song "Auld Lang Syne" was partially written by Robert Burns and published after his death in 1796.

A	*híres*	*"Auld*	*Lang*	*Syne"*	*("Régóta*	
the	famous	"Auld	Lang	Syne"	("for.a.long.time	
már")	*című*	*dalt*	*részben*	*Robert*	*Burns*	
already")	entitled	song	part-INE	Robert	Burns	
írta,		*és*	*halála*		*után,*	
write-PAST-3SG-OBJ		and	death-3SGPOSS		after	
1796-ban	***került***		***kiadásra***.			
1796-INE	get-PAST-3SG		publication-SUB			

This qualitative analysis of data may be fruitfully applied in contrastive linguistics, (machine) translation and cross-language information retrieval.

4. The usability of the corpus and the database

The corpus created can have an important role in training and testing algorithms implemented for identifying light verb constructions. Furthermore, several NLP applications like information extraction or modality detection can profit from the corpus and the multilingual database can be utilized in machine translation and multilingual applications.

4.1. Syntactic parsing

Traditionally, the identification of multiword expressions is based on syntactic information, that is, it follows syntactic parsing, e.g. Martens and Vandeghinste (2010) make use of dependency trees when identifying syntactically motivated multiword expressions. However, Wehrli et al. (2010) argue that collocations can highly contribute to the performance of the parser since many parsing ambiguities can be excluded if collocations are known and treated as one syntactic unit. At an early phase of parsing, it is checked whether the terms to be attached bear the lexical feature [+partOfCollocation] and if the combination of those terms can be found in the collocational database, the corresponding parse tree is prioritized over other possible derivations. For the nominal component is a special argument of the verb – they form one complex predicate –, their special relation should also be recognized by the parser. A database containing light verb constructions enables the parser to identify light verb constructions and to assign a proper syntactic analysis to them, which can be later exploited by higher-level applications such as information extraction.

4.2. Information extraction/retrieval

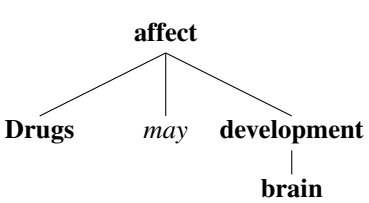
In information extraction, the proper identification of the predicate of the sentence plays an important role since it is the predicate that conveys core information on the event described and it is the arguments of the predicate that provide additional information on the circumstances and participants of the event. For these reasons, it is essential that the nominal component should be attached to the light verb in order that they can form a complex predicate and all the other arguments belong to the complex predicate. Thus, it can be assured that the sentence *Stan has fallen in love with Wendy* describes an event of *falling in love* with two participants (*Stan*, *Wendy*) and not simply *falling* with two participants (*Stan*, *Wendy*) and a location (*love*).

The database created can also be exploited in (cross-language) information retrieval. The verbal counterparts of most light verb constructions are provided in the database hence it can be seen as a multilingual list of synonyms, which can be made use of when matching the user's query to documents written in different languages: e.g. for the query *participate*, English and Hungarian documents containing *take part* or *részt vesz* (lit. part-ACC take) can also be retrieved.

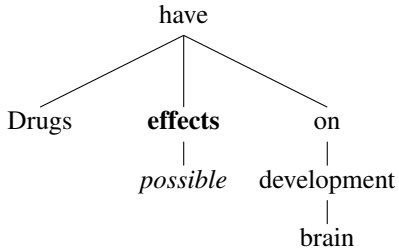
The list can also be used for extending wordnets. Only the most frequent light verb constructions were included in the Hungarian WordNet (Miháltz et al., 2008). However, in the Princeton Wordnet (Miller et al., 1990), the typical tendency is that the synset contains the verbal counterpart as a literal, which is defined by a light verb construction (e.g. *advise:1; counsel:1* 'give advice to'). Matching the elements of the bilingual list with existing synsets makes it possible to automatically extend synsets with synonyms that have not been included.

4.3. Modality detection

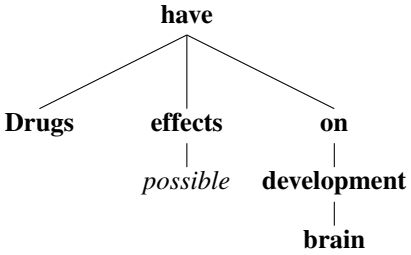
Current modality detectors mostly use syntactic features in order to determine what is in the scope of negation or speculation (Kilicoglu and Bergler, 2009; Farkas et al., 2010). If an element is negated/speculated, all its dependents are negated/speculated (as it is usually defined in negation/uncertainty detection systems). It entails that the whole proposition is under speculative scope because the main verb of the sentence is modified by the auxiliary *may* and all the other elements in the sentence are dependents of the main verb (the cue is italicized and the (extended) scope is bold):



However, in the following it is only the nominal component *effects* that is modified by the adjective. If the rules for detecting the scope of speculation are observed, speculation cannot be extended to the verb hence to the whole proposition since the verb is not an argument of *effects* (i.e. the modified element):



Apparently, the speculation scopes of the two sentences differ from each other. However, this problem can be overcome if it is recalled that the noun and the verb together form a complex predicate. In this given case, it is straightforward that the predicate should also include the verb *have* and the noun *effects* (if the verb and the nominal component were kept separated, the subject and the prepositional complement could not be in the scope of speculation since they are not a dependent of solely the noun). Thus, if the nominal component is modified by a speculative element, the scope is extended to the verbal component in the first step, then to the dependents of the verb as well:



In other words, either the verb or the nominal component is modified by a speculative element, they are treated in the same way: the other arguments of the verb or rather the light verb construction are also included in the scope of speculation, which is plausible from an applicational point of view: sentences with the same propositional content are treated uniformly. Our corpus and database can contribute to the identification of such constructions, in this way, the accuracy of modality detection can also be improved.

4.4. Machine translation

In the field of machine translation, lists of multiword expressions are of high significance. Since multiword expressions cannot be usually translated word-by-word from the source language to the target language, it is essential to include them in the dictionary. By integrating our database of English–Hungarian light verb constructions into a machine translation system, the quality of the translation is expected to improve.

Statistical machine translation relies heavily on word alignment: source words and target words are mapped to each other in parallel sentences, usually found in parallel corpora. However, previously known multiword expressions can enhance word alignment as it is emphasized in e.g. Okita et al. (2010). In the SzegedParalellFX corpus, which is manually aligned on the sentence level, light verb constructions can be used as anchors for automatic word alignment. These features can be exploited in statistical machine translation systems and the annotation of light verb constructions would most probably have a beneficial effect on translating light verb constructions even if bilingual lists

of multiword expressions are not integrated into the system. Statistical data on co-occurrence frequencies can also be used when automatically translating light verb constructions without bilingual lists or manually aligned corpora. For instance, *make a decision* and *take a decision* are both perfectly sound constructions in English. However, when translating the expressions word by word to Hungarian, the result would be *döntést tesz* and *döntést vesz*. Based on frequency data in large corpora, the possibility of translating either *take a decision* as *döntést vesz* or *make a decision* as *döntést tesz* is very low, thus, they are very improbable translation pairs. On the other hand, *döntés* "decision" co-occurs with a relatively high frequency with *hoz* "bring" hence *döntést hoz* would be probably judged by the system as the best candidate for translating these expressions into Hungarian.

4.5. Applications outside NLP

Other fields of linguistics can also profit from the corpus and the database. For instance, lexicographers can integrate the database into dictionaries while researchers of contrastive linguistics can also draw some conclusions concerning the differences between English and Hungarian data. Finally, the database and results of the qualitative analysis of data can be applied in language teaching as well.

5. Conclusions

In this paper, the first English–Hungarian parallel corpus annotated for light verb constructions was presented. From this corpus, light verb constructions were collected, producing a database that contains 703 light verb constructions in Hungarian and 727 in English. The annotated corpus and the database are available under the Creative Commons license at `http://www.inf.u-szeged.hu/rgai/mwe`.

The quantitative and qualitative comparison of data between domains and the two languages revealed interesting facts and tendencies that might be fruitfully applied in several fields of theoretical and applied linguistics besides NLP applications. We firmly believe that our corpus can also contribute to the NLP research on multiword expressions, more specifically, to the development and evaluation of algorithms aiming at detecting light verb constructions.

Acknowledgments

This work was supported in part by the National Innovation Office of the Hungarian government within the framework of the project MASZEKER. The author wishes to thank the annotators of the corpus for their devoted efforts.

6. References

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic Extraction of Arabic Multiword Expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 19–27, Beijing, China. Coling 2010 Organizing Committee.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 1–8, Morristown, NJ, USA. ACL.

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.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Silvie Cinková and Veronika Kolářová. 2005. Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Mária Šimková, editor, *Insight into Slovak and Czech Corpus Linguistics*, pages 113–139. Veda Bratislava, Slovakia.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 41–48, Morristown, NJ, USA. ACL.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22, Marrakech, Morocco.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12, Uppsala, Sweden. ACL.

Nicole Grégoire. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2):23–39.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 2–7, Portland, Oregon, USA. ACL.

Heiki-Jaan Kaalep and Kadri Muischnek. 2008. Multi-Word Verbs of Estonian: a Database and a Corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 23–26, Marrakech, Morocco.

Halil Kilicoglu and Sabine Bergler. 2009. Syntactic Dependency Based Heuristics for Biological Event Extraction. In *Proceedings of the BioNLP Workshop Companion Volume for Shared Task*, pages 119–127.

Brigitte Krenn. 2008. Description of Evaluation Resource – German PP-verb data. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 7–10, Marrakech, Morocco.

Éric Laporte, Elisabete Ranchhod, and Anastasia Yanna-

copoulou. 2008. Syntactic variation of support verb constructions. *Lingvisticae Investigationes*, 31(2):173–185. DOI: 10.1075/li.31.2.04lap.

Scott Martens and Vincent Vandeghinste. 2010. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 85–88, Beijing, China. Coling 2010 Organizing Committee.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An Interim Report. In Adam Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7. ACL.

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.

István Nagy T., Veronika Vincze, and Gábor Berend. 2011. Domain-Dependent Identification of Multiword Expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 622–627, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Tsuyoshi Okita, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010. Multi-word expression-sensitive word alignment. In *Proceedings of the 4th Workshop on Cross Lingual Information Access*, pages 26–34, Beijing, China. Coling 2010 Organizing Committee.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword Expressions in the wild? The mwetoolkit comes in handy. In *Coling 2010: Demonstrations*, pages 57–60, Beijing, China. Coling 2010 Organizing Committee.

Paul Rayson, Scott Songlin Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1-2):1–5.

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.

R. Mahesh K. Sinha. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46, Singapore. ACL.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*, pages 49–56, Trento, Italy. ACL.

Krisztina Tóth, Richárd Farkas, and András Kocsor. 2008. Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. *Acta Cybernetica*, 18(3):463–478.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China. Coling 2010 Organizing Committee.

Yuancheng Tu and Dan Roth. 2011. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA. ACL.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 25–32, Morristown, NJ, USA. ACL.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118, Beijing, China. Coling 2010 Organizing Committee.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011a. Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 116–121, Portland, Oregon, USA. ACL.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011b. Multiword Expressions and Named Entities in the Wiki50 Corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Eric Wehrli, Violeta Seretan, and Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 28–36, Beijing, China. Coling 2010 Organizing Committee.

Sina Zarrieß and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30, Singapore. ACL.

Keywords: light verb constructions, Englishâ€“Hungarian parallel corpus, multilinguality. 1. Introduction. In natural language processing (NLP), one of the most chal-lenging tasks is the proper treatment of multiword expres-sions (MWEs).Â In this paper, we describe SzegedParalellFX, the rst Englishâ€“Hungarian parallel corpus annotated for light verb constructions. We believe that the corpus can contribute to the research on multiword expressions and more speci-cally, to the development of algorithms aiming at detecting light verb constructions. The structure of the paper is as follows. First, related cor-pora and related work on the NLP treatment of multiword expressions are presented. Then the corpus is described to-gether with annotation principles. Light verb constructions consist of a verbal and a nominal component, where the noun preserves its original meaning while the verb has lost it (to some degree). They are syntactically flexible and their meaning can only be partially computed on the basis of the meaning of their parts, thus they require special treatment in natural language processing. For this purpose, the first step is to identify light verb constructions.Â Light verb constructions in the SzegedParalellFX English--Hungarian parallel corpus. In Proceedings of LREC'12. 53. Veronika Vincze , JÃ¡nos Csirik, Hungarian corpus of light verb constructions, Proceedings of the 23rd International Conference on Computational Linguistics, p.1110-1118, August 23-27, 2010, Beijing, China. 54.