Hungarian Corpus of Light Verb Constructions

Veronika Vincze
University of Szeged
Department of Informatics
vinczev@inf.u-szeged.hu

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

Abstract

The precise identification of light verb constructions is crucial for the successful functioning of several NLP applications. In order to facilitate the development of an algorithm that is capable of recognizing them, a manually annotated corpus of light verb constructions has been built for Hungarian. Basic annotation guidelines and statistical data on the corpus are also presented in the paper. It is also shown how applications in the fields of machine translation and information extraction can make use of such a corpus and an algorithm.

1 Introduction

In this paper, we report a corpus containing light verb constructions in Hungarian. These expressions are neither productive nor idiomatic and their meaning is not totally compositional (the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent), as it can be seen in the examples from different languages shown below. Since their meaning is the same, only literal translations are provided:

- English: *to give a lecture*, *to come into bloom*, *the problem lies (in)*
- German: *halten eine Vorlesung* to hold a presentation, *in Blüte stehen* in bloom to stand, *das Problem liegt (in)* the problem lies (in)
- French: *faire une présentation* to make a presentation, *être en fleur* to be in bloom, *le problème réside (dans)* the problem resides (in)
- Hungarian: *előadást tart* presentation-ACC holds, *virágba borul* bloom-ILL falls, *probléma rejlik (vmiben)* problem hides (in sg)

Several terms like *complex verb structures*, *support verb constructions* or *light verb constructions* have been used[1] for these constructions in the literature (Langer, 2004). In this paper, the term *light verb constructions* will be employed.

The structure of the paper is as follows. First, the importance of the special NLP treatment of light verb constructions is emphasized in section 2. The precise identification of such constructions is crucial for the successful functioning of NLP applications, thus, it is argued that an algorithm is needed to automatically recognize them (section 4). In order to facilitate the development of such an algorithm, a corpus of light verb constructions has been built for Hungarian, which is presented together with statistical data in section 5. Finally, it is shown how NLP applications in the fields of machine translation and information extraction can profit from the implementation of an algorithm capable of identifying light verb constructions (section 6).

2 Light verb constructions in NLP

In natural language processing, one of the most challenging tasks is the proper treatment of col-

[1] There might be slight theoretical differences in the usage of these terms – e.g. semantically empty support verbs are called *light verbs* in e.g. Meyers et al. (2004a), that is, the term *support verb* is a hypernym of *light verb*. However, these differences are not analyzed in detail in this paper.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1110–1118,
Beijing, August 2010

locations, which term comprises light verb constructions as well. Every multiword expression is considered to be a collocation if its members often co-occur and its form is fixed to some extent (Siepmann, 2005; Siepmann, 2006; Sag et al., 2001; Oravecz et al., 2004; Váradi, 2006). Collocations are frequent in language use and they usually exhibit unique behaviour, thus, they often pose a problem to NLP systems.

Light verb constructions deserve special attention in NLP applications for several reasons. First, their meaning is not totally compositional, that is, it cannot be computed on the basis of the meanings of the parts of the collocation and the way they are related to each other. Thus, the result of translating the parts of the collocation can hardly be considered as the proper translation of the original expression. Second, light verb constructions (e.g. *make a mistake*) often share their syntactic pattern with other constructions such as literal verb + noun combinations (e.g. *make a cake*) or idioms (e.g. *make a meal*), thus, their identification cannot be based on solely syntactic patterns. Third, since the syntactic and the semantic head of the construction are not the same – the syntactic head being the verb and the semantic head being the noun –, they require special treatment when parsing. It can be argued that they form a complex verb similarly to phrasal or prepositional verbs (as reflected in the term complex verb structures). Thus, it is advisable to indicate their special syntacto-semantic relationship: in dependency grammars, the new role QUASI-ARGUMENT might be proposed for this purpose.

3 Related work

Light verb constructions – as a subtype of multiword expressions – have been paid special attention in NLP literature. Sag et al. (2001) classify them as a subtype of lexicalized phrases and flexible expressions. They are usually distinguished from productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other hand: e.g. Fazly and Stevenson (2007) use statistical measures in order to classify subtypes of verb + noun combinations and Diab and Bhutada (2009) developed a chunking method for classifying multiword expressions.

Identifying multiword expressions in general and light verb constructions in particular is not unequivocal since constructions with similar syntactic structure (e.g. verb + noun combinations) can belong to different subclasses on the productivity scale (i.e. productive combinations, light verb constructions and idioms). That is why well-designed and tagged corpora of multiword expressions are invaluable resources for training and testing algorithms that are able to identify multiword expressions. For instance, Grégoire (2007) describes the design and implementation of a lexicon of Dutch multiword expressions. Focusing on multiword verbs, Kaalep and Muischnek (2006; 2008) present an Estonian database and a corpus and Krenn (2008) describes a database of German PP-verb combinations. The Prague Dependency Treebank also contains annotation for light verb constructions (Cinková and Kolářová, 2005) and NomBank (Meyers et al., 2004b) provides the argument structure of common nouns, paying attention to those occurring in support verb constructions as well. On the other hand, Zarrieß and Kuhn (2009) make use of translational correspondences when identifying multiword expressions (among them, light verb constructions). A further example of corpus-based identification of light verb constructions in English is described in Tan et al. (2006).

Light verb constructions are considered to be semi-productive, that is, certain verbs tend to co-occur with nouns belonging to a given semantic class. A statistical method is applied to measure the acceptability of possible light verb constructions in Stevenson et al. (2004), which correlates reasonably well with human judgments.

4 Identifying light verb constructions

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.

The recognition of light verb constructions cannot be solely based on syntactic patterns for other (productive or idiomatic) combinations may exhibit the same verb + noun scheme (see section

2). However, in agglutinative languages such as Hungarian, nouns can have several grammatical cases, some of which typically occur in a light verb construction when paired with a certain verb. For instance, the verb *hoz* 'bring' is a transitive verb, that is, it usually occurs with a noun in the accusative case. On the other hand, when it is preceded or followed by a noun in the sublative or illative case (the typical position of the noun in Hungarian light verb constructions being right before or after the verb[2]), it is most likely a light verb construction. To illustrate this, we offer some examples:

vizet hoz

water-ACC bring

'to bring some water'

zavarba hoz

trouble-ILL bring

'to embarrass'

The first one is a productive combination (with the noun being in the accusative form) while the second one is a light verb construction. Note that the light verb construction also has got an argument in the accusative case (syntactically speaking, a direct object complement) as in:

Ez a megjegyzés mindenkit zavarba hozott.

this the remark everyone-ACC trouble-ILL bring-PAST-3SG

'This remark embarrassed everybody.'

Thus, the presence of an argument in the accusative does not imply that the noun + verb combination is a light verb construction. On the other hand, the presence of a noun in the illative or sublative case immediately preceding or following the verb strongly suggests that a light verb instance of *hoz* is under investigation.

Most light verb constructions have a verbal counterpart derived from the same stem as the noun, which entails that it is mostly deverbal nouns that occur in light verb constructions (as in *make/take a* ***decision*** compared to ***decide*** or ***döntést*** *hoz* vs. ***dönt*** in Hungarian). The identification of such nouns is possible with the help of a morphosyntactic parser that is able to treat derivation as well (e.g. hunmorph for Hungarian (Trón et al., 2005)), and the combination of a possible light verb and a deverbal noun typically results in a light verb construction.

Thus, an algorithm that makes use of morphosyntactic and derivational information and previously given lists can be constructed to identify light verb constructions in texts. It is important that the identification of light verb constructions precedes syntactic parsing, for the noun and the verb in the construction form one complex predicate, which has its effects on parsing: other arguments belong not solely to the verb but to the complex predicate.

To the best of our knowledge, there are no corpora of light verb constructions available for Hungarian. That is why we decided to build such a corpus. The corpus is described in detail in section 5. On the basis of the corpus developed, we plan to design an algorithm to automatically identify light verb constructions in Hungarian.

5 The corpus

In order to facilitate the extraction and the NLP treatment of Hungarian light verb constructions, we decided to build a corpus in which light verb constructions are annotated. The Szeged Treebank (Csendes et al., 2005) – a database in which words are morphosyntactically tagged and sentences are syntactically parsed – constitutes the basis for the annotation. We first selected the subcorpora containing business news, newspaper texts and legal texts for annotation since light verb constructions are considered to frequently occur in these domains (see B. Kovács (1999)). However, we plan to extend the annotation to other subcorpora as well (e.g. literary texts) in a later phase. Statistical data on the annotated subcorpora can be seen in Table 1.

5.1 Types of light verb constructions

As Hungarian is an agglutinative language, light verb constructions may occur in various forms.

[2]In a neutral sentence, the noun is right before the verb, in a sentence containing focus, it is right after the verb.

	sentences	words
business news	9574	186030
newspapers	10210	182172
legal texts	9278	220069
total	29062	582871

Table 1: Number of sentences and words in the annotated subcorpora

For instance, the verbal component may be inflected for tense, mood, person, number, etc. However, these inflectional differences can be easily resolved by a lemmatizer. On the other hand, besides the prototypical noun + verb combination, light verb constructions may be present in different syntactic structures, that is, in participles and infinitives and they can also undergo nominalization. These types are all annotated in the corpus texts since they also occur relatively frequently (see statistical data in 5.3). All annotated types are illustrated below.

- **Noun + verb combination <verb>**

  bejelentést tesz

  announcement-ACC makes

  'to make an announcement'

- **Participles <part>**

  - Present participle
    életbe lépő (intézkedés)
    life-ILL stepping (instruction)
    '(an instruction) taking effect'
  - Past participle
    csődbe ment (cég)
    bankrupt-ILL gone (firm)
    '(a firm) that went bankrupt'
  - Future participle
    fontolóra veendő (ajánlat)
    consideration-SUB to be taken (offer)
    '(an offer) that is to be taken into consideration'
  - Infinitive
    forgalomba hozni
    circulation-ILL bring-INF
    'to put into circulation'

- **Nominalization <nom>**

  bérbe vétel

  rent-ILL taking

  'hiring'

Split light verb constructions, where the noun and the verb are not adjacent, are also annotated and tagged. In this way, their identification becomes possible and the database can be used for training an algorithm that automatically recognizes (split) light verb constructions.

5.2 Annotation principles

Corpus texts contain single annotation, i.e. one annotator worked on each text. Light verb constructions can be found in between XML tags <FX> </FX>. 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 developed for identifying Hungarian light verb constructions (Vincze, 2008).

The annotation process was carried out manually on the syntactically annotated version of the Szeged Treebank, thus, phrase boundaries were also taken into consideration when marking light verb constructions. Since the outmost boundary of the nominal component was considered to be part of the light verb construction, in several cases adjectives and other modifiers of the nominal head are also included in the construction, e.g.:

*<FX>**nyilvános** ajánlatot tesz</FX>*

public offer-ACC make

'to make a public offer'

In the case of participles, NP arguments may be also included (although in English, the same argument is expressed by a PP):

*<FX>**Nyíregyházán** tartott ülésén</FX>*

Nyíregyháza-SUP hold-PPT session-3SGPOSS-SUP

'at its session held in Nyíregyháza'

Constructions with a nominal component in the accusative case can be nominalized in two ways in Hungarian, as in:

szerződést köt

contract-ACC bind

'to make a contract'

<FX>szerződéskötés</FX>

contract+bind-GERUND

'making a contract'

<FX>adásvételi szerződések megkötése</FX>

sale contract-PL PREVERB-bind-GERUND-3SGPOSS

'making of sales contracts'

Both types are annotated in the corpus.

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 above reasons, a single light verb construction manifests in several different forms in the corpus. However, each occurrence is manually paired with its prototypical (i.e. bare noun + verb) form in a separate list, which is available at the corpus website.

5.3 Statistics on corpus data

The database contains 3826 occurrences of 658 light verb constructions altogether in 29062 sentences. Thus, a specific light verb construction occurs 5.8 times in the corpus on average. However, the participle form *irányadó* occurs in 607 instances (e.g. in *irányadó kamat* 'prime rate') due to the topic of the business news subcorpus, which may distort the percentage rates. For this reason, statistical data in Table 2 are shown the occurrences of *irányadó* excluded.

	verb	part	nom	split	total
business news	565 58.6%	270 28%	90 9.3%	40 4.1%	965 25.2%
newspapers	458 59.3%	192 24.9%	55 7.1%	67 8.7%	772 20.2%
legal texts	640 30.7%	504 24.1%	709 33.9%	236 11.3%	2089 54.6%
total	1663 43.5%	966 25.2%	854 22.3%	236 9%	3826 100%

Table 2: Subtypes of light verb constructions in the corpus

It is revealed that although it is verbal occurrences that are most frequent, the percentage rate of participles is also relatively high. The number of nominalized or split constructions is considerably lower (except for the law subcorpus, where their number is quite high), however, those together with participles are responsible for about 55% of the data, which indicates the importance of their being annotated as well.

As for the general frequency of light verb constructions in texts, we compared the number of verb + argument relations found in the Szeged Dependency Treebank (Vincze et al., 2010) where the argument was a common noun to that of light verb constructions. It has turned out that about 13% of verb + argument relations consist of light verb constructions. This again emphasizes that they should be paid attention to, especially in the legal domain (where this rate is as high as 36.8%). Statistical data are shown in Table 3.

	V + argument	LVC
business news	9524	624 (6.6%)
newspapers	3637	539 (14.8%)
legal texts	2143	889 (36.8%)
total	15574	2052 (13.2%)

Table 3: Verb + argument relations and light verb constructions

The corpus is publicly available for re-

[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.

search and/or educational purposes at www.inf.u-szeged.hu/rgai/nlp.

6 The usability of the corpus

As emphasized earlier, the proper treatment of light verb constructions is of primary importance in NLP applications. In order to achieve this, their identification is essential. The corpus created can function as the training database for the implementation of an algorithm capable of recognizing light verb constructions, which we plan to develop in the near future. In the following, the ways machine translation and information extraction can profit from such a corpus and algorithm are shortly presented.

6.1 Light verb constructions and machine translation

When translating collocations, translation programs face two main problems. On the one hand, parts of the collocation do not always occur next to each other in the sentence (split collocations). In this case, the computer must first recognize that the parts of the collocation form one unit (Oravecz et al., 2004), for which the multiword context of the given word must be considered. On the other hand, the lack (or lower degree) of compositionality blocks the possibility of word-by-word translation (Siepmann, 2005; Siepmann, 2006). However, a (more or less) compositional account of light verb constructions is required for successful translation (Dura and Gawrońska, 2005).

To overcome these problems, a reliable method is needed to assure that the nominal and verbal parts of the construction be matched. This requires an algorithm that can identify light verb constructions. In our corpus, split light verb constructions are also annotated, thus, it is possible to train the algorithm to recognize them as well: the problem of split collocations can be eliminated in this way.

A comprehensive list of light verb constructions can enhance the quality of machine translation – if such lists are available for both the source and the target language. Annotated corpora (especially and most desirably, parallel corpora) and explanatory-combinatorial dictionaries[4] are possible sources of such lists. Since in foreign language equivalents of light verb constructions, the nominal components are usually literal translations of each other (Vincze, 2009), by collating the corresponding noun entries in these lists the foreign language variant of the given light verb construction can easily be found. On the other hand, in order to improve the building of such lists, we plan to annotate light verb constructions in a subcorpus of SzegedParalell, a Hungarian-English manually aligned parallel corpus (Tóth et al., 2008).

6.2 Light verb constructions and information extraction

Information extraction (IE) seeks to process large amounts of unstructured text, in other words, to collect relevant items of information and to classify them. Even though humans usually overperform computers in complex information processing tasks, computers also have some obvious advantages due to their capacity of processing and their precision in performing well-defined tasks.

For several IE applications (e.g. relationship extraction) it is essential to identify phrases in a clause and to determine their grammatical role (subject, object, verb) as well. This can be carried out by a syntactic parser and is a relatively simple task. However, the identification of the syntactic status of the nominal component is more complex in the case of light verb constructions for it is a quasi-argument of the verb not to be confused with other arguments (Alonso Ramos, 1998). Thus, the parser should recognize the special status of the quasi-argument and treat it in a specific way as in the following sentences, one of which contains a light verb construction while the other one a verbal counterpart of the construction:

*Pete **made a decision** on his future.*

*Pete **decided** on his future.*

[4]Explanatory combinatorial dictionaries are essential for relation descriptions (up to the present, only fractions of the dictionary have been completed for Russian (Mel'čuk and Žolkovskij, 1984) and for French (see Mel'čuk et al. (1984 1999)), besides, trial entries have been written in Polish, English and German that contain the relations of a certain lexical unit to other lexemes given by means of lexical functions (see e.g. Mel'čuk et al. (1995)). These dictionaries indicate light verb constructions within the entry of the nominal component.

In the sentence with the verbal counterpart, the event of deciding involves two arguments: *he* and *his future*. In the sentence with the light verb construction, the same arguments can be found, however, it is unresolved whether they are the arguments of the verb (*made*) or the nominal component (*decision*). If a precise syntactic analysis is needed, it is crucial to know which argument belongs to which governor. Nevertheless, it is still debated if syntactic arguments should be divided between the nominal component and the verb (see Meyers et al. (2004a) on argument sharing) and if yes, how (Alonso Ramos, 2007).

For the purpose of information extraction, such a detailed analysis is unnecessary and in general terms, the nominal component can be seen as part of the verb, that is, they form a complex verb similarly to phrasal or prepositional verbs and this complex verb is considered to be the governor of arguments. Thus, the following data can be yielded by the IE algorithm: there is an event of **decision-making**, **Pete** is its subject and it is about **his future** (and not an event of **making** with the arguments **decision**, **Pete** and **his future**). Again, the precise identification of light verb constructions can highly improve the performance of parsers in recognizing relations between the complex verb and its arguments.

7 Conclusion

In this paper, we have presented the development of a corpus of Hungarian light verb constructions. Basic annotation guidelines and statistical data have also been included. The annotated corpus can serve as a training database for implementing an algorithm that aims at identifying light verb constructions. Several NLP applications in the fields of e.g. machine translation and information extraction may profit from the successful integration of such an algorithm into the system, which we plan to develop in the near future.

Acknowledgements

This work was supported in part by the National Office for Research and Technology of the Hungarian government within the framework of the project MASZEKER.

The authors wish to thank György Szarvas for his help in developing the annotation tool and Richárd Farkas for his valuable comments on an earlier draft of this paper.

References

Alonso Ramos, Margarita. 1998. *Etude sémantico-syntaxique des constructions à verbe support*. Ph.D. thesis, Université de Montréal, Montreal, Canada.

Alonso Ramos, Margarita. 2007. Towards the Synthesis of Support Verb Constructions. In Wanner, Leo, editor, *Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk*, pages 97–138, Amsterdam / Philadelphia. Benjamins.

B. Kovács, Mária. 1999. A funkcióigés szerkezetek a jogi szaknyelvben [Light verb constructions in the legal terminology]. *Magyar Nyelvőr*, 123(4):388–394.

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

Csendes, Dóra, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged TreeBank. In Matousek, Václav, Pavel Mautner, and Tomás Pavelka, editors, *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005*, Lecture Notes in Computer Science, pages 123–132, Berlin / Heidelberg, September. Springer.

Diab, Mona and Pravin Bhutada. 2009. Verb Noun Construction MWE Token Classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 17–22, Singapore, August. Association for Computational Linguistics.

Dura, Elżbieta and Barbara Gawrońska. 2005. Towards Automatic Translation of Support Verbs Constructions: the Case of Polish robic/zrobic and Swedish göra. In *Proceedings of the 2nd Language & Technology Conference*, pages 450–454, Poznań, Poland, April. Wydawnictwo Poznańskie Sp. z o.o.

Fazly, Afsaneh and Suzanne Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Grégoire, Nicole. 2007. Design and Implementation of a Lexicon of Dutch Multiword Expressions. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Kaalep, Heiki-Jaan and Kadri Muischnek. 2006. Multi-Word Verbs in a Flective Language: The Case of Estonian. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*, pages 57–64, Trento, Italy, April. Association for Computational Linguistics.

Kaalep, Heiki-Jaan 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, June.

Krenn, Brigitte. 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, June.

Langer, Stefan. 2004. A Linguistic Test Battery for Support Verb Constructions. *Lingvisticae Investigationes*, 27(2):171–184.

Mel'čuk, Igor and Aleksander Žolkovskij. 1984. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna, Austria.

Mel'čuk, Igor, André Clas, and Alain Polguère. 1995. *Introduction à lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve, France.

Mel'čuk, Igor, et al. 1984–1999. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I–IV.* Presses de l'Université de Montréal, Montreal, Canada.

Meyers, Adam, Ruth Reeves, and Catherine Macleod. 2004a. NP-External Arguments: A Study of Argument Sharing in English. In Tanaka, Takaaki, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 96–103, Barcelona, Spain, July. Association for Computational Linguistics.

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

Oravecz, Csaba, Károly Varasdi, and Viktor Nagy. 2004. Többszavas kifejezések számítógépes kezelése [The treatment of multiword expressions in computational linguistics]. In Alexin, Zoltán and Dóra Csendes, editors, *MSzNy 2004 – II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 141–154, Szeged, Hungary, December. University of Szeged.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. 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.

Siepmann, Dirk. 2005. Collocation, colligation and encoding dictionaries. Part I: Lexicological Aspects. *International Journal of Lexicography*, 18(4):409–444.

Siepmann, Dirk. 2006. Collocation, colligation and encoding dictionaries. Part II: Lexicographical Aspects. *International Journal of Lexicography*, 19(1):1–39.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. 2004. Statistical Measures of the Semi-Productivity of Light Verb Constructions. In Tanaka, Takaaki, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 1–8, Barcelona, Spain, July. Association for Computational Linguistics.

Tan, Yee Fan, 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, April. Association for Computational Linguistics.

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

Trón, Viktor, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. 2005. hunmorph: Open Source Word Analysis. In *Proceedings of the ACL Workshop on Software*, pages 77–85, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Váradi, Tamás. 2006. Multiword Units in an MT Lexicon. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*, pages 73–78, Trento, Italy, April. Association for Computational Linguistics.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Vincze, Veronika. 2008. A puszta köznév + ige komplexumok státusáról [On the status of bare common noun + verb constructions]. In Sinkovics, Balázs, editor, *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, pages 265–283, Szeged, Hungary. University of Szeged.

Vincze, Veronika. 2009. Főnév + ige szerkezetek a szótárban [Noun + verb constructions in the dictionary]. In Váradi, Tamás, editor, *III. Alkalmazott Nyelvészeti Doktorandusz Konferencia*, pages 180–188, Budapest. MTA Nyelvtudományi Intézet.

Zarrieß, Sina 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, August. Association for Computational Linguistics.