

# **Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses**

SUMMARY OF PHD THESIS

Veronika Vincze

University of Szeged

August 2011

Supervisor: Károly Bibok, PhD



University of Szeged  
Doctoral School in Linguistics  
Ph.D. Programme in Theoretical Linguistics



# 1 Introduction

In this thesis, semi-compositional bare noun + verb constructions are examined from a theoretical point of view and computational linguistic issues are also discussed. The research mostly focuses on Hungarian, however, data from other languages such as English, Russian or Spanish are also taken into consideration wherever appropriate.

Noun + verb constructions do not form a unified category, since, on the one hand, there are productive structures such as *újságot olvas* (newspaper-ACC reads) ‘to read a newspaper’ or *levelet ír* (letter-ACC writes) ‘to write a letter’. On the other hand, idiomatic expressions such as *csütörtököt mond* (Thursday-ACC says) ‘to fail to work’ and *lépre csal* (comb-SUB entices) ‘to toll’ can also be found. However, besides these constructions, there exist some expressions that are neither productive nor idiomatic but whose meaning is not totally compositional. For this latter type, examples from English and Hungarian are shown in (1). Since their meaning is the same, only glosses are provided:<sup>1</sup>

(1) (a) English:

*to give a lecture*  
*to come into bloom*  
*a possibility emerges*

(b) Hungarian:

*előadást tart*  
 presentation-ACC holds

*virágba borul*  
 bloom-ILL falls

*lehetőség nyílik*  
 possibility opens

Several terms have been used for these constructions in the literature (see e.g. Dobos (1991; 2001), Langer (2004)). The most common ones are as follows – all containing the verbal component within the term: in German, they are called *Funktionsverbgefüge* (function verb constructions)<sup>2</sup>, in English, *complex verb structures*, *support verb constructions* or *light verb*

<sup>1</sup>Nouns in the nominative case and verbs in the subjective conjugation are not marked distinctively in glosses.

<sup>2</sup>In German literature, constructions where the nominal component is the subject are not traditionally considered to be *Funktionsverbgefüge*.

*constructions* can be found,<sup>3</sup> in French, *constructions à verbe support* (support verb constructions) is usually used, whereas *costruzioni a verbo supporto* (support verb constructions) in Italian, *construções com verbo suporte* (support verb constructions) or *construções com verbo leve* (light verb constructions) in Portuguese. On the other hand, we can find terms like *opisatel'nye vyraženija* (descriptive expressions) in Russian, which do not include the verbal component (see also Hungarian).

In Hungarian, these constructions also have several names: *körülíró szerkezetek* (periphrastic constructions) in Sziklai (1986), *leíró kifejezések* (descriptive expressions) in Dobos (1991), and *funkcióigés szerkezetek* (function verb constructions) following Keszler (1992), however, the somewhat pejorative term *terpeszkedő szerkezetek* (“sprawling” constructions) occurs in the Hungarian Purists’ Dictionary (Grétsy and Kemény, 1996, p. 571) and in recent specialized articles as well (for instance, Heltai and Gósy (2005) focus on the effects of sprawling constructions on linguistic processing).

As can be seen from the examples given above, most names (except for Russian and several Hungarian terms) used for these constructions contain only one component of the construction, namely, the verbal component, suggesting that it is the verbal component that forms the head of the construction. However, since the verbal component functions only as the syntactic head of the construction – the nominal component being the semantic head of the construction (see below) –, it is perhaps better not to include any of the two components in the name of the construction. On the other hand, pejorative terms and those referring to the apparently periphrastic nature<sup>4</sup> of the construction should also be avoided. That is why the term *semi-compositional constructions* will be henceforth used for this type of common noun + verb constructions, following Langer (2004).

## 2 Motivation

Semi-compositional constructions are of dual nature. On the one hand, they are made of two syntactic parts (a nominal and a verbal component), on the other hand, they form one semantic unit. This duality determines their linguistic features on all layers of grammar:

---

<sup>3</sup>There might be slight theoretical differences in the usage of these terms – for instance, semantically empty support verbs are called *light verbs* in e.g. Meyers et al. (2004), that is, the term *support verb* is a hypernym of *light verb*.

<sup>4</sup>We argue that semi-compositional constructions add important aspects to the action, i.e. they do not only circumscribe the situation expressed by their verbal counterpart.

- 
- the syntactic and the semantic head of the construction do not coincide, the verb being the syntactic head and the noun being the semantic head (double headedness);
  - the meaning of the construction as a whole cannot be totally predicted on the basis of the meaning of their parts (semi-compositionality);
  - semantic components of the verb and the noun or classes of nouns must partially overlap in order to form one unit (semi-productivity);
  - semi-compositional constructions can be placed among productive bare noun + verb constructions and idioms with respect to their productivity and compositionality (scalability);
  - semi-compositional constructions may have a verbal counterpart, which is typically derived from the same root as the nominal component and has a similar meaning (variativity).

The duality of semi-compositional construction is related to other theoretical questions. First, can a one-to-one correspondence be assumed between the syntactic and semantic representation of a phrase, in other words, is there **homomorphism** between the two levels? Second, how is the notion **compositionality** understood: is it absolute (i.e. a phrase is compositional or not) or is it relative (i.e. a phrase can be more compositional than another phrase)?

Homomorphism between the level of syntax and semantics equals to the notion of **strong compositionality**, i.e. the meaning of the parts of a phrase and their syntactic relation **fully** determine the meaning of the whole phrase, cf. Hale (1997). However, not all linguistic units are compositional in the above sense: for instance, the meaning of idioms cannot be computed from the meaning of their parts and their syntactic relation, thus, there is no homomorphism between syntax and semantics in the case of idioms. In the thesis, it will be argued that besides phrases fully corresponding to the compositionality principle and phrases showing a total lack of compositionality there are other phrases that correspond to the compositionality principle to some degree. Thus, **semi-compositionality** is understood in the following way: the meaning of the phrase can be determined on the basis of the meaning of its parts and their syntactic relation **to some degree** (i.e. it is not totally compositional and not totally idiomatic).

The motivation of the research described in this thesis is twofold. First, theoretical aims include:

- examining to what extent semi-compositional constructions are similar to productive constructions or idioms;

- detailed syntactic and semantic analysis of semi-compositional constructions as a unit within several theoretical frameworks (such as generative and dependency grammars), involving data from different languages;
- analysis of syntactic and semantic relations between the two components (i.e. the noun and the verb) of the construction;
- comparing the syntactic and semantic features of semi-compositional constructions and their verbal counterparts.

Second, the treatment of semi-compositional constructions in computational linguistics is analyzed in detail. Special emphasis is put on the adaptation of theoretical results reached to computational linguistics (NLP) applications.

The main purpose of this thesis is that theoretical results be applied to the greatest extent possible in several fields of natural language processing, thus making a direct connection from theory to practice. This aspect constitutes the major link between the two main parts of the thesis.

### **3 Research questions**

In this section, a more detailed presentation of research questions follows that are to be answered in this thesis.

1. (a) What is the relation between semi-compositional constructions, productive constructions and idioms?  
(b) How can the relation of the constructions and their verbal counterparts be described?  
(c) What are the syntactic features of semi-compositional constructions?  
(d) How can the syntactic relation of the two components of the construction be described?  
(e) What are the semantic features of semi-compositional constructions?  
(f) How can the semantic relation of the two components of the construction be described?  
(g) What lexical representation can be assumed for semi-compositional constructions?

2. To what extent are the results of this research language-dependent or language-independent?
3. How can the theoretical results reached be applied in fields of computational linguistics, namely
  - (a) in the automatic identification of semi-compositional constructions;
  - (b) in word sense disambiguation;
  - (c) in information extraction and retrieval;
  - (d) in machine translation?

## 4 Methodology

In order to answer the above questions, language data from Hungarian and English corpora were analyzed. In this way, theoretical claims can be supported or rejected by empirical data. To gain an ample amount of language data, semi-compositional constructions were annotated in texts from various domains. Annotation guidelines were based on earlier theoretical results, however, the data collected from the corpora served as a basis for drawing theoretical conclusions on the behavior of semi-compositional constructions. Thus, a nice interplay of theory and practice can be observed in the data collecting methodology of this thesis.

For methodological purposes, we constructed three corpora of semi-compositional constructions. The Szeged Treebank (Vincze and Csirik, 2010), the SzegedParalell corpus (Vincze et al., 2010) and the Wiki50 corpus (Vincze et al., 2011) were annotated for semi-compositional constructions, thus yielding a database that consists of 1524 Hungarian and 827 English constructions. The author supervised the annotation work and also participated in the annotation process. The corpora created can be used for collecting linguistic data from realistic language use and for training and testing NLP applications aiming at identifying semi-compositional constructions in natural language texts.

## 5 Theoretical results

In the first part of the thesis, semi-compositional constructions were analyzed from several theoretical aspects, the results of which are now briefly presented.

## 5.1 Semi-compositional constructions as a subtype of multiword expressions

Semi-compositional constructions are a subtype of multiword expressions (see e.g. Sag et al. (2002)). They were contrasted to other types of multiword expressions based on several dimensions. It was indicated that they exhibit **lexical and semantic idiosyncrasy** and they are **syntactically flexible** and **semantically decomposable**. These features influence further theoretical and empirical investigations on the subject.

## 5.2 The status of semi-compositional constructions

The status of semi-compositional constructions was discussed as compared to productive constructions and idioms that share their syntactic structure (i.e. they also consist of a bare common noun and a verb). For that purpose, we constructed a test battery that distinguishes subtypes of bare common noun + verb constructions. It was revealed that the **continuum of bare common noun + verb constructions** can be characterized by the parameters of **compositionality** and **productivity**. On the basis of test results, constructions can be divided into three main groups: productive constructions, semi-compositional constructions and idioms. Semi-compositional constructions can be opposed to productive constructions and idioms on the basis of **variativity** and **omitting the verb** while semi-compositional constructions and idioms – as opposed to productive constructions – are instances of semantically idiosyncratic multiword expressions. Furthermore, semi-compositional constructions can be divided into subgroups. However, there is **no sharp and distinct boundary** in between groups on the scale since belonging to a (sub)group is not determined by a dichotomy of the either-or type: the place of the construction on the scale is rather a question of degree and scalability. The test results also indicate that the group of idioms is more diverse than it was earlier believed (cf. Kiefer (1990 1991), Komlósy (1992)): certain constructions have more idiom-like features than others among those that were traditionally seen as idioms.

## 5.3 Verbal counterparts of semi-compositional constructions

Most semi-compositional constructions have a verbal counterpart that contains the same stem as the nominal component. In the thesis, Hungarian and English semi-compositional constructions were contrasted with their verbal counterparts on the one hand and with their other language



equivalents on the other hand. It was shown that the usage of the construction and its verbal counterpart is not always in a perfect overlap due to **differences in the argument structure** (e.g. *to lay a charge on sy* vs. *to charge sy with sg*) or **in aspect and Aktionsart** (e.g. the construction *forgalomba hoz* (circulation-ILL brings) ‘to put into circulation’ bears an inchoative Aktionsart whereas the verbal counterpart *forgalmaz* ‘to circulate’ does not). As opposed to productive constructions (Kiefer, 2007), semi-compositional constructions do **not inherently have progressive aspect**. It was also revealed that interlingual differences can be accounted for differences in the argument structure and in contrast with Hungarian, the English construction can bear progressive aspect. Finally, it was also argued that the **acceptability** of semi-compositional constructions is a **matter of degree and scale**, similarly to their status on the continuum of bare common noun + verb constructions. As for NLP applications, information retrieval can profit from these conclusions and the interlingual results of this chapter can be exploited in machine translation.

#### 5.4 The syntax of semi-compositional constructions

The syntactic analyses of semi-compositional constructions were provided in the frameworks of constituency and dependency grammars based on the original analyses found in Hale and Keyser (2002), Newson et al. (2006), Larson (1988) and Alonso Ramos (1998). Special attention was paid to the alternations in the argument structure and the derivational processes between the construction and its verbal counterpart. It was emphasized that in each theoretical framework, **arguments** of the constructions **belong to the noun on a deep level** whereas some of them **move to the verb on the surface level**. However, the exact rules of argument transfer (i.e. which argument should move to the verb) are construction-specific. Although semi-compositional constructions show phrasal properties, they behave as one unit on the level of semantics. Thus, we also offered an alternative proposal for regarding semi-compositional constructions as one **complex predicate**, which eliminates the question of argument transfer and has also important consequences in NLP applications, e.g. in information extraction.

#### 5.5 The semantics of semi-compositional constructions

The semantics of semi-compositional constructions was analyzed within the framework of Meaning–Text Theory (see e.g. Mel’čuk et al. (1995)). The syntactic-semantic relation between the nominal component and the verb can be formalized by using lexical functions. For the anal-

ysis, we selected semi-compositional constructions containing one of the four most frequent verbs (*ad* ‘give’, *vesz* ‘take’, *hoz* ‘bring’ and *tesz* ‘do’) in the Hungarian database. It was revealed that there are **semantic correlations between the noun, the verb and the lexical function**, e.g. nouns denoting speech acts are usually paired with the verbs *ad* ‘give’, *tesz* ‘do’ or *hoz* ‘bring’. Aktionsart and aspectual information can also be formalized with lexical functions. Furthermore, lexical semantic relations were also discussed: there are semi-compositional constructions that are **synonymous** with each other on the one hand (such as *bajba jut – bajba kerül* ‘to get into trouble’) and there are **conversive** pairs of semi-compositional constructions (i.e. constructions describing the same situation from the view of another participant) on the other hand (such as *bérbe ad – bérbe vesz* ‘to rent/to hire’). These results can be applied in word sense disambiguation and in machine translation.

## 5.6 The lexical representation of semi-compositional constructions

Four possible ways of the lexical representation of semi-compositional constructions were presented in the thesis, with respect to their advantages and disadvantages from both theoretical and empirical aspects. After discussing the problems concerning the identification of the head of semi-compositional constructions and the features of electronic and paper-based dictionaries, methods were illustrated with mini dictionary entries for the verbs *köt* ‘bind’ and *hoz* ‘bring’ and their nominal components. From the methods discussed, the one of **listing the construction in the entries of both the noun and the verb** proved to be the most efficient from both a theoretical and an empirical aspect although most of the existing dictionaries list the constructions in the entry of the verb (Pusztai, 2003; Bárczi and Országh, 1959 1962) or in the entry of the noun (collocational dictionaries or explanatory combinatorial dictionaries (Mel’čuk and Žolkovskij, 1984; Mel’čuk et al., 1984 1999). Finally, it was also shown that **in the Hungarian WordNet** semi-compositional constructions are treated as **separate lexical units**. The way semi-compositional constructions are represented in the dictionary has influences on information extraction, word sense disambiguation and machine translation.

## 6 Computational linguistic results

In the second part of the thesis, it was shown how different fields of natural language processing deal with semi-compositional constructions. Special emphasis was put on the practical adap-

---

tation of theoretical results. In the following, thesis results related to computational linguistics are briefly summarized.

## 6.1 The automatic identification of semi-compositional constructions

For the automatic identification of semi-compositional constructions, different methods – statistical, rule-based and hybrid models – were discussed. The questions of how and when to identify multiword expressions in general and semi-compositional constructions in particular were raised. With regard to semi-compositional constructions, we implemented **rule-based and machine learning based methods** that are able to identify semi-compositional constructions in English texts as opposed to earlier applications that classify verb + object instances as idiomatic or not (Bannard, 2007; Cook et al., 2007; Tan et al., 2006; Tu and Roth, 2011). Our results suggest that **shallow morphological information** (such as lemmas, stems, POS-tags and suffixes) is sufficient to detect semi-compositional constructions in texts while **syntactic features** improve the performance of the system. It was also shown that with a well-designed **domain-specific list of light verb candidates**, competitive results can be achieved on any domain, especially if enhanced with syntactic features. Hence their identification can take place in a post-processing step after syntactic parsing, thus, the output of the latter can be further exploited in higher-level applications such as information extraction and machine translation.

## 6.2 Semi-compositional constructions in word sense disambiguation

For the word sense disambiguation of semi-compositional constructions, well-defined senses are necessary for both the nominal component and the verb. On the basis of lexicographic considerations, **light verb senses of the verb** can be distinguished from other senses of the verb. As for the ordering of NLP applications, it was argued that WSD algorithms can perform better if semi-compositional constructions are known to them, that is, the detection of multiword expressions precedes word sense disambiguation. **Morphosyntactic information** can also enhance WSD, and it was shown that the **reduction of senses** and **more precise sense definitions** can lead to a higher accuracy in both manual and automatic word sense disambiguation.

### 6.3 Semi-compositional constructions in information extraction and information retrieval

The treatment of semi-compositional constructions was discussed in three fields of information extraction: semantic frame mapping, semantic role labeling and modality detection. By exploiting **lexical functions**, we proposed a more dense and more compact way of **representing semantic frames** containing semi-compositional constructions than the ones used in Farkas et al. (2004) and the strategy of using **wordnet synsets** in the lexical representation of such constructions was also presented. Concerning the effects of regarding semi-compositional constructions as **one complex predicate**, it was shown how **semantic role labeling** and **modality detection** can benefit from this theoretical achievement. Finally, it was argued that our **bilingual list** of semi-compositional constructions and their verbal counterparts can enhance the recall of (cross-language) **information retrieval** systems and can be used in extending wordnets.

### 6.4 Semi-compositional constructions and machine translation

The lack of total compositionality results in difficulties of translating multiword expressions in general and semi-compositional constructions in particular (Dura and Gawrońska, 2005). However, **lexical functions** can formalize the relation between the verbal and the nominal component of semi-compositional constructions and there are **correlations between the semantic type of the noun and the verb**, which can be exploited in machine translation applications and dictionary building. Since machine translation methods need a large dictionary containing entries on multiword expressions, our **bilingual database** of semi-compositional constructions and their verbal counterparts can prove useful. On the other hand, the method of representing semi-compositional constructions as one unit proposed by us can also contribute to machine translation. It was also emphasized that statistical machine translation can make use of the **SzegedParalellFX** corpus annotated for semi-compositional constructions.

## 7 Results of interlingual analyses

In the thesis, we paid special attention to compare our results achieved for Hungarian to those described in the literature for other languages. The **test battery** presented in the thesis can be applied to English and Hungarian as well, thus, a similar scale can be sketched for noun

+ verb constructions in both languages. The **basic syntactic features** of semi-compositional constructions are **language-independent** and **semantic correlations between the type of the noun and the verb** also exist in several languages as illustrated with Russian and Hungarian examples, where the verbs *davat' / ad* ‘give’ are usually paired with nouns denoting speech acts. However, there might be **differences concerning syntactic structure, argument structure, aspect or Aktionsart** between a Hungarian construction and its other language equivalent.

As for natural language processing, the **automatic identification** of semi-compositional constructions requires **highly language specific** methods since the very same construction might be present in heterogeneous surface forms in agglutinative languages such as Hungarian, which is less characteristic of morphologically poor languages (such as English). On the other hand, the treatment of semi-compositional constructions in **word sense disambiguation** and **information extraction / retrieval** is mainly **language independent**. Finally, **interlingual differences** can be overtly exploited in **machine translation**.

## 8 Conclusions and future work

In this thesis, semi-compositional constructions were analyzed from the aspects of theoretical linguistics and natural language processing. The main motivation behind the thesis was to apply the theoretical results achieved to the empirical field of natural language processing. The connection between theoretical results and NLP fields is now summarized:

- being situated in between productive constructions and idioms, the automatic identification of semi-compositional constructions cannot be solely based on syntactic patterns;
- the recognition of nouns derived from verbal stems (i.e. the verbal counterpart of semi-compositional constructions) can enhance the identification of semi-compositional constructions;
- interlingual differences gained from contrasting English and Hungarian semi-compositional constructions and their verbal counterparts can be applied in machine translation;
- the bilingual list of semi-compositional constructions and their verbal counterparts can enhance the performance of machine translation and information retrieval systems;
- treating semi-compositional constructions as one complex predicate on the level of syntax can be fruitful in semantic role labeling and modality detection;

	identification	WSD	IE	IR	MT
status	•				
verbal counterparts	•			•	•
syntax			•		
semantics		•			•
lexical representation		•	•		•

Table 1: Connections between topics of the thesis

- lexical functions are able to formalize the semantic relations between the two components of semi-compositional constructions, which can be exploited in dictionary building for machine translation systems;
- correlations between the semantic type of the noun and the verb can be applied in word sense disambiguation;
- the method proposed for the lexical representation of semi-compositional constructions can be integrated into applications of semantic frame mapping, word sense disambiguation and machine translation.

Table 1 also visualizes the interrelationships among the theoretical linguistic and computational linguistic chapters of the thesis.

The theoretical investigations on semi-compositional constructions also highlighted that linguistic units do not necessarily coincide on different layers of grammar (namely, on the level of syntax and semantics): for instance, the phrase *to take a decision* can be separated into syntactic parts (e.g. the nominal component can function as an answer for a question) although it bears the same meaning as *to decide*, which cannot be divided into syntactic parts. The distinction of the notions **syntactic** and **semantic** actants in the Meaning–Text Theory (Mel’čuk, 2004a; Mel’čuk, 2004b) also indicates that there is no necessary homomorphism between syntax and semantics. The notion of **semi-compositionality** and the **lack of syntax-semantics homomorphism** entail that semi-compositional constructions have a dual nature: they can be seen as a unit and as consisting of two parts. However, in order to apply theoretical results to the empirical field of computational linguistics to the greatest extent possible, both solutions were presented for the NLP applications investigated in the thesis whenever it was possible.

This duality has different impacts on different fields of theoretical and computational linguistics. They consist of a verb and a nominal component, which has influences on the following

issues:

- their semi-productivity and semi-compositionality is determined on the basis of the relation of the two parts;
- the modifiability of the noun suggests that the construction has an internal syntactic structure;
- semantic correlations can be found between the verb and the noun;
- they can be conversives of other semi-compositional constructions (the meaning of their verbal component is considered);
- they can be represented in the lexicon as appearing in the entry of the verb, noun or both;
- their automatic identification is based on finding noun + verb combinations;
- in word sense disambiguation, the verb can be attributed a special light verb sense;
- in transfer-based machine translation, transfer rules regulate the mapping of constructions which have different syntactic structures in different languages.

In the following points, their being one unit is emphasized:

- they can have verbal counterparts with a similar meaning;
- the argument structure of the constructions and their verbal counterparts might differ;
- there might be differences concerning aspect or Aktionsart between constructions and their verbal counterparts;
- they are complex predicates;
- they can be synonymous with other semi-compositional constructions (their meaning as a unit is the same);
- they can be represented in the lexicon as one unit;
- they can be treated as one lexical unit in word sense disambiguation;
- in semantic frame mapping, semantic role labeling and modality detection, they are seen as one syntactic unit (complex predicate);

- in information retrieval, their variability with verbal counterparts can be exploited;
- they are regarded as one unit in statistical machine translation.

Table 2 offers a summary of these points.

	<b>parts</b>	<b>unit</b>
<b>Status</b>		
semi-productivity	•	
semi-compositionality	•	
<b>Verbal counterpart</b>		
differences in argument structure		•
differences in aspect/Aktionsart		•
<b>Syntax</b>		
modifiability of noun	•	
complex predicate		•
<b>Semantics</b>		
correlations between noun and verb	•	
synonymy		•
conversion	•	
<b>Lexical representation</b>		
entry of verb	•	
entry of noun	•	
separate entry		•
entry of verb and noun	•	
<b>Automatic identification</b>	•	
<b>Word sense disambiguation</b>		
one meaning as lexical unit		•
light verb sense	•	
<b>Information extraction</b>		
semantic frame mapping		•
semantic role labeling		•
modality detection		•
<b>Information retrieval</b>		•
<b>Machine translation</b>		
transfer-based methods	•	
statistical methods		•

Table 2: The dual nature of semi-compositional constructions and fields of the thesis

As future work, both theoretical questions and NLP challenges are to be answered. For instance, the database of semi-compositional constructions can be turned into a database formalized with the help of lexical functions. The analysis of lexical semantic relations between the nominal component and the verb should be extended to constructions containing other verbs as well. Distinctions between the senses of light verbs in different semi-compositional constructions should be made in order to enhance dictionary building and lexicological work: lexicons



and dictionaries should be constructed where semi-compositional constructions are included in the entries of both the verb and the noun. With regard to natural language processing, algorithms aiming at identifying semi-compositional constructions should be improved and developed and for this reason, more language resources should be created that can serve as training and testing databases. Solutions that can improve the performance of NLP applications with regard to the treatment of semi-compositional constructions should be found not only in the fields discussed in this thesis (i.e. word sense disambiguation, information extraction and retrieval and machine translation) but also in fields yet unexplored. Hopefully, all these research tasks will be performed in the near future at the Research Group on Artificial Intelligence at the University of Szeged.

## References

- Alonso Ramos, Margarita. 1998. *Etude sémantico-syntaxique des constructions à verbe support*. Ph.D. thesis, Université de Montréal, Montreal, Canada.
- Bannard, Colin. 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, pp. 1–8, Morristown, NJ, USA. Association for Computational Linguistics.
- Bárczi, Géza; Országh, László (eds.). 1959–1962. *A magyar nyelv értelmező szótára [The Explanatory Dictionary of the Hungarian Language]*. Akadémiai Kiadó, Budapest.
- Cook, Paul; Fazly, Afsaneh; Stevenson, Suzanne. 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, pp. 41–48, Morristown, NJ, USA. Association for Computational Linguistics.
- Dobos, Csilla. 1991. *Leíró kifejezések az orosz jogi szaknyelvben [Descriptive expressions in the Russian legal language]*. Ph.D. thesis, University of Debrecen, Debrecen, Hungary.
- Dobos, Csilla. 2001. *A funkcióigés szerkezetek vizsgálata (különös tekintettel az orosz jogi szaknyelvre) [An analysis of function verb constructions (with special emphasis on Russian legal language)]*. Ph.D. thesis, University of Debrecen, Debrecen, Hungary.
- Dura, Elżbieta; Gawrońska, Barbara. 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*, pp. 450–454, Poznań, Poland, April. Wydawnictwo Poznańskie Sp. z o.o.
- Farkas, Richárd; Konczer, Kinga; Szarvas, György. 2004. Szemantikuskeret illesztés és az IE-rendszer automatikus kiértékelése [Semantic frame mapping and the automatic evaluation of the IE system]. In Alexin, Zoltán; Csendes, Dóra (eds.), *II. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 49–53, Szeged, Hungary, December. University of Szeged.

- Grétsy, László; Kemény, Gábor. 1996. *Nyelvművelő kéziszótár [Purists' dictionary]*. Auktor Könyvkiadó, Budapest.
- Hale, Kenneth; Keyser, Samuel J. 2002. *Prolegomenon to a Theory of Argument Structure*. MIT Press, Cambridge.
- Hale, Bob. 1997. Grundlagen section 64. *Proceedings of the Aristotelian Society*, XCVII:243–261.
- Heltai, Pál; Gósy, Mária. 2005. A terpeszkedő szerkezetek hatása a feldolgozásra [The effect of sprawling constructions on processing]. *Magyar Nyelvőr*, 129:470–487.
- Keszler, Borbála. 1992. A mai magyar nyelv szófaji rendszere [The system of parts of speech in contemporary Hungarian]. In Kozocsa, Sándor Géza; Laczkó, Krisztina (eds.), *Emlékkönyv Rácz Endre hetvenedik születésnapjára*, pp. 131–139, Budapest.
- Kiefer, Ferenc. 1990–1991. Noun Incorporation in Hungarian. *Acta Linguistica Hungarica*, 40(1–2):149–177.
- Kiefer, Ferenc. 2007. *Jelentésemélet [Semantics]*. Corvina, Budapest.
- Komlósy, András. 1992. Régensek és vonzatok [Governors and arguments]. In Kiefer, Ferenc (ed.), *Strukturális magyar nyelvtan. I. Mondattan*, pp. 299–527, Budapest. Akadémiai Kiadó.
- Langer, Stefan. 2004. A Linguistic Test Battery for Support Verb Constructions. *Linguisticae Investigationes*, 27(2):171–184.
- Larson, Richard. 1988. On the double object construction. *Linguistic Inquiry*, 19:335–391.
- Mel'čuk, Igor; Žolkovskij, Aleksander. 1984. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach, Vienna, Austria.
- Mel'čuk, Igor; Clas, André; Polguère, Alain. 1995. *Introduction à lexicologie explicative et combinatoire*. Duculot, Louvain-la-Neuve, France.
- Mel'čuk et al., Igor. 1984–1999. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I–IV*. Presses de l'Université de Montréal, Montreal, Canada.
- Mel'čuk, Igor. 2004a. Actants in semantics and syntax I: Actants in semantics. *Linguistics*, 42(1):1–66.
- Mel'čuk, Igor. 2004b. Actants in semantics and syntax II: Actants in syntax. *Linguistics*, 42(2):247–291.
- Meyers, Adam; Reeves, Ruth; Macleod, Catherine. 2004. NP-External Arguments: A Study of Argument Sharing in English. In Tanaka, Takaaki; Villavicencio, Aline; Bond, Francis; Korhonen, Anna (eds.), *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 96–103, Barcelona, Spain, July. Association for Computational Linguistics.
- Newson, Mark; Hordós, Marianna; Pap, Dániel; Szécsényi, Krisztina; Tóth, Gabriella; Vincze, Veronika. 2006. *Basic English Syntax With Exercises*. Bölcsész Konzorcium, Budapest.

- 
- Pusztai, Ferenc (ed.). 2003. *Magyar értelmező kéziszótár [The Concise Dictionary of the Hungarian Language]*. Akadémiai Kiadó, Budapest.
- Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan. 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)*, pp. 1–15, Mexico City, Mexico.
- Sziklai, Lászlóné. 1986. Terpeszkednek vagy körülírnak? [Do they sprawl or do they circumscribe?]. *Magyar Nyelvőr*, 110:268–273.
- Tan, Yee Fan; Kan, Min-Yen; Cui, Hang. 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*, pp. 49–56, Trento, Italy, April. Association for Computational Linguistics.
- Tu, Yuancheng; Roth, Dan. 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*, pp. 31–39, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Vincze, Veronika; Csirik, János. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1110–1118, Beijing, China, August. Coling 2010 Organizing Committee.
- Vincze, Veronika; Felvégi, Zsuzsanna; R. Tóth, Krisztina. 2010. Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban [Semi-compositional constructions in the SzegedParalell English–Hungarian parallel corpus]. In Tanács, Attila; Vincze, Veronika (eds.), *MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 91–101, Szeged, Hungary, December. University of Szeged.
- Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

## Publications related to the thesis

1. Vincze, Veronika 2007. A félig kompozicionális szerkezetek gépi fordításainak lehetőségéről [On possible ways of automatically translating semi-compositional constructions]. In: Váradi, Tamás (ed.): *I. Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Budapest, MTA Nyelvtudományi Intézet, pp. 207–218.
2. Vincze, Veronika 2008. A puszta köznév + ige komplexumok státusáról [On the status of bare common noun + verb constructions]. In: Sinkovics, Balázs (ed.): *LingDok 7. Nyelvész-doktoranduszok dolgozatai*. Szeged, Szegedi Tudományegyetem, pp. 279–297.
3. Vincze, Veronika; Szarvas, György; Almási, Attila; Szauter, Dóra; Ormándi, Róbert; Farkas, Richárd; Hatvani, Csaba; Csirik, János 2008. Hungarian word-sense disambiguated corpus. In: *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco.
4. Vincze, Veronika 2009. On the Machine Translatability of Semi-Compositional Constructions. In: Váradi, Tamás (ed.): *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból – Selected Papers from the 1st Applied Linguistics PhD Conference*. Budapest, MTA Nyelvtudományi Intézet, pp. 166–178.
5. Vincze, Veronika 2009. Angol–magyar főnév + ige szerkezetek és igei párjaik [English–Hungarian noun + verb constructions and their verbal counterparts]. In: Váradi, Tamás (ed.): *II. Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Budapest, MTA Nyelvtudományi Intézet, pp. 112–122.
6. Vincze, Veronika 2009. Félig kompozicionális szerkezetek a Szeged Korpuszban [Semi-compositional constructions in the Szeged Corpus]. In: Tanács, Attila; Szauter, Dóra; Vincze, Veronika (eds.): *VI. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, pp. 390–393.
7. Vincze, Veronika 2009. *Előadást tart* vs. *előad*: főnév + ige szerkezetek igei variánsai [To give a lecture vs. to lecture: verbal counterparts of noun + verb constructions]. In: Sinkovics, Balázs (ed.): *LingDok 8. Nyelvész-doktoranduszok dolgozatai*. Szeged, Szegedi Tudományegyetem, pp. 265–278.

8. Vincze, Veronika 2009. Főnév + ige szerkezetek a szótárban [Noun + verb constructions in the dictionary]. In: Váradi, Tamás (ed.): *III. Alkalmazott Nyelvészeti Doktorandusz Konferencia*. Budapest, MTA Nyelvtudományi Intézet, pp. 180–188.
9. Vincze, Veronika 2010. Félig kompozicionális főnév + ige szerkezetek a számítógépes nyelvészetben [Semi-compositional noun + verb constructions in natural language processing]. In: Gecső, Tamás; Sárdi, Csilla (eds.): *Új módszerek az alkalmazott nyelvészeti kutatásban*. Budapest, Tinta Könyvkiadó, pp. 327–332.
10. Vincze, Veronika; Csirik, János 2010. Hungarian corpus of light verb constructions. In: *Proceedings of COLING 2010*, Beijing, China, pp. 1110–1118.
11. Vincze, Veronika, Felvégi, Zsuzsanna, R. Tóth, Krisztina 2010. Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban [Semi-compositional constructions in the SzegedParalell English–Hungarian parallel corpus]. In: Tanács, Attila; Vincze, Veronika (eds.): *MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged, Szegedi Tudományegyetem, pp. 91–101.
12. Vincze, Veronika 2011. Mi fán terem a főnév + ige szerkezet? [From what tree can you harvest noun + verb constructions?]. In: Gécseg, Zsuzsanna (ed.): *LingDok 10. Nyelvészeti doktoranduszok dolgozatai*. Szeged, Szegedi Tudományegyetem, pp. 225–243.
13. Vincze, Veronika; Nagy T., István; Berend, Gábor 2011. Detecting noun compounds and light verb constructions: a contrastive study. In: *Proceedings of the ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Portland, Oregon, USA, Association for Computational Linguistics, pp. 116–121.
14. Nagy T., István; Vincze, Veronika; Berend, Gábor 2011. Domain-dependent identification of multiword expressions. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.
15. Vincze, Veronika; Nagy T., István; Berend, Gábor 2011. Multiword expressions and named entities in the Wiki50 corpus. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria.

	ALKNYELVDOK1 2007	LINGDOK7 2008	LREC 2008	ALKNYELVDOK2 2009	LINGDOK8 2009	MSZNY 2009	ALKNYELVDOK3 2009	APPLINGPHD 2009	ALKNYELV 2010	COLING 2010	MSZNY 2010	LINGDOK10 2011	MWE 2011	RANLP 2011a	RANLP 2011b
Corpora						•				•	•			•	
Status		•													
Verbal counterparts				•	•										
Syntax												•			
Semantics	•							•							
Lexical representation							•								
Automatic identification									•	•			•		•
Word sense disambiguation			•						•						
Information extraction and retrieval									•						
Machine translation	•							•	•						

Table 3: The author's publications and fields of the thesis

## Conference presentations related to the thesis

1. Funkcióigés szerkezetek vizsgálata lexikai függvények segítségével [Analyzing light verb constructions with the help of lexical functions]. 9th National Conference of PhD Students of Linguistics, Szeged, 18 November 2005.
2. A funkcióigés szerkezetekről [On light verb constructions]. University of Szeged, Faculty of Arts, Conference of PhD Students of Linguistics, Szeged, 21 June 2006.
3. A főnév + ige szerkezetek státusáról [On the status of noun + verb constructions]. 10th National Conference of PhD Students of Linguistics, Szeged, 1 December 2006.
4. A félig kompozicionális szerkezetek gépi fordításainak lehetőségeiről [On possible ways of automatically translating semi-compositional constructions]. 1st Applied Linguistics PhD Conference, Budapest, 2 February 2007.
5. Félig kompozicionális szerkezetek és kollokációs láncok [Semi-compositional constructions and collocational chains]. University of Szeged, Faculty of Arts, Conference of PhD Students of Linguistics, Szeged, 20 June 2007.

6. *Előadást tart vs. előad*: főnév + ige szerkezetek igei variánsai [*To give a lecture vs. to lecture*: verbal counterparts of noun + verb constructions]. 11th National Conference of PhD Students of Linguistics, Szeged, 6 December 2007.
7. Angol–magyar főnév + ige szerkezetek és igei párjaik [English–Hungarian noun + verb constructions and their verbal counterparts]. 2nd Applied Linguistics PhD Conference, Budapest, 8 February 2008.
8. (with György Szarvas, Attila Almási, Dóra Szauter, Róbert Ormándi, Richárd Farkas, Csaba Hatvani and János Csirik) Hungarian word-sense disambiguated corpus. 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 30 May 2008.
9. A főnév + ige szerkezetek lexikai reprezentációjáról [On the lexical representation of noun + verb constructions]. 12th National Conference of PhD Students of Linguistics, Szeged, 3 December 2008.
10. Főnév + ige szerkezetek a szótárban [Noun + verb constructions in the dictionary]. 3rd Applied Linguistics PhD Conference, Budapest, 6 February 2009.
11. Félig kompozicionális főnév + ige szerkezetek a számítógépes nyelvészetben [Semi-compositional noun + verb constructions in natural language processing]. Novel Methods in Applied Linguistics Research, Budapest, 27 October 2009.
12. Mi fán terem a főnév + ige szerkezet? [From what tree can you harvest noun + verb constructions?]. 13th National Conference of PhD Students of Linguistics, Szeged, 25 November 2009.
13. Félig kompozicionális főnév + ige szerkezetek a Szeged Korpuszban [Semi-compositional constructions in the Szeged Corpus]. 6th Conference on Hungarian Computational Linguistics, Szeged, 3 December 2009.
14. (with János Csirik) Hungarian corpus of light verb constructions. COLING 2010, Beijing, China, 24 August 2010.
15. (with Zsuzsanna Felvégi and Krisztina R. Tóth) Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban [Semi-compositional constructions

in the SzegedParalell English–Hungarian parallel corpus]. 7th Conference on Hungarian Computational Linguistics, Szeged, 2 December 2010.

16. (with István Nagy T. and Gábor Berend) Detecting noun compounds and light verb constructions: a contrastive study. *ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland, Oregon, USA, 23 June 2011.
17. (with István Nagy T. and Gábor Berend) Domain-dependent identification of multiword expressions. *Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria, 12 September 2011.
18. (with István Nagy T. and Gábor Berend) Multiword expressions and named entities in the Wiki50 corpus. *Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria, 12 September 2011.