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

Veronika Vincze

University of Szeged

August 2011

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

OF THE UNIVERSITY OF SZEGED

Supervisor: Károly Bibok, PhD

University of Szeged

Doctoral School in Linguistics

Ph.D. Programme in Theoretical Linguistics

Contents

List of Tables

List of Figures

List of Abbreviations

ABL	ablative
ACC	accusative
ADE	adessive
ADJ	adjective
ADV	adverb
ALL	allative
ART	article
CAU	causal-final
CAUS	causative
COMP	complementizer
DAT	dative
DEL	delative
ELA	elative
EN	English
ESS	essive
ÉKsz.	The Concise Dictionary of the Hungarian Language
ÉrtSz.	The Explanatory Dictionary of the Hungarian Language
FX	semi-compositional construction
GEN	genitive
GS	gold standard
HU	Hungarian
IE	information extraction
ILL	illative
INE	inessive

INS	instrumental
IR	information retrieval
MT	machine translation
MWE	multiword expression
NE	named entity
NLP	natural language processing
NOM	nominalized semi-compositional construction
NP	noun phrase
OBJ	objective conjugation
PART	semi-compositional construction in the form of a participle
PART (in glosses)	participle
PASS	passive
PL	plural
POS	part of speech
POSS	possessive
PP	prepositional phrase
PREP	preposition
PRON	pronoun
SAU	sentence alignment unit
SG	singular
SPLIT	split semi-compositional constructions
SUB	sublative
SUP	superessive
TER	terminative
TRANS	translative
UTAH	Uniform Theta-role Assigment Hypothesis
WSD	word sense disambiguation
XML	Extended Markup Language

Preface

In this thesis, a subtype of multiword expressions, namely, semi-compositional constructions will be analyzed from the perspectives of theoretical and computational linguistics. Multiword expressions are lexical items with spaces that exhibit peculiarities on several levels of grammar: for instance, their parts cannot be substituted by another word similar in meaning (lexical idiosyncracy) or their meaning cannot be computed from the meaning of their parts and their combinatorial rules (semantic idiosyncracy). However, in many cases they are similar to productive and compositional phrases on the surface level. These features make it difficult to offer a linguistically apt analysis that can give an account of their special features and natural language processing applications also encounter problems when dealing with multiword expressions.

Semi-compositional constructions are common noun + verb combinations where the meaning of the construction is mostly determined by the noun. In other words, the noun functions as the semantic head of the construction, on the other hand, the syntactic head is the verb. Their structure is similar to productive or idiomatic noun + verb combinations, they nevertheless behave differently from those two classes with respect to their syntax and semantics, which provides a rich soil for linguistic investigation.

The main objective of this thesis is to provide analyses of semi-compositional constructions at several levels of grammar (i.e. syntax, semantics and lexicology) that can be fruitfully exploited in their NLP treatment (automatic identification and several applications such as word sense disambiguation, information extraction and retrieval and machine translation). In this way, it will be demonstrated how theoretical linguistic results can be applied in an empirical field of study (i.e. in natural language processing). This aspect of the thesis constitutes a direct link from theory to practice.

Veronika Vincze

Szeged, August 2011

Acknowledgements

First of all, I would like to thank my supervisor, Károly Bibok for his guidance and for his useful comments and remarks on my work. He has had a crucial role in turning my interest to computational linguistics, for which I can never be grateful enough.

I would also like to thank my PhD classmates for making the years of our PhD studies at the University of Szeged unforgettable and for our vivid (gastro)linguistic sessions. In alphabetical order: Péter Nádasdi, Katalin Nagy, Magdolna Ohnmacht and Balázs Szilárd.

My thanks go to János Csirik for letting me work at the inspiring Human Language Technology Group. I am grateful to my colleagues for creating a challenging and inspiring atmosphere at our department, which led me to address a variety of new tasks and interesting challenges throughout the years. I am indebted to my computer scientist colleagues – especially to Richárd Farkas and György Szarvas –, whom I could always turn to with computational issues. I would also like to thank my linguist colleagues – especially Attila Almási and Ágnes Klausz – for their help in linguistic issues and in annotating the corpora used in this research. Their efforts were indispensable in realizing the projects described here.

Last but not least, I would like to thank my parents, my grandparents and my sister for their constant love and support and for believing in me from the beginnings. As a way of expressing my gratitude, I would like to dedicate this thesis to them.

Part I

Background

Chapter 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

(1.1) *újságot* *olvas*
newspaper-ACC reads
'to read a newspaper'

or

(1.2) *levelet* *ír*
letter-ACC writes
'to write a letter'

On the other hand, idiomatic expressions such as

(1.3) *csütörtököt* *mond*
Thursday-ACC says
'to fail to work'

and

(1.4) *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 different languages are shown in (1.5). Since their meaning is the same, only glosses are provided:[1]

(1.5) (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

(c) German:

eine *Vorlesung* *halten*
a presentation to.hold

in *Blüte* *stehen*
in bloom to.stand

es *gibt* *eine* *Möglichkeit*
it gives a possibility

(d) French:

faire *une* *présentation*
to.make a presentation

être *en* *fleur*
to.be in bloom

l'occasion *se* *présente*
the.possibility itself presents

[1]Nouns in the nominative case and verbs in the subjective conjugation are not marked distinctively in glosses.

(e) Russian:

čitat' *doklad*
to.read presentation

pokryt'sja *cvetami*
to.cover.itself bloom-INS

predstavljaetsja *vozmožnost'*
to.emerge.itself possibility

(f) Spanish:

dar *una* *conferencia*
to.give a presentation

dar *flores*
to.give bloom-PL

la *ocasión* *se* *presenta*
the possibility itself presents

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)[2], in English, *complex verb structures*, *support verb constructions* or *light verb constructions* can be found,[3] 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)

[2]In German literature, constructions where the nominal component is the subject are not traditionally considered to be *Funktionsverbgefüge* – thanks are due to György Scheibl for pointing out this issue.

[3]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. (2004a), that is, the term *support verb* is a hypernym of *light verb*.

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 Chapters 6 and 7) –, 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[4] of the construction should also be avoided. That is why the term *semi-compositional constructions*[5] will be henceforth used for this type of common noun + verb constructions, following Langer (2004).

1.1 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:

- 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 noun + verb constructions and idioms with respect to their productivity and compositionality (scalability);

[4] In Chapter 5 we will 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.

[5] In Hungarian, they are called *félig kompozicionális szerkezetek*, which term serves as the base for the abbreviation *FX* (i.e. a shorter form of FKSZ) used in tables and figures throughout this thesis.

- 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).

These questions will be investigated in this thesis in detail.

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.

1.2 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?

1.3 Methodology

In order to answer the above questions, language data from Hungarian and English corpora will be 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: the test battery described in Chapter 4 was exploited in the annotation process. However, the data collected from the corpora serve as a basis for drawing theoretical conclusions on the behavior of semi-compositional constructions. Differences in corpus annotation can also lead to the clarification and reformulation of annotation guidelines (and the test battery). Annotating parallel data in two languages also made it possible to focus on interlingual differences. Thus, a nice interplay of theory and practice can be observed in the data collecting methodology of this thesis, which is described in detail in Chapter 3.

1.4 The structure of the thesis

The thesis has the following structure. First, a general overview of multiword expressions which semi-compositional constructions are a subtype of is presented (Chapter 2) and the corpora and database that constitute the language material behind this research are described in detail (Chapter 3). The remainder of the thesis is divided into two main parts. First, theoretical questions are discussed (Part II) concerning (1) the status of semi-compositional constructions with regard to constructions with a similar syntactic structure (Chapter 4), (2) their relation to their verbal counterparts (Chapter 5), (3) the syntactic and semantic features of constructions (Chapters 6 and 7, respectively) and (4) their lexical representation (Chapter 8). Second, computational linguistic issues are highlighted (Part III) with special emphasis on the automatic identification of semi-compositional constructions (Chapter 9) and their treatment in word sense disambiguation (Chapter 10), information extraction and retrieval

(Chapter 11) and machine translation (Chapter 12). The thesis concludes with a summary of results and possible ways of future work are also suggested (Chapter 13). Finally, in the Appendices, lists of the most frequent semi-compositional constructions and their verbal counterparts are provided.

1.5 The relation of former publications and thesis topics

As portions of this thesis have previously appeared in several papers by the author, it seems reasonable to summarize which results were achieved by the author in which publication, which is visually represented in Table 1.1.

Vincze (2009c), Vincze and Csirik (2010), Vincze et al. (2010a) and Vincze et al. (2011b) describe the methodology and annotation guidelines for annotating semi-compositional constructions. The author's main contributions were designing the methodology and annotation principles of corpus building and supervising the annotation work, besides she also participated in annotating and checking the data (see Chapter 3).

In Vincze (2008a), the test battery for distinguishing semi-compositional constructions, productive constructions and idioms is described (see Chapter 4).

In Vincze (2009b), semi-compositional constructions are compared to their verbal counterparts while in Vincze (2009a) an English–Hungarian contrastive study is carried out on semi-compositional constructions and their verbal counterparts (see Chapter 5).

In Vincze (2011), syntactic analyses of semi-compositional constructions are provided in generative and dependency grammar frameworks (see Chapter 6).

In Vincze (2007) and Vincze (2009e), semantic relations between the nominal component and the verb are briefly presented (see Chapter 7).

In Vincze (2009d), several methods for the lexical representation of semi-compositional constructions are contrasted and one of them is proposed to be applied in further lexicological work (see Chapter 8).

In Vincze (2010), germs of ideas on the computational linguistic treatment of semi-compositional constructions are designed, which are now elaborated in Chapters 9, 10, 11 and 12.

In Vincze and Csirik (2010), some hints on the automatic identification of semi-compositional constructions were also provided and in Vincze et al. (2011a), methods and results on identifying semi-compositional constructions in the Wiki50 corpus are presented, which are

	Chapter									
	3	4	5	6	7	8	9	10	11	12
ALKNYELVDOK1 2007 (Vincze, 2007)					•					•
LINGDOK7 2008 (Vincze, 2008a)		•								
LREC 2008 (Vincze et al., 2008)								•		
ALKNYELVDOK2 2009 (Vincze, 2009a)			•							
LINGDOK8 2009 (Vincze, 2009b)			•							
MSZNY 2009 (Vincze, 2009c)	•									
ALKNYELVDOK3 2009 (Vincze, 2009d)						•				
APPLINGPHD 2009 (Vincze, 2009e)					•					•
ALKNYELV 2010 (Vincze, 2010)							•	•	•	•
COLING 2010 (Vincze and Csirik, 2010)	•						•			
MSZNY 2010 (Vincze et al., 2010a)	•									
LINGDOK10 2011 (Vincze, 2011)				•						
MWE 2011 (Vincze et al., 2011a)							•			
RANLP 2011 (Vincze et al., 2011b)	•									
RANLP 2011 (Nagy T. et al., 2011)							•			

Table 1.1: The author's publications and chapters of the thesis

adapted to the SzegedParalellFX corpus in Nagy T. et al. (2011). The author was responsible for providing the linguistic background and for defining some of the features and adaptation techniques applied in the experiments (see Chapter 9).

In Vincze et al. (2008), the construction of a WSD corpus for Hungarian is described, which corpus serves now as the starting point of experiments presented in Chapter 10. The author's contributions to this paper were designing some of the annotation principles and annotating the database.

In Vincze (2007) and Vincze (2009e), some ways of supporting the machine translation of semi-compositional constructions are described (see Chapter 12).

Chapter 2

Semi-compositional constructions as a subtype of multiword expressions

2.1 Introduction

In natural language processing (NLP), one of the most challenging tasks is the proper treatment of multiword expressions (MWEs). Multiword expressions are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002). To put it differently, they are lexical items that contain spaces or "idiosyncratic interpretations that cross word boundaries" (see also Siepmann (2005), Siepmann (2006). Multiword expressions are frequent in language use and they usually exhibit unique and idiosyncratic behavior, thus, they often pose a problem to NLP systems (on the treatment of MWEs in Hungarian, see Oravecz et al. (2004) and Váradi (2006)).

In this chapter, multiword expressions are characterized in general. Special attention is paid to their idiosyncratic features together with the problems concerning their NLP treatment, then a possible classification of MWEs is offered. At the end of the chapter, the characteristics of semi-compositional constructions as a subtype of multiword expressions are analyzed in detail and it is also indicated what theoretical and practical consequences these features can result in.

2.2 Idiosyncratic features of multiword expressions

Multiword expressions show some idiosyncratic features on several levels of grammar, i.e. lexical, syntactic, semantic and pragmatic idiosyncracy can be observed. Certain MWEs cannot be characterized with such types of idiosyncracy since they are only subject to statistical idiosyncracy: the co-occurrence of their parts is of significantly high frequency. Such types of MWEs are called collocations (Sag et al., 2002). It is important to note, however, that not all idiosyncratic features are valid for each MWE – there can be MWEs that are idiosyncratic only from a semantic perspective but show ordinary syntactic behavior and vice versa.

In this section, each type of MWE idiosyncracy is illustrated with examples from both English and Hungarian.

2.2.1 Lexically idiosyncratic multiword expressions

Lexical idiosyncracy refers to the fact that the parts of a given MWE cannot be substituted by another word of similar meaning (Oravecz et al., 2005) without losing its meaning (or sometimes its grammaticality). This phenomenon is also called non-substitutability (see Manning and Schütze (1999)). For instance, the verbal component of semi-compositional constructions cannot be replaced by another word: we cannot say **do a mistake* instead of *make a mistake* or as for idioms, *kick the pail* and *kick the bucket* – or *fűbe harap* (grass-ILL bites) 'to die' and *pázsitba harap* (lawn-ILL bites) 'to bite into the lawn'– do not mean the same thing.

This feature might be useful in the automatic identification of MWEs since the relative frequency of the co-occurrence of *do* and *a mistake* is smaller than that of *make* and *a mistake*, which is indicative of the latter being a MWE (or at least a collocation or an institutionalized phrase, see 2.2.5).

2.2.2 Syntactically idiosyncratic multiword expressions

Syntactic idiosyncracy means that the syntax of the MWE in question exhibits some peculiarity, that is, its syntactic properties do not follow from the syntactic properties of its lexical parts. An example of a syntactically idiosyncratic MWE is *kerek perec* (round pretzel) 'overtly, unambiguously', which consists of an adjective and a noun but the whole expres-

sion functions as an adverb. An English example is *all of a sudden*, which also functions as an adverb, however, it is composed of a pronoun (*all*), a preposition (*of*), an article (*a*) and an adjective (*sudden*). Thus, multiple syntactic idiosyncracies can be identified with regard to this MWE:

- within the construction it is only an article and an adjective that follow each other, which is deviant in terms of English syntax (note that it would be a perfectly normal sequence if a noun was to follow) – **a sudden* vs. *a sudden change*
- the construction as a whole is an adverb while none of its parts is an adverb – PRON + PREP + ART + ADJ = ADV

This latter feature reflects one important difficulty for NLP applications, namely, the identification of multiword expressions. For each higher level of processing (e.g. syntactic or semantic analysis), it should be known that this sequence of words forms one unit and its part-of-speech (POS)-tag is ADV. Thus, this sequence of words should be merged into one lexical unit at an early phase of text processing in order to treat it as an adverb (more on this issue in Chapter 9).

2.2.3 Semantically idiosyncratic multiword expressions

The semantic idiosyncracy of MWEs refers to the fact that they are not (totally) compositional: their meaning cannot be computed solely on the basis of the meaning of their parts and their connection. The most typical examples are idioms, where the meaning of the whole construction has usually nothing to do with the meaning of its parts as in *to be on cloud nine* 'to be very happy'. In the meaning of the idiom, neither a 'cloud' nor any number can be identified. The compositionality of MWEs can be measured with different metrics (Bu et al., 2010; Pecina, 2010).

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

(2.1) *to spill the beans*

'to reveal a secret'

beans = 'secret'

spill = 'reveal'

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

vesz = 'understand'

lap = 'message'

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

The above distinction may have interesting implications for word sense disambiguation. If the parts of a MWE can be attributed a special distinct meaning, the question arises whether this meaning should be added to the sense inventory of the given word or not, in other words, to decompose its meaning or not. With the addition of these extra meanings, the interpretation of such decomposed MWEs may be facilitated, which might prove useful for higher level tasks such as information extraction or machine translation. On the other hand, non-decomposable MWEs should be treated as one unit on every level of processing – and their meaning should be encoded in the lexicon – since in this case, it is impossible to divide the expression into words among which the meaning of the whole expression could be distributed. Thus, to investigate the possibilities and consequences of different NLP treatment of non-decomposable and decomposable MWEs is a highly interesting question to answer.

2.2.4 Pragmatically idiosyncratic multiword expressions

Pragmatically idiosyncratic multiwords expressions are usually used only in a specific situation or under certain circumstances. For instance, the phrase *How do you do?* occurs

most often when introducing oneself to someone or it is highly improbable that two people meeting at 8pm would say *Good morning* (instead of *Good evening*).

Pragmatically idiosyncratic MWEs may be used in other – not prototypical – situations, however, in such cases their usage invokes irony or other instances of figurative speech, e.g. *Good morning* uttered at 8pm might convey the meaning 'it is high time to realize something'. The automatic recognition of such conversational implicatures is desirable for developing more efficient man–machine dialogue systems.

2.2.5 Statistically idiosyncratic multiword expressions

The parts of statistically idiosyncratic MWEs occur together with a significantly high probability, nevertheless, they do not show any syntactic or semantic peculiarities. These MWEs are called collocations (Sag et al., 2002). An example for statistically idiosyncratic MWEs is *black and white* or its literal Hungarian equivalent *fekete-fehér* – the phrases *white and black* or *fehér-fekete* would have exactly the same meaning, however, in the overwhelming majority of cases, *black and white* and *fekete-fehér* are used.

Institutionalized phrases are related to collocations: their meaning is compositional and in principle, each part of them could be substituted by another lexical item with the same meaning, however, combinations yielded in this way (i.e. anti-collocations (Pearce, 2001)) are hardly attested. For instance, *first lady* could be called *prime lady* but the latter term is typically not applied. Another example is that what is called *return ticket* in British English is called *round-trip ticket* in American English, thus, there can be regional differences in the usage and acceptability of collocations and MWEs too.

2.3 A classification of multiword expressions

Multiword expression can be divided into several groups based on the parts of speech of their components or based on their syntactic and semantic behavior. In the following, such classifications are briefly presented.

2.3.1 Compounds

A compound is a lexical unit that consists of two or more elements that exist on their own. Orthographically, compounds may include spaces (*high school*, *kútba esés* (well-ILL falling) 'failure') or hyphens (*well-known*, *időjárás-jelentés* 'weather forecast') or none of them (*headmaster*, *iskolaigazgató* (school.director) 'headmaster').

Compounds can be divided into the following categories (Sag et al., 2002; Kim, 2008):

- nominal compounds: *killer whale*, *iron maiden*, *középiskola* (middle.school) 'high school', *fülhallgató* (ear.listener) 'earphone'
- adjectival compounds: *red haired*, *Roman Catholic*, *nagyotmondó* (big-ACC.telling) '(someone) telling tall tales', *délszláv* (south.Slavic) 'South Slavic'
- adverbial compounds: *above all*, *at most*, *csakazértis* (just.that-CAU.too) 'just because', *dehogy* (but.that) 'not at all'
- prepositional compounds: *next to*, *out of*
- conjunctions: *so that*, *in order to*, *nehogy* (not.that) 'in order not to'

Some compounds might show syntactic peculiarities (e.g. the plural of *attorney general* is *attorneys general* instead of **attorney generals*). Their semantic interpretation, however, might differ from case to case: for instance, *sunflower oil* is made **from** sunflower seeds but *baby oil* is made **for** babies. In other cases, the meaning of the phrase is not related to the meaning of its parts at all such as in *commonplace*. Thus, the semantic relation between the parts of the compound is not straightforward to determine.

2.3.2 Verb–particle constructions and verbs with prefixes

Verb–particle constructions (also called phrasal verbs or phrasal–prepositional verbs) are combined of a verb and a particle/preposition (see e.g. Kim (2008)). They can be adjacent (as in *put off*) or separated by an intervening object (***turn*** *the light* ***off***). Their meaning can be compositional, i.e. it can be computed from the meaning of the preposition and the verb (*lie down*) or non-compositional (*do in* 'kill').

In Hungarian, verbs can have verbal prefixes, which can be separated from the verb for syntactic reasons (compare *bejön* (in.comes) 'he enters' and *nem jön be* (not comes in) 'he

does not enter'). In this case, the meaning of the verb with the prefix is compositional, however, other examples such as *berúg* (in.kick) 'to get drunk' are idiomatic.

2.3.3 Idioms and proverbs

An idiom is a MWE whose meaning cannot (or can only partially) be determined on the basis of its components (Sag et al., 2002; Nunberg et al., 1994). Idioms can be decomposable and non-decomposable (see 2.2.3). Although most idioms behave normally as morphology and syntax are concerned, i.e. they can undergo some morphological change (e.g. verbs are inflected in a normal way as in *He spills/spilt the beans*), their semantics is totally unpredictable.

Proverbs express some important facts that are thought to be true by most people. Proverbs usually take the same form and show no morphological change (i.e. they are fixed expressions). Some examples are: *It's no use crying over spilt milk* or *The early bird catches the worm* or their Hungarian equivalents *Eső után köpönyeg* (rain after cloak) or *Ki korán kel, aranyat lel* (who early gets.up gold-ACC finds).

2.3.4 Determinerless prepositional phrases

Determinerless PPs are made up of a preposition and a singular noun (without a determiner) (Kim, 2008) – in this way, they are syntactically marked – and they usually function as an adverbial modifier (*in turn*, *in part*, *on TV*, *by car* etc.).

2.3.5 Light verb constructions

Light verb constructions[1] consist of a nominal and a verbal component where the noun is usually taken in one of its literal senses but the verb usually loses its original sense to some extent. 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)). Some examples are offered here: *to give a lecture – előadást tart* (presentation-ACC holds), *to come into bloom – virágba borul* (bloom-ILL falls).

[1]This is one common name typically found in the literature for those that are called semi-compositional constructions in this thesis, cf. Chapter 1.

2.3.6 Other types of multiword expressions

There are other types of MWEs that do not fit into the above categories (some of them are listed in Jackendoff (1997)). Such expressions are multiword named entities, which can be composed of any words or even characters and their meaning cannot be traced back to their parts. For instance, *Ford Focus* refers to a car and has nothing to do with the original meaning of *ford* or *focus*.

Another group of MWEs is formed of foreign phrases such as *status quo*, *c'est la vie* and *ad hoc*. Although they are composed of perfectly meaningful parts in the original language, these words do no exist on their own in English and in Hungarian hence it is impossible to derive their meaning from their parts and the expression must be stored as a whole.

More complex and longer MWEs are quotations ("May the Force be with you"), lyrics of songs, clichés and commonplaces (*That's life*) and proverbs are also similar to them in that they are longer MWEs and are not changeable (see Jackendoff (1997) for details).

2.4 The syntactic behavior of multiword expressions

As for the syntactic behavior of multiword expressions, they can be grouped as fixed, semi-fixed and flexible phrases (Sag et al., 2002). Fixed expressions do not exhibit a syntactic variability: they cannot be inflected and cannot be modified etc. (This feature is mentioned as *non-modifiability* in Manning and Schütze (1999).) Determinerless prepositional phrases and proverbs are good examples of such expressions.

Semi-fixed expressions are flexible to a certain degree: for instance, verbs in non-decomposable idioms can undergo inflection and noun compounds can have a plural form.

Syntactically flexible expressions show a higher degree of syntactic variability. For instance, the two components of light verb constructions and verb-particle combinations might not even be adjacent, they can take arguments and the nominal component of the light verb construction can be modified.

2.5 Features of semi-compositional constructions

As this thesis pays special attention to light verb constructions (i.e. semi-compositional constructions), their characteristics are briefly summarized below.[2] It is also to be discussed what consequences these features can have in theoretical analyses and practical applications.

Concerning the idiosyncratic features, semi-compositional constructions exhibit lexical and semantic idiosyncracy (to some extent). As for the first one, the verbal component of the construction cannot be substituted by another verb with a similar meaning: instead of *make a decision* we cannot say **do a decision*. On the other hand, the change of the noun for a semantically similar word does not yield the agrammaticality of the construction: *make a contract* and *make a treaty* are both acceptable constructions. Finally, it should also be mentioned that there seem to be systematic cases where two semi-compositional constructions share all of their meaning components but their verbal components differ, for instance:

(2.2) *segítséget nyújt / ad*
help-ACC offers / gives
'to offer help'

With regard to semantic idiosyncracy, the meaning of semi-compositional constructions can (at least partially) be computed from the meanings of their parts and the way they are connected. Although it is the noun that conveys most of the meaning of the construction, the verb itself cannot be seen as semantically bleached (see e.g. Apresjan (2004), Alonso Ramos (2004), Sanromán Vilas (2009)) since it also adds important aspects to the meaning of the construction. For instance, (2.3) and (2.4) do not mean the same though they describe the same situtation:

(2.3) *segítséget ad*
help-ACC gives
'to offer help'

(2.4) *segítséget kap*
help-ACC receives
'to receive help'

[2] In Chapter 4, the characteristics of semi-compositional constructions will be contrasted to those of productive noun + verb combinations and idioms.

However, it is interesting to examine whether semi-compositional constructions are decomposable or not. On the one hand, the noun occurs in its usual sense (or in one of its usual senses), on the other hand, the verb typically takes a more abstract meaning of 'doing something' or 'performing some action' rather than keeping its original sense. On the basis of this, the meaning of the semi-compositional construction can be 'doing something that is encoded in the meaning of the noun', thus, semi-compositional constructions can be considered as decomposable MWEs.

Semi-compositional constructions are syntactically flexible, that is, they can manifest in various forms: the verb can be inflected, the noun can occur in its plural form and the noun can be modified. The nominal and the verbal component may not be adjacent in the sentence as in:

(2.5) *Ő hozta tegnap a jó döntést.*
he bring-PAST-3SG-OBJ yesterday the good decision-ACC
'It was him who made the good decision yesterday.'

The theoretical implications of the above features are the following. First, the lexical representation and the semantic analysis of semi-compositional constructions should be able to account for the fact that it is the noun that has primary importance in determining the meaning of the construction. Second, the sense of the verb should also be kept in mind and should not be neglected. Third, the syntactic analysis of semi-compositional constructions should be able to treat the modifiability of the constructions.

The above facts have some consequences for the NLP treatment of semi-compositional constructions as well. The syntactic modifiability makes the automatic identification of semi-compositional constructions difficult, especially in agglutinative languages such as Hungarian (see Chapter 9). The question of determining the sense of the verb within the construction will be analyzed in detail in Chapter 10. Lexical and semantic idiosyncracy can also affect the machine translation of the constructions: the nominal component being the semantic center of the construction seems to be constant across languages in the case of parallel constructions, hence it can be translated literally whereas the verb can be determined only lexically, i.e. in dictionaries (see Chapter 12).

2.6 Summary of results

In this chapter, multiword expressions and their idiosyncratic features have been presented. The following points have been emphasized:

- semi-compositional constructions have been described as a subtype of MWEs;
- idiosyncratic, syntactic and semantic features of MWEs have been discussed;
- a classification of MWEs has been provided;
- features of semi-compositional constructions have been highlighted;
- implications of the above have been indicated in theoretical and practical issues.

In both the theoretical and the computational linguistic analyses found in this thesis, these results will be paid distinguished attention.

Chapter 3

Data collection and the corpora analyzed

3.1 Introduction

In this chapter, the methods of data collection and corpus building are presented. The motivation behind corpus building is discussed, then corpus texts selected for manual annotation are shortly described. The process of annotation is illustrated with examples and finally, some statistical data are offered on the corpora.

3.2 The motivation behind corpus building

In our research, we voted for making use of empirical data from authentic texts because with this methodology, theoretical hypotheses and statements can be empirically supported or refuted on the basis of realistic language use. However, to the best of our knowledge, no adequate resources have been constructed on Hungarian and English semi-compositional constructions, i.e. there have been no data available on the spot. Thus, we decided to build such corpora. The motivation behind corpus building is twofold, reflecting the double aiming of this thesis: first, an annotated corpus can be automatically transformed into a database of semi-compositional constructions (i.e. a huge set of real-language examples can be quickly gathered from the corpus) and second, the annotated corpus can be utilized in several applications in natural language processing and it can be used as a benchmark database to compare the performance of different methods for detecting semi-compositional constructions.

In data collection, we aimed at gathering data from as many domains as possible, thus, we did not commit ourselves to work in only one domain since we believe that a compari-

son of data from multiple domains can effectively enhance research on semi-compositional constructions and it can help create domain-specific rules for their treatment from an applicational point of view.

In corpus building, theory and practice intertwine to a great extent. The annotation principles are primarily based on theoretical issues (see 3.4) while empirical data are used for supporting, rejecting or modifying theoretical claims. Thus, the interaction of empirical data and theoretical claims through annotation offers a nice methodological example for considering both theory and practice in linguistic analysis.

When selecting the corpora for annotation, the following aspects were taken into consideration:

- **variety of domains**

  In order to examine semi-compositional constructions from a linguistic spectrum as wide as possible, multiple domains are required to be included in the corpus. In this way, more types of semi-compositional constructions are expected to be found (we suppose that different constructions will occur in different domains such as legal or literary texts, cf. *perbe von* (sue-ILL draws) 'to sue' vs. *pofont ad* (slap-ACC gives) 'to slap in the face').

- **variety of language registers**

  Official and non-official types of texts can also differ in the number of semi-compositional constructions they contain (e.g. fields of official language use – e.g. texts of laws and decrees – are claimed to contain more semi-compositional constructions (B. Kovács, 1999)) and they may also contain different types of semi-compositional constructions.

- **size**

  In order to collect an appropriate size of data for linguistic analysis, the size of the corpus to be annotated has to be precisely estimated: it should not be too small because it would raise the risk of not containing enough data. On the other hand, the corpus should not be too big either since the time needed for annotation has to be also taken into account when planning corpus building.

- **multilinguality**

If texts to be annotated are available in several languages, interlingual comparisons can be easily carried out. Moreover, annotating data in multiple languages makes it possible to automatically extract parallel constructions from all annotated texts, thus yielding a multilingual database, which can be exploited in e.g. dictionary construction, machine translation or cross language information retrieval.

- **previous annotation**

  If the texts selected for corpus building contain some previous annotation, they can be fruitfully exploited in the analysis of data, moreover, NLP applications can also profit from the multi-layered annotations within one text.

With these aspects in mind, we finally selected two corpora for annotation and data collection, namely, the Szeged Treebank (Csendes et al., 2005) and the SzegedParalell corpus (Tóth et al., 2008; Vincze et al., 2010a). Both corpora contain texts from several domains, which makes it possible to compare the frequency of semi-compositional constructions across domains. Since one of the corpora is a parallel corpus, i.e. it comprises texts in two languages (English and Hungarian), observations on data from different languages can also be carried out, thus multilingual aspects are also emphasized in the research and some conclusions can also be drawn from the viewpoint of contrastive linguistics. We also built a corpus of English Wikipedia articles which are annotated for several types of multiword expressions – thus for semi-compositional constructions as well – and named entities (Vincze et al., 2011b). These corpora will be presented in detail in 3.5.

3.3 Types of semi-compositional constructions

Semi-compositional constructions may occur in various forms due to their syntactic flexibility (see Chapter 2). For instance, the verbal component may be inflected or the nominal component can be in the plural, etc. However, these inflectional differences can be easily resolved by a lemmatizer. On the other hand, besides the prototypical noun + verb combination in Hungarian and the verb + noun combination in English, semi-compositional constructions may be present in different syntactic structures, that is, in participles (this category including present, past and future participles, adverbial participles and infinitives) and they can also

undergo nominalization.[1] These types are all annotated in the corpus texts since they also occur relatively frequently (see statistical data in 3.5). All annotated types are illustrated below.

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

  bejelentést *tesz*
  announcement-ACC makes
  'to make an announcement'

  to take a look

- **Participles <part>**

  – Present participle

  életbe *lépő* *(intézkedés)*
  life-ILL stepping (instruction)
  '(an instruction) taking effect'

  decision-making (process)

  – Past participle

  csődbe *ment* *(cég)*
  bankrupt-ILL gone (firm)
  '(a firm) that went bankrupt'

  photos taken

  – Future participle

  fontolóra *veendő* *(ajánlat)*
  consideration-SUB to.be.taken (offer)
  '(an offer) that is to be taken into consideration'

  steps to be taken

  – Adverbial participle

  figyelembe *véve*
  account-ILL taking
  taking into account

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

coming into effect

– Infinitive

forgalomba *hozni*
circulation-ILL to.bring
'to put into circulation'

- **Nominalization <nom>**

bérbe *vétel*
rent-ILL taking
'hiring'

service provider

Split semi-compositional constructions, where the noun and the verb are not adjacent, are also annotated. In this case, the nominal and the verbal component were distinctively marked since other words and phrases might intervene in between the two parts of the construction (e.g. ***figyelembe*** *kellett neki* ***venni*** (consideration-ILL must-PAST3SG he-DAT to.take) 'he had to take it into consideration' or *a* ***decision*** *has been recently* ***made***). In this way, their identification becomes possible and the database can be used for training an algorithm that automatically recognizes (split) semi-compositional constructions.

3.4 Annotation principles

Corpus texts contain single annotation[2], i.e. one annotator worked on each text. The annotation process was supervised by the author and she also participated in annotating the corpora. In the Szeged Treebank, semi-compositional constructions can be found in between XML tags <FX> </FX>. On the other hand, texts from the SzegedParalellFX and Wiki50 corpora 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 semi-compositional construction or not, annotators were suggested to make use of a test battery developed for identifying Hungarian semi-compositional constructions (Vincze, 2008a), which is described in Chapter 4 and was adapted to English.

[2]A sample set of sentences in the SzegedParalellFX and Wiki50 was annotated by all annotators in order to check inter-annotator agreement, see below.

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 semi-compositional constructions. Since the outmost boundary of the nominal component was considered to be part of the semi-compositional construction, in several cases adjectives and other modifiers of the nominal head are also included in the construction, e.g.:

(3.1) *<FX>**nyilvános*** *ajánlatot* *tesz</FX>*
public offer-ACC makes
'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):

(3.2) *<FX>**Nyíregyházán*** *tartott* *ülésén</FX>*
Nyíregyháza-SUP held session-3SGPOSS-SUP
'at its session held in Nyíregyháza'

Constructions with a nominal component in the accusative case (3.3) can be nominalized in two ways in Hungarian, as in (3.4) and (3.5):

(3.3) *szerződést* *köt*
contract-ACC bind
'to make a contract'

(3.4) *<FX>szerződéskötés</FX>*
contract+bind-DERIV.SUFFIX
'making of a contract'

(3.5) *<FX>adásvételi* *szerződések* *megkötése</FX>*
sale contract-PL PREVERB-bind-DERIV.SUFFIX-3SGPOSS
'making of sales contracts'

Both types are annotated in the corpus.

Besides the prototypical occurrences of semi-compositional 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:

[3]As opposed to other languages where prototypical semi-compositional 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 semi-compositional construction. However, annotators were told to annotate such cases as well.

(3.6) *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 semi-compositional 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 (`http://www.inf.u-szeged.hu/rgai/mwe`) and in Appendices A and B, the most frequent ones are listed.

In SzegedParalellFX, both English and Hungarian semi-compositional constructions were annotated following the above annotation principles. The same annotator worked on both the source and the target language versions of each text. Since the SzegedParalell corpus does not contain syntactic annotation, no phrase boundaries were taken into account when annotating the texts: the smallest expression containing the semi-compositional construction was marked, which means that no adjectives and determiners of nominal components were included if the noun occurred right before the verb.

In the Wiki50 corpus, the same annotation principles were followed, however, nominalized forms of semi-compositional constructions were considered as compound nouns. Due to annotation principles, no text spans were classified as belonging to two MWE classes at the same time, thus, they were not annotated as semi-compositional constructions.

3.5 The corpora

In the following sections, the corpora Szeged Treebank, SzegedParalell and Wiki50 are described in detail together with some statistical data.

3.5.1 The Szeged Treebank

The Szeged Treebank is a morphosyntactically tagged and syntactically annotated database, which is available in both constituency-based (Csendes et al., 2005) and dependency-based (Vincze et al., 2010b) versions. In the corpus, each word is assigned all its possible morphosyntactic tags and lemmas and the appropriate one is selected according to the context.

When selecting texts for the corpus, the main criterion was that they should be thematically representative of different text types, thus they should derive from largely different

genres. As a result, the created corpus contained texts from six genres, each comprising circa 200,000 words. Naturally, the mentioned quantity (1.2 million word entries + punctuation marks) is still insufficient to cover an entire written language, but due to its variability, it serves as a good training database for machine learning algorithms and as a reference material for different research applications in the future. With this size, however, the Szeged Treebank is the biggest manually annotated Hungarian corpus. The genres included in the Szeged Treebank are the following:

- business news (short – 1 or 2 sentence long – pieces of news from the archive of the Hungarian News Agency)
- newspaper articles (excerpts from three daily papers (*Népszabadság*, *Népszava* and *Magyar Hírlap*) and one weekly paper (*HVG*))
- legal texts (excerpts from laws on economic enterprises and authors' rights)
- fiction (three novels: Jenő Rejtő: *Piszkos Fred, a kapitány* (Dirty Fred, the Captain), Antal Szerb: *Utas és holdvilág* (Journey by Moonlight) and George Orwell: *1984*)
- computer-related texts (excerpts from Balázs Kis: *Windows 2000 manual book* and some issues of the *ComputerWorld: Számítástechnika* magazine)
- short essays of 14-16-year-old students

Statistical data on the corpus can be seen in Table 3.1.

	Sentences	**Words**	**Punctuation marks**
Business news	9,574	186,030	25,712
Newspapers	10,210	182,172	32,880
Legal texts	9,278	220,069	33,515
Fiction	18,558	185,436	47,990
Computer texts	9,759	175,991	31,577
Student essays	24,720	278,497	59,419
Total	82,099	1,228,195	231,093

Table 3.1: Number of sentences, words and punctuation marks in the Szeged Treebank

In order to be able to compare the frequency of semi-compositional constructions in different domains, all texts in the corpus are annotated. The corpus contains 6734 occurrences of 1215 semi-compositional constructions altogether in 82,099 sentences. Thus, a specific

	<verb>	<part>	<nom>	<split>	**total**
Business news	565	697	90	40	1392
	40.6%	50.1%	6.5%	2.9%	20.7%
Newspapers	458	197	55	67	777
	58.9%	25.4%	7.1%	8.6%	11.5%
Legal texts	641	679	710	241	2271
	28.2%	29.9%	31.3%	10.6%	33.7%
Fiction	567	61	5	93	726
	78.1%	8.4%	0.7%	12.8%	10.8%
Computer texts	429	126	85	76	716
	59.9%	17.6%	11.9%	10.6%	10.6%
Student essays	582	122	9	139	852
	68.3%	14.3%	1.1%	16.3%	12.7%
Total	3242	1882	954	656	6734
	48.1%	27.9%	14.2%	9.7%	100%

Table 3.2: Subtypes of semi-compositional constructions in the Szeged Treebank

semi-compositional construction occurs 5.54 times in the corpus on average. Statistical data on the corpus are shown in Table 3.2.

It is revealed that although it is the 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 more than half of the data, which indicates the importance of their being annotated as well.

As for the general frequency of semi-compositional constructions in texts, we compared the number of verb + argument relations found in the Szeged Dependency Treebank (Vincze et al., 2010b) where the argument was a common noun to that of semi-compositional constructions. It has turned out that about 2.66% of verb + argument relations consist of semi-compositional constructions on average. However, there are differences among domains: in legal texts, this rate is twice as much as in business news.

Since a verb typically has more than one argument (except for intransitive verbs), we also calculated the rate of verbs and semi-compositional constructions. It is revealed that again, verbs in legal texts are most probable to occur in semi-compositional constructions while business news are also above average in this respect. However, student essays seem to contain less semi-compositional constructions as compared to the number of verbs.

The average number of semi-compositional constructions was also calculated with respect to the number of sentences in each subcorpus. In this case, all semi-compositional

constructions were considered, not only those with internal argument structure as in the previous cases. Legal texts abound in semi-compositional constructions on the sentence level (24.48%), their number is also high in business news compared to other subcorpora, however, in fiction and in student essays they cannot be frequently found. This again emphasizes that there are domains (especially law and ecomonics) where they should be paid distinctive attention. Statistical data are shown in Table 3.3.

Subcorpus	**V + arg**	**FX**	**%**	**V**	**FX**	**%**	**Sent**	**FX**	**%**
Business news	31,579	642	2.03	16,913	642	3.80	9574	1392	14.54
Newspaper texts	25,355	584	2.30	20,751	584	2.81	10,210	777	7.61
Legal texts	23,618	1002	4.24	15,557	1002	6.44	9278	2271	24.48
Fiction	22,435	710	3.16	34,805	710	2.04	18558	726	3.91
Computer texts	23,412	561	2.4	19,192	561	2.92	9759	716	7.34
Student essays	35,029	797	2.28	58,702	797	1.36	24,720	852	3.45
Total	161,428	4296	2.66	165,920	4296	2.59	82,099	6734	8.2

Table 3.3: The rate of semi-compositional constructions with regard to verb + argument relations, number of verbs and number of sentences

The corpus is publicly available for research and/or educational purposes under the Creative Commons license at `www.inf.u-szeged.hu/rgai/mwe`.

3.5.2 The SzegedParalell corpus

The English–Hungarian parallel corpus contains texts selected on the basis of grammatical and translational criteria. Sentences representing the grammar of the given language (usually taken from language books) and authentic texts are both included in the parallel corpus, thus, the balance is maintained between artificially constructed and real language structures. The corpus contains texts from the following domains:

- **language book sentences**: This subcorpus comprises sentences from language books that were compiled for language learners revising for a language exam.
- **texts on the European Union**: This subcorpus comprises texts collected from the `https://europa.eu.int` website. Topics range from the history of the European Union to the monetary system of the EU.
- **bilingual magazines**: This subcorpus consists of texts from the bilingual in-flight magazine, *Horizon Magazine* of Malév (Hungarian Airlines) and texts from a bilingual newspaper on real estates (*Resource Ingatlan Info*) in an easy-to-understand language.

Topics range from interviews with famous people through culture to the presentation of famous and interesting cities in the former, and from logistic centers through infrastructure developments to legal issues concerning real estates in the latter.

- **literature**: Modern literary works were mainly collected from the Hunglish corpus (Halácsy et al., 2005) and the Hungarian Electronic Library (`http://www.mek.hu`).

- **miscellaneous texts**: Some short texts (e.g. recipes) were also collected from the internet and included in the corpus.

Both paragraph and sentence alignment were checked and corrected manually, yielding in this way the first manually checked English–Hungarian parallel corpus.

Contrary to the Szeged Treebank, not the whole SzegedParalell corpus was annotated for semi-compositional constructions. The annotated parts will be called SzegedParalellFX, which includes only three novels from the literature subcorpus since it is primarily texts written in an official style (e.g. texts on economics, law or texts from magazines) that are expected to contain a number of semi-compositional constructions (as our preliminary results on monolingual data indicated). However, we selected the three novels carefully: the source language of two of them (Mark Twain: *The Man That Corrupted Hadleyburg* and Jonathan Swift: *Gulliver's Travels*) is English and the source language of Frigyes Karinthy: *Tanár úr kérem* (Please, Sir!) is Hungarian. With this choice 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 (*Gulliver's Travels* was published in 1726, *The Man That Corrupted Hadleyburg* in 1900 and *Tanár úr kérem* in 1916,[4] thus, differences between earlier and contemporary language use (found in magazine texts or language books) might also be revealed).

Some data on texts selected for annotation can be seen in Table 3.4.

The total number and the number of the subtypes of semi-compositional constructions are presented in Table 3.5. The first number in each cell refers to the English data and the second to the Hungarian data, respectively.

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

Subcorpus	Number of SAUs
Language book sentences	3496
EU	1518
Bilingual magazines	5320
Literature	3232
Miscellaneous	695
Total	14261

Table 3.4: Data on SzegedParalellFX

Subcorpus	VERB	PART	NOM	SPLIT	Total
Language book sent.	106/62	4/10	13/3	8/10	131 (3.7%)/85 (2.4%)
EU	132/158	30/76	24/32	41/29	227 (15%)/295 (19.4%)
Bilingual magazines	281/330	48/86	16/23	55/38	400 (7.5%)/477 (9%)
Literature	222/196	13/11	5/0	96/40	336 (10.4%)/247 (7.6%)
Miscellaneous	4/7	1/0	0/0	1/1	6 (0.8%)/8 (1.2%)
Total	745/753	96/183	58/58	201/118	1100 (7.71%)/1112 (7.8%)

Table 3.5: Semi-compositional constructions in SzegedParalellFX

The number of English and Hungarian semi-compositional constructions is approximately the same, thus, approximately the same percentage of sentence alignment units contains a semi-compositional construction (see Table 3.5).

However, it does not entail that each semi-compositional construction has an equivalent in the other language – in other words, there are constructions that are included only in one of the two parts of the corpus.

In the Hungarian part of the corpus, there are 1112 occurrences of 578 semi-compositional constructions in 14,261 sentence alignment units, thus, a specific semi-compositional construction occurs 1.92 times in the corpus on average. With regard to the English data, 587 semi-compositional constructions occur altogether 1100 times (1.87 times each on average).

As for the types of semi-compositional 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 main verb. Concerning the category PART, the presence of 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:[5]

(3.7) *az emberi jogokba* ***vetett*** *hit*
the human right-PL-INE cast-PAST-PART belief
a belief in human rights

In order to compare the difficulty of annotating semi-compositional 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 (GS) annotation. The inter-annotator agreement rates are presented in Table 3.6. Agreement rate was calculated on two levels: first, it was only considered whether the given semi-compositional construction was marked (i.e. no type was taken into account). On this level, the average agreement rate among the annotations was 78.15% on English data and 74.23% on Hungarian data (agreement rates are given in F-measure[6]). Second, the type of the semi-compositional 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. On this stricter level of measurement, the average agreement rates were 64.79% and 71.18% on English and Hungarian data, respectively. On Level 2, the metrics Jaccard index[7] and κ-measure[8] were also calculated.

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 on an advanced level, however, the latter was not their mother tongue. On the other hand, it is interesting to see that on 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.

[5]The example contains the semi-compositional construction *hitet vet* (belief-ACC cast) 'to believe', which was judged as ungrammatical by some speakers. However, it does occur in corpus texts – though not frequently – in participial and verbal forms as well therefore it is considered to be a genuine example.

[6]F-measure is the harmonic mean of precision and recall. It measures the agreement rate of two annotations with respect to the identification of target classes.

[7]The Jaccard index measures similarity of annotated examples and is defined as the size of the intersection divided by the size of the union of the sets of annotated examples.

[8]κ measures the agreement between two raters while taking into account the agreement by chance.

ENGLISH		**Precision**	**Recall**	**F-score**	**Jaccard**	**κ-measure**
GS vs. Annotator 1	Level 1	0.7793	0.7457	0.7621		
	VERB	0.7647	0.6566	0.7065	0.5462	0.5649
	PART	0.4118	0.5600	0.4746	0.3111	0.4145
	NOM	0.4000	0.2222	0.2857	0.1667	0.2515
	SPLIT	0.6774	0.7000	0.6885	0.5250	0.6506
	Level 2	0.5635	0.5347	0.5388	0.3872	0.4704
GS vs. Annotator 2	Level 1	0.7352	0.6940	0.7140		
	VERB	0.6847	0.7677	0.7238	0.5672	0.5679
	PART	0.7222	0.5200	0.6047	0.4333	0.5739
	NOM	0.8182	0.5000	0.6207	0.4500	0.6019
	SPLIT	0.4231	0.3667	0.3929	0.2444	0.3283
	Level 2	0.6620	0.5386	0.5855	0.4237	0.5180
GS vs. Annotator 3	Level 1	1.0000	0.7672	0.8683		
	VERB	1.0000	0.8283	0.9061	0.8283	0.8469
	PART	1.0000	0.7200	0.8372	0.7200	0.8211
	NOM	1.0000	0.5000	0.6667	0.5000	0.6485
	SPLIT	1.0000	0.7667	0.8679	0.7667	0.8512
	Level 2	1.0000	0.7037	0.8195	0.7037	0.7919
Total average	Level 1	0.8381	0.7356	0.7815		
	Level 2	0.7418	0.5923	0.6479	0.5049	0.5934
HUNGARIAN						
GS vs. Annotator 1	Level 1	0.6344	0.7254	0.6769		
	VERB	0.6187	0.8113	0.7020	0.5409	0.5433
	PART	0.7551	0.6727	0.7115	0.5522	0.6607
	NOM	0.8421	0.8000	0.8205	0.6957	0.8098
	SPLIT	0.5000	0.5714	0.5333	0.3636	0.5010
	Level 2	0.6790	0.7139	0.6919	0.5381	0.6287
GS vs. Annotator 2	Level 1	0.7256	0.6393	0.6797		
	VERB	0.7835	0.7170	0.7488	0.5984	0.6226
	PART	0.5556	0.3636	0.4396	0.2817	0.3456
	NOM	0.8750	0.3500	0.5000	0.3333	0.4804
	SPLIT	0.5417	0.6190	0.5778	0.4063	0.5441
	Level 2	0.6889	0.5124	0.5665	0.4049	0.4982
GS vs. Annotator 3	Level 1	1.0000	0.7705	0.8704		
	VERB	1.0000	0.7453	0.8541	0.7453	0.7680
	PART	1.0000	0.8727	0.9320	0.8727	0.9140
	NOM	1.0000	0.8000	0.8889	0.8000	0.8802
	SPLIT	1.0000	0.7143	0.8333	0.7143	0.8205
	Level 2	1.0000	0.7831	0.8771	0.7831	0.8456
Total average	Level 1	0.7867	0.7117	0.7423		
	Level 2	0.7893	0.6698	0.7118	0.5754	0.6575

Table 3.6: Inter-annotator agreement rates on the SzegedParalellFX corpus

However, differentiating between types of semi-compositional constructions (i.e. annotating on 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 Hungarian data than on English. Her moderate performance on 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 English data was better than on Hungarian data. Annotator 3 had the most experience in annotating linguistic corpora, which manifested in perfect precision[9]. Thus, in her case, annotation errors were related only to recall[10]. In other words, she failed to recognize some instances of semi-compositional constructions in text, but the text spans she marked were indeed semi-compositional constructions.

3.5.3 The Wiki50 corpus

When constructing the Wiki50 corpus, 50 random articles were selected from the English Wikipedia. The only selectional criterion applied was that each article should consist of at least 1000 words and they should not contain lists, tables or other structured texts (i.e. only running texts were included). In the corpus, several types of multiword expressions and named entities (NEs) were manually annotated. The corpus contains 114,570 tokens in 4350 sentences. Table 3.7 summarizes the number of occurrences and the number of unique phrases (i.e. no multiple occurrences are counted here) for each annotated category.

The corpus contains 368 occurrences of 287 semi-compositional constructions in 4350 sentences. Fifteen articles were annotated by all the annotators of the corpus and the agree-

[9]Precision is the number of correct results divided by the number of all returned results. It measures how many of the predicted elements actually belong to the target class.

[10]Recall is the number of correct results divided by the number of results that should have been returned. It measures how many of the target class elements the system is able to identify.

Category	Occurrence	Unique phrases
Noun compound	2929	2405
Adjectival compound	78	60
Verb–particle combination	446	342
Semi-compositional construction	368	338
Idiom	19	18
Other MWE	21	17
MWEs total	3861	3180
Person	4093	1533
Organization	1498	893
Location	1558	705
Miscellaneous NE	1827	952
NEs total	8976	4083

Table 3.7: Identified occurrences of categories in the Wiki50 corpus

ment rate concerning semi-compositional constructions was 70.69% (F-measure), 0.5467 (Jaccard) and 0.698 (κ-measure). The F-measure is somewhat lower than that of the English data in SzegedParalellFX, however, it should be noted that the annotation task was more complex in this corpus since other MWE categories and NEs were also to be annotated, which may have resulted in lower recall (i.e. the annotators were prone to overlook certain instances of semi-compositional constructions).

3.6 The database

The language data used in this research are composed of the union of the annotated corpora described in this chapter. Altogether, it contains 1524 semi-compositional constructions in Hungarian and 827 in English.

The ten most frequent semi-compositional constructions are listed in Table 3.8.

It can be seen that the pairs *döntést hoz – make/take a decision* and *részt vesz – take part* can be found among the most frequent semi-compositional constructions in both languages. However, interlingual comparisons should be made only cautiously due to the difference in the amount of data for the two languages.

The most frequent Hungarian and English verbal components are presented in Tables 3.9 and 3.10 (TREEBANK refers to data from the Szeged Treebank, PARALELL to those from the SzegedParalellFX corpus, WP50 to those from the Wiki50 corpus and UNION to those

	FX (HU)	Occurrence	FX (EN)	Occurrence
1.	irányt ad	608	take place	36
2.	részt vesz	278	make a decision	29
3.	szerződést köt	217	take part	26
4.	forgalomba hoz	205	play a role	25
5.	nyilvánosságra hoz	197	take care	18
6.	eszébe jut	191	take a decision	16
7.	sor kerül	182	make a remark	14
8.	figyelembe vesz	167	take a look	13
9.	határozatot hoz	153	give order	11
10.	döntést hoz	136	take seat	11

Table 3.8: The most frequent semi-compositional constructions

from the merged datasets) and in Figures 3.1 and 3.3, respectively.

	UNION	#	TREEBANK	#	PARALELL	#
1.	ad	1526	ad	1424	vesz	121
2.	vesz	916	vesz	795	ad	105
3.	hoz	836	hoz	775	tesz	84
4.	tesz	557	tesz	468	hoz	61
5.	köt	374	köt	348	kerül	53
6.	kerül	357	kerül	304	tart	52
7.	jut	259	jut	241	nyújt	41
8.	tart	247	tart	197	kap	36
9.	nyújt	187	lép	164	köt	26
10.	lép	179	nyújt	146	áll	24
11.	áll	169	áll	145	ér	22
12.	kap	163	kap	128	jut	18
13.	végez	122	végez	111	nyílik	17
14.	folytat	113	folytat	105	lép	16
15.	ér	100	ér	77	játszik	16

Table 3.9: The most frequent Hungarian verbal components

Figure 3.1 illustrates that the Hungarian verbal components in the unified dataset and in the Szeged Treebank show a Zipf-like distribution: Zipf's law states that in a corpus, the frequency of any word is inversely proportional to its rank in the frequency table. Thus, the most frequent word occurs twice as much as the second most frequent word, three times as much as the third most frequent word etc. (Zipf, 1949; Manning and Schütze, 1999). Note that the second and third most frequent verbal components almost share their frequency. However, Hungarian data from the SzegedParallelFX are more balanced (see also Figure 3.2), which might be due to the considerably smaller size of the dataset.

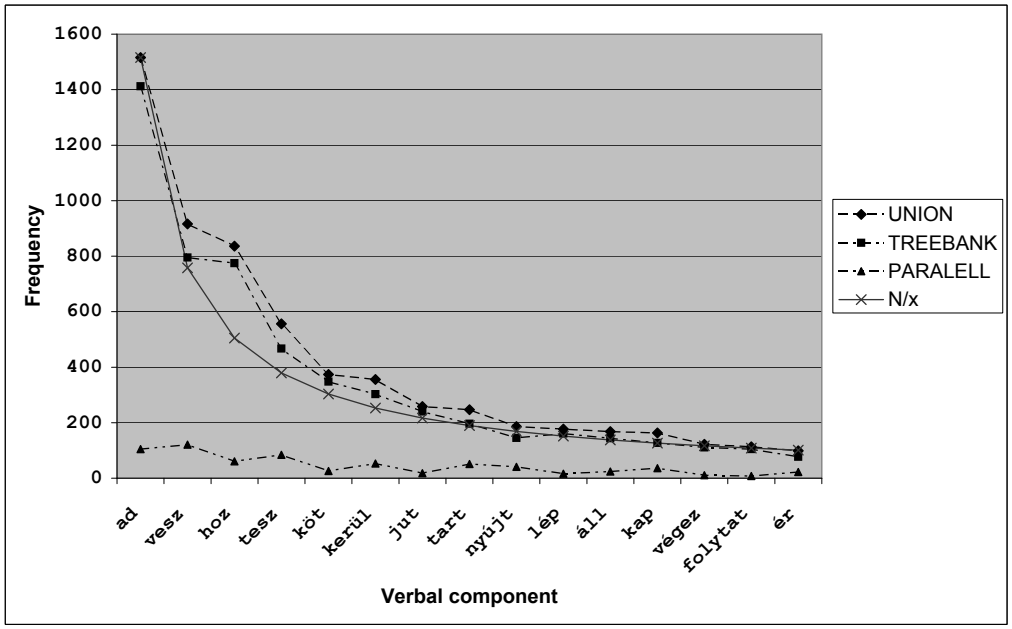
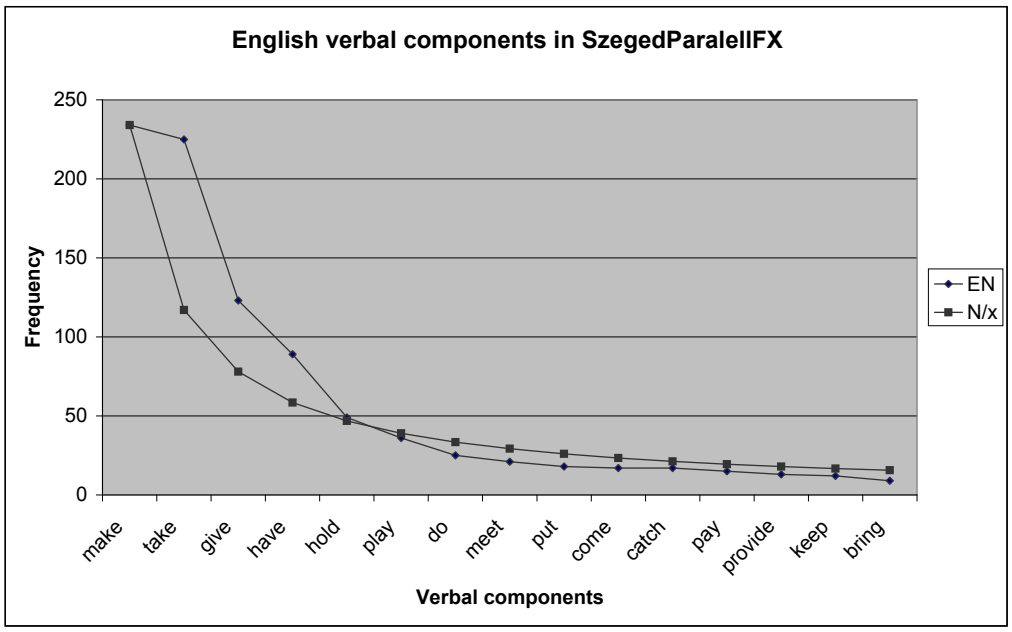


Figure 3.1: The most frequent Hungarian verbal components

	UNION		**PARALELL**		**WP50**	
1.	make	295	make	234	make	61
2.	take	290	take	225	have	29
3.	give	144	give	123	hold	28
4.	have	118	have	89	give	20
5.	hold	78	hold	49	play	13
6.	play	49	play	36	commit	10
7.	do	29	do	25	draw	9
8.	meet	29	meet	21	meet	8
9.	put	26	put	18	put	8
10.	come	23	catch	17	bring	6
11.	draw	17	come	17	come	6
12.	catch	17	pay	15	go	6
13.	commit	16	provide	13	gain	5
14.	pay	16	keep	12	do	4
15.	keep	15	bring	9	express	4

Table 3.10: The most frequent English verbal components



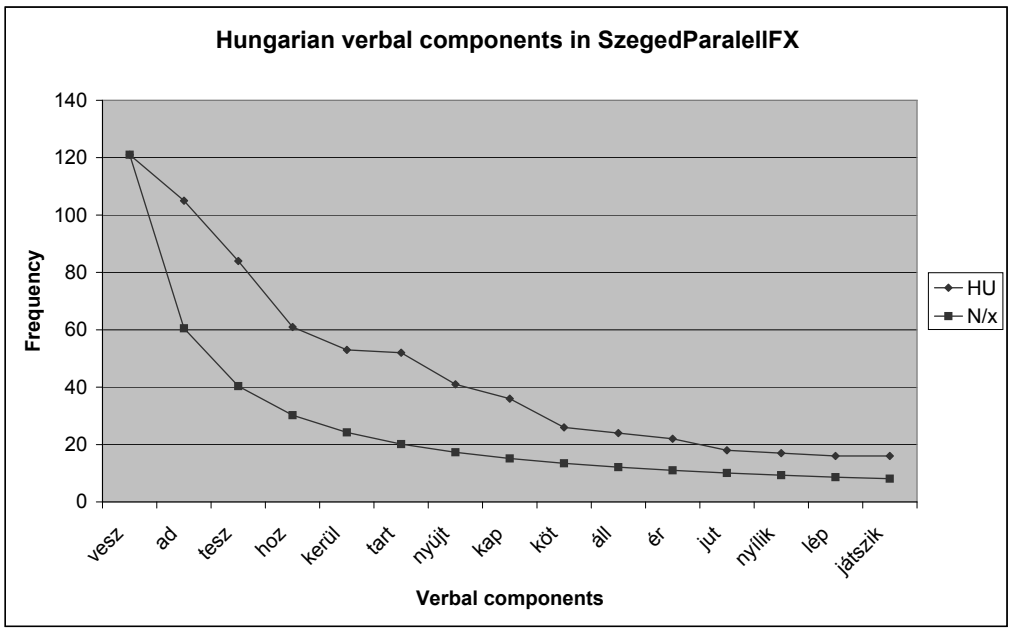


Figure 3.2: Distribution of the most frequent verbal components in SzegedParalellFX

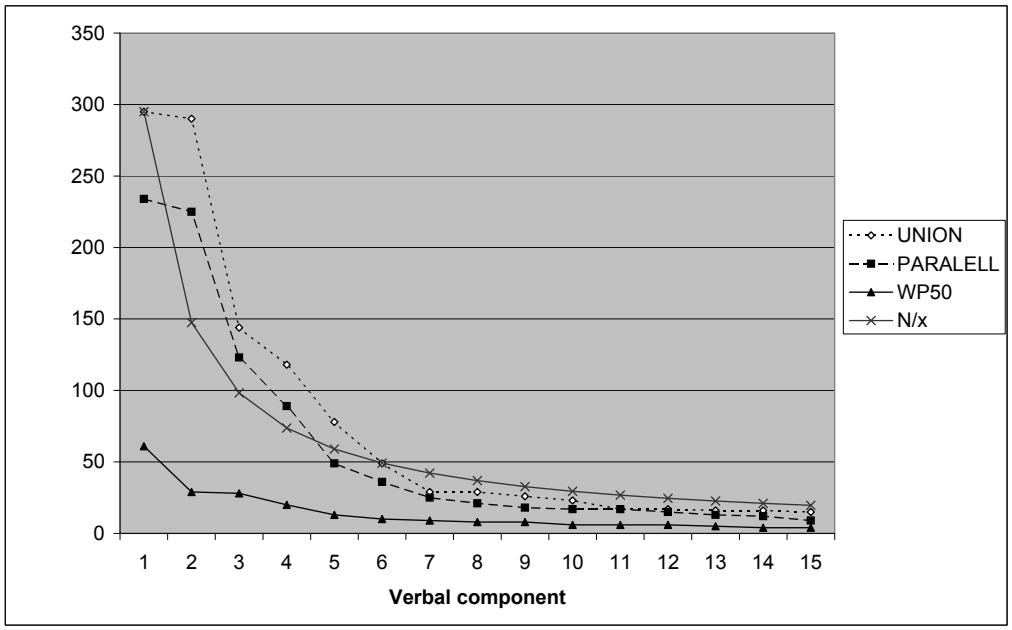


Figure 3.3: The most frequent English verbal components

The most frequent English verbal components are provided in Table 3.10 and in Figure 3.3 (see also Figure 3.2). With the first two verbal components sharing their frequency, this distribution also exhibits a Zipf curve (denoted by N/x). Again, if the tendencies observed in the smaller and the bigger dataset are compared, it can be seen that in the smaller English dataset, namely, Wiki50, the data are more balanced just like it was the case in the Hungarian datasets. Thus, it might be suggested that the size of the annotated corpora has a considerable effect on observing Zipf-like distribution.

3.7 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/last* 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 contain the most semi-compositional constructions (legal texts in the Szeged Treebank and texts on the European Union in the SzegedParalellFX). However, in language book sentences there are hardly any 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 semi-compositional constructions in a much bigger rate than their Hungarian counterparts. This is especially true for *Gulliver's Travels* (16% of sentence alignment units contain a semi-compositional construction, which is the highest rate in the English subcorpus). On the other hand, the distribution of semi-compositional constructions in the literary texts of the Szeged Treebank is similar to that of the English data (11.6% and 10.4%), which might indicate that the difference between the texts in the SzegedParalellFX is due to the novels chosen and not to more general differences between the two languages. It must also be admitted that *Gulliver's Travels* was published in 1726, thus it reflects the early 18th century language use, which might also be a reason for the difference between English and Hungarian literary texts. However, it might be too hasty to conclude that the English language of that period contained more semi-compositional constructions than contemporary English – more substantial research in historical linguistics is needed to investigate this issue.

When examining the matching of semi-compositional constructions across languages, it can be found that in newspaper texts and texts about 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 applies one). Nevertheless, this tendency does not hold for literary texts. First, English literary texts contain much more semi-compositional 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 semi-compositional construction is not a construction (or not even a verb), for instance, in the following example from *Gulliver's Travels*:

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

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

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 semi-compositional 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 as in this sentence from a magazine:

(3.9) I don't usually moan or **make any special requests**.

Nem	*vagyok*	*nyűgös,*	***nincsenek***	*extra*	*kívánságaim.*
not	be-1SG	moody	have.not-1SG	special	request-PL

A special case of the above is when a verbal counterpart – typically derived from the same root as the nominal component of the construction (see Chapter 5) – occurs in the other language (the example comes from a text on the European Union):

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

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

A final interesting fact is that English passive constructions are frequently paired with semi-compositional constructions including the verb *kerül* 'get', e.g. in this sentence from a magazine:

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

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

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

3.8 Summary of results

In this chapter, the methods of corpus building for data collection were presented. Results of this chapter include:

- the corpora Szeged Treebank, SzegedParalell and Wiki50 were annotated for semi-compositional constructions;
- from these corpora, semi-compositional constructions were collected and lemmatized;
- a database was produced that contains 1524 semi-compositional constructions in Hungarian and 827 in English.

The annotated corpora 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, applied and computational linguistics. In the following chapters, this dataset will serve as the basis for all analyses.

Part II

Theoretical questions

Chapter 4

The status of bare noun + verb constructions

4.1 Introduction

In this chapter, the status of bare noun + verb constructions is discussed. First, features of semi-compositional constructions are compared to those of idioms on the one hand and to those of productive constructions on the other hand. Second, different subtypes of semi-compositional constructions are characterized as well.

If a verb is preceded by a bare noun, this may yield a productive construction, an idiom or a semi-compositional construction. If these constructions are examined in more detail, it is revealed that they do not behave similarly, which leads to their belonging to different groups. Thus, in this chapter, the following research question is put under investigation (see Chapter 1):

- What is the relation of semi-compositional constructions, productive constructions and idioms?

In 4.2 previous research on bare noun + verb constructions is presented (with main focus on Hungarian data), then in 4.3 possible relations between the verb and its arguments are discussed. In 4.4 and 4.5, a test battery is presented which can serve as a base for systematic categorization of these constructions and finally in 4.6 the picture provided by this test battery is analyzed.

4.2 Related work on bare noun + verb constructions

Bare noun + verb constructions have been analyzed in earlier literature as well. In this section, earlier classification of bare noun + verb constructions is presented, which is later contrasted to our grouping.

Komlósy (1992) classifies them into four categories. The first group contains idioms or idiom-like constructions, with the following characteristics:

- They contain more than one morphologically whole word.
- Their words occur in a syntactic position that corresponds to their morphological structure.
- Relying on general semantic rules, their meaning cannot be computed from meanings that their parts can have outside the construction.

The following phrases can be considered as idioms:

(4.1) *fűbe* *harap*
grass-ILL bites
'to die'

tüzet *okád*
fire-ACC vomits
'to shout'

tőrbe *csal*
dagger-ILL entices
'to deceit'

csütörtököt *mond*
Thursday-ACC says
'to fail to work'

The second group includes constructions where the verb has a common noun argument. The meaning of the unit formed by the verb and its common noun argument is not compositional for it is used only in a narrower sense than the compositionally derived one. However, the construction cannot be considered as an idiom since the noun preserves its original meaning and the core meaning of the verb also plays an important role in computing the meaning of the construction. Phrases like

(4.2) *fát* *vág*
wood-ACC cuts
'to chop wood'

kórházba *visz*
hospital-ILL takes
'to take to hospital'

moziba *megy*
cinema-ILL goes
'to go to the cinema'

iskolába *jár*
school-ILL goes
'to attend school'

intézetbe *küld*
school-ILL sends
'to send to school'

belong to this group. In Komlósy's examples, the meaning of the construction differs systematically from the meaning computed from the meaning of its parts: the construction refers to reaching some conventionalized goal.

Idiom-like constructions form the third group:

(4.3) *fejbe* *csap*
head-ILL hits
'to hit on the head'

orrba *vág*
nose-ILL punches
'to punch in the nose'

vállon *csíp*
shoulder-SUP bites
'to bite in the shoulder'

hason *szúr*
stomach-SUP stabs
'to stab in the stomach'

hátba *vág*
back-ILL punches
'to hit on the back'

Semantic restrictions hold for both the verbal and the nominal components of these phrases: in the above examples, the noun refers to a part of the body and the verb denotes a physical contact. Within the construction, both components preserve their core meaning (or one of their core meanings).

The fourth group contains set phrases the semantic head of which is the noun while the verb is responsible only for the construction being a verbal element as a whole. Expressions of this type are:

(4.4) *alkalom* *nyílik* *vmire*
opportunity opens sg-SUB
'an opportunity opens for sg'

lehetőség *kínálkozik* *vmire*
possibility offers sg-SUB
'a possibility emerges for sg'

módot *ad* *vmire*
way-ACC gives sg-SUB
'to provide an opportunity for sg'

okot *ad* *vmire*
reason-ACC gives sg-SUB
'to give reason for sg'

Kiefer (1990 1991) pays special attention to units formed by a determinerless bare noun in the accusative case and a verb when analyzing noun incorporation in Hungarian. These units form two groups according to him. First, productive constructions such as

(4.5) *újságot* *olvas*
newspaper-ACC reads
'to read a newspaper'

levelet *ír*
letter-ACC writes
'to write a letter'

zenét *hallgat*
music-ACC listens
'to listen to music'

házat *épít*
house-ACC builds
'to build a house'

can be paired with a corresponding free construction with an accusative noun with a definite article and the verb in the objective conjugation, for instance:

(4.6) *olvassa* *az* *újságot*
read-3SGOBJ the newspaper-ACC
'he is reading a newspaper'

Second, idiomatic expressions can also be divided into two subgroups according to their ability to have a corresponding free construction. Some idiomatic expressions cannot have such construction: e.g. in the case of

(4.7) *igazat* *mond*
truth-ACC tells
'to tell the truth'

there is no

(4.8) **mondja* *az* *igazat*
tell-3SGOBJ the truth-ACC
'he is telling the truth'

Note that the English equivalent of the construction contains the definite article, however, its usage is ungrammatical in Hungarian.

According to Kiefer, other phrases of this type include:[1]

(4.9) *csipkét* *ver*
lace-ACC hits
'to make lace'

gazdát *cserél*
owner-ACC changes
'to change hand'

[1]In our view, *veri a csipkét* (hit-3SG-OBJ the lace) 'she is making lace' is grammatical, thus, it belongs to the second subgroup.

szerepet *cserél*
role-ACC changes
'to change roles'

tapasztalatot *cserél*
experience-ACC changes
'to change experience'

erőt *vesz*
strength-ACC takes
'to overcome'

elégtételt *vesz*
revenge-ACC takes
'to take revenge'

The second subgroup contains phrases that have a corresponding free structure, however, only with a literal meaning.[2] Some examples:

(4.10) *gyökeret* *ver*
root-ACC beats
'to strike root'

gyereket *vár*
child-ACC expects
'to expect a child'

férjet *keres*
husband-ACC searches
'look for a husband'

csipkét *ver*
lace-ACC hits
'to make lace'

In each of these constructions the verb incorporates the noun and the construction denotes a typical or institutionalized action. However, the corresponding free construction can only have a literal meaning, e.g.:

[2]In our view, *csipkét ver* and *veri a csipkét* can be both used in the institutionalized sense (i.e. 'to make lace').

(4.11) *veri* *a* *gyökeret*
beats-3SG-OBJ the root
'to beat a root'

Kiefer and Ladányi (2000) examine nominal verbal modifiers. According to their analysis, the noun cannot be modified (for instance, it can have no adjectival modifier) while the modified object is a dependent of the verb (as an object). Nouns do not have references on their own. The complex verb formed by the noun and the verb gets easily lexicalized, that is, its meaning is less predictable: in the constructions

(4.12) *ajánlatot* *tesz*
offer-ACC does
'to make an offer'

vizsgát *tesz*
exam-ACC does
'to take an exam'

esküt *tesz*
oath-ACC does
'to make an oath'

the verb *tesz* 'do' occurs in the same form, however, the three actions differ from each other. Similarly, the constructions

(4.13) *gazdát* *cserél*
owner-ACC changes
'to change hand'

szerepet *cserél*
role-ACC changes
'to change roles'

tapasztalatot *cserél*
experience-ACC changes
'to change experience'

refer to three kinds of "change". Finally, if the noun gets a determiner, the meaning of the construction changes, what is more, the verb gets a prefix many times:

(4.14) *??teszi* *az* *ajánlatot*
do-3SGOBJ the offer-ACC
'he is making an offer'

cf.

megtette *az* *ajánlatot*
PREVERB-do-PAST-3SGOBJ the offer-ACC
'he made the offer'

Kiefer (2003) analyzes two types of verbal modifiers, namely, bare nominal verbal modifiers and verbal prefixes. According to him, bare nominal verbal modifiers are always arguments of the verb, e.g. in

(4.15) *újságot* *olvas*
newspaper-ACC reads
'to read a newspaper'

újságot is an argument of the verb *olvas* and they can never be a referential expression, that is, they cannot refer to a specific object. For this reason, in the example

(4.16) *német* *újságot* *olvas*
German newspaper-ACC reads
'he is reading a German newspaper'

újságot cannot be a verbal modifier. Verbs with determinerless objects cannot be nominalized as a construction hence the form

(4.17) **újságot* *olvasás*
newspaper-ACC reading
'reading a newspaper'

is not attested since the (deverbal) noun cannot give an accusative case – in contrast to this, the form

(4.18) *újságolvasás*
newspaper.reading
'reading of a newspaper'

is grammatical. In some cases, however, nominalization is possible: first, the verb is nominalized then it takes the noun as its argument:

(4.19) *(moziba)* *jár*
cinema-ILL goes
'to go to the cinema'

járás
going

moziba *járás*
cinema-ILL going
'going to the cinema'

(4.20) *(vízbe)* *ugrik*
water-ILL jumps
'to jump into the water'

ugrás
jump

vízbe *ugrás*
water-ILL jump
'jump into the water'

Viszket (2004) classifies bare nouns occurring in subject or object position. In her first group, lexicalized constructions occur where the meaning of the construction is not totally compositional:

(4.21) *gondot* *visel*
care-ACC takes
'to take care'

bakot *lő*
he-goat-ACC shoots
'to blunder'

aranyat *ér*
gold-ACC is.worth
'it is worth gold'

köszönetet *mond*
thanks-ACC says
'to say thank you'

Test	**Vague action verb**	**True light verb**
Passivization	YES	NO
WH-movement	YES	NO
Pronominalization	YES	NO
Indefinite NP	NO	YES
NP stem is identical to a verb	NO	YES
Differences compared to verbal counterpart	NO	YES
Examples	make an inspection	give a groan

Table 4.1: True light verbs and vague action verbs in English

Other examples involve compositional constructions where the noun is lexically restricted as in:

(4.22) *vendég* *érkezik*
guest arrives
'there arrives a guest'

fát *vág*
wood-ACC cuts
'to chop wood'

A third group of lexicalized constructions includes examples such as *homok ment a szemébe* (sand go-PAST-3SG the eye-3SGPOSS-ILL) 'sand went into his eyes', meaning that something is somewhere and something must be done because of this.

The second main group of bare NPs can be the arguments of verbs (1) denoting creation, e.g. *tó keletkezett* (lake was.created) 'a lake has been created', (2) having a Beneficiary/Goal argument such as *szarvast lő* (deer-ACC shoots) 'to shoot a deer' and (3) denoting a progressive action (with special intonation pattern) e.g. *levelet ír* (letter-ACC writes) 'to write a letter'. In this group, constructions can include any NPs, that is, NPs without any semantic or lexical restrictions. In her third group, constructions with habitual meaning can be found: *inget hord* (shirt-ACC wears) 'to wear a shirt'.

As for English, Kearns (2002) distinguishes between two subtypes of what is traditionally called light verb constructions. True light verb constructions such as *to give a wipe* or *to have a laugh* and vague action verbs such as *to make an agreement* or *to do the ironing* differ in some syntactic and semantic properties and can be separated by various tests, e.g. passivization, WH-movement, pronominalization etc. as shown in Table 4.1.

4.3 On the possible relations between the verb and its arguments

As it was shown above, earlier research on bare common noun + verb constructions is not based on a unified typology although each classification contains several groups of those constructions and they are also characterized by different features. For instance, Komlósy (1992) describes four groups – idioms, verbs with a common noun argument, idiom-like constructions and set phrases –, Kiefer (1990 1991) distinguishes productive and idiomatic constructions, and two subgroups within the latter, Kiefer and Ladányi (2000) analyze the relation of nominal verbal modifier and the verb while Kiefer (2003) examines bare nominal verbal modifiers and their relation to the verb. From all these data it seems unequivocal that bare noun + verb constructions do not form a unified group: different constructions belonging to this group exhibit different semantic and syntactic features and the relation between the noun and the verb also varies.

If the relation between a verb and its nominal arguments is examined in detail, it must be mentioned that participants of a situation denoted by a verb are traditionally divided into two groups: arguments (complements) or adjuncts – the diversity of terms applied for these phenomena is discussed in detail in Mel'čuk (2004a). Gábor and Héja (2006) distinguish complements and adjuncts by using criteria based on compositionality and productivity. Thus, an adjunct must be productive in relation to a verb class defined by a specific metapredicate and its meaning, the meaning of the verb and the meaning of the case suffix must determine the meaning of the constituent in a compositional way. On the other hand, the presence of complements does not depend on verb classes and the case suffix does not have a distinct meaning. In this way, adjuncts can be primarily characterized by compositionality.

According to Kálmán (2006), the traditional distinction of arguments and adjuncts relies on binarity (a participant is either an argument or an adjunct) and autonomy of syntax. However, tests that are to determine the status of the participant are usually semantic by nature, that is, defining arguments and adjuncts cannot be solely based on syntactic criteria. If governors and their governed constituents are understood as a construction, it can be seen that there are many subtypes: constructions differ from each other with regard to their productivity, compositionality, transparency and the cohesion between the two parts of the construction. According to these criteria, constructions can be placed on a scale. The degrees

of opacity and productivity are connected: the more cohesive and opaque a construction is, the more productively it can be used. Concerning the relation of the governor and its governed constituent, it can be either symmetric or asymmetric. If their relation is symmetric, none of them is able to predict the element that signals their relationship. An asymmetric relation holds between a semantically bleached relation marker and the other element, in this case the cohesion of the relation is gradual. Within this framework, argument is an umbrella term used for several cohesive asymmetric relations.

Both Gábor and Héja (2006) and Kálmán (2006) emphasize the role of productivity and opacity when examining verbs and their arguments. Based on these data, it seems advisable to analyze the relation between the components of bare common noun + verb constructions by using these two characteristics. With these characteristics in mind, a test battery is designed and presented in this chapter with which bare common noun + verb constructions can be described in more detail, with a primary focus on Hungarian data, however, it can be easily adapted to other languages. When constructing syntactic and semantic tests, features of bare common noun + verb constructions described in earlier Hungarian literature (Komlósy, 1992; Kiefer, 1990 1991; Kiefer and Ladányi, 2000; Kiefer, 2003; Viszket, 2004), in Kearns (2002) (for English) and in Langer (2004) (for French, English and German) are also considered.

4.4 Tests for classifying bare common noun + verb constructions

In order to categorize bare common noun + verb constructions in Hungarian, the following test battery proves to be useful. Constructions illustrating test results were mostly selected from the examples mentioned in Kiefer (1990 1991), Komlósy (1992) and Viszket (2004). These tests were also used in the annotation of the corpora described in Chapter 3, and basically the same test battery was applied to both Hungarian and English constructions (salient differences are distinctively marked in the description of the respective tests).

Test 1: the test of the WH-word

If we ask a question on the nominal component, can the nominal component function as a grammatical answer? (Compare Kearns's (2002) WH-movement.)

a) Újságot olvas. Mit olvas? Újságot. 'He reads a newspaper. What does he read? (A) newspaper.'

b) Iskolába jár. Hova jár? Iskolába. 'He goes to school. Where does he go? To school.'

c) Ház épül. Mi épül? Ház. (house is.built) 'A house is being built. What is being built? A house.'

d) Csütörtököt mond. Mit mond? *Csütörtököt. (Thursday-ACC says) 'It fails to work. What is he saying? *Thursday.'

e) Lépre csal valakit. Hova csal? *Lépre. (comb-SUB entices somebody) 'He tolls someone. Where does he entice him? *Comb.'

f) Ebcsont beforr[3]. Mi forr be? *Ebcsont. (dog.bone heals) 'The wound inflicted will soon be healed. What will be healed? *Dog bone.'

Test 2: the test of the free corresponding construction

Can the nominal component be added an article in a way that we get a free corresponding construction? A slight change in meaning is tolerable now as in Kiefer (1990 1991). In English, this test examines whether the noun can bear a definite article since in many semi-compositional constructions, an indefinite article can be found (e.g. *take a chance* or *make an effort*).[4]

a) az újságot olvassa 'he is reading the newspaper' (not in the conventionalized sense)

b) az iskolába jár 'he is going to the school' (not in the conventionalized sense)

c) a ház épül 'the house is being built'

d) *a fittyet hányja (the fig-ACC scatter-3SGOBJ) 'to snap someone's finger'

e) *a lépre csal (the comb-SUB entices) 'to toll'

f) *az ebcsont beforr (the dog.bone heals) 'The wound inflicted will soon be healed.'

Test 3: the test of plural

Can the nominal component be pluralized in a way that the construction remains grammatical? (A slight change in meaning is again tolerable.) If the noun happens to be in the plural in the original construction as in *tűkön ül* (needle-PL-SUP sits) 'to be anxious', the

[3]Although in this example the verb contains a prefix, we selected this bare noun + verb combination for testing since there are hardly any instances of idiomatic bare noun + verb constructions where the noun is in the subject position.

[4]Although this test is related to definiteness effect (see e.g. Maleczki (2008)), it is still able to reveal differences between constructions with the same internal structure and to show the degree of lexical fixedness of the construction.

question is whether it can occur in the singular form.

a) újságokat olvas 'he reads newspapers' (not in the conventionalized sense)

b) iskolákba jár 'he goes to (different) schools'

c) házak épülnek 'houses are being built'

d) *csütörtököket mond (Thursdays says) 'to fail to work'

e) *lépekre csal (comb-PL-SUB entices) 'to toll'

f) *ebcsontok beforrnak (dog.bone-PL heal-3PL) 'The wounds inflicted will soon be healed.'

Test 4: the test of negation

Can the nominal component be negated by the construction *egy... sem* 'none of'? In English, the determiner *no* is used in this test.

a) egy újságot sem olvas 'he does not read any of the newspapers'

b) egy iskolába sem jár 'he does not go to any school'

c) egy ház sem épül 'no house is being built'

d) *egy fittyet sem hány (one fig-ACC not scatters) 'he does not snap any of his fingers'

e) *egy lépre sem csal (one comb-SUB not entices) 'not to toll'

f) *egy ebcsont sem forr be' (one dog.bone not heals) 'The wound inflicted will not be healed.'

Test 5: the test of the possessor

Can the possessor be marked on the nominal component (with or without an article)? In English, it is checked whether the nominal component can have a possessive determiner.

a) (az) újságját olvassa 'he is reading his newspaper' (slight change in meaning)

b) (az) iskolájába jár 'he goes to his school' (slight change in meaning)

c) (a) háza épül 'his house is being built'

d) *tüzét okádja / *a tüzét okádja ((the) fire-3SGPOSS-ACC vomits) 'to shout'

e) *lépére csal / *a lépére csal ((the) comb-3SGPOSS-SUB entices) 'to toll'

f) *ebcsontja beforr / *az ebcsontja beforr ((the) dog.bone-3SGPOSS heals) 'The wound inflicted will soon be healed.'

Test 6: the test of the attributive

Can the nominal component be added an attributive in a way that the construction remains grammatical?

a) német újságot olvas 'he reads a German newspaper' (not in the conventionalized sense)

b) német iskolába jár 'he goes to a German school' (not in the conventionalized sense)

c) nagy ház épül 'a big house is being built'

d) *perzselő tüzet okád (burning fire-ACC vomits) 'to shout'

e) *szőke hajba kap (blond hair-ILL grabs)[5] 'to grab blonde hair' (grammatical only in the literal sense)

f) *törékeny ebcsont beforr (fragile dog.bone heals) 'The wound inflicted will soon be healed.'

Test 7: the test of coordination (zeugma)

Can another nominal component be coordinated to the noun in a way that the construction remains grammatical and with the zeugma effect?

a) újságot és könyvet olvas 'he reads a newspaper and a book'

b) iskolába és kórházba jár 'he goes to school and hospital'

c) ház és uszoda épül 'a house and a swimming pool are being built'

d) *kígyót-békát és farkast kiált (*kígyót-békát kiált rá* (snake-ACC-frog-ACC shouts he-SUB 'to call him names' and *farkast kiált* wolf-ACC shouts 'to cry wolf') 'he calls him names and he cries wolf'

e) *tőrbe és lépre csal (*tőrbe csal* dagger-ILL entices 'to deceit' and *lépre csal* comb-SUB entices 'to toll') 'he deceits and he tolls'

f) *ebcsont és rosszcsont beforr (*ebcsont beforr* (dog.bone heals) 'The wound inflicted will soon be healed.' and *rosszcsont* (bad.bone) 'rascal') 'The wound inflicted will soon be healed and rascals will be good.'

Test 8: nominalization of the verb

Can the verb be nominalized with preserving the original form of the noun? (This test is not relevant in English since nouns do not have case suffixes.)

a) *újságot olvasás 'reading of a newspaper'

[5] *hajba kap* (hair-ILL grabs) 'to start arguing'. Our original example of *lépre csal* cannot be modified by an adjective that is why the example has been changed.

b) iskolába járás ‘going to school’

c) (not applicable)

d) *csütörtököt mondás (Thursday-ACC saying) ‘failing to work’

e) lépre csalás (comb-SUB enticing) ‘tolling’

f) (not applicable)

Test 9: nominalization of the construction

Can the construction itself be nominalized? (For grammatical reasons, Tests 8 and 9 cannot be both applied at the same time.) In English, this test examines whether the construction has a gerund form.

a) újságolvasás ‘reading a newspaper’

b) *iskolajárás (school.going) ‘going to school’

c) ?házépülés (house.building) ‘building a house’

d) *csütörtökmondás (Thursday.saying) ‘failing to work’

e) *lépcsalás (comb.enticing) ‘tolling’

f) *ebcsontbeforrás (dog.bone.healing) ‘healing a wound’

Test 10: the test of the participle – 1

Can a participle be derived from the construction with preserving the original form of the noun? (The same applies to this test in English as to Test 8.)

a) újságot olvasó ‘(someone) reading a newspaper’

b) iskolába járó ‘(someone) going to school’

c) épülő ház ‘a house being built’

d) csütörtököt mondó (Thursday-ACC saying) ‘(something) failing to work’

e) lépre csaló (comb-SUB enticing) ‘(someone) tolling’

f) *beforró ebcsont (being.healed dog.bone) ‘a wound that is being healed’

Test 11: the test of the participle – 2

Can a participle be derived from the construction?

a) újságolvasó ‘(someone) reading a newspaper’

b) *iskolajáró (school.going) ‘(someone) going to school’

c) (not applicable)

d) *csütörtökmondó (Thursday.saying) '(something) failing to work'

e) *lépcsaló (comb.enticing) '(someone) tolling'

f) (not applicable)

Test 12: the test of variativity

Can a verb (derived from the same root as the nominal component) substitute the construction? As opposed to Kearns (2002), who uses this test in a slightly different version in English (the nominal component coincides with a verb or not), we allow verbs that are derived from the same root as the verb but not necessarily coincide with the noun (e.g. *decide* and *decision*).

a) *újságozik (the verb *újságol* derived from the root *újság* is lexicalized with the meaning 'to share the news')

b) *iskolázik

c) *házazik (the verbs *házal* and *házasodik* derived from *ház* are lexicalized with the meanings 'to go from house to house and sell goods' and 'to marry', respectively)

d) *csütörtöközik / *csütörtököl

e) *lépezik[6]

f) *ebcsontozik

Test 13: changing the verb

Can a synonym verb substitute the original verb in a way that the construction preserves its meaning?

a) ?levelet alkot 'he is creating a letter' (instead of *levelet ír* 'he is writing a letter')

b) ?iskolába megy 'he goes to school' (acceptable when not in the institutionalized sense)

c) ?ház készül 'a house is being created'

d) *csütörtököt szól (Thursday-ACC tells) 'to fail to work'

e) *lépre hív (comb-SUB call) 'to toll'

f) *ebcsont beheged (dog.bones heals) 'The wound inflicted will soon be healed.'

[6] *Lép* 'comb' is homonymous with the verb *lép* 'to step', however, their meaning is totally different.

Test 14: omitting the verb

When omitting the verb (e.g. in a possessive construction where the original subject/object becomes the possessor), can the original action be reconstructed? In other words, is the meaning of the noun related to the action the construction denotes? Aspectual differences are disregarded now.

a) *Lajos újságja 'Lewis's newspaper' (in the sense 'Lewis is reading a newspaper')

b) *Lajos iskolája 'Lewis's school' (in the sense 'Lewis goes to school')

c) *Lajos háza 'Lewis's house' (in the sense 'Lewis house is being built')

d) *Lajos tüze 'Lewis's fire' (in the sense 'Lewis is shouting')

e) *Lajos lépe 'Lewis's comb' (in the sense 'the enticement of Lewis')

f) *Lajos ebcsontja 'Lewis's dog bone' (in the sense 'Lewis's wound is being healed')

On the basis of the above-mentioned tests, two groups of bare common noun + verb constructions can be separated. As for the examples found in a), b) and c) most of the tests were applicable with a grammatical result. Constructions belonging to this group will be called productive constructions following Kiefer (1990 1991). However, none of the tests provides a grammatical result for the examples in d), e) and f) – except for Test 10 (i.e. one of the tests of the participle). Constructions for which all the tests (except for Test 10) give an ungrammatical result will be called idioms.

Productive constructions and idioms can be unambiguously distinguished on the basis of their grammatical behavior as it is shown in Table 4.2 (YES stands for the grammatical result of the test while NO for the ungrammatical result).

The table illustrates well that each test provides a grammatical result for productive constructions while an ungrammatical one for idioms except for tests in italics where they exhibit the same behavior. Among the examples of Kiefer (1990 1991), Komlósy (1992) and Viszket (2004), the following (4.23) are considered as productive constructions based on the test results too:

(4.23) *újságot* *olvas*
newspaper-ACC reads
'to read a newspaper'

moziba *megy*
cinema-ILL goes
'to go to the cinema'

Test	**Productive construction**	**Idiom**
WH-word	YES	NO
Article	YES	NO
Plural	YES	NO
Negation	YES	NO
Possessor	YES	NO
Attributive	YES	NO
Coordination / zeugma	YES	NO
Nominalization (V)	NO	YES/NO
Nominalization (construction)	YES	NO
Participle – 1	YES	YES
Participle – 2	YES	NO
Variativity	NO	NO
Changing the verb	YES	NO
Omitting the verb	NO	NO
Examples	újságot olvas iskolába jár ház épül	csütörtököt mond lépre csal ebcsont beforr

Table 4.2: Test results for productive constructions and idioms

iskolába *jár*
school-ILL goes
'to attend school'

levelet *ír*
letter-ACC writes
'to write a letter'

zenét *hallgat*
music-ACC listens
'to listen to music'

házat *épít*
house-ACC builds
'to build a house'

ház *épül*
house builds
'a house is being built'

vendég *érkezik*
guest arrives
'there arrives a guest'

fát *vág*
wood-ACC cuts
'to chop wood'

Most of the constructions in this group describe a conventionalized action. The constructions are semantically transparent, their meaning can be easily computed on the basis of the meaning of the verb, the common noun and the case suffix (i.e. their compositionality is high) hence their productivity is also of high degree – this is why they are called productive (Jackendoff, 2010). (On the correspondence between the degree of compositionality and productivity see e.g. Kálmán (2006)).

Some idioms can be seen in (4.24):

(4.24) *tüzet* *okád*
fire-ACC vomits
'to shout'

tőrbe *csal*
dagger-ILL entices
'to deceit'

csütörtököt *mond*
Thursday-ACC says
'to fail to work'

gyökeret *ver*
root-ACC beats
'to strike root'

lépre *csal*
comb-SUB entices
'to toll'

hajba *kap*
hair-ILL grabs
'to start arguing'

bakot *lő*
he-goat-ACC shoots
'to blunder'

ebcsont *beforr*
dog.bone PREVERB-heals
'The wound inflicted will soon be healed.'

Idioms are semantically opaque, relying only on the meaning of the parts of the construction the meaning of the whole construction cannot be computed hence their productivity is of very low degree.

4.5 Further test results

For another group of bare common noun + verb constructions, some of the tests give a grammatical result while other tests give an ungrammatical one. Here follow some illustrative examples for the test results:

Test 1: the test of WH-word

a) Előadást tart. Mit tart? Előadást. (presentation-ACC holds) 'He has a presentation. What does he have? (A) presentation.'

b) Igényt tart. Mit tart? *Igényt. (claim-ACC holds) 'He establishes a claim. What does he establish? (A) claim.'

c) Életre kel. Mire kel? *Életre. (life-SUB raises) 'He revives. What does he do? He revives.'

d) Virágba borul. Mibe borul? *Virágba. (bloom-ILL falls) 'It comes into bloom. Where does it come? (Into) bloom.'

e) Lehetőség adódik. Mi adódik? Lehetőség. (possibility give-PASS-3SG) 'A possibility emerges. What emerges? (A) possibility.'

f) Sor kerül valamire.[7] Mi kerül? *Sor. (turn gets sg-SUB) 'The time has come for sg. What has come? Time.'

Test 2: the test of the free corresponding construction

a) *az előadást tartja / az előadást megtartja / éppen az előadást tartja (the presentation-ACC holds) 'he is having the presentation' (the construction is grammatical with a preverb or in the progressive aspect)

[7]Besides the given one, this expression can have several translations in English, e.g. 'to take place' or 'to be carried out' but it can also be translated as a passive construction (*something is done*).

b) *az igényt tartja (the claim-ACC holds) ‘to establish the claim’

c) *az életre kel (the life-SUB raises) ‘to revive’

d) *a virágba borul (the bloom-ILL falls) ‘to come into bloom’

e) *a lehetőség adódik (the possibility give-PASS-3SG) ‘the possibility emerges’

f) *a sor kerül (the turn gets) ‘the time has come’

Test 3: the test of the plural

a) előadásokat tart (presentation-PL-ACC holds) ‘to have presentations’

b) *igényeket tart (claim-PL-ACC holds) ‘to establish claims’

c) *életekre kel (life-PL-SUB raises) ‘to revive’

d) *virágokba borul (bloom-PL-ILL falls) ‘to come into bloom’

e) lehetőségek adódnak (possibility-PL give-PASS-3PL) ‘possibilities emerge’

f) *sorok kerülnek (turn-PL get-3PL) ‘the time has come’

Test 4: the test of negation

a) egy előadást sem tart (one presentation-ACC not holds) ‘not to have any presentation’

b) *egy gyanút sem fog (one suspect-ACC not takes) ‘not to smell a rat’ [8]

c) *egy életre sem kel (one life-PL-SUB not raises) ‘not to revive’

d) *egy virágba sem borul (one bloom-ILL not falls) ‘not to come into bloom’

e) egy lehetőség sem adódik (one possibility not give-PASS-3SG) ‘no possibility emerges’

f) *egy sor sem kerül (one turn not gets) ‘the time has not come’

Test 5: the test of the possessor

a) (az) előadását tartja ((the) presentation-3SGPOSS-ACC hold-3SGOBJ) ‘to have his presentation’

b) *igényét tartja (claim-3SGPOSS-ACC hold-3SGOBJ) ‘to establish his claim’

c) *életére kel (life-3SGPOSS-SUB raises) ‘to revive’

d) *virágába borul (bloom-3SGPOSS-ILL falls) ‘to come into its bloom’

e) lehetősége adódik (possibility-3SGPOSS give-PASS-3SG) ‘he has the possibility’

f) *sora kerül (turn-3SGPOSS gets) ‘his time has come’ (ungrammatical in Hungarian)

[8] *Igény* is an uncountable noun in the construction that is why the example has been changed.

Test 6: the test of the attributive

a) érdekes előadást tart (interesting presentation-ACC holds) 'to have an interesting presentation'

b) kizárólagos igényt tart (exclusive claim-ACC holds) 'to establish his exclusive claim'

c) új életre kel (new life-SUB raises) 'to come to a new life'

d) *fehér virágba borul (white bloom-ILL falls) 'to come into white bloom'

e) rendkívüli lehetőség adódik (extraordinary possibility give-PASS-3SG) 'an extraordinary possibility emerges'

f) *hirtelen sor kerül (sudden turn gets) 'the time has suddenly come'

Test 7: the test of coordination

a) *előadást és igényt tart (presentation-ACC and claim-ACC holds) 'to have a presentation and to make a claim'

b) *bérbe és figyelembe vesz (*bérbe vesz* rent-ILL takes 'to hire' and *figyelembe vesz* consideration-ILL takes 'to take into consideration') 'to hire and take into consideration

c) ?lehetőség és alkalom adódik (*lehetőség adódik* possibility give-PASS-3SG 'a possibility emerges' and *alkalom adódik* opportunity give-PASS-3SG 'an opportunity emerges') 'an opportunity and a possibility emerge'

Test 8: nominalization of the verb

a) *előadást tartás (presentation-ACC holding) 'holding of a presentation'

b) *igényt tartás (claim-ACC holding) 'establishment of a claim'

c) életre kelés (life-SUB raising) 'reviving'

d) virágba borulás (bloom-ILL falling) 'coming into bloom'

e) (not applicable)

f) (not applicable)

Test 9: nominalization of the construction

a) előadástartás (presentation.holding) 'holding a presentation'

b) *igénytartás (claim.holding) 'establishing a claim'

c) *életkelés (life.raising) 'reviving'

d) *virágborulás (bloom.falling)'coming into bloom'

e) *lehetőségnyílás (possibility.opening) 'emerging of a possibility'

f) *sorkerülés (turn.getting) ‘coming of the time’

Test 10: the test of the participle – 1

a) előadást tartó (presentation-ACC holding) ‘(someone) holding a presentation’

b) igényt tartó (claim-ACC holding) ‘(someone) establishing a claim’

c) életre kelő (life-SUB raising) ‘(someone) reviving’

d) virágba boruló (bloom-ILL falling) ‘(something) coming into bloom’

e) (meg)nyíló lehetőség (opening possibility) ‘emerging possibility’ (it is more acceptable with a preverb)

f) *kerülő sor (getting turn) ‘coming time’

Test 11: the test of the participle – 2

a) ajánlattevő (offer.making) ‘(someone) making an offer’[9]

b) *igénytartó (claim.holding) ‘(someone) establishing a claim’

c) *életkelő (life.raising) ‘(someone) reviving’

d) *virágboruló (bloom.falling) ‘(something) coming into bloom’

e) (not applicable)

f) (not applicable)

Test 12: the test of variativity

a) előad ‘to present’

b) igényel ‘to claim’

c) éled ‘to revive’

d) virágzik ‘to blossom’

e) lehetséges[10] ‘possible’

f) megtörténik ‘to happen’

Test 13: changing the verb

a) ajánlatot ad / tesz (rule-ACC gives / does) ‘to make an offer’

b) igényt tart / támaszt (claim-ACC holds / supports) ‘to establish a claim’

[9]The participle form of our original example of *előadást tart* (i.e. *előadástartó*) is lexically blocked by the nominalized participle derived from the verb from which the nominal component is derived , i.e. *előadó* ‘speaker’.

[10]In this case, it is not a verb but an adjective that conveys the same meaning.

c) életre kel / ?kap (life-SUB raises / gets) ‘to revive’

d) virágba borul / *dől (bloom-ILL falls) ‘to come into bloom’

e) lehetőség nyílik / adódik / kínálkozik (possibility opens / offer-PASS-3SG / give-PASS-3SG) ‘a possibility emerges’

f) sor kerül / ??jut (turn gets) ‘the time has come’

Test 14: omitting the verb

a) Lajos előadása ‘Lewis’s presentation’ (in the sense of ‘Lewis is having a presentation’)

b) Lajos igénye ‘Lewis’s claim’ (in the sense of ‘Lewis establishes a claim’)

c) Lajos élete ‘Lewis’s life’ (differences concerning Aktionsart are now discarded)

d) a fa virága ‘the bloom of the tree’

e) Lajos lehetősége ‘Lewis’s possibility’ (in the sense of ‘Lewis is having a possibility’)

f) *a szavazás sora ‘time for voting’ (in the sense ‘the time has come for voting’)

The test results show that this group is less uniform than productive constructions or idioms: the applicability of tests varies and it is usually constructions in groups a), c) and e) that behave similarly as opposed to constructions in b), d) and f). This is the reason why this group of bare common noun + verb constructions are called *semi-compositional constructions* since they can be placed in the middle of a scale at the extremities of which productive constructions and idioms are situated.

Semi-compositional constructions can be divided into subgroups depending on their test results. Constructions in the first subgroup are more similar to productive constructions since they share more features with productive constructions than with idioms. This group contains:

(4.25) *előadást* *tart*
presentation-ACC holds
‘to have a presentation’

parancsot *ad*
order-ACC gives
‘to give an order’

döntést *hoz*
decision-ACC brings
'to make a decision'

bejelentést *tesz*
announcement-ACC does
'to make an announcement'

intézkedést *tesz*
measure-ACC does
'to take measures'

órát *ad*
lesson-ACC gives
'to give a lesson'

órát *vesz*
lesson-ACC takes
'to take lessons'

módot *ad*
way-ACC gives
'to provide an opportunity'

életre *kel*
life-SUB raises
'to revive'

lehetőség *nyílik*
possibility opens
'a possibility emerges'

In this group, the meaning of the constructions can be relatively easily computed on the basis of the meaning of their parts, however, it must be highlighted that the noun plays much greater role in this than the verb – the latter being responsible only for the verbal nature of the construction, which is proved by the fact that the verb can be replaced in the construction and the action can be reconstructed even if omitting the verb (see Tests 13 and 14).

Constructions belonging to the second subgroup are considered to be more closely related to idioms based on the test results. Such constructions are the following:

(4.26) *virágba* *borul*
bloom-ILL falls
'to come into bloom'

igénybe *vesz*
demand-ILL takes
'to take up'

igényt *tart*
claim-ACC holds
'to claim'

áruba *bocsát*
ware-ILL sends
'to start to sell'

gyanút *fog*
suspicion-ACC takes
'to smell a rat'

tetten *ér*
act-SUP catches
'to catch in the act'

csapra *ver*
tap-SUB beats
'to tap the barrel'

figyelembe *vesz*
consideration-ILL takes
'to take into consideration'

sor *kerül*
turn gets
'the time has come'

The productivity and compositionality of these constructions are less high than those of the first subgroup: e.g. the change of the verb is not possible (Test 13), which signals the greater coherence of the components.

Semi-compositional constructions that occurred at least 3 times in the corpora (438 in number, altogether 6523 occurrences in the corpus and they account for about 83°% of all

Type	**# of types**	%	**# of tokens**	**% in the corpus**
Productive-like FX	262	59.8%	3544	45.2%
Idiom-like FX	176	40.2%	2979	38%

Table 4.3: Classes of semi-compositional constructions

occurrences) were classified as more productive-like or more idiom-like semi-compositional constructions. Table 4.3 shows their frequency according to the above classes and with respect to the total number of semi-compositional constructions.

Table 4.4 shows the test results for productive constructions, idioms and semi-compositional constructions altogether. It can be also seen that the more similar a construction is to idioms, the fewer tests give a grammatical result. However, it is very important to emphasize that these (sub)groups cannot be distinguished sharply from each other: there is no way to classify semi-compositional constructions into unambiguously defined groups. The subgroups should be understood as fuzzy sets among which the boundaries are uncertain. In this way, grouping cannot be absolute either: it cannot be determined which tests should provide a grammatical result in order to classify the given construction into a specific group. It can, however, be definitely claimed that the more tests give grammatical results, the more productive and transparent the construction is. The position of a semi-compositional construction is relative on the scale between productive constructions and idioms: it can be placed only compared to the two extremities of the scale. Thus, the construction *döntést hoz* 'to take a decision' is more productive than *igényt tart* 'to claim' since the former shares more test results with productive constructions than with idioms. This scale is applicable to constructions where the noun occurs in (1) the accusative case (object); (2) an oblique case and (3) the nominative case (subject) as illustrated in the last rows of Table 4.4.

The tests indicate a continuum at the ends of which idioms and productive constructions are situated and in the middle of the scale, semi-compositional constructions can be found, which also can be divided into two more or less well-defined subgroups. However, test results do not only indicate a ternary opposition among productive constructions, semi-compositional constructions and idioms but a binary opposition as well between productive constructions and idioms on the one hand and semi-compositional constructions on the other hand. Thus, productive constructions and idioms can be contrasted with the help of two tests that provide ungrammatical results for idioms and productive constructions, however,

Test	**Productive**	**FX**		**Idiom**
		productive-like	**idiom-like**	
WH-word	YES	YES	NO	NO
Article	YES	YES	NO	NO
Plural	YES	YES	NO	NO
Negation	YES	YES	NO	NO
Possessor	YES	YES	NO	NO
Attributive	YES	YES	NO	NO
Coordination	YES	NO	NO	NO
Nominalization (V)	NO	NO	YES	NO
Nominalization (FX)	YES	YES	NO	NO
Participle – 1	YES	YES	YES	YES
Participle – 2	YES	YES	NO	NO
Variativity	NO	YES	YES	NO
Changing the verb	YES	YES	NO	NO
Omitting the verb	NO	YES	YES	NO
Examples	(1) újságot olvas (2) iskolába jár (3) ház épül	előadást tart életre kel lehetőség adódik	igényt tart virágba borul sor kerül	csütörtököt mond lépre csal ebcsont beforr

Table 4.4: Test results for semi-compositional constructions

grammatical results for semi-compositional constructions: these are the tests of variativity (Test 12) and omission of the verb (Test 14). If either of these two tests provides a grammatical result, it is a sufficient (although not necessary) condition for considering the given construction as semi-compositional. On the other hand, productive constructions, semi-compositional constructions and idioms can also be opposed by exploiting all tests in the test battery. Within the category of semi-compositional constructions, constructions can also be classified as being more similar to productive constructions or idioms, or being in the middle of the continuum.

Constructions can also be contrasted depending on their productivity and compositionality: while productive constructions are totally productive (and compositional), this is not true for semi-compositional constructions and idioms. This opposition can also be illustrated with the notion of semantically idiosyncratic multiword expressions as discussed in Chapter 2 (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002; Siepmann, 2005; Siepmann, 2006; Guenthner and Blanco, 2004). In this sense, semantically idiosyncratic multiword expressions are contrasted to productive constructions.

Our scale is similar to that of Jackendoff (2010), on which constructions are placed on the basis of their productivity: productive, semi-productive and non-productive constructions

are also listed. He also argues that syntax and semantics are not parallel, thus, there is no homomorphism between the two levels. Productivity and compositionality are in this way interrelated, suggesting that semi-compositional constructions are semi-productive as well, thus, new semi-compositional constructions can be created in the language by obeying certain semantic constraints (see Chapter 7).

One of the tests (namely, Test 2) examines whether the bare noun + verb construction has a corresponding free structure and Kiefer (1990 1991) and Viszket (2004) also mention this test as an indicator of the construction belonging to certain classes. The internal structure of this free structure is verb + definite article + noun (e.g. *tartja az előadást* (hold-3SGOBJ the presentation-ACC) 'he is having a presentation' – this construction is progressive in Hungarian). However, this structure is unrestricted in Hungarian: each transitive verb can occur in this syntactic pattern, even those that cannot occur with a bare noun in preverbal position (except for pragmatic reasons, cf. Viszket (2004)): *mossa a lámpát* (wash-3SGOBJ the lamp-ACC) 'he is washing the lamp' is acceptable while **lámpát mos* (lamp-ACC washes) 'to wash a lamp' is not. In this way, totally productive and compositional constructions are those that are called "corresponding free structures", i.e. follow the verb + definite article + noun pattern without any restrictions on the semantic content of their parts and their meaning is totally compositional[11]. The group of bare noun + verb constructions called productive in this thesis is still productive but to a lesser degree than free structures for there are certain restrictions on their meaning (e.g. they denote institutionalized actions, cf. Kiefer (1990 1991)).

As for English, the same tests can be applied and in this way, a similar scale can be sketched for verb + noun constructions. However, it is interesting to note that Hungarian constructions and their English equivalents can occupy different positions on the scale, for instance:

(4.27) *gyanút* *fog*
suspicion-ACC takes
'to smell a rat'

where the English construction is an idiom while the Hungarian one is a semi-compositional construction (being more similar to idioms).

[11]However, there are ambiguous examples which can be understood literally and idiomatically such as *tartja a hátát* (hold-3SGOBJ the back-3SGPOSS-ACC) 'to hold his back' or 'to take responsibility' and *bedobja a törölközőt* (PREVERB-throw-3SGOBJ the towel-ACC) 'to throw the towel into sg' or 'to give up'. The point here is that these constructions can have a literal meaning too.

4.6 Comparing the earlier classifications to the groups formed by the test results

The classification found in Kiefer (1990 1991) is interpreted now in this way: productive constructions are basically understood in the same way in this thesis as well and this category embraces the same constructions. Within Kiefer's idiomatic group, constructions without a corresponding free construction are considered as idioms according to the test results as well. However, constructions with a corresponding free construction do not behave as idioms in all aspects: Tests 6, 9 and 11 all yield a grammatical result, which is rather a characteristics of productive constructions. Thus, it seems that it is justifiable to make a distinction between the two groups, which can be supported by the test results.

Although productive constructions in Kiefer (1990 1991) can be considered as instances of incorporation, semi-compositional constructions differ from them in several aspects. First, they are not institutionalized actions: while *iskolába jár* (school-ILL goes) 'to go to school' means that someone regularly goes to school in order to study, *tanácsot ad* (advice-ACC gives) 'to give advice' can be used in every situation where a piece of advice is given, not necessarily by an official counselor. Second, as opposed to actions with incorporated nouns that are progressive (Kiefer, 2007), semi-compositional constructions can be perfective and they can express Aktionsart as well (see Chapter 5). Thus, semi-compositional constructions are not instances of incorporation: they are situated on the same scale of noun + verb constructions, however, they do not occupy the same position.

Constructions with a bare common noun argument found in Komlósy (1992) are classified as productive constructions according to the tests discussed in this thesis, however, there can be differences among productive constructions as well since not all of the tests give a grammatical result for each example. Set phrases are considered as semi-compositional constructions (either belonging to the subgroup similar to productive constructions (*módot ad* 'to offer a possibility') or to the semi-compositional subgroup situated in the middle of the scale (*lehetőség kínálkozik* 'a possibility opens'). His idiom-like expressions and idioms exhibit an interesting picture since they do not behave uniformly with regard to the test results. There are constructions for which it is possible to nominalize the construction or the verb (*hátbavágás* 'hitting on the back', *hason szúrás* 'stabbing in the stomach' from his idiom-like expressions or *tűzokádás* fire.vomiting 'shouting', *számonkérés* account-SUP.asking 'calling

to account' etc. from his idioms) but for others it is not (**csütörtökmondás* Thursday.saying, **gyökérverés* root.beating – they are considered as idioms in his classification). This might suggest that idioms do not form a uniform group, in other words, idiomacity is also gradual: some constructions are more idiom-like than others. As productive constructions behave similarly, it can be stated that neither idioms nor productive constructions can be seen as a uniform class since there are more typical (i.e. more tests yield the result expected for the class) and less typical (i.e. less tests yield the expected result) instances belonging to those categories.

Viszket's (2004) first group, that is, the one including lexicalized constructions includes examples that are classified as productive, semi-compositional and idiomatic in this thesis. She considers incorporated constructions as being lexicalized as well, that is why constructions such as *fát vág* (wood-ACC cuts) 'to chop wood' are listed here although she separates them from not totally compositional constructions. Thus, her classification is similar to our distinction of semantically idiosyncratic multiword expressions and productive constructions.

With the help of the test battery, bare common noun + verb constructions can be placed on a scale where (totally) productive constructions form one extremity of the scale, idioms (in all aspects) form the other end of the scale and in the middle there are constructions that behave similarly to productive constructions in several aspects but in other aspects they are rather like idioms. Kearns's two groups (Kearns, 2002) can also be determined on the basis of this scale: her true light verb constructions are semi-compositional constructions that behave similarly to idioms, whereas her vague action verbs are semi-compositional constructions that are more similar to productive constructions.

On the scale, a ternary opposition can be observed among productive constructions, semi-compositional constructions and idioms. Binary oppositions are also present: (1) semi-compositional constructions differ in several aspects from productive constructions and idioms, (2) semi-compositional constructions and idioms behave similarly regarding compositionality, i.e. they are not totally compositional as opposed to productive constructions and (3) within semi-compositional constructions there are constructions that are more similar to productive constructions or to idioms.

4.7 Summary of results

In this chapter, bare common noun + verb constructions were classified according to their compositionality and productivity. The following points were highlighted:

- the continuum of bare common noun + verb constructions can be characterized by the parameters of compositionality and productivity;
- constructions are to be placed in the continuum by applying a test battery: the more tests yield a grammatical result, the more compositional and the more productive the construction is;
- constructions can be divided into three main groups – productive constructions, semi-compositional constructions and idioms – with no sharp boundary between them;
- semi-compositional constructions can be further divided into subgroups, however, there is no sharp and distinct boundary between them;
- semi-compositional constructions can be opposed to productive constructions and idioms on the basis of variativity and omitting the verb;
- semi-compositional constructions and idioms – as opposed to productive constructions – are instances of semantically idiosyncratic multiword expressions;
- the test results also indicate that the group of idioms is more diverse than it was earlier believed: certain constructions have more idiom-like features than other constructions traditionally seen as idioms.

The results of this chapter suggest that the traditional classification of bare common noun + verb constructions can be preserved only with the restriction that the groups are not absolute but relative. In other words, it is not the either-or dichotomy that determines the place of the construction on the scale but it is rather a question of degree and scalability.

Chapter 5

Verbal counterparts of semi-compositional constructions

5.1 Introduction

In this chapter, the relationship of semi-compositional constructions and their verbal counterparts is to be examined in detail. Special attention is paid to examples in the Hungarian and English languages.

Most semi-compositional constructions have a verbal counterpart that originates from the same stem as the nominal component (and in Hungarian, it can have a preverb as well), e.g.

(5.1) *esküt* *tesz*
oath-ACC makes
'to make an oath'

megesküszik

'to swear'

(5.2) *döntést* *hoz*
decision-ACC brings
'to take a decision'

dönt

'to decide'

(5.3) *to have a walk*

to walk

(5.4) *to make an offer*

to offer

However, the construction and the verb do not always share all of their characteristics, thus, sometimes they cannot be considered as equivalents. The research questions to be answered in this chapter are as follows:

- To what extent do the constructions and their verbal counterparts correspond to each other within one language?
- To what extent does a Hungarian construction or verbal counterpart correspond to its English equivalent?

The structure of this chapter is the following. First, the relation of semi-compositional constructions and their verbal counterparts is presented. Then syntactic alternations observed between the constructions and their verbal counterparts are analyzed through English and Hungarian examples and differences in aspect and Aktionsart are also discussed. Intralingual and interlingual differences are shown as well: cases when an English construction and its Hungarian translation cannot be considered equivalents are analyzed. The chapter concludes with a summary of results achieved.

5.2 The relation of semi-compositional constructions and verbal counterparts

In this section, earlier literature on the acceptability of semi-compositional constructions are shortly presented, which is followed by a general characterization of the relation between semi-compositional constructions and their verbal counterparts.

5.2.1 Views on the acceptability of semi-compositional constructions

Some of the phrases used for semi-compositional constructions in (earlier) literature create the impression that these constructions unnecessarily extend the wording of the thought. Some of the Hungarian terminology that suggest this view are offered here: *körülíró szerkezetek* (periphrastic constructions) in Sziklai (1986), *leíró kifejezések* (descriptive expressions) in Dobos (1991), and, mainly, the somewhat pejorative term *terpeszkedő szerkezetek*

(sprawling constructions), which occurs in the Hungarian Purists' Dictionary (Grétsy and Kemény, 1996) and in recent specialized articles as well, for instance, Heltai and Gósy (2005) focus on the effects of sprawling constructions to linguistic processing. What is more, the Purists' dictionary (Grétsy and Kemény, 1996) defines "sprawling constructions" as the connection between abstracts nouns and a semantically bleached verb, which are considered to be incorrect: their usage reflects the lack of responsibility and the schematic thinking of the speaker. According to the authors, sprawling constructions are often vague and insincere hence they undermine mutual trust in conversation. The comparison of semi-compositional constructions and their verbal counterpart based on language data may provide well-founded arguments for or against this approach.

5.2.2 Semi-compositional constructions vs. verbal counterparts

Most semi-compositional constructions have a verbal[1] counterpart that originates from the same root as the nominal component[2], e.g.

(5.5) *döntést* *hoz*
decision-ACC brings
'to take a decision'

dönt

'to decide'

(5.6) *letartóztatást* *foganatosít*
arrest-ACC carries.into.effect
'to arrest'

letartóztat

'to arrest'

(5.7) *felelősséget* *vállal*
responsibility-ACC undertakes
'to accept responsibility'

felel (vmiért)

'be responsible (for sg)'

[1]In certain cases, the corresponding counterpart is not a verb but an adjective as in *lehetőség nyílik* 'a possibility emerges' – *lehetséges* 'possible'. However, these cases do not seem to be frequent.

[2]Kearns (1998) considers the verbal counterpart of a semi-compositional construction as a portmanteau or incorporated form of the latter whereas the construction itself is a more basic form.

(5.8) *to take into consideration*
to consider

(5.9) *to make a decision*
to decide

(5.10) *to give an answer*
to answer

Thus, their relationship can be seen as an instance of variance and not synonymy: while variants exhibit formal similarity, synonyms completely differ from each other with regard to their form (Lőrincz, 2004). However, sometimes there is no morphological connection between the semi-compositional construction and a verb with exactly the same meaning, which cases can be considered as an instance of synonymy:

(5.11) *kihirdet*

nyilvánosságra *hoz*
publicity-SUB brings
'to publish'

It is partly because of this correspondence that purists had a low opinion of such constructions as they were considered inaccurate and avoidable (Dobos, 2001; B. Kovács, 2000; B. Kovács, 1999; Grétsy and Kemény, 1996). In certain cases this view is justifiable since e.g. the following construction sounds strange from a stylistic point of view, the accepted Hungarian form being *bevásárol*:

(5.12) *?bevásárlást* *eszközöl*
shopping-ACC carries.out
'to do shopping'

On the other hand, there are examples when only the semi-compositional construction is acceptable. First, not every construction has a verbal counterpart: in Hungarian, there is no **házkutat*, only *házkutatást tart* (search.of.premises-ACC holds) 'to conduct search of premises' or in English there is no **to waygive*, only *to give way*. Second, the syntactic structure of the sentence may also play an important role: if the nominal component is modified, it cannot usually be substituted by an adverb in the sentence with the verbal counterpart: *to give useful advice – *to advise usefully*. Third, in some cases the meaning of

the verb derived from the same root as the nominal component differs from the meaning of the semi-compositional construction hence they cannot be paired. For instance, none of the verbs derived from the noun *kéz* 'hand' (*kezel* 'to treat, to shake hands', *kezez* 'to touch the ball with hand (in sports)') shares its meaning with semi-compositional constructions containing the nominal component *kéz* 'hand':

(5.13) *kézbe* *vesz*
hand-ILL takes
'to take in hand'

kézbe *ad*
hand-ILL gives
'to give into sy's hand'

kézben *tart*
hand-INE holds
'to handle, to control'

Similarly, in English *to give a present* and *to present* may describe two different actions as only one of the meanings of *to present* coincides with the meaning of the semi-compositional construction. However, the semi-compositional constructions to be analyzed in this chapter do have a verbal counterpart that is morphologically related to the construction and their basic meaning is the same.

5.3 Differences between semi-compositional constructions and their verbal counterparts

In this section, syntactic alternations between the semi-compositional construction and its verbal counterpart are discussed.

5.3.1 Syntactic alternations

Syntactic alternation or argument structure alternation have been studied for a long time in linguistics, especially in the English language (Fillmore, 1968; Fillmore, 1977; Levin, 1993; Beavers, 2006). The definition of syntactic alternation based on Kiefer (2007) is provided

here: if a sentence with the syntactic structure A is grammatical with a given verb, then the sentence with the syntactic structure B is also grammatical with the same form of the verb. Here is an example:

(5.14) *Béla bólintott.*
Béla nod-PAST3SG
'Béla nodded.'

(5.15) *Béla bólintott a fejével.*
Béla nod-PAST3SG the head-3SGPOSS-INS
'Béla nodded with his head.'

In (5.15) there appears a noun in the instrumental case next to the verb unchanged.

Syntactic alternation can be understood in a broader sense as well: the two verbs in the sentences might slightly differ, however, there must be some morphological connection between them (e.g. to one of them a preverb or a verbal suffix is added). Some examples are offered here:

(5.16) *Péter szénát **rak** a szekérre.*
Peter hay-ACC loads the cart-SUB
'Peter is loading hay onto the cart.'

(5.17) *Péter **megrakja** a szekeret szénával.*
Peter PREVERB-load-3SGOBJ the cart-ACC hay-INS
'Peter is loading the cart with hay.'[3]

(5.18) *Pisti **elmozdította** a követ.*
Steve PREVERB-move-PAST-3SGOBJ the stone-ACC
'Steve moved the stone.'

(5.19) *A kő **elmozdult**.*
the stone PREVERB-move-PAST3SG
'The stone moved.'[4]

In (5.16) the noun occurred in the accusative case while in (5.17) it is in the instrumental case and the noun in sublative occurs in the accusative next to the verb with the preverb. The object of (5.18) becomes the subject in (5.19) and in the latter sentence, the Agent is not present at all.

[3]The preverb *meg* is added to the verb and the conjugation switched to the definite paradigm.
[4]The derivational suffixes change here.

5.3.2 Syntactic alternations between semi-compositional constructions and their verbal counterparts

The above classification can be extended to semi-compositional constructions and their verbal counterparts as well, that is, it can be investigated whether there are alternations in the argument structure comparing that of the verbal counterpart and the construction.

Let us have a look at the following examples.

(5.20) *Zoli búcsút vett Sárától.*
Zoli farewell-ACC take-PAST3SG Sara-ABL
'Zoli said farewell to Sara.'

(5.21) *Zoli elbúcsúzott Sárától.*
Zoli PREVERB-say.farewell-PAST3SG Sara-ABL
'Zoli said farewell to Sara.'

(5.22) *Az öreg hölgy látogatást tett mindenkinél.*
the old lady visit-ACC make-PAST3SG everyone-ADE
'The old lady paid a visit to everyone.'

(5.23) *Az öreg hölgy mindenkit meglátogatott.*
the old lady everyone-ACC PREVERB-visit-PAST3SG
'The old lady visited everyone.'

In (5.21) and (5.20) the argument structure of the verbal counterpart and the construction is the same: they have a nominal argument in the ablative case. On the other hand, in (5.22) the argument is in the adessive case while in (5.23) in the accusative case, that is, the argument structure has changed.

B. Kovács (1999) discusses the substitutability of semi-compositional constructions and the possible changes in the argument structure in the Hungarian legal language. She describes the following cases:

- there is no change in argument structure:

  döntést hoz vmiről
  decision-ACC brings sg-DEL
  'to take a decision on sg'

  dönt vmiről

  'to decide on sg';

- the argument structure changes
  - the arguments already present are reorganized:

    különbséget *tesz* *vmik* *között*
    distinction-ACC does sg-PL between
    'to make a distinction between sg'

    vmiket *megkülönböztet*
    sg-PL-ACC distinguishes
    'to distinguish sg';

  - a new argument emerges:

    bejelentést *tesz*
    announcement-ACC does
    'to make an announcement'

    bejelent vmit

    'to announce sg';

- the construction cannot be substituted by a verbal counterpart.

In the second case, the example of *bejelentést tesz* (announcement-ACC does) 'to make an announcement' is not correct since it can have an argument (*vmiről* sg-DEL 'about sg'), which fulfills the same function as the accusative argument of the verbal counterpart. In the last group the non-substitutability can be traced back to the presence of a modifier before the noun, which cannot be converted to an adverbial modifier in the sentence with the verbal counterpart or to the fact that no verb can be derived from the nominal component (see above). Finally, it is also possible that the substitution cannot be carried out because there is only one proper way to express the meaning in the legal language.

The above grouping can be applied to semi-compositional constructions and their verbal counterparts from a general (i.e. not only legal) domain, moreover, the classification is revised and extended in the following.

The first group contains constructions and verbs where the substitution does not involve any change in the argument structure. Some examples are listed here:

(5.24) *Imre* *sétát* *tett.*
Imre walk-ACC make-PAST3S
'Imre had a walk.'

(5.25) *Imre sétált.*
Imre walk-PAST3SG
'Imre walked.'

(5.26) *Pali döntést hozott az ügyről.*
Paul decision-ACC bring-PAST3SG the case-DEL
'Paul made a decision on the case.'

(5.27) *Pali döntött az ügyről.*
Paul decide-PAST3SG the case-DEL
'Paul decided on the case.'

In (5.24) and (5.25), neither the verb nor the construction have any arguments (except for the subject) while in (5.26) and (5.27) the Hungarian examples contain arguments in the delative case (*ügyről* case-DEL) whereas the English sentences include a PP argument with the preposition *on*. Other examples:

(5.28) *fürdőt vesz*
bath-ACC takes
'to take a bath'

fürdik

'to bathe'

(5.29) *perbe fog vkit*
sue-ILL takes sy-ACC
'to sue sy'

perel vkit

'to sue sy'

(5.30) *to draw breathe – to breathe*

With regard to the second group (i.e. where the substitution involves change in the argument structure), B. Kovács's subclasses should be modified to some extent since the examples provided by her in different subclasses share the feature that both the construction and the verbal counterpart contain an argument in the accusative case as in:

(5.31) *Nem tudott különbséget tenni az ikrek között.*
not can-PAST3SG distinction-ACC to.make the twin-PL between
'He could not make a distinction between the twins.'

(5.32) *Nem tudta megkülönböztetni az ikreket.*
not can-PAST-3SGOBJ PREVERB-to.distinguish the twin-PL-ACC
'He could not distinguish the twins.'

(5.33) *Bejelentést tett a jövő heti kirándulás részleteivel kapcsolatban.*
announcement-ACC do-PAST3SG the next week-DERIV.SUFFIX trip detail-3SGPOSS-PL-INS relation-INE
'He made an announcement on the details of next week's trip.'

(5.34) *Bejelentette a jövő heti kirándulás részleteit.*
PREVERB-announce-PAST-3SGOBJ the next week-DERIV.SUFFIX trip detail-3SGPOSS-PL-ACC
'He announced the details of next week's trip.'

In (5.32) and (5.34) the PP argument of the construction occurs in the accusative case as an argument of the verbal counterpart, thus, they are considered to belong to the same subgroup in the new classification, namely, the one where another argument occurs in the accusative case with the verbal counterpart instead of the the nominal component of the construction.

The second subgroup contains examples where the arguments of the construction occur in different cases next to the verbal counterpart:

(5.35) *A kocsiért János egy egyszobás lakást adott cserébe Péternek.*
the car-CAUS John a one.room-DERIV.SUFFIX flat-ACC give-PAST3SG change-ILL Peter-DAT
'For the car, John gave a studio to Peter in return.'

(5.36) *János elcserélte Péterrel az egyszobás lakást a kocsira.*
John PREVERB-change-PAST-3SGOBJ Peter-INS the one.room-DERIV.SUFFIX flat-ACC the car-SUB
'John changed the studio for a car with Peter.'

In (5.35) the Beneficient occurs in the dative case whereas in (5.36) in the instrumental case, where the verbal component has no argument in the dative. In other words, the number and thematic role of the arguments coincide in the case of the construction and the verbal counterpart, however, their grammatical case or preposition might differ. This subgroup resembles the second group of B. Kovács. Other examples:

(5.37) *bosszút* *áll* *vkin* *vmiért*
revenge-ACC stands sy-SUP sg-CAU
'to take revenge on sy for sg'

megbosszul vkin vmit

'to revenge sg on sy'

(5.38) *parancsot* *ad* *vkinek* *vmire*
order-ACC gives sy-DAT sg-SUB
'to give order to sy for sg'

parancsol vkinek vmit

'to order sy sg'

(5.39) *to take notice of sg*

to notice sg

(5.40) *to put one's trust in sy*

to trust sy

bizalmát helyezi vkibe

bízik vkiben[5]

In the third subgroup new arguments emerge with the verbal counterpart, which are not present in the semi-compositional construction. This subgroup could be found in B. Kovács's classification, however, her example was not correct. Some examples are offered here:

(5.41) *Az* *asszony* *ajándékot* *adott* *a* *lányának.*
the woman present-ACC give-PAST3SG the daughter-3SGPOSS-DAT
'The woman gave a present to her daughter.'

(5.42) *Az* *asszony* *megajándékozta* *a* *lányát* *egy*
the woman PREVERB-gift-PAST-3SGOBJ the daughter-3SGPOSS-DAT a
gyűrűvel.
ring-INS
'The woman gifted a ring to her daughter.'

(5.43) *Az* *asszony* *megajándékozta* *a* *lányát.*
the woman PREVERB-gift-PAST-3SGOBJ the daughter-3SGPOSS-DAT
'The women gifted her daughter.'

[5]These examples are translational equivalents.

(5.44) *A pszichológus tanácsot adott Máriának.*
the psychologist advice-ACC give-PAST3SG Mary-DAT
'The psychologist gave advice to Mary.'

(5.45) *A pszichológus óvatosságot tanácsolt Máriának.*
the psychologist carefulness-ACC advise-PAST3SG Mary-DAT
'The psychologist advised Mary to be careful.'

(5.46) **A pszichológus tanácsolt Máriának.*
the psychologist advise-PAST3SG Mary-DAT
'The psychologist advised Mary.'

In (5.41) the exact nature of the gift is not revealed but it can be present as an optional argument in the instrumental case next to the verbal counterpart (compare (5.42) and (5.43)). In (5.44) the content of the advice remains hidden while it must be overt next to the verb in (5.45) (compare (5.46)). In fact, it is again another argument of the verb that takes the accusative case instead of the nominal component, however, this example does not belong to our first subgroup since the object of the verbal counterpart is not an argument of the semi-compositional construction. That is, it is a new argument of the verbal counterpart. As opposed to this, in the first subgroup it is only the already existing arguments that are rearranged.

The fourth subgroup comprises instances of diathesis, that is, the subject in the semi-compositional construction becomes the object of the verbal counterpart. In English, these constructions are typically translated with a passive sentence (compare data on *kerül* 'get' in Chapter 3).

(5.47) *Az előadás megrendezésre fog kerülni.*
the performance stage-DERIV.SUFFIX-SUB will to.get
'The performance will be held.'

(5.48) *Az előadást meg fogják rendezni.*
the performance-ACC PREVERB will-3PLOBJ to.stage
'They will hold the performance.'

(5.49) *Géza jutalomban részesült.*
Géza reward-INE get-PAST3SG
'Géza got a reward.'

(5.50) *Gézát megjutalmazták.*
Géza-ACC PREVERB-reward-PAST-3PLOBJ
'They rewarded Géza.'

Semi-compositional constructions in the third group cannot be substituted by a verb, not even by changing their argument structure. Some examples:

(5.51) *A rendőrség házkutatást tartott a lakásán.*
the police search.of.premises-ACC hold-PAST3SG the flat-3SGPOSS-SUP
'The police conducted a search of premises in his flat.'

(5.52) **A rendőrség házkutatott a lakásán.*
the police premises.search-PAST3SG the flat-3SGPOSS-SUP

(5.53) *Elsőbbséget adott a villamosnak.*
way-ACC give-PAST3SG the tram-DAT
'He gave way to the tram.'

(5.54) **Elsőbbségezte a villamost.*
waygive-PAST3SG the tram-ACC
'He waygave the tram.'

It can be seen that (5.52) and (5.54) contain ungrammatical verb forms, which indicates that the intended meaning can be expressed only through a semi-compositional construction.

The above grouping of semi-compositional constructions is summarized in Table 5.1 with some examples.

Concerning the above grouping, total synonymy between the verbal counterpart and the construction can be assumed only in cases (1), (2a) and (2b). Here the number and the thematic role of the arguments coincide – though their case suffix (or preposition) might be different. As for the other groups, the number and quality of the arguments might differ: in (2c) new arguments appear with the verbal counterpart, which results in the fact that the meaning of the construction and its verbal counterpart is not exactly the same: it is mostly the verb that offers more details of the event. In the case of (2d), the agent of the action is present (at least in the form of an indefinite subject bound by an existential quantifier) next to the verbal counterpart whereas it is not expressed in the semi-compositional construction – thus, here again the verbal counterpart is more detailed about the event. In the third group, the construction cannot be substituted by a verb, thus, there is no point in discussing the question of synonymy.

If the verbal component occurring in different groups are compared, it turns out that the verbal component itself cannot determine the group the construction belongs to. For instance, the verb *ad* 'give' occurs in (2a), (2b) and (2c) as well though the constructions it occurs in are different. However, diathesis is characteristic only of certain verbs, typically *kerül* 'get', *részesül* 'receive', *lel* 'find' etc.

Group		Features	Examples
(1)		the argument structure of the construction and the verbal counterpart is the same	*döntést hoz vmiről – dönt vmiről*; *perbe fog vkit – perel vkit*; *fürdőt vesz – fürdik*; *to draw breathe - to breathe*
(2)		the argument structure of the construction and the verbal counterpart is not the same	
	a)	instead of the nominal component another argument bears the accusative case	*parancsot ad vkinek vmire – megparancsol vkinek vmit*; *bosszút áll vkin vmiért – megbosszul vkin vmit*; *to take notice of sg - to notice sg*
	b)	some arguments of the verbal component have a case that is not present in the construction	*cserébe ad vkinek vmit vmiért – elcserél vkivel vmit vmiért*
	c)	the verbal counterpart has new (obligatory) arguments	*ajándékot ad vkinek – megajándékoz vkit (vmivel)*; *büntetést ad vkinek – megbüntet vkit (vmivel)*; *tanácsot ad vkinek – tanácsol vkinek vmit*; *to lay a charge on sy - to charge sy with sg*
	d)	diathesis	*megrendezésre kerül vmi – vki megrendez vmit / megrendeznek vmit*; *jutalomban részesül vki – vki megjutalmaz vkit / megjutalmaznak vkit*; *vigaszt lel vki – vki megvigasztal vkit / megvigasztalnak vkit*
(3)		the construction cannot be substituted by a verbal counterpart	*házkutatást tart; bosszút foral vki ellen; to give way*

Table 5.1: Semi-compositional constructions and their substitutability with their verbal counterparts

5.3.3 Interlingual differences

If the English and Hungarian constructions are contrasted, in many cases they behave in a parallel way, i.e. differences between the constructions and the verbal counterparts are similar in both languages. However, some cases can be found when the two corresponding constructions have different arguments in the two languages. For instance, *választ ad vkinek vmire* (answer-ACC gives sy-DAT sg-SUB) has two arguments while its English equivalent (*to give an answer for sg*) has only one. Similarly, in *példát mutat vkinek vmiben* (example-ACC show sy-DAT sg-INE) has two arguments while *to set an example for sy* has only one.

Another interesting interlingual difference is that there are cases when in one of the languages there is a semi-compositional construction which has only a verbal counterpart in the other language (i.e. there is no parallel construction). For instance:

(5.55) *perbe* *fog*
sue-ILL takes
'to sue'

tervbe *vesz*
plan-ILL takes
'to plan'

tornázik

'to do exercise'

Furthermore, examples from SzegedParalellFX can also be gathered to illustrate that semi-compositional constructions can have multiple translational equivalents in the other language, e.g. *döntést hoz* (decision-ACC brings) can be translated as *to make a decision*, *to take a decision* or *to decide*, and, on the other hand, *taking a decision* is sometimes translated simply by *decision* as in:

(5.56) *In taking this decision, the European Union was not simply increasing its surface area and its population.*

Ez a döntés jóval többet jelentett az Európai Unió
this the decision many more-ACC mean-PAST3SG the European Union
méretének és lakossága számának
size-3SGPOSS-GEN and population-3SGPOSS number-3SGPOSS-GEN
növelésénél.
increase-3SGPOSS-ADE

These examples suggest that the same semantic unit can be expressed by several grammatical devices (such as semi-compositional constructions, verbs or nouns) within the same language or in different languages, which could be further investigated in contrastive linguistic research or (machine) translation studies.

5.4 Differences concerning aspect or Aktionsart

When examining the substitutability of semi-compositional constructions with their verbal counterparts, the similarities (or differences) of their semantic features must be also considered. For instance, they may differ concerning their aspectual features – this can be indicated by their compatibility with certain time adverbial phrases (cf. Kiefer (2006; 2007)).

5.4.1 Analysis of Hungarian data

Kiefer (2007) argues that Hungarian complex predicates such as *újságot olvas* (newspaper-ACC reads) 'to read a newspaper' or *levelet ír* (letter-ACC writes) 'to write a letter' are inherently progressive since they are compatible with certain adverbials such as *javában* 'at its height'. Let us examine whether this claim holds for semi-compositional constructions. Here are some examples:

(5.57) *Pontosan délben kezelésbe vette az orvos.*
exactly noon-INE treatment-ILL take-PAST-3SGOBJ the doctor
'The doctor started to give him a treatment exactly at noon.'

(5.58) **Pontosan délben kezelte az orvos.*
exactly noon-INE treat-PAST-3SGOBJ the doctor
'The doctor treated him exactly at noon.'

(5.59) *Az orvos két hétig kezelte.*
the doctor two week-TER treat-PAST-3SGOBJ
'The doctor was treating him for two weeks.'

(5.60) **Az orvos két hétig kezelésbe vette.*
the doctor two week-TER treatment-ILL take-PAST-3SGOBJ
'The doctor started to give him a treatment for two weeks.'

With the time adverbial phrase *pontosan délben* 'exactly at noon' referring to a specific point in time only the semi-compositional construction is compatible – the verbal component

cannot tolerate this adverbial phrase: (5.58) is not acceptable as a neutral sentence.[6] On the other hand, a time adverbial referring to a longer period of time (*két hétig* 'for two weeks') can occur only with the verbal counterpart, thus, it seems that the verbal counterpart describes an event lasting for a longer period of time thus it is progressive while the semi-compositional construction expresses a momentary action, the beginning of the action, that is, it has inchoative Aktionsart (Kiefer, 2003, p. 293), (Kiefer, 2006, p. 169–170). In Hungarian, Aktionsarts are usually inherently paired with (perfective) aspect (Kiefer, 2007), which is the case here as well: (5.57) is perfective.

Another example:

(5.61) *Az élelmiszeripari cég csak három hónapig forgalmazza új*
the food.industrial firm only three month-TER circulate-3SGOBJ new
ízesítésű üdítőitalát.
flavored beverage-3SGPOSS-ACC
'The food manufacturing firm will be circulating its new flavored beverage only for three months.'

(5.62) **Az élelmiszeripari cég csak három hónapig hozza*
the food.industrial firm only three month-TER bring-3SGOBJ
forgalomba új ízesítésű üdítőitalát.
circulation-ILL new flavored beverage-3SGPOSS-ACC
'The food manufacturing firm will be putting into circulation its new flavored beverage only for three months.'

(5.63) *Az élelmiszeripari cég csak három nap múlva hozza*
the food.industrial firm only three day-TER in bring-3SGOBJ
forgalomba új ízesítésű üdítőitalát.
circulation-ILL new flavored beverage-3SGPOSS-ACC
'The food manufacturing firm will put into circulation its new flavored beverage only in three days.'

(5.64) **Az élelmiszeripari cég csak három nap múlva forgalmazza új*
the food.industrial firm only three day-TER in circulate-3SGOBJ new
ízesítésű üdítőitalát.
flavored beverage-3SGPOSS-ACC
'The food manufacturing firm will circulate its new flavored beverage only in three days.'

Similarly to the previous example, the semi-compositional construction can only tolerate a time adverbial referring to a point in time (*három nap múlva* 'in three days') while the verbal counterpart co-occurs with a durative time adverb (*három hónapig* 'for three months'),

[6] The sentence is acceptable in the progressive aspect.

which yields that it is progressive. The construction has inchoative Aktionsart and perfective aspect.

Examples for another phenomenon:

(5.65) *Véleményét* *javában* *hangoztatta.*
opinion-3SGPOSS-ACC good-3SGPOSS-INE sound-PAST-3SGOBJ
'He was sounding his opinion at its height.'

(5.66) **Véleményének* *javában* *adott* *hangot.*
opinion-3SGPOSS-DAT good-3SGPOSS-INE give-PAST3SG sound-ACC
'He was sounding his opinion at its height.'

(5.67) *Délután* *3-tól* *5-ig* *pofozta* *a* *fiát.*
afternoon 3-ABL 5-TER slap-PAST-3SGOBJ the son-3SGPOSS-ACC
'He was slapping his son from 3pm to 5pm.'

(5.68) **Délután* *3-től* *5-ig* *adott* *pofont* *a* *fiának.*
afternoon 3-ABL 5-TER give-PAST3SG slap-ACC the son-3SGPOSS-DAT
'He was slapping his son in the face from 3pm to 5pm.'

From these examples it can be seen that the time adverbials (*órákig* 'for hours' and *délután 3-tól 5-ig* 'from 3pm to 5pm') can only co-occur with verbal counterparts and they are not compatible with semi-compositional constructions. (5.65) and (5.67) describe events of slapping and expressing that are continuously repeated for some hours, that is, the sentences with the verbal counterparts possess iterative/frequentative[7] Aktionsart and progressive aspect (Kiefer, 2006, p. 150–162). However, the corresponding constructions do not have this meaning component: they describe a unitary action which is perfective.

In the previous examples the Aktionsart of the verbal counterpart and the construction differed from each other. However, there exist some morphological tools with the help of which the construction and its verbal counterpart share the same Aktionsart feature: e.g. a preverb can be added to the verbal counterpart, cf. Kiefer (2006; 2007):

(5.69) *A* *gazda* *perbe* *fogta* *a* *szomszédját*
the farmer sue-ILL take-PAST-3SGOBJ the neighbor-3SGPOSS-ACC
szeptember *utolsó* *napján,* *mert* *az* *birtokháborítást*
September last day-3SGPOSS-SUP because he trespass-ACC
követett *el.*
commit-PAST3SG PREVERB
'The farmer sued his neighbor on the last day of September for he committed trespass.'

[7]The difference being is that iterative Aktionsart involves regular repetitions of the action while frequentative means repetitions at irregular intervals.

(5.70) *?A gazda perelte a szomszédját szeptember utolsó*
the farmer sue-PAST-3SGOBJ the neighbor-3SGPOSS-ACC September last
napján, mert az birtokháborítást követett el.
day-3SGPOSS-SUP because he trespass-ACC commit-PAST3SG PREVERB
'The farmer was suing his neighbor on the last day of September for he committed trespass.'

(5.71) *A gazda beperelte a szomszédját*
the farmer PREVERB-sue-PAST-3SGOBJ the neighbor-3SGPOSS-ACC
szeptember utolsó napján, mert az birtokháborítást
September last day-3SGPOSS-SUP because he trespass-ACC
követett el.
commit-PAST3SG PREVERB
'The farmer sued his neighbor on the last day of September for he committed trespass.'

On the basis of (5.69) the semi-compositional construction has inchoative Aktionsart (and perfective aspect). The preverbless verbal counterpart (*perel* 'sue') cannot be used in the same syntactic environment, thus, it is not inchoative. The verb *beperel* 'sue', however, can co-occur with the time adverbial referring to a point and the time adverbial denotes the starting point of the interval of the action (the beginning of the event of suing) hence it is inchoative (Kiefer, 2006, p. 169–170).

Based on the examples found in the database, it seems that pairs of semi-compositional constructions and their verbal counterparts can be classified into two groups as far as aspect and Aktionsart are concerned. First, the construction has inchoative Aktionsart and perfective aspect: it refers to the beginning of the action (unitary action) while the verbal counterpart describes a longer action (progressive aspect). The following semi-compositional constructions belong to this group:

(5.72) *kezelésbe vesz*
treatment-ILL takes
'to start to give treatment'

nyilvántartásba vesz
register-ILL takes
'to register'

forgalomba hoz
circulation-ILL takes
'to put into circulation'

figyelembe *vesz*
consideration-ILL takes
'to take into consideration'

gyanúba *hoz/fog*
suspicion-ILL brings/takes
'to suspect'

vizsgálat *alá* *vesz*
investigation under takes
'to put under investigation'

iskolába *ad*
school-ILL gives
'to send to school'

perbe *fog*
sue-ILL takes
'to sue'

Their verbal counterparts are also provided here:

(5.73) *kezel* 'to treat', *nyilvántart* 'to keep registered', *forgalmaz* 'to circulate', *figyel* 'to consider', *gyanúsít/gyanakszik* 'to suspect', *vizsgál* 'to investigate', *iskoláztat* 'to educate', *perel* 'to sue'

In order to preserve Aktionsart, a preverb can also be added to the verb in some cases as it is a typical morphological means in Hungarian to express Aktionsart and/or perfective aspect (Kiefer, 2007):

(5.74) *meggyanúsít* 'to suspect', *beiskoláz* 'to send to school', *beperel* 'to sue'

Typical verbal components occurring in constructions belonging to this group are *fog* 'take' and *vesz* 'take'. They apparently possess inchoative Aktionsart, which is also supported by the fact that their sense definitions found in dictionaries (e.g. The Concise Dictionary of the Hungarian Language, henceforth ÉKsz.) also contain the meaning component 'begin':

> **vesz**: Használni kezd, a megnevezett változás megkezdődik 'to begin to use, the change mentioned starts'. (ÉKsz. 1488–1489)
>
> **fog**: Belekezd vmibe 'to begin to do sg'. (ÉKsz. 415)

In the second group, the verbal counterparts express repetition, that is, they have iterative or frequentative Aktionsart while the constructions describe a single (unitary) perfective action. Some of the semi-compositional constructions that belong to this group are listed below:

(5.75) *kivételt* *tesz*
exception-ACC does
'to make an exception'

pofont *ad*
slap-ACC gives
'to slap in the face'

hangot *ad*
sound-ACC gives
'to give his opinion'

verést *ad*
beat-ACC gives
'to beat'

Their verbal counterparts are the following:

(5.76) *kivételez* 'to except', *pofoz* 'to slap', *hangoztat* 'to sound', *ver* 'to beat'

The appearance of a preverb again results in the fact that the construction and its verbal counterpart with a preverb have the same aspect or Aktionsart. Some verbs with preverb:

(5.77) *megpofoz / felpofoz* 'to slap in the face', *megver* 'to beat'

With regard to differences of aspect and Aktionsart, it can be concluded that if there are differences, the semi-compositional construction and its verbal counterpart cannot be seen as synonyms. The above classification is also summarized in Table 5.2 and it is also revealed that semi-compositional constructions tend to have perfective aspect as opposed to productive complex predicates, which have progressive aspects (cf. Kiefer (2007)).

FX	Verb	Verb with preverb	Example
perfective inchoative unitary	progressive	perfective inchoative	*perbe fog – perel – beperel*
perfective unitary	progressive iterative/frequentative	perfective unitary	*verést ad – ver – megver*

Table 5.2: Differences in aspect and Aktionsart between semi-compositional constructions and their verbal counterparts

5.4.2 Analysis of English data

As it was presented, Hungarian semi-compositional constructions and their verbal counterparts can be classified into two groups concerning aspect and Aktionsart: either the construction is inchoative or the verbal counterpart has an iterative meaning. However, the English language uses different tools for expressing aspect and Aktionsart: in English, there is no Aktionsart to be expressed morphologically (Kiefer, 2006, pp. 186–187), on the other hand, the progressive aspect can be expressed through morphological devices (*be* + present participle) and it can occur on its own (i.e. without any time adverbial or time clause) (Freed, 1979; Kiefer, 2006; Kiefer, 2007, p. 102). Thus, in English, aspect is primarily relevant on the sentence level and it is not lexically fixed (Verkuyl, 1972): the progressive and perfective aspects are expressed with complex verb forms and they can also co-occur. Progressive or Continuous forms have progressive aspect, Perfect forms have perfective aspect while Perfect Progressive forms have both aspects. Since every English verb or verbal construction (except for stative verbs) can occur in each form – at least in principle –, aspectual differences between a semi-compositional construction and its verbal counterpart are hard to find. However, there are some differences involving Aktionsart:

(5.78) The company put this product into circulation two days ago.

(5.79) *The company circulated this product two days ago.

(5.80) The company circulated this product for two months.

(5.81) *The company put this product into circulation for two months.

The time adverbial referring to a specific point (*two days ago*) can only co-occur with the semi-compositional construction whereas the verbal component is compatible only with

the durative time adverbial (*for two months*). As it was already argued in the analysis of the Hungarian examples, the construction refers to a specific moment, the beginning of the action while the verbal counterpart describes a longer process: the construction is inchoative while the verbal counterpart is not. These results are in line with Wierzbicka (1982), who claims that as opposed to *have a V* frames, the *take a V* frame[8] describes events that are unitary and the action has a specific moment as a starting point.

The verbal counterpart can have iterative Aktionsart as well – although these cases seem to be far less frequent than in Hungarian:

(5.82) The teacher gave Joe priority several times during the schoolyear.

(5.83) ?The teacher prioritized Joe several times during the schoolyear.

The verbal counterpart with iterative Aktionsart describes a repetitive action, thus, its meaning contains the component 'regularly' and this is why it is less compatible with the adverb *several times*. However, it seems that the construction can co-occur with this adverbial, which reflects that the iterative Aktionsart is not dominant in its meaning.

5.4.3 Comparing English and Hungarian results

The comparison of English and Hungarian examples reveals that in both languages there are cases when the meanings of the semi-compositional construction and its verbal counterpart do not completely overlap due to some differences in their Aktionsart. This phenomenon is, however, more frequent in Hungarian than in English. On the other hand, in both languages there are inherently inchoative constructions or those that describe a single action.

In some cases, the Hungarian construction and its English equivalent do not correspond to each other in every respect. If dictionary entries are concerned, the pairs *to take revenge* vs. *bosszút áll* and *to take into consideration* vs. *figyelembe vesz* are equivalents, respectively, however, their usage is somewhat different as it is illustrated below:

(5.84) The chairman was taking all the suggestions into consideration when deciding on the new budget.

[8] Wierzbicka (1982) argues that the "nominal" component of the construction is in fact a verb, however, we consider them as nouns.

(5.85) **Az elnök éppen figyelembe vett minden*
the chairman just consideration-ILL take-PAST3SG all
javaslatot, amikor az új költségvetésről döntött.
suggestion-ACC when the new budget-DEL decide-PAST3SG

(5.86) She was taking revenge on her ex-boyfriend in those days.

(5.87) *??Azokban a napokban éppen bosszút állt a volt*
that-PL-INE the day-PL-INE just revenge-ACC stand-PAST3SG the ex
barátján.
boyfriend-3SGPOSS-SUP

Each sentence has a progressive aspect – in English, it is the verb form that carries this piece of grammatical information while in Hungarian it is the adverb *éppen* 'just' and the word order. However, the Hungarian progressive forms are not (totally) acceptable as opposed to their English counterparts – thus, they cannot be considered as totally equivalent because of such aspectual differences. As discussed above, semi-compositional constructions are perfective in Hungarian, thus, they do not tolerate progressive aspect at the same time. However, in English, perfective and progressive aspects can co-exist (see e.g. Perfect Continuous verb forms), which leads to interlingual differences concerning the usage of translational equivalents.

5.5 The acceptability of semi-compositional constructions

Based on the above results, semi-compositional constructions cannot be always matched to a verbal counterpart. First, sometimes the given meaning can only be expressed by a construction and second, they might differ in style or frequency. Thus, they cannot be seen as variants, only lexical units (derived from the same root) with similar (but not identical) meanings. In this way, the claim found in the Purists' dictionary (Grétsy and Kemény, 1996) cannot be supported on the basis of language data.

However, certain semi-compositional constructions seem to be more marked than others. In other words, the acceptability of semi-compositional constructions is also a matter of degree and scale (similarly to their productivity and compositionality, see Chapter 4). While constructions such as

(5.88) *tanácsot ad*
advice-ACC gives
'to give advice'

or

(5.89) *bosszút* *áll*
revenge-ACC stands
'to take revenge

are perfectly sound, others are not unequivocally judged as acceptable (in the corpora, they rarely occur and if so, they can be typically found in the economic or legal domains, which abound in semi-compositional constructions). Such constructions are:

(5.90) *kiadásra* *kerül*
publication-SUB gets
'to get published'

beruházást *eszközöl*
investment-ACC performs
'to invest'

intézkedést *foganatosít*
instruction-ACC carries.out
'to carry out an instruction'

nyereséget *realizál*
profit-ACC realizes
'to produce profit'

These verbal components are semantically more bleached than the others, being situated at the farther end of the scale of emptiness (Alonso Ramos, 2004; Sanromán Vilas, 2009). Following Meyers et al. (2004a) they can be called light verbs as opposed to support verbs[9] – hence the meaning of the construction more heavily relies on the meaning of the noun, which is also underlined by the fact that it is only deverbal nouns that occur together with this group of verbs. It is also an interesting question to examine whether the length of the verb has an effect on the acceptability of semi-compositional constructions: it seems that the longer the verb is, the less frequent and more marked the construction is.[10] However, this claim needs further investigation.

[9] *Eszközöl* 'perform' and *foganatosít* 'carry out' have only a light verb sense, i.e. they do not occur outside a semi-compositional construction.

[10] We would like to thank Zsuzsanna Gécseg for pointing out this issue.

5.6 Summary of results

In this chapter, the relation of semi-compositional constructions and their verbal counterparts were analyzed. Emphasis was put on the following:

- 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;
- the construction and its verbal counterpart are not always in a perfect overlap in style or frequency;
- differences between the construction and its verbal counterpart can be traced back to either differences in the argument structure or in aspect and Aktionsart;
- as opposed to productive constructions that typically bear progressive aspect, semi-compositional constructions typically have perfective aspect in Hungarian;
- interlingual differences can be accounted for by differences in the argument structure or in aspect and Aktionsart;
- in contrast with Hungarian, the English construction can bear progressive and perfective aspect at the same time;
- the acceptability of semi-compositional constructions is a matter of degree and scale.

The results of this chapter can be applied in language teaching (both in teaching the mother tongue and foreign languages) and they can also have impact on information retrieval and machine translation (see Chapters 11 and 12).

Chapter 6

The syntax of semi-compositional constructions

6.1 Introduction

In the previous chapter, semi-compositional constructions were placed in a continuum of bare noun + verb constructions based on their productivity and compositionality. In this chapter, semi-compositional constructions are discussed from a syntactic point of view. Analyses are provided in the frameworks of constituency and dependency grammars and special attention is paid to argument structure. With regard to the research questions listed in Chapter 1, the following ones are to be discussed here:

- What are the syntactic features of semi-compositional constructions?
- How can the syntactic relation of the two components of the construction be described?

The chapter is structured as follows. First, semi-compositional constructions are characterized from a syntactic point of view, then general questions related to argument structure are discussed. Some possible analyses of semi-compositional constructions are presented in a generative framework, which is followed by an analysis in a dependency framework. In this way, it is possible to compare analyses in a framework using non-terminal symbols (generative grammar) and in another one without any abstract nodes (dependency grammar). The chapter concludes with the comparison of the analyses discussed.

6.2 Syntactic features of semi-compositional constructions

Komlósy (1992) discusses semi-compositional constructions among bare common noun + verb constructions. According to his analysis, the scheme preverb + verb serves as a pattern for idioms with the structure preverbless verb + argument since the syntactic behavior of the two constructions are similar to each other as in:

(6.1) *tőrbe* *csal*
dagger-ILL entices
'to deceit'

(6.2) *csütörtököt* *mond*
Thursday-ACC says
'to fail to work'

He also classifies the nominal components of semi-compositional constructions as elements behaving similarly to preverbs. In his system, semi-compositional constructions are considered as idiom-like constructions because they share their syntactic pattern with idioms mentioned above but semantically they preserve the meaning of their components to some degree.

In his view, there is no typical governor-argument relationship between the two components since the syntactic dependent (the nominal component) is a non-referring expression, however, it is not a semantic dependent of the syntactic head: semantically, it is the noun that heads the construction. The syntactic dependent has the head as its semantic argument, with which it forms a semantic unit. This is also confirmed by the fact that the whole construction bears one stress.

The semantic head of the construction is the argument (that is, the nominal component) while the governor (the verb) is responsible for the verbal nature of the construction. If the construction has some arguments, semantically they belong to the noun and syntactically to the verb.

As for the part of speech of the verbal component, Keszler (1994) claims that function verbs (*funkcióigék*) form an intermediate category in between normal verbs and auxiliaries. Her view is echoed in Lengyel (1999; 2000), who considers them as copula-like elements occurring in complex predicates. The verbal component in the construction enables the noun to function as a predicate, e.g. the verb *végez* 'carry out' fulfills a similar role in the construction *javítást végez* (repair-ACC carry out) 'to repair' than the copula *volt* (be-PAST-3SG) 'was'

next to the predicative adjective in *ügyes volt* (smart be-PAST-3SG) 'he was smart'. However, they differ in certain aspects: the copula attaches to a noun or adjective without any suffix (or in the nominative case) (e.g. *ügyes* 'smart') while the verbal component forms one unit with a noun with a suffix (e.g. *javítást* repair-ACC) which was originally an argument of the verb. On the other hand, the verbal component cannot fulfill the role of predicate on its own – just like verbs co-occurring with elements behaving similarly to preverbs (Komlósy, 1992) – only together with the nominal component. The relation between the two components is in between morphological constructions (formed by a conceptual word and a word with only morphological function) and syntagmatic relations (Keszler, 2000). In this way, the nominal component and the verb form the predicate of the sentence together hence a closer syntactic relationship is presumed between the verb and the nominal component than between the verb and its other arguments (see 6.7).

6.3 Issues related to the argument structure

The analysis of the argument structure of semi-compositional constructions deserves special attention from two aspects. First, it is worth examining the distribution of the arguments of the construction among the verbal and the nominal component. Second, the comparison of the argument structure of a semi-compositional construction and its verbal counterpart can also offer some interesting conclusions.

First, the distribution of the arguments of the construction is discussed. In principle, arguments co-occurring with the construction may belong either to the verb or to the noun. Let us start with an example:

(6.3) *parancsot* *ad* *vkinek* *vmire*
order-ACC gives sy-DAT sg-SUB
'to order sy to do sg'

The construction has two arguments: *vkinek* (somebody-DAT) and *vmire* (something-SUB). If the two components of the construction are analyzed on their own, the following are revealed: the verb *ad* 'give' originally has two arguments (*vkinek vmit* somebody-DAT something-ACC) while the noun has one (*vmire* something-SUB). Since within the construction, it is the nominal component that occurs in an accusative form, the object position of the verb *ad* 'give' is now fulfilled. With regard to the original arguments of the components of the construction

it can be assumed that among the arguments of the construction, *vmire* (something-SUB) is more closely related to the noun while *vkinek* (somebody-DAT) is more closely related to the verb.

However, earlier literature has not reached a consensus about the question whether arguments should be distributed between the verbal and the nominal component (see Meyers et al. (2004a)).

According to Meyers et al. (2004a), verbs occurring in semi-compositional constructions have at least two arguments (NP_1 and XP_1) where XP_1 is an argument of the head of NP_1. For instance, in the sentence *John took a walk* the verb *took* shares one of its arguments (*John*) with its other argument (*walk*) that is, *John* is an argument of the verbal and the nominal component at the same time. The argument is shared between a higher-level predicate (P_1) and a lower-level predicate (P_2) in such a way that the phrase the head of which is P_2 is an argument of P_1. The authors distinguish semi-compositional constructions according to their similarity to raising verbs or control verbs. In semi-compositional constructions similar to raising verbs (e.g. *seem*, *appear* etc.), P_1 is only a holder of the features modality, tense etc., but is considered to be semantically bleached. The relationship of the higher-level predicate and the shared argument is only superficial, not semantic. On the other hand, the verb gives a semantic role to the shared argument in semi-compositional constructions similar to control verbs (e.g. *want*, *agree* etc.). Thus, in the sentence *John made an attack*, *John* has simply the role ATTACKER since the verb *made* – being similar to raising verbs – does not assign a thematic role to it. However, in *John attempted an attack John* is both the ATTACKER and the ATTEMPTOR. Sometimes it is not unequivocal whether the given semi-compositional construction is similar to control verbs or raising verbs: in *John gave them a standing ovation*, it is undecided whether *John* has the role GIVER or not. In our dataset, most of the verbs are semantically bleached, thus, they are not able to assign a thematic role to the arguments, which means that they are similar to raising verbs.

The argument structure of the semi-compositional construction and its verbal counterpart may coincide as in *döntést hoz vmiről* (decision-ACC brings sg-DEL) ‘to make a decision on sg’ – *dönt vmiről* ‘to decide on sg’ or may differ as well (*tanácsot ad vkinek* (advice-ACC gives sy-DAT) ‘to give advice’ – *tanácsol vkinek vmit* ‘to advice sg to sy’). Since the nominal component and the verbal counterpart are often derived from the same root as in the examples above, the syntactic analysis should also take account of the derivational processes between the noun and the verbal counterpart (if any) and it should also explain alternations in the

argument structure.

6.4 Analyses within a generative framework

In this section, some analyses of semi-compositional constructions will be discussed from a generative grammar point of view.

6.4.1 Former analyses

Gracia i Sole (1986) analyzes semi-compositional constructions as complex predicates:

(6.4) [VP [V' [V + N(P)] ...] ...]

Thus, V' is formed by merging a verbal head (V) and a noun phrase (NP) or a nominal head (N).

The analysis of verbal constructions with a verbal modifier found in Kiefer and Ladányi (2000) is similar to the previous one – they assume a complex verb with the following structure:

(6.5) $[V_0\ [X_0 + V_0]]$

Here there are a verbal head and another head within the verbal head (V_0) – in the case of semi-compositional constructions the other head is a nominal head, thus, according to this analysis, semi-compositional constructions have the following syntactic structure:

(6.6) $[V_0\ [N_0 + V_0]]$

Grimshaw and Mester (1988) examine semi-compositional constructions in Japanese. The verbal components of the constructions are semantically bleached and are not capable of assigning thematic roles: if they are not phonologically empty, they are responsible only for bearing tense and agreement. The semantic arguments of the noun are transferred to the verb the argument structure of which is empty – this phenomenon is called argument transfer. Thus, arguments present in the construction are licensed by the argument structure of the nominal component. The following constraints are hypothesized for argument transfer in Japanese:

- besides the subject, at least one argument must be outside the NP;
- the subject is always outside the NP;
- if the nominal component assigns the roles Theme and Goal and the Theme is outside the NP, the Goal also must be outside the NP.

Here we offer one example with the verb *suru* 'to do' (its past tense, *shita* occurs here):

(6.7) *John-wa murabito-ni [[ookami-ga kuru-to]-no keikoku]-o shita.*
John-topic villager-DAT wolf-NOM come-COMP-GEN warning-ACC suru
'John warned the villagers that the wolf was coming.'

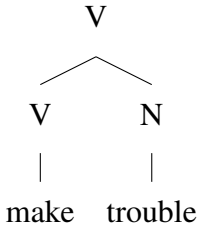
The argument transfer takes place in the following way:

(6.8) keikoku (Agent, Goal, Theme) + suru ( ) <ACC> -> keikoku (Theme) + suru (Agent, Goal)

As a result, the semantic arguments of the noun appear in the complement positions of the verb.

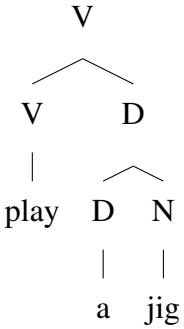
Hale and Keyser (2002) analyze V+N constructions in English as follows:

(6.9)



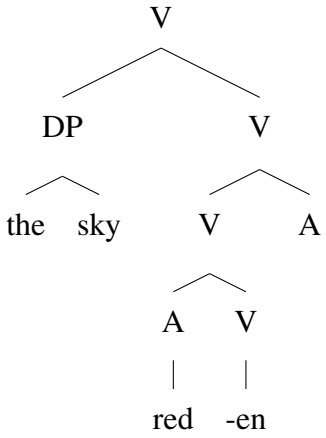
Thus, a verbal and a nominal head form a verbal head. If the nominal component is accompanied by an article, the construction is structured in this way:

(6.10)



Unergative verbs are derived in their analysis in the following way: the object / complement of the transitive verb is incorporated into an abstract verbal head, in which a derivational suffix may also be found in certain cases. As the following example shows, *red* was originally a complement of the verb but later it was incorporated into the verbal head where it merged with the verbal derivational suffix *-en*. The sentence *The sky reddened* can be derived as follows (ignoring now the tense):

(6.11)

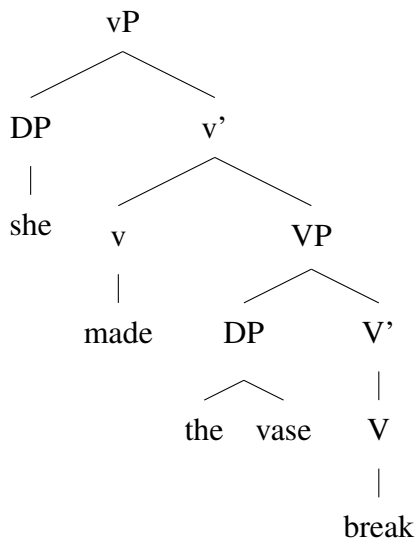


From the complement position A, the root *red* moved into the verbal head where it merged with the derivational suffix *-en*, yielding the form *redden*.

6.4.2 A possible analysis

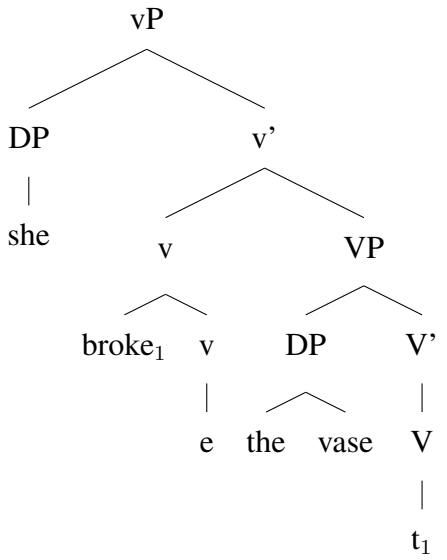
Let the starting point of our analysis be the causative construction in English since this construction also contains two verbal elements, similarly to semi-compositional constructions, where the nominal component is often deverbal. On the basis of the analysis assuming two VP shells (Newson et al., 2006; Larson, 1988; Chomsky, 1995) the following structure can be assigned to the causative construction:

(6.12)



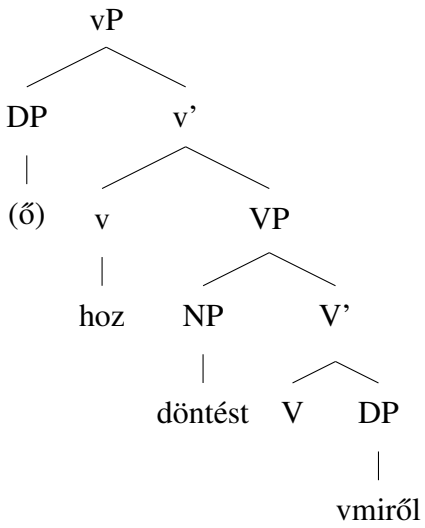
The causative verb (light verb) can be found in v while the main verb in position V. If the causative verb is not present, the main verb moves to position v where it merges with the invisible light verb (denoted by e):

(6.13)



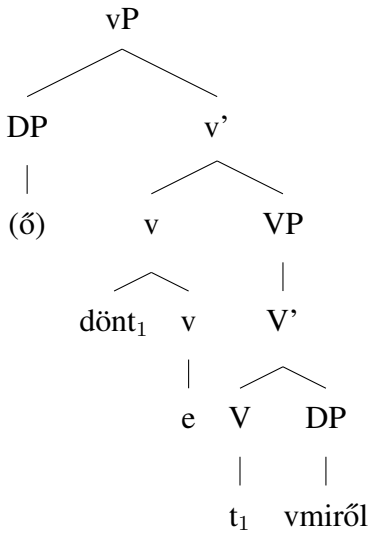
If we apply this to Hungarian semi-compositional constructions, the following structure can be assumed for them:

(6.14) *döntést* *hoz* *vmiről*
decision-ACC bring sg-DEL
'to make a decision on sg'



In this case, the verbal component *hoz* 'bring' is in position v, however, apparently there is no other verbal element in the sentence. The verbal counterpart of this structure can be derived in this way:

(6.15)



The main verb *dönt* 'decide' is generated in V, however, it moves to v. In its original position (V) it assigns a thematic role to the DP (*vmiről*).

This approach nevertheless suffers from several disadvantages. First, in (6.14) the position of the main verb, that is, position V is empty. In this way, the assigment of thematic roles might be problematic: neither *döntést* (decision-ACC) nor *vmiről* (sg-DEL) can be assigned a thematic role since the head position is empty.

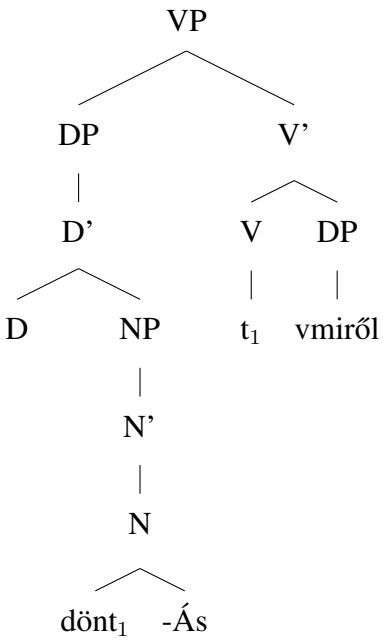
As a solution for the above problem morphology can be applied. It should be recognized that in the construction *döntést hoz* (decision-ACC brings) 'to make a decision', the root *dönt* 'decide' is present as well (just like in the verbal counterpart): the nominal component in the

construction is also derived from this root. On the level of morphology, *döntés* ‘decision’ is derived from the root *dönt* ‘decide’ by the process of deverbal derivation:

(6.16) $[\text{dönt}]_V$ + Ás $\rightarrow$ $[\text{döntés}]_N$

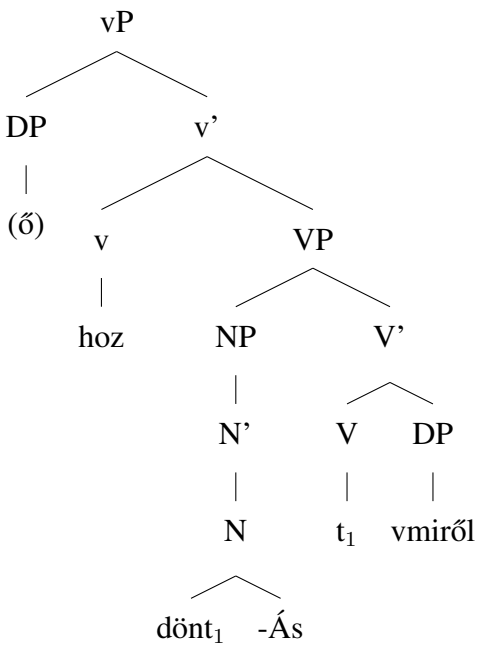
The construction *döntés vmiről* ‘decision on sg’ can be assumed the following syntactic analysis:

(6.17)



The root *dönt* ‘decide’ can be originally found in position V from where it moves to the head of the NP, where it merges with the nominal derivational suffix *-Ás*. The original argument of *dönt* ‘decide’ (*vmiről* sg-DEL) is still preserved in the nominal form as well. Based on this, the following syntactic analysis can be assumed for the semi-compositional construction *döntést hoz* (decision-ACC brings) ‘to make a decision’:

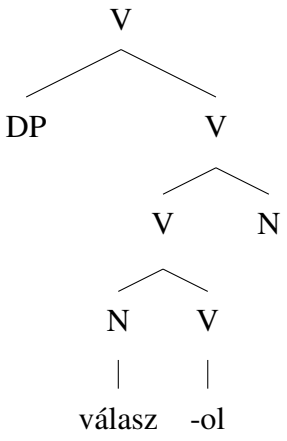
(6.18)



In the construction, the root *dönt* 'decide' moves from position V to position N including the nominal derivational suffix *-Ás*, where it merges with the suffix, yielding the nominal component of the semi-compositional construction. Position v, on the other hand, is occupied by the verbal component. This derivation can explain both the origin of the deverbal nominal component and the same argument structure of the verbal counterpart and the construction.

However, if the verbal counterpart is derived from the nominal component (e.g. *választ ad* (answer-ACC gives) 'to give an answer' – *válaszol* 'to answer'), the above derivation fails since the nominal component cannot be derived from a verb, hence the assignment of the thematic roles of the arguments in the construction is also problematic. These cases can be derived in accordance with Hale and Keyser's analysis of *redden*:

(6.19)



In this example, the noun *válasz* 'answer' was originally the complement of the verb, however, it moved to the verbal head and merged with the verbal derivational suffix.

If the analyses based on Hale and Keyser (2002) and Larson (1988), Chomsky (1995) and Newson et al. (2006) are compared, the following conclusions can be drawn. According to the first one (Hale and Keyser, 2002), the semi-compositional construction is yielded by lexical replacement and the verbal counterparts derived from nouns are considered as secondary. If there is no verbal counterpart, the analysis based on Hale and Keyser (2002) can be applied (i.e. lexical replacement takes place). On the other hand, the latter analysis (based on Larson (1988), Chomsky (1995) and Newson et al. (2006)) offers a proper derivation of constructions with a deverbal nominal component, however, the treatment of verbal counterpart derived from the noun remains problematic. From the above, it can be concluded that apparently no unified analysis can be proposed for the two constructions and separate analyses are needed for them. However, later investigations can shed more light on this phenomenon.

6.4.3 Alternation in the argument structures

Comparing our analyses offered in 6.4.1 and 6.4.2, the following can be stated about the alternation in the argument structure. If the argument structure of the semi-compositional construction is in complete overlap with that of the verbal counterpart, the analysis based on Larson (1988), Chomsky (1995) and Newson et al. (2006) can offer a satisfactory explanation for it (see (6.18)). If the argument structure of the construction differs from that of the verbal counterpart, there are more possibilities. First, in certain cases two arguments of the verbal counterpart are "merged" into a postposition with two arguments (e.g. the postposition *között* '(in) between' can have two coordinated DPs as argument, cf. Tóth (2007)): thus, the arguments of *megkülönbözteti x-et y-tól* 'distinguish x from y' manifest as one PP in *különbséget tesz x és y között* 'make a distinction between x and y'.

In other cases, the construction and its verbal counterpart have different arguments on the surface level (see Chapter 5): in the example *büntetést ad vkinek – megbüntet vkit vmivel* (punishment-ACC gives sy-DAT – PREVERB-punish sy-ACC sg-INS) 'to give a punishment to sy – to punish sy with sg' the participant occurring in the object position (*vkit*) of the verbal counterpart can be undoubtedly identified with the participant in the dative (*vkinek*) case in the construction. Thus, their thematic role must be the same hence they must occupy the same syntactic position in harmony with UTAH (the same thematic role can be assigned to two elements if and only if they occur in the same syntactic position (Baker, 1988)):

however, one of them occurs in VP, Spec and the other one in the complement position of V. This apparent contradiction can be resolved if we recall that the nominal component is not considered as an argument of the verb (Komlósy, 1992), it only occupies a position in the argument structure, thus, the object position (i.e. VP, Spec) is not available for the Patient hence it occupies the next one available (the complement of V) in accordance with the case hierarchy (Comrie, 1976).

Whenever the argument structure changes, it is typically the case of the nominal component that is taken by another argument of the verbal counterpart. This is especially true for the accusative case: the accusative case of the nominal component is given to the next argument in the thematic hierarchy. The prominence of the arguments is determined by the hierarchy below (following É. Kiss (2002, p. 38)):

Agent/Experient > Beneficiary > Theme > Goal > Instrumental > Locative

Let us illustrate this with an example:

(6.20) *A* *főnök* *jutalmat* *adott* *a* *titkárnőnek* *a* *sok*
the boss reward-ACC give-PAST3SG the secretary-DAT the many
túlóráért.
extra.hour-CAUS
'The boss gave a reward to the secretary for her many extra hours.'

(6.21) *A* *főnök* *megjutalmazta* *a* *titkárnőt* *a* *sok*
the boss reward-PAST-3SGOBJ the secretary-ACC the many
túlóráért.
extra.hour-CAUS
'The boss rewarded his secretary for her many extra hours.'

In (6.20) the Beneficiary (*a titkárnőnek – the secretary*) is in the dative since the accusative case is not available. However, in (6.21) it is the thematically most prominent Beneficiary that takes the accusative case.

6.5 Dependency grammars

The dependency tree format differs from the constituent tree format inasmuch as every node in the tree corresponds to a word (or a morpheme) in the sentence. On the top of the sentence tree a virtual root node can be found in certain dependency grammars to which words in the sentence are subordinated, that is, no abstract nodes can be found apart from the root node

(if any). Every word in the sentence is strictly subordinated to another one: a word can only have one superordinate, however, there can be several words below a node, e.g. all the arguments of a verb fall under the verb node. Nodes in the dependency tree can have diverse relations, usually tagged to denote the nature of the particular relation.

Tesnière's book (Tesnière, 1959) is considered to be the first dependency grammar, which lays the foundations of the theory. According to his famous metaphor, the verb is the central element of the sentence, which "expresses a whole little drama": the arguments of the verb are the actors, which Tesnière calls actants. Consequently, in a sentence subordinated and superordinated elements are integrated into a unit.

Koutny and Wacha (1991) and Prószéky et al. (1989) give a summary of a dependency grammar for Hungarian and the authors briefly outline their morpheme-based dependency grammar. In their model, morphemes are the basic constituents of dependency trees since in agglutinative languages not (only) words but morphemes too are capable of expressing different grammatical relations. This solution facilitates mapping between dependency trees of different types of languages because the node of e.g. the auxiliary *may* in English corresponds to the node of the morpheme *-hAt* in the Hungarian tree. This procedure may greatly enhance the efficiency of dependency grammar-based translation systems.

The dependency analysis for semi-compositional constructions is presented on the basis of Alonso Ramos's work (1998; 2007). The framework she applies is the Meaning–Text Theory (see e.g. Mel'čuk et al. (1995), Mel'čuk (2004b; 2004a)).

The Meaning–Text Model is made up from representational levels. Except for the semantic level, each level is divided into two sublevels (a deep and a surface level), thus there are seven levels altogether. On each level, the deep level emphasizes the semantic differences while the surface level focuses on formal differences.

The deep syntactic level represents the syntactic structure of the sentence as a dependency tree. There are generalized lexemes in the nodes of the tree and they can be classified into four groups: (1) semantically full lexemes, (2) fictive lexemes (e.g. the symbol for the pronoun expressing indefinite subject), (3) idioms and (4) lexical functions (see the example of **Oper** below). The edges of the tree represent deep syntactic relations: relations between the predicate and its actants (they are marked with Roman numerals: I–VI), attributive relations marking several types of modifiers (ATTR), coordination (COORD) and quasi-coordination (QUASI-COORD), appenditive relation marking elements outside the structure of the sentence (e.g. exclamations, addressing terms) (APPEND) and reported speech.

On the surface syntactic level, there is also a dependency tree, however, all the lexemes that occur within the phrase in the given language are present in it. The edges of the tree reflect the syntactic relations of the given language.

In the model, it is the lexical function **Oper** (among others) that expresses the relation between the nominal component and the verb the meaning of which is 'to do X, to have X or to be in the state X' (Apresjan, 2004) (see Chapter 7 for details). On the deep syntactic level, the lexical function stands for the verb and it is replaced by the specific lexical item (i.e. the verbal component) on the surface syntactic level. Thus, on the deep syntactic level the noun (C_0) has arguments (or rather actants, following the terminology of the theoretical framework).

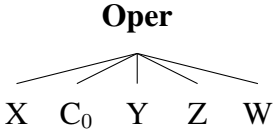
Mel'čuk's (1988; 2003; 2004b) dependency grammar emerged within the Meaning-Text Theory. In this framework, dependency appears as a linear relation between words. On the deep syntactic level, he assumes twelve relation types, out of which six exist between the verb and its various arguments (actants) and the other relations designate coordination and diverse modifying roles. The heart of Mel'čuk's dependency grammar is that it interprets coordination as a kind of subordination: the conjunction is connected to the first member of coordination and the other member(s) of the coordination are connected to the latter with a special (COORD) relation. Another peculiarity of this approach is that in certain cases this grammar permits the insertion of nodes denoting abstract, that is, phonetically non-overt linguistic elements into the dependency tree: such is the case with the copula in Russian (and in Hungarian as well) in third person singular, present tense, which does not become overt in the sentence phonetically still it is there on an abstract level since it becomes manifest in past and future tenses.

6.5.1 The distribution of actants on the deep syntactic level

The distribution of the actants on the deep syntactic level can be carried out in the following way (Alonso Ramos, 2007). If the keyword (the nominal component, denoted by C_0) has four semantic actants, the actants can be distributed in four different ways in principle.

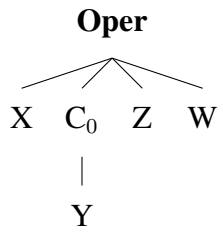
First, it can happen that all the four actants are connected to the verb:

(6.22)



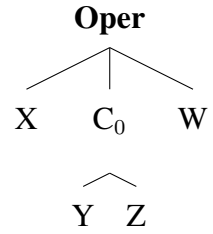
Second, the keyword has only one actant, and the other three are connected to the verb:

(6.23)



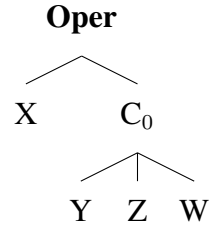
Third, both the keyword and the verb have two actants:

(6.24)



Fourth, the verb has one actant while the keyword has three:

(6.25)



However, on the surface syntactic level, it is preferably the option described in (6.24) that manifests in Spanish, see 6.5.2.

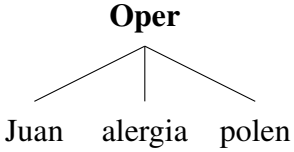
If the noun has three actants, two possibilities are offered. Let our examples be the sentences in (6.26) and (6.27).

(6.26) *Juan le tiene alergia al polen.*
John for.him has allergy for.the pollen
'John suffers from pollen allergy.'

(6.27) *Juan sufre de alergia al polen.*
John suffers from allergy for.the pollen
'John suffers from pollen allergy'.

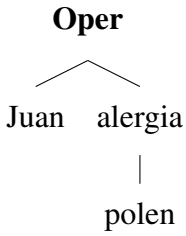
In the first case, each actant belongs to the verb:

(6.28)



In the second one, *polen* ‘pollen’ belongs to *alergia* ‘allergy’:

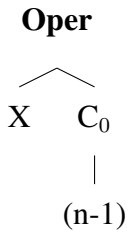
(6.29)



On the surface syntactic level a verb substitutes the lexical function **Oper**: in Spanish it is selected from *tener* ‘to possess’ and *sufrir* ‘to suffer’. *Tener* ‘to possess’ has three actants whereas *sufrir* ‘to suffer’ has two – in this way, the surface syntactic structure is determined by the selection of the verb.

However, on the deep syntactic level there is no choice: only one of the theoretical possibilities can be realized. Thus, **Oper** will have two actants since the relation between *polen* ‘pollen’ and **Oper** is not linguistically motivated (as opposed to the relation between *polen* ‘pollen’ and *alergia* ‘allergy’ – *polen* ‘pollen’ can be seen as an argument of *alergia* ‘allergy’). Thus, the default deep syntactic representation of semi-compositional constructions is the following (C_0 represents the nominal component and n stands for the number of actants):

(6.30)



6.5.2 The distribution of actants on the surface syntactic level

The distribution of arguments between the noun and the verb can also be accounted for within a dependency framework. There are certain tendencies observed in Spanish for the distribution of actants (Alonso Ramos, 2007). The starting point is the number of the semantic actants of the noun and some of them are transferred to the verb following the tendencies described below.

If the nominal component has one semantic actant, the verb has two actants on the surface syntactic level (compare 6.23):

(6.31) *Juan ha dado un paseo.*
John has given a walk
‘John had a walk.’

The nominal component manifests as the subject in the surface syntactic level, and the other actant of the verb is the nominal component itself.

If the noun has two semantic actants, the first one belongs to the verb (more precisely, it functions as its subject) and the second one may belong to either the verb or the noun in the surface syntactic structure. In the first case, it is connected to the noun *miedo* 'fear' with the preposition *de* similarly to (6.29):

(6.32) *Juan tiene miedo de María.*
John has fear of Mary
'John fears Mary.'

In the second case, it is connected to the verb with the preposition *a* 'for', at the same time, the dative pronoun *le* appears before the verb, signaling the relation of the actant to the verb with another grammatical device (cf. (6.29)):

(6.33) *Juan le tiene miedo a María.*
John for.him has fear for Mary
'John fears Mary.'

If the noun has three semantic actants, the first one and the third one belong to the verb while the second one to the noun in the surface syntactic structure:

(6.34) *Juan ha hecho una oferta de dinero a María.*
John has made an offer of money for Mary
'John made a financial offer to Mary.'

The first actant is the subject, the actant of the noun is connected to the noun with the preposition *de* 'of' and the third one is connected to the verb with the preposition *a* 'for'.

If the noun has four semantic actants, the first, third and fourth one belong to the verb and the second one to the noun in the surface syntactic structure:

(6.35) *Por este coche, el secretario hizo un pago de cien mil*
for this car the secretary made a payment of hundred thousand
pesetas a Juan.
peseta-PL for John
'For this car, the secretary made a payment of a hundred thousand pesetas to John.'

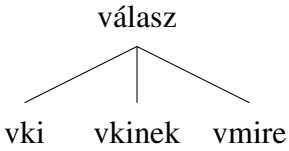
The first actant functions as the subject, the second one is connected to the noun with the preposition *de* 'of', the third and the fourth ones are connected to the verb with the prepositions *a* 'for' and *por* 'for', respectively.

A significant part of the argument alternations between the construction and its verbal counterpart can also be accounted for within a generative framework (see 6.4.1 and 6.4.2). This question was not examined by Alonso Ramos (1998), however, it would be interesting to see what a dependency grammar has to offer. The derivational processes between the nominal component and the verbal counterpart are also worth examining within a dependency framework. In the following, these issues will be analyzed.

6.5.3 Semi-compositional constructions and their verbal counterparts

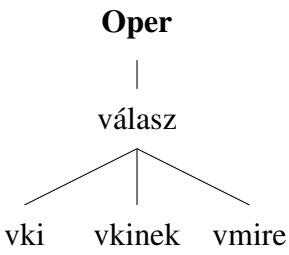
The starting point of the analysis offered by Alonso Ramos (1998) is the nominal component, which has several actants. There are two ways for verbalizing the meaning conveyed by the nominal component: first, a verb can be derived from the noun (derivation) and second, a light verb can be attached to the noun, yielding a semi-compositional construction.[1] Let us now examine the noun *válasz* 'answer'. As a semantic predicate (see Mel'čuk (2004a)), it has three actants: *vki* 'someone', *vkinek* 'to someone' and *vmire* 'for sg'.

(6.36)



If this noun is verbalized, it can be formalized on the deep syntactic level by applying the verbal lexical function **Oper**:

(6.37)

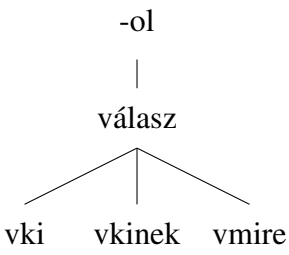


This structure can be transformed into two different surface structures by substituting **Oper** with either a suffix or a light verb. First, the verb derived from it is *válaszol* 'to answer' and second, the corresponding semi-compositional construction is *választ ad* (answer-ACC gives) 'to give an answer'. In a morpheme-based dependency grammar (for Hungarian, see

[1] As Bolshakov and Gelbukh (1998) point out, the verbal derivational suffix attached to the noun has exactly the same role as the light verb.

Prószéky et al. (1989) and Koutny and Wacha (1991)), *válaszol* ‘to answer’ can be derived as follows:[2]

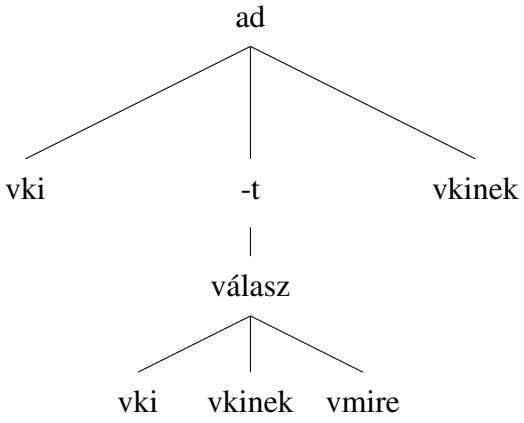
(6.38)



That is, the noun *válasz* ‘answer’ is attached to a verbal derivational suffix.

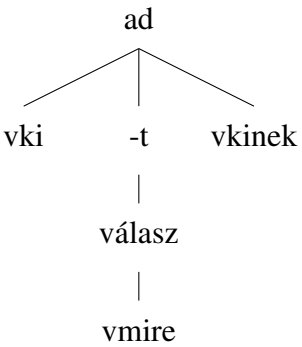
In the second case, *válasz* ‘answer’ is paired with the verb *ad* ‘give’ in order to produce a semi-compositional construction:

(6.39)



In this case, both *válasz* ‘answer’ and *ad* ‘give’ have an argument in the dative. When forming the semi-compositional construction, function composition is applied: following the scheme X/Y + Y/Z = X/Z, the dative argument of the nominal component fulfills the position of the dative argument. Thus, the semi-compositional construction will have the actants *vki* ‘someone’, *vkinek* ‘to someone’ and *vmire* ‘for sg’.

(6.40)

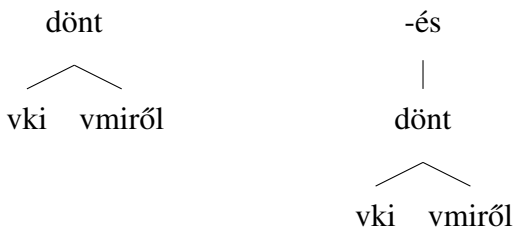


[2]The suffixes of the actants are not segmented from the root in the following examples for the sake of simplicity.

In the construction, the nominal component functions as the semantic predicate, which means that the noun gives semantic roles to its arguments on the semantic level (i.e. *vki* is the one who answers, *vkinek* is the one who asked a question and *vmire* is the question). Since there is no verbal element in the deep syntactic structure (only a lexical function being present and the verb being materialized later, i.e. only in the surface structure), it cannot assign any role to its arguments.

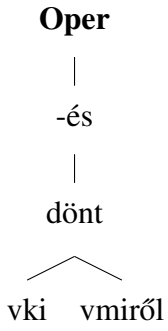
If the nominal component of the construction is derived from the verbal counterpart (e.g. *dönt* 'to decide' and *döntést hoz* (decision-ACC bring) 'to take a decision'), the following analysis can be provided. The starting point in this case is the verb *dönt* 'decide', a semantic predicate, which then undergoes nominalization by adding a nominal suffix:

(6.41)



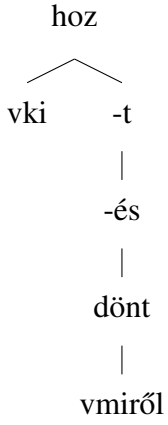
If the noun *döntés* 'decision' with two actants (*vki* 'someone' and *vmiről* 'on sg') is verbalized, it is first added the lexical function **Oper** on the deep syntactic level:

(6.42)



On the surface syntactic level, it is then paired with the verb *hoz* 'bring' in order to yield a semi-compositional construction:

(6.43)



On the deep syntactic level, actants belong to the nominal component whereas on the surface syntactic level, actants move to the verb in each case. In this construction, *dönt* assigns semantic roles to its actants on the semantic level.

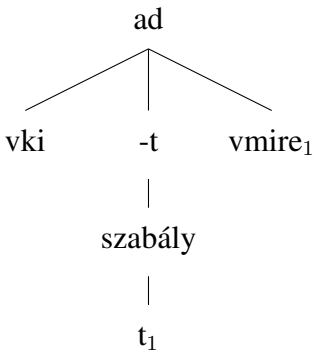
In the above cases, the argument structure of the verbal counterpart and the semi-compositional construction was the same. Let us examine another pair where the arguments bear different grammatical cases on the surface:

(6.44) *szabályt* *ad* *vmire*
rule-ACC gives sg-SUB
'to give a rule for sg'

szabályoz *vmit*
rules sg-ACC
'to rule sg'

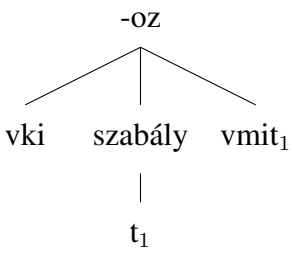
The semi-compositional construction contains the subject, the nominal component in the accusative case and its actant in the sublative case:

(6.45)



However, the verbal component has the subject and one actant in the accusative case:

(6.46)



From the above examples it can be seen that in both cases, there is one noun in the accusative case, in (6.45) it is the nominal component whereas in (6.46) it is the actant. This phenomenon is in accordance with the tendency that whenever a verb has only one argument (except for the subject), it is usually in the accusative case (see 6.4.3).

6.6 Comparing the analyses offered

In this chapter semi-compositional constructions have been discussed from a syntactic point of view. Several analyses have been provided in the frameworks of constituency and dependency grammars as well as argument structure and the distribution of actants / arguments have been put under special investigation. The comparison of the analyses offered can shed light on the following issues.

Most analyses pay attention to argument structure. Grimshaw and Mester (1988) explain the transfer of the arguments of the noun to the verb with the rules and constraints of argument transfer, Meyers et al. (2004a) apply the term argument sharing while Alonso Ramos (1998) describes typical tendencies for the distribution of actants on the surface syntactic level. Basically, all the analyses discussed distinguish between a deep and a surface representation of the construction: on the deep level the noun has arguments, which are later shared with the verb (Meyers et al., 2004a) or transferred to the verb (Grimshaw and Mester, 1988; Alonso Ramos, 1998) – the same phenomenon is called differently in different approaches.

Similar consequences were drawn on the basis of data from different languages, e.g. Japanese (Grimshaw and Mester, 1988), Spanish (Alonso Ramos, 1998), English (Meyers et al., 2004a), and in Hungarian as well.

Different analyses provide different solutions for the derivation of verbal counterparts and nominal components. Constructions with a deverbal nominal component can be derived from the verbal counterpart in our analysis within a generative framework (6.4.2) while Hale and Keyser (2002) derive denominal verbal counterparts from the nominal components – the two analyses jointly are able to explain both ways of derivation. On the other hand, analyses within the framework of dependency grammar can start from both the nominal component and the verb (as it was shown through some Hungarian examples).

It would be also necessary to harmonize the results achieved in different theoretical frameworks or analyses. A successful example for harmonizing phrase structure grammars and dependency grammars is offered by the Szeged Treebank, which originally contained phrase structure trees but they were converted into dependency trees applying certain conversion rules and restrictions – the automatic conversion was followed by a manual checking phase (Alexin, 2007; Vincze et al., 2010b). Rules and constraints on argument transfer, and argument sharing / the distribution of actants (Grimshaw and Mester, 1988; Meyers et al.,

2004a; Alonso Ramos, 1998) should be unified, even independently of theoretical frameworks. It must be mentioned that on the deep syntactic level, all arguments belong to the noun whereas on the surface syntactic level, some of them move to the verb. These rules of dependency grammar seem to be parallel to the movements described in the framework of generative grammars (cf. also raising and control verbs) hence the idea of the arguments belonging to the noun can be seen as a common (theory-independent) background and the exact rules of argument transfer (i.e. which argument should move to the verb) are construction-specific.

6.7 Semi-compositional constructions as complex predicates: an alternative to argument sharing

Although semi-compositional constructions are made of two parts, namely, the nominal component and the verb, thus, they show phrasal properties, it can be argued that from a semantic point of view they form one unit. First, many semi-compositional constructions have a verbal counterpart with the same meaning (see Chapter 5). Second, there are meanings that can only be expressed through a semi-compositional construction (e.g. *házkutatást tart* (search.of.premises-ACC hold) 'to conduct search of premises' in Hungarian). Third, there are languages that abound in verb + noun constructions or multiword verbs (Hindi (Sinha, 2009; Sinha, 2011), Bengali (Das et al., 2010), Estonian (Kaalep and Muischnek, 2006; Kaalep and Muischnek, 2008; Muischnek and Kaalep, 2010), Persian (Mansoory and Bijankhan, 2008)): verbal concepts are mostly expressed by combining a noun with a light verb (Mansoory and Bijankhan, 2008).

On the other hand, there are views that the relationship between the verbal and the nominal component is not that of a normal argument. For instance, Chomsky (1981, p.37) calls *advantage* a quasi-argument of *take* in the idiom *take advantage of*.[3] Alonso Ramos (1998) proposes the role of quasi-object for the nominal component of certain Persian constructions (which she later adopts to Spanish) with the following properties:

- they must be adjacent to the verb,
- they cannot have a determiner,

[3] In our view, *take advantage of* is a semi-compositional construction rather than an idiom.

- no modification is possible for them,
- they cannot undergo passivization,
- they cannot be substituted by a pronoun.

According to her, this relationship holds between parts of more idiomatic constructions. This is in accordance with Chomsky's usage of the term *idiom*. However, the term *quasi-argument* might be extended to signal the relationship between the verbal and the nominal components of less idiomatic semi-compositional constructions as well since they behave as a semantic unit, forming one complex predicate.[4] An advantage of this solution is that the question of argument transfer is eliminated: it is the construction that has arguments and not the verb or the noun on their own. The semantic analysis of the construction also becomes easier in this way since the syntactic structure reflects more truly the semantic relations. Finally, higher-level NLP applications can also profit from this solution because the identification of semi-compositional constructions can be enhanced in this way (Chapter 9), which has impact on e.g. information extraction (Chapter 11) and machine translation (Chapter 12).

6.8 Summary of results

In this chapter, syntactic analyses of semi-compositional constructions were provided in the frameworks of constituency and dependency grammars and special attention was paid to the alternations in the argument structure and the derivational processes between the construction and its verbal counterpart.

The following points were highlighted:

- syntactic features of semi-compositional constructions were discussed;
- syntactic analyses within generative and dependency frameworks were provided;
- issues of argument transfer, argument sharing and the distribution of actants were presented;

[4]With the above extension, some of the properties can be also violated in the case of certain semi-compositional constructions: e.g. some nominal components can be modified.

- an alternative representation was also proposed in order to treat semi-compositional constructions as complex predicates.

The syntactic representation of semi-compositional constructions is not only a question of theory – it has important implications for practical applications as well, which will be discussed later in the chapters on natural language processing (Part III).

Chapter 7

The semantics of semi-compositional constructions

7.1 Introduction

In this chapter, the semantics of semi-compositional constructions are discussed in detail within the framework of Meaning–Text Theory. The research questions to be answered are the following (cf. Chapter 1):

- What are the semantic features of semi-compositional constructions?
- How can the semantic relation of the two components of the construction be described?

Semantic features of the constructions and correlations between the verb and the semantic type of the noun are analyzed with the help of lexical functions, which are presented first. A thorough semantic analysis of data on constructions containing the four most frequent verbs is carried out. Finally, questions of aspect, Aktionsart, synonymy and conversion are also discussed.

7.2 Lexical functions

The theory of lexical functions was born within the framework of Meaning-Text Theory (the model is described in detail in e.g. Mel'čuk (1974; 1989; 1996; 1998; 2004a), Mel'čuk et al. (1984; 1995; 1984 1999) and Wanner (1997; 2007). The most important theoretical

innovation of this model is the theory of lexical functions, which is universal: with the help of lexical functions, all semantic relations between lexemes of a given language can be described. Although this theory has been thoroughly applied to different languages such as Russian, French, English or German, it has been rarely adapted to Hungarian: so far, it is only the applicability of **Magn** that has been studied (Répási and Székely, 1998; Székely, 2003).

Lexical functions have the form **f**(x) = y, where f is the lexical function itself, x stands for the argument of the function and y is the value of the function. The argument of the lexical function is a lexeme, while its value is another lexeme or a set of lexemes. A given lexical function always expresses the same semanto-syntactic relation, that is, the relation between an argument and the value of the lexical function is the same as the relation between another argument and value of the same lexical function. Thus, lexical functions express semantic relations between lexemes. In the case of syntagmatic lexical functions, these relations hold between expressions that are not totally compositional, that is, they must be learnt (Mel'čuk et al., 1995).

7.2.1 Verbal lexical functions

Since this thesis focuses on semi-compositional constructions, i.e. a combination of a noun and a verb, verbal lexical functions are presented in detail (Mel'čuk et al., 1995).

In the case of **Oper**, the keyword of the lexical function is the object. The subject of the sentence is the first actant of the verb. For instance:

(7.1) **Oper**$_1$ (*lehetőség*) = [~et] ad[1]

(7.2) **Oper**$_1$ (*engedély*) = [~t] ad[2].

The keyword of the lexical function, the name of the situation (*engedély* 'permission') can be paraphrased with the construction *vki (1) engedélyez vkinek (2) vmit (3)* 'sy (1) permits sg (3) to sy (2)' where actants are signaled with numbers. The first actant of the situation functions as the subject in the semi-compositional construction that is why the lexical function **Oper** is indexed with 1. If the second actant of the verb is the subject – i.e. the lexical function **Oper**$_2$ is applied to the same keyword –, another semi-compositional construction is yielded:

[1] *lehetőséget ad* (opportunity-ACC gives) 'to provide an opportunity'
[2] *engedélyt ad* (permission-ACC gives) 'to give permission'

(7.3) **Oper**$_2$ (*engedély*) = [~t] kap[3].

Diathesis is often characteristic of verbs being the values of **Oper**$_2$: *kerül* 'get', *részesül* 'receive', *lel* 'find' etc. (cf. Chapter 5).

If the keyword is the subject of the semi-compositional construction, and its object is the second argument of the verb, the lexical function **Func**$_2$ is applied:

(7.4) **Func**$_2$ (*lehetőség*) = [~] nyílik [vmire][4]

In the case of **Labor**$_{ij}$, the i-th participant of the situation is the first (deep) actant (and subject) of the semi-compositional construction while the j-th participant of the situation is the second actant and first complement of the verb and the name of the situation is the third deep actant and the second complement of the verb. For instance:

(7.5) **Labor**$_{12}$ (*ebéd*) = eszik [vmit ~re][5]

In other words: the first actant of the expression *vki (1) ebédel vmit (2)* 'sy (1) has sg (2) for lunch' denoting the situation described by the keyword is the subject in the semi-compositional construction, its second actant is the object and the name of the situation is the second complement of the verbal lexical function.

The lexical functions **Real**, **Fact** and **Labreal** are syntactic equivalents of **Oper**, **Func** and **Labor**, respectively, however, they differ in their meaning: they refer to the fulfilment of the requirements encoded in the meaning of the keyword.

Other lexical functions are related to causativity: **Caus** means the causing of the situation, **Perm** permits that the situation exists and **Liqu** eliminates the situation. The beginning of an event is encoded by **Incep**, its continuation is referred to by **Cont** whereas **Fin** expresses the cessation of a situation. These groups of lexical functions usually attach to other lexical functions from the group of **Oper**, **Func** or **Labor**.

Some examples are provided here for applying different lexical functions for the same keyword:

(7.6) **IncepLabor**$_{12}$ (*kapcsolat*) = [~ba] lép[6]

[3] *engedélyt kap* (permission-ACC gets) 'to get permission'
[4] *lehetőség* nyílik vmire (opportunity opens sg-SUB) 'an opportunity emerges for sg'
[5] *eszik vmit ebédre* (eats sg-ACC lunch-SUB) 'to have sg for lunch'
[6] *kapcsolatba lép* (connection-ILL steps) 'to connect'

(7.7) **ContOper**$_1$ (*kapcsolat*) = tart [~ot][7]

(7.8) **FinOper**$_1$ (*kapcsolat*) = megszakítja a [~ot][8]

These examples illustrate that differences in Aktionsart can also be encoded by lexical functions.

7.2.2 Lexical functions and semi-compositional constructions

Research on the relationship of semi-compositional constructions and lexical functions has been rarely conducted. One of the few exceptions is Reuther (1996), who focuses on three Russian support verbs – *vesti* 'lead', *provodit'* 'conduct', and *proizvodit'* 'manufacture'. He uses the Russian explanatory-combinatorial dictionary (Mel'čuk and Žolkovskij, 1984) as a source of data. In his examples, the relation between the nominal component and the support verb can be formalized with the help of **Oper**$_1$. He concludes that nominal components can be divided into definite semantic groups in the case of all the three verbs: for instance, the support verb *provodit'* 'conduct' occurs together with nouns denoting an organized social activity or a complex procedure.

Apresjan (2004) examines Russian verbal constructions that can be related to different lexical functions. He claims that there is a correlation between the given lexical function and, on the one hand, the type of the predicate, and, on the other hand, the semantic type of the nominal component of the construction. Since the meaning of the lexical function of **Oper** is given as follows: "delat' X, imet' X ili byt' v sostojanii X" 'to do X, to have X or to be in the state X' (Apresjan, 2004, p. 6), the values of **Oper**$_1$ will be such verbs whose meaning contains the element 'do' according to the definition.

Studies on Russian material suggest that there is a correlation between the verb and the semantic type of the noun on the one hand, and between the lexical function and the type of the verb on the other hand (see Apresjan (2009) on the general law of semantic agreement). In the following, we will examine whether these relations hold for Hungarian semi-compositional constructions as well.

For analysis, semi-compositional constructions containing one of the four most frequent verbal components (*ad* 'give', *vesz* 'take', *hoz* 'bring', *tesz* 'do' (see Chapter 3)) have been selected for Hungarian. Data on the above verbs are provided in Table 7.1.

[7] *kapcsolatot tart* (connection-ACC holds) 'to keep in touch'

[8] *megszakítja a kapcsolatot* (PREVERB-cease-3SGOBJ connection-ACC) 'to cease the connection'

Verb	Number of FX	Number of nouns
ad	1526	164
vesz	916	50
hoz	836	51
tesz	557	67

Table 7.1: Data on verbs *ad* 'give', *vesz* 'take', *hoz* 'bring' and *tesz* 'do'

7.3 Data analysis

The corpus data can be described with the help of standard and complex lexical functions. The relations between the verbs and the nouns are formalized in terms of lexical functions, that is, it is revealed which verb is the value of which lexical function in the case of a specific noun. Thus, correlations between verbs and semantic classes of nouns can be formulated: namely, special semantic classes co-occur with certain light verbs.

7.3.1 Constructions with the verb *ad* 'give'

The most frequent verbal component in the corpus was *ad* 'give'. The entry of the verb **ad**[1] contains seven groups of senses[9] in The Explanatory Dictionary of the Hungarian Language (henceforth ÉrtSz.) (Bárczi and Országh (1959 1962), I. 19–22), out of which group VII is of primary importance with regard to semi-compositional constructions. Section VII.1. contains several semi-compositional constructions where *ad* co-occurs with nouns denoting actions in the accusative case and the meaning of the verb is "Azt a cselekvést végzi, amire a fn-i tárgy utal" 'to perform the action encoded in the nominal object'.

The second edition of ÉKsz. (Pusztai, 2003) offers sixteen meanings of the verb *ad* 'give'. Senses 8 and 14 are related to semi-compositional constructions: "Létrehoz, eredményez" 'to create, to yield' and "A megnevezett cselekvést végzi, teljesíti" 'to perform the action mentioned'.

The nominal components of semi-compositional constructions containing this verb mostly occurred in the object position (in 150 cases): it entails that in such cases, $\mathbf{Oper}_1$ is applied. Recalling Apresjan's (2004) definition of the semantic content of $\mathbf{Oper}_1$ ("to do X, to have X or to be in the state X"), a parallel can be drawn between this definition and the one found in ÉrtSz. ('to perform the action encoded in the nominal object'). Both definitions comprise

[9] **Ad**[2] used as a Latinate preposition is not considered here.

the component 'do action X', which reveals that the semantic content of the lexical function **Oper**$_1$ and its value (now the verb *ad* 'give') (at least partially) overlap.

For instance, *támogatást ad* (support-ACC gives) 'to give support' can be formalized as:

(7.9) **Oper**$_1$ (*támogatás*) = [~t] ad

In seven cases, the nominal component is in the illative case and the rest of the nominal components bears another oblique case (e.g. *férjhez ad* (husband-ALL gives) 'to cause to marry').

If the semantic characteristics of nouns occurring with semi-compositional constructions described by **Oper**$_1$ containing the verb *ad* 'give' are examined, it is revealed that their meaning is mostly abstract. Many of them contain the derivational suffix *-Ás*, which refer to some kind of action or activity, for instance:

(7.10) *ismertetést* *ad*
review-ACC gives
'to give a review'

megbízást *ad*
assignment-ACC gives
'to give an assignment'

Other deverbal nouns also occur in the dataset referring to some kind of action, e.g.:

(7.11) *engedélyt* *ad*
permission-ACC gives
'to give permission'

Nominal components occurring with *ad* can be classified into well-defined semantic groups. The semantic groups of nouns being the keywords of **Oper**$_1$ are listed here:

- nouns denoting speech acts or verbally performed actions (*biztosíték* 'caution', *definíció* 'definition', *információ* 'information', *magyarázat* 'explanation', *részletezés* 'specification', *tájékoztatás* 'informing', *útmutatás* 'guidance', *válasz* 'answer', *vélemény* 'opinion')

  These types of nouns are often paired with the verb *davat'* 'give' in Russian (Apresjan, 2004), the Hungarian translation of which is 'ad'.

- nouns denoting permission (*engedély* 'permission', *felmentés* 'release', *haladék* 'moratorium', *hozzájárulás* 'contribution', *jogosítás* 'legalization', *szabály* 'rule')

  In the language of lexical functions, their meaning contains the permissive lexical function **Perm** (the meaning of which is 'to permit that the given situation exists').

- nouns denoting a possibility (*alkalom* 'opportunity', *lehetőség* 'possibility', *mód* 'way')

- nouns denoting social events (*koncert* 'concert', *est* 'evening')

- nouns denoting virtues (*remény* 'hope', *erő* 'strength', *szeretet* 'love', *bátorság* 'courage')

- nouns denoting financial resources (*hitel* 'credit', *fedezet* 'cover', *kölcsön* 'loan')

Semi-compositional constructions with nouns in an oblique case can be described by **Labor**$_{12}$. Some of the nominal components in the illative case are related to the transfer of possession/usage (*bér* 'rent', *használat* 'usage', *kölcsön* 'loan'), and they often occur with the verb *vesz* 'take' as well (cf. 7.3.2). It should be mentioned that there are some nominal components that cannot be classified according to the above categorization, however, the categories can be seen as tendencies that show some similarity with other languages (cf. Russian).

7.3.2 Constructions with the verb *vesz* 'take'

The ÉrtSz. lists six groups of senses of **vesz**[1] (VII. 361-365).[10] Concerning semi-compositional constructions, it is group III that merits special attention since definitions similar to the ones mentioned at the entry of *ad* can be found there:

> "Vmely tevékenységre kiszemel v. vmely tevékenység céljára használni kezd [...] vmely cselekvésre, munkája v. értelmi tevékenysége körébe vonja; elkezdi rajta v. általa azt a cselekvést, amelyre a fn-i határozó utal" 'to select sg for an action or to start using sg for the purpose of an activity [...] to involve sg in an action, work or mental activity; to start performing the action denoted by the nominal adverbial construction' (VII. 364).

The sense given in IV.1. is also comparable: "Egyszer v. rendszeresen használ vmit" 'to use sg once or regularly'.

[10] **Vesz**[2] 'to waste, to be lost' is not considered here.

The ÉKsz. lists nineteen senses of **vesz**[1] (1451–1452), out of which senses 13 and 15 contain the definition relevant for our purposes: "vmire kiszemel, használni kezd [...] Vmely tevékenység körébe von, vmit művel vele [...] Vmit csinál, cselekszik" 'to select sg for sg, to start using sg [...] to involve sg in an activity, to do sg with sg [...] to do sg' and "Cselekvés eredményeképpen kap, ill. létrehoz vmit" 'to get or create sg as a result of an action'.

An important feature of constructions containing the verb *vesz* 'take' is that they refer to the beginning of the event, that is, the action described by the noun exhibits inchoative Aktionsart (see Chapter 5). This meaning component is represented by the lexical function **Incep**, which means 'to start the event'. Thus, here again the definition of the verb and the lexical function shares a semantic component.

Most of the semi-compositional constructions formed with the verb *vesz* 'take' require the presence of a subject and an object in the sentence and they fulfill the requirements encoded in the meaning of the noun (21 nominal components in the illative case and 8 in other oblique cases). Among lexical functions, it is **Labreal** that expresses such relation.

As mentioned above, the beginning of the action/activity is specially emphasized, thus, the lexical function **Incep** referring to the beginning of the event serves as the other component of the complex lexical function:

(7.12) **IncepLabreal**$_{12}$ (*számítás*) = [~ba] vesz[11]

The nominal components form semantic groups here as well:

- nouns denoting transfer of possession or usage (*bér* 'rent', *birtok* 'possession', *igény* 'claim', *tulajdon* 'possession')
- nouns related to accounting (*nyilvántartás* 'register', *leltár* 'inventory')
- nouns with the meaning 'considering something' (*figyelem* 'consideration', *számítás* 'account', *tekintet* 'consideration')

Other 18 nominal components occur in the accusative case next to the verb *vesz* 'take'. Their meaning typically does not contain the elements 'to begin' and 'to fulfill', thus, it is the lexical function **Oper** that attaches the nominal component to its verb. However, if the semi-compositional constructions belonging to this group are analyzed more deeply, it can

[11] *számításba vesz* (account-ILL takes) 'to take into account'

be revealed that in some cases it is the first actant of the situation that is the subject of the construction (*lendület* 'impetus', *búcsú* 'leave', *zuhany* 'shower' etc.), applying the lexical function $\mathbf{Oper}_1$, while in other cases, the subject of the construction is the second actant of the situation (*példa* 'example', *táncóra* 'dance class' etc.), i.e. $\mathbf{Oper}_2$ is used (conversion of $\mathbf{Oper}_1$, for the notion of conversion, see 7.6.2). This can be formalized with the help of the lexical function **Conv** (subscripts denote the reorganization of actants):

(7.13) $\mathbf{Conv}_{21}(\mathbf{Oper}_1(\textit{példa}))$ = [~t] vesz

As for the keywords of $\mathbf{Oper}_2$, most of them occur with the verb *ad* 'give' in the corpora as well, forming pairs such as:

(7.14) *órát* *ad*
lesson-ACC gives
'to give a lesson'

órát *vesz*
lesson-ACC takes
'to take lessons'

They differ only in the fact that either the first or the second actant of the situation functions as the subject of the construction.

7.3.3 Constructions with the verb *hoz* 'bring'

33 nominal components paired with the verb *hoz* 'bring' function as the object of the verb hence they are connected to the verb with the lexical functions **Oper** or **Real**. The following criterion might be applied to choose between the two possibilities: if another semi-compositional construction can be found the meaning of which is to fulfill the requirements encoded in the meaning of the noun, then the components of the given construction are connected to each other with **Oper**. For example:

(7.15) $\mathbf{Oper}_1$ (*ítélet*) = [~et] hoz[12]

cf.

(7.16) $\mathbf{Real}_2$ (*ítélet*) = végrehajtja [az ~et][13]

[12] *ítéletet hoz* (verdict-ACC brings) 'to make a verdict'
[13] *végrehajtja az ítéletet* (realize-3SGOBJ the verdict-ACC) 'to realize a verdict'

With this verb, the nouns *döntés* 'decision' and *határozat* 'verdict' frequently occurred, which – on the other hand – were also paired with the verb *végrehajt* 'realize' as well. If the two instances are compared, the difference between the two lexical functions is revealed. In the **Oper** construction only the creation of the decision is encoded (i.e. the decision is born the same moment) while in the **Real** construction the decision created before is realized or fulfilled.

(7.17) $\mathbf{Oper}_1$ (*döntés*) = [~t] hoz[14]

(7.18) $\mathbf{Real}_2$ (*döntés*) = végrehajtja [a ~t][15]

On the basis of the above examples, all the constructions mentioned above can be described with the help of $\mathbf{Oper}_1$. As for the dictionary entry of *hoz* 'bring', there are three groups of meaning in the ÉrtSz. (III. 350–352). In section II.4., semi-compositional constructions are referred to: "állandósult szókapcsolatokban, kül. a *hoz*-zal kapcs. névszóból alkotható igével egyszerűbben kifejezhető jelentésben" 'in fixed expressions, esp. in the sense more simply expressed by the verb derived from the noun connected to *hoz*'. The nominal components can be grouped in the following way:

- nouns denoting speech acts or verbally performed actions
  - nouns with the meaning 'decision' (*döntés* 'decision', *határozat* 'verdict', *intézkedés* 'measure' etc.)
  - nouns with the meaning 'rule, obligation' (*szabály* 'rule', *törvény* 'law', *rendelet* 'decree' etc.)

  This semantic group is in overlap with one group of keywords in 7.3.1: similarly to them, these nouns also denote speech acts or verbally performed actions. Thus, nouns belonging to this semantic group can be attached to two verbs as well, namely, *ad* 'give' and *hoz* 'bring', however, the choice between the two is lexically restricted. There are few nouns that occur with both verbs in the dataset (e.g. *szabály* 'rule'), however, this is a rare phenomenon and is considered as exceptional (see also 7.6.1).
- nouns denoting (positive or negative) change (*fordulat* 'turn', *baj* 'trouble', *siker* 'success', *változás* 'change' etc.)

[14] *döntést hoz* (decision-ACC brings) 'to make a decision'
[15] *végrehajtja a döntést* (realize-3SGOBJ the decision-ACC) 'to realize a decision'

- nouns denoting financial profit (*profit* ‘gain’, *bevétel* ‘income’, *nyereség* ‘profit’ etc.)

Keywords of the verb *hoz* can also appear in positions other than the object. In certain constructions they require a subject and an object as well hence they are described by the lexical functions **Labreal** or **Labor** (the difference between them corresponds to the one between **Oper** and **Real**, see above):

(7.19) *forgalomba* *hoz*
circulation-ILL brings
‘to put into circulation’

mozgásba *hoz*
motion-ILL brings
‘to put in motion’

nyilvánosságra *hoz*
publicity-SUB brings
‘to publish’

összhangba *hoz*
accordance-ILL brings
‘to bring into harmony’

tudomására *hoz*
knowledge-3SGPOSS-SUB brings
‘to acquaint’

In these examples, an agent must act in order to create the action described by the keyword (in other words, to initiate a process). That is why the lexical functions describing these constructions are somewhat complex: they contain the element **Caus** and **Incep** denoting causation and beginning of an event, respectively. For instance:

(7.20) **CausIncepLabreal**$_{12}$ (*tudomás*) = [~ára] hoz[16].

Group III.4. in the ÉrtSz. also contains semi-compositional constructions in which the sense of *hoz* is ‘to cause, to yield’. Among the eleven senses given in ÉKsz., number 7, 8 and 9 mention these constructions. While sense 7 does not offer a common definition, senses 8 and 9 are the following: “Okoz, ill. szerez, eredményez [...] megalkot, létrehoz” ‘to cause, to

[16] *tudomására hoz* (knowledge-3SGPOSS-SUB brings) ‘to acquaint’

yield [...] to create, to produce'. The element 'cause' in the definition corresponds to the **Caus** part within the complex lexical function.

Constructions described with **CausIncepLabreal** all cause that something starts to be in the state denoted by the noun, however, no further generalizations can be made because of the diverse nature of the nominal components.

7.3.4 Constructions with the verb *tesz* 'do'

The last verb to be analyzed was *tesz* 'do'. 60 nominal components in the semi-compositional constructions occurred as the object in the sentence whereas the subject of the sentence is the first actant of the situation. Thus, it is the lexical function $\mathbf{Oper}_1$ that connects the verb and the noun:

(7.21) $\mathbf{Oper}_1$ (*említés*) = [~t] tesz[17]

As for the verb *tesz*, the ÉrtSz. provides the following sense definitions relevant for our purposes in group I (VI. 653–657):

> 1. <vmit, vmely cselekvést> végrehajt, cselekszik, csinál 'to perform, to do <sg, some action>' 2. <cselekvést jelentő fn-i tárggyal> vmit tesz: végzi, végrehajtja azt a cselekvést, amelyre a tárgy utal [...] cselekvésével érvényre juttat '<with an object denoting an action> to do sg: to perform the action the object refers to [...] to reveal sg by action' 3. <vmely hatást v. állapotot> cselekvésével v. meglétével okoz, létrehoz (VI. 653) 'to cause, produce <some effect or state> by acting or existing'.

Among the ÉKsz. definitions, the second one illustrates the sense of semi-compositional constructions (1338–39):

> 2. A jelzett mozgást, cselekvést végrehajtja [...] a jelzett dolgot rendszerint hivatalos formában teszi 'to perform the motion, action mentioned [...] to do the mentioned thing usually in an official form'.

The verb *tesz* 'do' corresponds to a semantic primitive (Apresjan, 2004; Apresjan, 2005), more specifically to the semantic primitive that constitutes a part of the meaning of verbs

[17] *említést tesz* (mention-ACC does) 'to mention'

being the value of $\mathbf{Oper}_1$, the natural language equivalent of which is 'do'. Thus, it is not surprising that this lexical function appears here as well.

The nominal components occurring with *tesz* 'do' can also be classified into well-defined semantic groups:

- nouns denoting speech acts or verbally performed actions (*ajánlat* 'offer', *bejelentés* 'announcement', *javaslat* 'proposal', *jelentés* 'report', *nyilatkozat* 'declaration', *panasz* 'complaint', *tájékoztatás* 'informing')

  Here we can draw a parallel with Russian data: some nouns denoting speech acts can also occur with the verb *delat'* 'do' (Apresjan, 2004).

- nouns denoting some kind of quality or qualification (*hitelesítés* 'authorization', *kivétel* 'exception', *különbség* 'difference')

- nouns denoting movement (*mozdulat* 'move', *séta* 'walk', *lépés* 'step')

- nouns with the meaning 'trial, trying' (*próba* 'try', *kísérlet* 'experiment', *erőfeszítés* 'effort')

It must be mentioned that there are seven nouns in an oblique case that occur with *tesz* 'do', for instance:

(7.22) *folyamatba* *tesz*
process-ILL does
'to start'

pénzzé *tesz*
money-TRANS does
'to convert into cash'

However, no generalizations can be made on their semantic type because of the sparsity of data.

7.4 Correlations between the noun, the verb and the lexical function

In the following, correlations between the noun, the verb and the lexical functions are presented.

7.4.1 Correlations between the semantic characteristics of the noun and the verb

On the basis of the semi-compositional constructions analyzed, some generalizations can be made according to which the semantic type of the noun is able to predict the verbal component to some extent in accordance with the general law of semantic agreement, which requires that at least one non-trivial meaning component is shared between the verb and the noun (Apresjan, 2009). Nouns denoting a possibility or permission are paired with the verb *ad* 'give'. Nouns denoting speech acts or verbally performed actions occur with *ad* 'give', *tesz* 'do' or *hoz* 'bring'. These generalizations – because of their predictive force – might be utilized in language teaching and in NLP applications (Chapters 10 and 12).

As it was presented earlier, nouns denoting speech acts or verbally performed actions can be connected to the above-mentioned three verbs, however, the choice among the three verbs is lexically bound: typically only one of the verbs can co-occur with a noun in a semi-compositional construction. However, there are some cases in the dataset when one noun co-occurs with two verbs, e.g.:

(7.23) *szabályt ad* - *szabályt hoz*
rule-ACC gives - rule-ACC brings
'to provide a rule'

(7.24) *nyilatkozatot ad* - *nyilatkozatot tesz*
declaration-ACC gives - declaration-ACC does
'to make a declaration'

The markedness (and frequency) of these pairs is nevertheless different: *nyilatkozatot tesz* (declaration-ACC does) 'to make a declaration' can be found 25 times in the dataset while its counterpart only 4 times.

7.4.2 Correlations between the senses of the verb and the semantic content of the lexical function

The comparison of the dictionary entries of verbs and the semantic content of lexical functions revealed that their semantic components exhibit some similarity, in accordance with the general law of semantic agreement (Apresjan, 2009). Verbs containing the semantic primitive 'do' (*ad* 'give', *hoz* 'bring' and *tesz* 'do') are values of $\mathbf{Oper}_1$ or $\mathbf{Labor}_{12}$. The meaning

of *vesz* 'take' includes the semantic element 'begin' hence it is connected to the nominal component by the lexical function **Incep**. The verb *hoz* 'bring' and the lexical function **Caus** share the meaning component 'cause'. In this way, the verb and the lexical function show a semantic match – they are related to the same semantic field. The lexical function determines from which set of verbs to select the appropriate one, however, it is the lexico-semantic features of the nominal component that are responsible for choosing the specific verb for the noun.

7.4.3 Correlations between the semantic characteristics of the noun and the lexical function

Nominal components can be classified into several semantic groups and these groups correlate with the applied lexical functions. Nouns denoting speech acts and verbally performed actions are connected to their verbs with the relation expressed by the lexical function **Oper** – the same holds for nouns whose meaning includes the permissive lexical function **Perm** as in *lehetőséget ad* (possibility-ACC gives) 'to offer a possibility'. If the noun denotes an event with a specific starting point, the complex lexical function contains **Incep**. Thus, there are correlations between the semantic characteristics of the noun and the lexical function applied.

7.5 Aspect and Aktionsart

In Chapter 5, interlingual and intralingual comparison of data has shed light on differences between the construction and its verbal counterpart (intralingual difference) on the one hand and it also has turned out that sometimes two constructions cannot be considered as total equivalents on the other hand (interlingual differences). These differences can be formalized with the help of lexical functions (Mel'čuk et al., 1995). For instance, the inchoative Aktionsart is expressed by the lexical function **Incep**:

(7.25) **CausIncepLabor**$_{12}$ (*circulation*) = put into [~]

(7.26) **CausIncepLabor**$_{12}$ (*forgalom*) = [~ba] hoz

In the formal description of the semantic relation between *put* and *into circulation* **Incep** can be found, so does in the description of the relation between *forgalomba* circulation-ILL and *hoz* 'bring' – in this way, the two constructions behave similarly.

When the English and the Hungarian constructions show some difference with respect to their behavior, the lexical functions applied also differ. We repeat examples (5.84) and (5.85) for convenience.

(7.27) The chairman was taking all the suggestions into consideration when deciding on the new budget.

(7.28) **Az elnök éppen figyelembe vett minden*
the chairman just consideration-ILL take-PAST3SG all
javaslatot, amikor az új költségvetésről döntött.
suggestion-ACC when the new budget-DEL decide-PAST3SG

The semi-compositional constructions can be analyzed in the following way:

(7.29) **IncepLabor**$_{12}$ (*figyelem*) = [~be] vesz

(7.30) **Labor**$_{12}$ (*consideration*) = take into [~]

In the English construction, there is no reference to the beginning of an event whereas there is in Hungarian (cf. (7.27) and (7.28)). Thus, in the English phrase there is no Aktionsart encoded this is why aspectual information can be attached to it (by morphological devices).

Examples (5.86) and (5.87) are also repeated here for convenience:

(7.31) She was taking revenge on her ex-boyfriend in those days.

(7.32) *??Azokban a napokban éppen bosszút állt a volt*
that-PL-INE the day-PL-INE just revenge-ACC stand-PAST3SG the ex
barátján.
boyfriend-3SGPOSS-SUP

To take revenge and *bosszút áll* (revenge-ACC stands) can be analyzed in the following way:

(7.33) **Real**$_1$(*bosszú*) = [~t] áll

(7.34) **Real**$_1$(*revenge*) = take [~]

Lexical functions of the **Real** family refer to the fulfillment of some requirement (Mel'čuk et al., 1995), thus, they can be regarded as the lexical function equivalent of perfective aspect. Since the Hungarian constructions *figyelembe vesz* (consideration-ILL takes) 'to take into consideration' and *bosszút áll* (revenge-ACC stands) 'to take revenge' inherently possess aspect and Aktionsart, moreover, in Hungarian, perfective and progressive aspects are incompatible, another aspect cannot be added to the construction. In (7.31) the perfective aspect is also present, however, the co-occurrence of progressive and perfective aspects is acceptable in English. This assumption can be nicely illustrated with lexical functions – **Cont** signals continuity (progressive aspect) (Mel'čuk et al., 1995, p. 142):

(7.35) **ContLabor**$_{12}$(*consideration*) = be taking into [~]

(7.36) **ContReal**$_1$(*revenge*) = be taking [~]

As opposed to English, Hungarian does not tolerate the connection of progressive or Aktionsart lexical functions such as **Cont** and **Incep** on the one hand and perfective lexical functions such as **Real** on the other hand. To sum up, aspect manifests differently in the verbal constructions of the two languages: in English, progressive and perfective aspects can co-exist while in Hungarian they cannot, which can be formalized by lexical functions as well.

7.6 Lexical semantic relations between semi-compositional constructions

Semi-compositional constructions – similarly to other lexical units – also exhibit synonymy and conversion, that is, they either have the same meaning (synonymy) or show systematic differences between their meanings (conversion). These lexico-semantic characteristics of semi-compositional constructions will be presented in this section.

7.6.1 Synonymy

Although it is lexically restricted which nominal component can be paired with which verb, there are instances of synonym semi-compositional constructions in the database. In other

words, the constructions share their nominal components but the verb is different, however, the meaning of the two constructions is the same (although there might be some differences in usage, style or frequency). In the following, some typical synonym pairs of verbal components are listed together with some examples:

- *ad–nyújt* 'give–provide'

  garanciát ad – garanciát nyújt 'to give warranty'

  szolgáltatást ad – szolgáltatást nyújt 'to provide a service'

  segítséget ad – segítséget nyújt 'to offer help'

- *ad–hoz* 'give–bring'

  szabályt ad – szabályt hoz 'to give a rule'

- *hoz–tesz* 'bring–do'

  intézkedést hoz – intézkedést tesz 'to take action'

- *jut–kerül* 'get'

  bajba jut – bajba kerül 'to get into trouble'

  csődbe jut – csődbe kerül 'to go bankrupt'

- *okoz-szerez* 'cause–obtain'

  meglepetést okoz – meglepetést szerez 'to surprise'

  örömet okoz – örömet szerez 'to make happy'

- *ad–szab* 'give–cut'

  határidőt ad – határidőt szab 'to put a deadline'

  irányt ad – irányt szab 'to give a direction'

In these constructions, the meaning of the two semi-compositional constructions can be considered the same, even if the verbal components are not synonyms when being used as main verbs (e.g. *ad* 'give' and *szab* 'cut'). This finding might offer an argument for having a separate entry (or a separate sense in the entry of the verb) in the dictionary for each verbal component (see Chapter 8 for details).

On the other hand, it should be mentioned with regard to synonymy that most semi-compositional constructions are synonymous with a verb. This issue was analyzed in detail in Chapter 5.

7.6.2 Conversion

In certain cases, the conversive pair of a semi-compositional construction can also be found in the database: they describe the same situation, however, the order (and the grammatical role) of the actants is different. Typical pairs of conversive verbal components are as follows:

- *ad–kap* 'give–receive'

  engedélyt ad – engedélyt kap 'to give/receive permission'

  haladékot ad – haladékot kap 'to give/receive a moratorium'

- *ad–vesz* 'give–take'

  bérbe ad – bérbe vesz 'to rent/to hire'

  használatba ad – használatba vesz 'to give for use/to start to use'

- *hoz–kerül/jut* 'bring–get'

  összhangba hoz – összhangba kerül 'to harmonize/to get into harmony'

  forgalomba hoz – forgalomba kerül 'to put/come into circulation'

Note that, however, in the traditional sense of conversion, only pairs where the thematic role of the actants is the same are considered as conversives (e.g. in the case of *bérbe ad – bérbe vesz* 'to rent/to hire', the thematic role of the subject is Agent. This is not true for pairs such as *engedélyt ad – engedélyt kap* 'to give/receive permission', where the subject of the *kap* construction is not agentive. However, in an extended sense of the word, these cases can also be called conversion.

Again, sometimes the main verb usage of the verbal components also reflects this relation (e.g. *ad* 'give' and *vesz* 'take' are conversives as main verbs as well), however, for other verbs it is only their occurrences in semi-compositional constructions that can be seen as conversives (e.g. *hoz* 'bring' and *kerül* 'get').

7.7 Summary of results

In this chapter, the semantic features of semi-compositional constructions were analyzed within the framework of the Meaning–Text Theory. Semi-compositional constructions containing one of the four most frequent verbs in the dataset were selected for analysis. Semantic correlations between the noun, the verb and the lexical function were presented in detail and findings were analyzed in multilingual context (i.e. they were compared to Russian data).

The following issues were emphasized throughout this chapter:

- lexical functions are responsible for connecting the verb and the nominal component of the construction;
- semantic correlations can be found between the semantic type of the noun, the verb and the lexical function;
- lexical functions are also able to formalize differences in aspect and Aktionsart;
- there are synonymous semi-compositional constructions;
- some constructions are conversives of each other.

The results of this chapter can be fruitfully applied in language teaching and in NLP applications such as word sense disambiguation and machine translation (see Chapters 10 and 12).

Chapter 8

The lexical representation of semi-compositional constructions

8.1 Introduction

In this chapter, the lexical representation of semi-compositional constructions is discussed. In terms of research questions, this involves the following one:

- What lexical representation can be assumed for semi-compositional constructions?

In order to select from the emerging theoretical possibilities, advantages and disadvantages of each solution are presented. The problems concerning the determination of the head of the construction are also shown and it is argued which way of lexical representation proves to be the most applicable from both a theoretical and an empirical aspect. Finally, as an example for an electronic lexical database, it is examined how the Hungarian WordNet treats semi-compositional constructions.

8.2 Questions of lexical representation

In this section, two issues related to the lexical representation of semi-compositional constructions are discussed – namely, how to determine the head of the construction and how to take into account the features of paper-based and electronic dictionaries.

8.2.1 The head of the construction

In order to determine the proper lexical representation of semi-compositional constructions, the head of the construction must be specified first. The construction comprises of two members – the nominal and the verbal component –, in terms of multiword expressions (see Chapter 2), the noun is the base and the verb is the collocate (Siepmann, 2005; Siepmann, 2006; Sag et al., 2002; Guenthner and Blanco, 2004). However, it is not unequivocal to determine which one is the head of the construction.

First, the syntactic head of semi-compositional constructions is the verb since it enables the construction to function as the predicate within the sentence and the features of tense, mood, voice, aspect, person, number are borne by the verb (Dobos, 2001). Besides, the noun is a syntactic argument of the verb.

Second, although syntactically the noun is a dependent of the verb, semantically the former functions as the head of the construction (Dobos, 2001). It is justified by the fact that the verbal counterpart of the semi-compositional construction is often derived from the same root as the noun (*to decide – to make a decision*) (Vincze, 2009b). On the other hand, the replacement of the noun results in a construction with a totally different meaning, e.g. *to make difference* and *to make peace*: although their verbal component is the same, the meaning of these constructions are not related at all. However, if the verb is replaced, the construction may sound somewhat strange but the intended meaning can still be deduced: *?to do an error* instead of *to make an error*.

It can be concluded from the above that 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 (see also Chapters 6 and 7). Thus, there is no unique head of the construction – we can speak only of semantic or syntactic head. This is another reason why the term *semi-compositional construction* is used in this thesis as opposed to terms that emphasize the central nature of the verb (e.g. *light verb construction* or *support verb construction*).

8.2.2 Traditional paper-based and electronic dictionaries

The appearance of intelligent dictionaries resulted in several novelties in the structure of dictionaries. Due to computational technologies, there is no need for a reference entry since the entry that encodes all necessary information on the headword can be reached in one step. Computational intelligent dictionaries also include a morphological module due to which not

only the base form (lemma) of the headword can be found but any of its inflected forms as well (e.g. with the query *zavar* 'trouble' the following expressions can be also listed as hits:

(8.1) *zavarba* *hoz*
trouble-ILL bring
'to embarrass'

zavarba *jön*
trouble-ILL come
'to feel embarrassed'

An illustrative example of tools enabling intelligent search in electronic corpora is called Mazsola (Sass, 2007; Sass, 2008; Sass, 2009), developed for the Hungarian National Corpus (Váradi, 2002). With its help, multiword expressions related to a given verb (i.e. institutionalized expressions, set phrases, idioms and semi-compositional constructions) can be collected and the query can also rely on morphosyntactic features of parts of the collocation. A dictionary containing the most frequent verbal collocations collected from the Hungarian National Corpus has been recently published (Sass et al., 2010).

Due to the technological innovations, it is also easier to find other types of multiword expressions such as idioms: in traditional paper-based dictionaries they occurred only once, namely, in the entry of the keyword – e.g. the idiom *kutyából nem lesz szalonna* (dog-ELA not becomes bacon) 'a leopard cannot change its spots' could be found in the entry of *kutya* 'dog' or *szalonna* 'bacon' but only in one of them. With the novel technique, however, it is possible to find this idiom no matter if the original query is *kutya* 'dog or *szalonna* 'bacon'.

In intelligent bilingual dictionaries the pair headword–entry is replaced by pairs such as headword–part of speech, headword–pronunciation, headword–sense etc. This greatly accelerates the process of searching and, on the other hand, the difference between the source and the target language also disappears: electronic bilingual dictionaries are able to provide the target language equivalents of the word in the source language and target language entries including a target language equivalent of the word in the source language are also listed. Thus, for the query *ló* 'horse', the words *horse*, *knight*, *pommel horse* are listed as hits in a Hungarian–English dictionary, which also refers to the connection between different senses of the Hungarian word (see Table 8.1), which information remains hidden in traditional dictionaries (Prószéky, 2004).

Traditional – paper-based – and electronic dictionaries are thus essentially different in their nature. It entails that when examining the possibilities of the lexical representation of

English	Hungarian
horse	**ló**
knight	lovag
	huszár
	ló (chess)
	lovaggá üt (verb)
pommel horse	**ló** (gymnastics)

Table 8.1: Senses of *ló*

semi-compositional constructions – especially when weighing the advantages and disadvantages of each method – it should be considered that lexicologists and dictionary users have different expectations towards paper-based and electronic dictionaries.

8.3 The possibilities of the lexical representation of semi-compositional constructions

In theory, semi-compositional constructions can have four different representations in the dictionary, which also reflects that the construction can be seen as either a unit or a construction consisting of two parts. The construction can occur:

- within the entry of the verbal component, that is, the verb is considered to be the head of the construction;
- within the entry of the nominal component, that is, the noun is considered to be the head of the construction;
- within a separate entry, that is, there is no head and the construction is seen as a separate lexical unit;
- both within the entry of the nominal and the verbal component (Alonso Ramos, 1998).

In the following, mini-entries are constructed on the basis of our database obeying the principles of each strategy separately (using the dictionary entries of ÉKsz. and the Hungarian WordNet (HuWN) (Miháltz et al., 2008) as patterns). With the help of such mini-entries, advantages and disadvantages of each strategy are discussed and it is also discussed how these strategies manifest in present-day dictionaries.

The most frequent verbal components occurring in semi-compositional constructions in the database are listed in Table 8.2 in terms of their frequency and the number of nominal components they co-occur with is also given for each verb.

Number	Verbal component	Percentage rate	Number of nominal components
1.	ad	22.66%	164
2.	vesz	13.6%	50
3.	hoz	12.41%	51
4.	tesz	8.27%	67
5.	köt	5.55%	23
6.	kerül	5.3%	70
7.	jut	3.85%	30
8.	tart	3.67%	64
9.	nyújt	2.78%	46
10.	lép	2.66%	17
11.	áll	2.51%	40
12.	kap	2.42%	71
13.	végez	1.81%	39
14.	folytat	1.68%	29
15.	ér	1.49%	14

Table 8.2: The most frequent verbal components

In the mini-entries to be built, semi-compositional constructions containing the verbs *hoz* 'bring' and *köt* 'bind' are listed. *Hoz* 'bring' is the third most frequent verbal component in the database. It co-occurs with 51 nominal components. Semi-compositional constructions in this group can be divided into four categories with regard to the meaning of the verbal component as described below.

Köt 'bind' is the fifth most frequent verbal component in the database and it attaches to 23 nominal components. However, among them there are compounds with the same second part (*szerződés* 'treaty') describing different types of treaties, thus they are counted as one. In this way, 15 nominal components will be paid attention to in the analysis. The generalized meaning of the verb can be paraphrased in the following way: 'to create a mutual relationship'.

In the following, mini-entries built for semi-compositional constructions containing these two verbs will illustrate the possible ways of lexical representation.

8.3.1 The verbal component as the head

If the verbal component is considered to be the head of the construction, the semi-compositional constructions are listed in the dictionary entry of the verb. Illustrating with the verb *köt* 'bind':

(8.2) **köt** (Adott cél érdekében) kapcsolatot, kölcsönös viszonyt létrehoz 'to create a mutual relationship (for a specific purpose)'.
Alkut ~; barátságot ~; békét ~; biztosítást ~; egyezményt ~; egyezséget ~; fogadást ~; házasságot ~; ismeretséget ~; kompromisszumot ~; megállapodást ~; szerződést ~; szövetséget ~; ügyletet ~; üzletet ~ 'bargain[1], friendship, peace, insurance, convention, agreement, bet, marriage, acquaintance, compromise, agreement, contract, ally, transaction, business'.

Thus, the definition of *köt* 'bind' as a verbal component, which is based on the entries of ÉKsz. and ÉrtSz., is followed by the list of the corresponding constructions – however, the entries of the nominal components do not mention either the verbal component or the semi-compositional constructions. ÉrtSz. (Bárczi and Országh, 1959 1962), ÉKsz. (Pusztai, 2003) and the Purists' dictionary (Grétsy and Kemény, 1996) belong to this group.

One of the advantages of this method is that the semi-compositional constructions containing the given verb are collected in the dictionary. Another advantage is that the verbal component occurs in the dictionary as a lexical unit with a distinct meaning which often differs from the ordinary meanings of the verb (see also 7.6.1). However, this strategy suffers from several disadvantages: the construction is encoded in the entry of the collocate. Collocational dictionaries typically list collocations and multiword expressions in the entry of their base (i.e. their semantic head), thus, it would be more convenient to follow that strategy here as well. Consequently, it can be hard for the dictionary user to find the construction in a (paper-based) dictionary, which is especially true in the case of multilingual dictionaries: while the nominal components are usually literal translations of each other (Vincze, 2009d), verbal components can highly differ in foreign language equivalents of semi-compositional constructions. For instance, the Hungarian semi-compositional construction *döntést hoz* (decision-ACC brings) corresponds to the English constructions *to make a decision* or *to take a decision*, however, the literal translation **to bring a decision* is not grammatical

[1]Only nominal components are translated here and they all are in the accusative case, which is not marked distinctively.

in English. With this strategy the dictionary user should search for all the occurrences of *decision* in the dictionary and then select the appropriate construction, however, this might be time-consuming in the case of paper-based dictionaries.

8.3.2 The nominal component as the head

If the nominal component is seen as the head of semi-compositional constructions, the constructions can be found in the dictionary entries of the nominal components. An example with some nominal components occurring with *köt* 'bind':[2]

(8.3) **biztosítás** Szerződés, mely alapján a biztosító káresemény bekövetkeztekor kárpótolja a biztosítottat. **~t köt**: ilyen szerződést létrehoz.

insurance Promise of reimbursement in the case of loss; paid to people or companies so concerned about hazards that they have made prepayments to an insurance company. **to get insurance**: to create such an insurance.

(8.4) **szövetség** Csoportoknak közös cél érdekében való együttműködése. **~et köt**: ilyen együttműködést hoz létre.

federation An organization formed by merging several groups or parties. **to form a ~**: to create such an organization.

(8.5) **üzlet** Adásvétel. **~et köt**: adásvételt bonyolít le.

business The volume of business activity. **to do ~**: to perform such activity.

In the mini-entries, the definition of the noun is followed by the semi-compositional constructions that contain the given noun, however, the entry of the verb *köt* 'bind' does not refer to the nouns co-occurring with it. Among existing dictionaries, monolingual English (Hornby and Wehmeier, 2002), Russian (Mel'čuk and Žolkovskij, 1984) and French (Mel'čuk et al., 1984 1999; L'Homme et al., 2007; Mével, 2004) dictionaries, bilingual (e.g. Hungarian–French (Eckhardt, 1992b), Hungarian–English (Magay and Országh, 2001b) and Portuguese–Hungarian (Király, 1993b)) dictionaries and specialized dictionaries (Iordanskaja and Paperno (1996) on the Russian and English collocations related to the human body or Székely (2003) on German and Hungarian intensifiers) for instance.

[2]Definitions and their English equivalents are imported from the Hungarian and Princeton WordNets, respectively (Miháltz et al., 2008; Miller et al., 1990).

The advantage of this method is that the semi-compositional constructions containing the given noun are collected in the dictionary. As it was shown above, when comparing semi-compositional constructions in different languages the verb – as a collocate – is generally unpredictable whereas the noun (the base) is given, that is, it can be literally translated. In this way, this strategy can be exploited in language learning and machine translation (Apresjan and Tsinman, 2002). On the other hand, listing the grammatical information on the usage of the verb and the construction in the entries of each nominal component multiple times leads to redundancy since the verbal component is not listed in the dictionary as a separate lexical meaning (despite the fact that its "semi-compositional" meaning may differ from that of its ordinary usage).

8.3.3 A separate entry

If it is assumed that the construction has no head, semi-compositional constructions are listed as separate entries in the dictionary. For example:

(8.6) **biztosítást köt** Olyan szerződést hoz létre, mely alapján a biztosító káresemény bekövetkeztekor kárpótolja a biztosítottat.

to get insurance To create a treaty on the basis of which the insurance company pays some money to the person having suffered a loss

szerződést köt Közös megegyezéssel írásban és törvényesen megállapodik.

to make a treaty To make a written and legal agreement.

szövetséget köt Bizonyos cél érdekében együttműködést hoz létre.

to create a federation To create cooperation for some purpose.

ügyletet köt Pénzügyi műveletet bonyolít le.

to do business To carry out a financial transaction.

üzletet köt Adásvételt bonyolít le.

to do business To perform business activity.

Definitions are composed of the definitions provided for the nominal and the verbal components in the following sections.

With this strategy, all the semi-compositional constructions including a given verb (or noun) are listed separately in the dictionary: neither the entry of the noun nor that of the verb mention them. Though this method might seem somewhat unusual, its advantage is that each semi-compositional construction is seen as a separate lexical unit, reflecting the semantic unity of the construction. This is in line with construction grammars (see e.g. Goldberg (1995)), where contents and forms are paired to form a construction with typically unpredictable meaning. The problem of determining the head of the construction is also eliminated in this way. However, there are disadvantages of this method. This representation is not able to signal that there are similar constructions: if they occur in alphabetical order (as it is typically the case in paper-based dictionaries), then constructions with the same nominal component are adjacent (at least in Hungarian where the canonical order of the constructions is noun + verb), however, constructions with the same verbal component and different nominal component may occur far from each other. Thus, their unified treatment becomes almost impossible and similarities in their meaning and usage remain hidden. On the other hand, this solution is not economical since it leads to the extension of the size of the paper-based and electronic dictionary (both the nominal and the verbal component occur several times). However, bigger size is less problematic in the case of electronic dictionaries and it is also easier to collect constructions with the same nominal/verbal component as opposed to paper-based dictionaries.

With regard to theoretical aspects, another disadvantage of the method is that the meaning of the construction is not completely independent of the meaning of its parts since they are placed in between productive constructions and idioms (see Chapter 4). It would be, however, plausible to include semi-compositional constructions in the entries of one of their components, which method is also enhanced by the fact that non-compositional idioms are typically listed in the entry of one of their parts. Due to the typical organizational principle found in paper-based dictionaries (i.e. the alphabetical order), the entries of semi-compositional constructions would follow the entry of their nominal components (in Hungarian, and that of the verbal component in English). It can be concluded that from both a theoretical and an empirical point of view it is more effective to have semi-compositional constructions listed in the entry of its components.

8.3.4 The construction occurs in the entries of both the nominal and the verbal component

According to the fourth possibility, semi-compositional constructions occur in the entries of both the nominal and the verbal component. The verbal component has a separate entry in the dictionary (or a distinct sense within the entry of the verb) and the entry of the nominal component includes every verb with which it can form a semi-compositional construction.

Although the semantic head of the construction is the nominal component, the meaning of the verb also contributes to the meaning of the construction (i.e. it cannot be considered as meaningless, cf. Apresjan (2004)): there is a difference between e.g. *giving an order* and *receiving an order* (see also 7.6.2). The meaning, usage and syntactic-semantic features of certain verbal components are preserved with a nominal component belonging to a given semantic class (Alonso Ramos, 1998), thus it can be argued that the verbal component has a separate dictionary entry (or at least a distinct meaning within the entry of the verb). This includes the definition of the verb, its syntactic features, information on its usage and the appropriate semantic classes of nouns with exceptional cases if any, i.e. there are no lists of nominal components if they can be covered by a single hypernym. More than one entry (or sense) can be assumed if necessary. The entry of the hypernym of the nominal components contains each verb with which they can form a semi-compositional construction, thus it proves to be sufficient to refer to the verb only in the entry of the hypernym since its hyponyms automatically inherit this feature.

This way of lexical representation is illustrated with the example of semi-compositional constructions containing *hoz* 'bring':

(8.7) **hoz**

1. <összeget> Termel. <sum> produce
2. <verbális cselekvést> Megalkot, eredményez. <verbal act> create, yield
3. <fordulatot> Másik állapotba juttat. <turn> cause that something starts to be in another state
4. <állapotba/ra> Adott állapotba juttat. <state> to cause that something starts to be in a specific state

The entry of *hoz* 'bring' includes the four light verb senses, which form a semi-compositional construction with the noun belonging to the given semantic class. Obviously, the other

(i.e. not light verb) senses of *hoz* 'bring' also occur in the dictionary, however, they are not significant for our research purposes hence they are omitted from the mini-entry.

Let us now turn to the mini-entries of some nominal components occurring with *hoz* 'bring'[3].

(8.8) **fordulat** Valami egyik fázisból vagy állapotból a másikba lépésének eseménye.
synonym: módosulás, változás
N+V: ~ot hoz 'to cause to be in another state'

{change, alteration, modification} An event that occurs when something passes from one state or phase to another.

szégyen Az a kínos érzés, hogy mások előtt nagyon kedvezőtlen színben tűntünk fel.
hypernym: állapot

{dishonor, dishonour} A state of shame or disgrace.
hypernym: state

állapot Valakinek, valaminek valamely időszakban jellemző létezési módja.
N+V: ~ba hoz, ~ra hoz

state The way something is with respect to its main attributes.

profit Vállalkozásból, adásvételből származó tiszta haszon.
hypernym: összeg

gain The amount by which the revenue of a business exceeds its cost of operating.
hypernym: sum

összeg Bizonyos mennyiségű pénz.
N+V: ~et hoz

sum A quantity of money.

Let us examine first the entry of *fordulat* 'change'. The definition is followed by synonyms of the headword (in accordance with wordnet building principles), then the semi-compositional constructions are listed which contain the headword (*fordulat* 'change'). In a given construction, not only the headword but its synonyms may also occur – in this way, constructions such as:

[3]Again, English equivalents are imported from the Princeton WordNet if available, thus, translations are not always word-by-word.

(8.9) *fordulatot hoz*
turn-ACC brings
'to change'

módosulást hoz
modification-ACC brings
'to modify'

változást hoz
change-ACC brings
'to change'

can be also obtained. In order to determine the meaning of the construction, the entry of the verbal component must be scrutinized, where the third sense in (8.7) will play an important role the nominal component being mentioned there. Thus, the meaning of the construction is: 'to get into another state'. Here is an example:

(8.10) *A perben hirtelen fordulatot hozott az új ügyész*
the sue-INE sudden turn-ACC bring-PAST3SG the new attorney.general
kinevezése.
appointment-3SGPOSS
'The appointment of the new attorney general resulted in a sudden turn in the lawsuit.'

The semi-compositional construction can be paraphrased in the following way:

(8.11) The lawsuit started to be in another state after the appointment of the new attorney general.

In our second example, the construction *szégyenbe hoz* (shame-ILL brings) 'put to shame' will be computed with the help of the mini-entries defined above. Within the entry of *szégyen* 'shame', there is no reference to a semi-compositional construction, which directs to the entry of the hypernym of the word (*állapot* 'state'). This entry apparently includes the information needed: it is paired with the verb *hoz* 'bring'. The meaning of the verbal component is given in its entry: 'to cause that something starts to be in a specific state'. In the present case, this state is shame.[4]

The construction *profitot hoz* (gain-ACC bring) 'to produce gain' can be yielded similarly to the previous example. The entry of *profit* 'gain' refers only to its hypernym (*összeg* 'sum'),

[4]It must be mentioned that there is a synonym construction *szégyent hoz* (shame-ACC brings) 'put to shame'. In this example, the nominal component is in the accusative case which might be suggestive of merging the two light verb senses of *hoz* 'bring' related to states, however, this claim needs further investigations.

however, in the entry of the hypernym, the verb *hoz* 'bring' can be found. The entry of *hoz* 'bring' includes the necessary data for choosing the appropriate sense for the construction: *profitot hoz* means 'to produce gain' since *profit* 'gain' belongs to the domain of *sum* (it is a hyponym of the latter). Some bilingual dictionaries that follow this strategy are the French–Hungarian (Eckhardt, 1992a), English–Hungarian (Magay and Országh, 2001a) and Hungarian–Portuguese (Király, 1993a) dictionaries.

An advantage of this method is that the entries of both components contain information on the other one, thus the construction can be found starting from either the verb or the noun. The verbal component is present in the dictionary as a separate lexical unit. The redundancy emerging when the nominal component is the head can be eliminated by applying hypernyms and by attributing a separate entry to the verb since – after having made the appropriate generalizations – it is not necessary to include the verb or the construction in the entry of every noun. Instead, it suffices to mention them in the entry of the hypernym or the verb. Consequently, this solution can offer help in language learning and language teaching for it is now easy to find the constructions in the dictionary and due to the generalizations made, the language learner is not only able to understand but to form the equivalent of the construction in another language. In electronic dictionaries and databases semi-compositional constructions can be easily listed in the entry of the collocate and the base as well, which is of great importance in the case of bilingual dictionaries and can enhance the performance of information extraction systems and machine translation systems (Apresjan and Tsinman, 2002; Apresjan et al., 2007; Vincze, 2007; Vincze, 2009e).

As a disadvantage of the solution it must be mentioned that this method does not prove economical in the case of paper-based dictionaries (Siepmann, 2005; Siepmann, 2006).

8.4 Comparing the methods

In the previous section different methods for the lexical representation of semi-compositional constructions were discussed together with their advantages and disadvantages. With regard to theoretical aspects, the fourth method seems to be the most promising, that is, it is advisable to mention the semi-compositional construction in the entry of both the verbal and the nominal component (or its hypernym). This solution merges the advantages of all the other methods since on the one hand, the verbal component occurs as a separate entry in the dictionary distinct from the other senses of the verb and, on the other hand, the collocation

can be found in the entry of the base. It was also shown in 8.2.1 that none of the components of the semi-compositional construction can be considered as being the absolute head of the construction – only syntactic and semantic heads can be identified. On the basis of this, it would be inappropriate to emphasize the role of any of the components in the lexical representation of the construction, which requirement is also observed by the fourth solution.

Besides theoretical advantages, empirical considerations also lead to the conclusion that the fourth method is the most convenient for encoding semi-compositional constructions in the dictionary. For the dictionary user it is easy to find the construction in a traditional dictionary starting from either the verb or the noun. However, it must be emphasized that from a practical point of view the question of lexical representation is somewhat less essential in the case of electronic dictionaries since constructions can be easily found there whichever method is applied – if the proper query is used. Due to the generalizations made on (the usage of) the constructions, dictionaries built in this way can be made use of in language teaching (concerning the mother tongue and foreign languages as well). Finally, NLP applications can also profit from databases constructed on similar principles.

However, no such dictionary has been constructed in practice: as it was shown above, most of the existing dictionaries list the constructions in the entry of the verb or in the entry of the noun (mainly collocational dictionaries or explanatory combinatorial dictionaries). Although in some bilingual dictionaries, some constructions occur in the entries of the noun and the verb as well, this practice is not consequent even within the given dictionary and dictionaries of the same language pair are also inconsistent, for instance, the English–Hungarian (Magay and Országh, 2001a) and the Hungarian–English (Magay and Országh, 2001b) dictionary do not belong to the same group (Vincze, 2008b). In order to create (paper-based or electronic) dictionaries where semi-compositional constructions are listed in the entries of both the noun and the verb, more theoretical research is needed on the semantic description of nouns belonging to a specific verbal component on the one hand and on the syntactic and semantic description of verbs on the other hand. Definitions of the verb senses found in semi-compositional constructions should also be given, which process culminates in constructing dictionary entries. Hopefully, this research will be carried out in the near future.

8.5 Semi-compositional constructions in the Hungarian WordNet

In the following, an example of electronic databases, namely, the Hungarian WordNet is presented and it is shown how semi-compositional constructions are included in the conceptual hierarchy.

Dictionaries are usually structured on the basis of word forms: words are (alphabetically) listed in the dictionary, and their meanings are given one after the other. However, the most innovative aspect of wordnets is that lexical information is organized in terms of meaning; that is, a synset (the basic unit of wordnets) contains words of the same part of speech which have approximately the same meaning. Thus, it is semantic relations, more specifically, synonymy that functions as the essential principle in the construction of wordnets (Miller et al., 1990). An example of a synset is the following:

(8.12) bicycle:1, bike:2, wheel:6, cycle:6

Literals (i.e. components of a synset) are numbered as a word can have several meanings and it is important to represent that a word is synonymous with other words in one given sense. Thus, *cycle* occurs in five other synsets, including:

(8.13) cycle:1, rhythm:3, round:2
cycle:2
cycle:3
Hertz:1, Hz:1, cycle per second:1, cycles/second:1, cps:1, cycle:4
cycle:5, oscillation:3

Synsets are connected to each other by means of semantic and lexical relations, yielding a hierarchical network of concepts. Semantic relations hold between concepts. In other words, not the forms but their meanings are related. Such relations include hyponymy and meronymy. On the other hand, lexical relations connect different word forms. For instance, synonymy, antonymy and different morphological relations belong to this group (Miller et al., 1990).

As for semi-compositional constructions, the Hungarian WordNet treats them as separate lexical units (cf. the third method of lexical representation described above), that is, they

behave as normal literals. When constructing the Hungarian Wordnet, wordnet builders were given special instructions to include the most frequent semi-compositional constructions in synsets (frequency data were estimated on the basis of the Hungarian National Corpus). They can be found in synsets[5] together with their verbal counterparts (see Chapter 5) as in[6]:

(8.14) POS: v

ID: ENG20-00777368-v

Synonyms: engedélyez:1, engedélyt ad:1

Definition: Hatóság vagy hivatal engedélyt megad.

POS: v

ID: ENG20-00777368-v

Synonyms: authorize:1, authorise:2, pass:24, clear:4

Definition: Grant authorization or clearance for.

Sometimes, there are more than one semi-compositional constructions within one synset, which entails that they are synonyms (see 7.6.1):

(8.15) POS: v

ID: ENG20-00862885-v

Synonyms: hálát ad:1, köszönetet mond:1, köszönetet nyilvánít:1, megköszön:1, köszön:1

Definition: Köszönetét fejezi ki valakinek valamiért.

POS: v

ID: ENG20-00862885-v

Synonyms: thank:1, give thanks:1

Definition: Express gratitude or show appreciation to.

In certain cases, the synset contains only one semi-compositional construction, that is, it is seen as a separate lexical unit (having one entry in the dictionary or rather forming one synset in the wordnet, compare 8.3.3):[7]

(8.16) POS: v

[5]Within the examples, POS denotes the part of speech of the literals, ID refers to the identification number of the synset, and literals and the definition of the concept denoted by them are also provided.

[6]Again, the corresponding English synsets are imported from the Princeton WordNet, thus, they do not always contain a semi-compositional construction.

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

ID: ENG20-00992244-v

Synonyms: szóba hoz:1

Definition: Szól róla, megemlíti.

POS: v

ID: ENG20-00992244-v

Synonyms: raise:19, bring up:6

Definition: Put forward for consideration or discussion.

Based on these examples, HuWN can be considered as a database in which semi-compositional constructions are treated as separate lexical entries. Since wordnets are electronic databases by nature, it is achievable to find semi-compositional constructions if the query includes either the nominal or the verbal component. However, as wordnets contain several lexical relations among synsets, it would prove useful to link the synset of the nominal and the verbal component to that of the semi-compositional construction, e.g. the relation *derivative* might connect them to each other, thus signaling their syntactic and semantic interrelatedness. This extension of relations between synsets would be fruitful in the sense that the synsets of the construction and its components would be directly connected hence they could inherently be matched without any further (query) steps. Hopefully, this work can be carried out in the future.

8.6 Summary of results

In this chapter, the lexical representation of semi-compositional constructions was discussed. The following points were emphasized:

- problems concerning the identification of the head of the construction were presented;
- four ways of lexical representations were illustrated with examples;
- theoretical and empirical advantages and disadvantages of each solution were discussed from the perspectives of paper-based and electronic dictionaries;
- the method of providing 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;

- as an example of electronic databases, the treatment of semi-compositional constructions in the Hungarian WordNet was presented.

These results can be employed in lexicography and language teaching. Besides, the lexical representation of semi-compositional constructions also has an impact on word sense disambiguation, information extraction and machine translation (see Chapters 10, 11 and 12).

Part III

Computational linguistic analyses

Chapter 9

The automatic identification of semi-compositional constructions

9.1 Introduction

In all kinds of NLP applications it is important to identify multiword expressions since they require special treatment (see Chapters 10, 11 and 12), especially because of their semantic features. However, the variability of multiword expressions may result in the fact that different types of multiword expressions require different treatment (i.e. a solution developed for one specific type of multiword expressions might not be applicable to another type of MWEs). In this chapter, existing solutions for identifying multiword expressions are presented. Special emphasis is put on semi-compositional constructions: our rule-based and machine learning based methods to identify English semi-compositional constructions are also described and some further suggestions are made in order to make their automatic identification easier. Thus, this chapter focuses on the identification of semi-compositional constructions while applications such as word sense disambiguation, information extraction and retrieval and machine translation are discussed in later chapters.

9.2 Related work on the automatic identification of multiword expressions

Recently, multiword expressions have been received special interest in the NLP research community (Rayson et al., 2010). Different subtypes of MWEs have been examined from several viewpoints, for instance, extraction from natural language texts, machine translation and their semantic interpretation (see e.g. Task 9 of the 2010 SemEval challenge (Butnariu et al., 2010)). In this section, the automatic identification of multiword expressions will be focused on, thus, methods used for extracting MWEs will be discussed.

There are several applications developed for identifying MWEs, which can be classified according to the methods they make use of (Piao et al., 2003; Dias, 2003). First, statistical models rely on word frequencies, co-occurrence data and contextual information in deciding whether a bigram or trigram (or even an n-gram, i.e. a sequence of words) can be labeled as a multiword expression or not. Such systems are used for several languages and several types of multiword expressions (e.g. Bouma (2010), Villavicencio et al. (2007)). The advantage of statistical systems is that they can be easily adapted to other languages and other types of multiword expressions. However, they are not able to identify rare multiword expressions. As Piao et al. (2003) emphasize, about 68% of multiword expressions occur only once or twice in their corpus. Similarly, in our corpora, about 71% of the Hungarian semi-compositional constructions and 57% of the English ones occur less than 3 times.

Some hybrid systems make use of both statistical and linguistic information as well, that is, rules based on syntactic or semantic regularities are also incorporated into the system (Dias, 2003; Bannard, 2007; Cook et al., 2007; Al-Haj and Wintner, 2010; Sinha, 2011). This results in better coverage of multiword expressions. On the other hand, these methods are highly language-dependent because of the amount of linguistic rules encoded, thus, it requires much effort to adapt them to different languages or even to different types of multiword expressions. However, the combination of different methods may improve the performance of MWE-extracting systems (Pecina, 2010).

Several MWE detectors make use of data from parallel corpora, which are based on word alignment. For instance, one-to-many alignment can be exploited: if a word corresponds to several words in the other language, it is highly probable that the other language equivalent can be considered as a multiword expression (see e.g. Caseli et al. (2009), Caseli et

al. (2010), Zarrieß and Kuhn (2009), Sinha (2009), Attia et al. (2010) or Haugereid and Bond (2011)). However, this method cannot identify multiword expressions that are aligned to another multiword expression in the other language. In other words, multiword expressions whose translational equivalent is a multiword expression cannot be extracted with this method.

Several features are used in identifying multiword expressions, which are applicable to different types of multiword expressions to various degrees. Co-occurrence statistics and POS-tags seem to be useful for all types of multiword expressions, for instance the tool `mwetoolkit` (Ramisch et al., 2010a) makes use of such features (which is illustrated through the example of identifying English compound nouns (Ramisch et al., 2010c; Ramisch et al., 2010b)). Morphological information can be also exploited in the case of e.g. semi-compositional constructions (in Hungarian, the suffix of the nominal component may imply that it forms a semi-compositional construction with the verb, see below). Syntactic patterns can be also applied in identifying more complex or syntactically more flexible multiword expressions (e.g. some idioms can be passivized, compare *Who let the cat out of the bag?* and *The cat was let out of the bag*). Lists can be also integrated into the systems: this method is essential for non-productive types of multiword expressions, e.g. multiword prepositions such as *in front of* or *out of* or multiword conjunctions such as *in order to* or *in case that*. However, for productive types such as compound nouns lists can also improve the performance of the system.

9.2.1 Corpora and databases

Identifying multiword expressions in general and semi-compositional constructions in particular is not unequivocal since constructions with similar syntactic structure (e.g. verb + noun combinations) may belong to different subclasses on the productivity scale (i.e. productive combinations, semi-compositional constructions and idioms, see Chapter 4). In order to identify multiword expressions in texts, well-designed and tagged corpora of multiword expressions are invaluable resources for training and testing algorithms. For instance, Grégoire (2007; 2010) presents a lexicon of Dutch multiword expressions (DuELME), where over 5000 Dutch MWEs are stored, using the parametrized equivalence class method (i.e. MWEs are grouped according to their syntactic pattern). A 1000-sentence database from the British National Corpus contains 345 noun compounds (Nicholson and Baldwin, 2008).

Kaalep and Muischnek (2006; 2008) describe an Estonian database and a corpus of multiword verbs (see also Muischnek and Kaalep (2010)) and Krenn (2008) developed a database of German PP-verb combinations. The Prague Dependency Treebank is also annotated for multiword expressions (Bejcek and Stranák, 2010), thus for semi-compositional constructions too (Cinková and Kolářová, 2005). For Portuguese, Hendrickx et al. (2010) created an annotated corpus of complex predicates (i.e. multiword verbs), and Sanches Duran et al. (2011) present a dictionary of Brazilian Portuguese complex predicates. NomBank (Meyers et al., 2004b) contains the argument structure of common nouns, paying attention to those occurring in semi-compositional constructions as well. Literal and idiomatic usages of English verb + noun combinations are annotated in the VNC-Tokens dataset (Cook et al., 2008). A further example of corpus-based identification of semi-compositional constructions in English is described in Tan et al. (2006). We also developed a corpus of 50 English Wikipedia articles, in which several types of multiword expressions (including semi-compositional constructions) were marked (Vincze et al. (2011b) and see also Chapter 3). As for Hungarian, we have developed an annotated corpus and a database containing semi-compositional constructions described in 3.5 (see also Vincze and Csirik (2010)).

9.2.2 Parallel corpora in identifying multiword expressions

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 to take 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 candidates are filtered according to some empirically defined patterns or frequency data. This method is also applied to the pediatrics domain (Caseli et al., 2009).

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.

Data gained from parallel corpora are also exploited in extending the Slovene wordnet with nominal multiword expressions (Vintar and Fišer, 2008).

Sass (2010) developed a method to extract multiword verbs from parallel corpora, which is based on a former algorithm to collect verbs and their arguments from texts (Sass, 2008). By aligning the verbs in parallel clauses, a complex verb is yielded to which arguments are ordered with tags denoting the language of the subcorpus it comes from. From these representations the original algorithm is able to detect the multiword verbs for each language of the parallel corpus, besides, cases when a multiword verb corresponds to a single word verb in the other language can be also extracted.

With regard to their NLP treatment, a database of semi-compositional constructions and an annotated corpus may be of great help in the automatic recognition of semi-compositional 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 multiword expressions and semi-compositional constructions (see 9.2.1), however, to the best of our knowledge, no parallel corpora have been annotated for semi-compositional constructions. With this motivation in mind, we have developed an English–Hungarian annotated parallel corpus of semi-compositional constructions, which is described in detail in 3.5. In this way, the performance of methods developed for detecting semi-compositional constructions in English and Hungarian can be compared within the same domain (i.e. on the same texts).

9.2.3 Identifying semi-compositional constructions

Semi-compositional constructions are usually distinguished from productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other hand in NLP literature too (compare the theoretical results of Chapter 4): for instance, 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.

Semi-compositional constructions deserve special attention in NLP applications for several reasons. First, their meaning cannot be computed on the basis of the meanings of the parts of the collocation and the way they are related to each other (lack of total compositionality). Thus, the result of translating their parts literally can hardly be considered as the proper translation of the original expression. Second, semi-compositional constructions (e.g. *make a mistake*) often share their syntactic pattern with literal verb + noun combinations (e.g. *make a cake*) or idioms (e.g. *make a meal*), which yields that their identification cannot be based on solely syntactic patterns. Third, since the syntactic and the semantic head of the construction are not the same (see Chapters 6 and 7 – the syntactic head being the verb and the semantic head being the noun –), they require special treatment when parsing. On the other hand, it should be mentioned that because of their semi-productivity (i.e. certain verbs tend to co-occur with nouns belonging to a given semantic class), it is possible to generate semi-compositional construction candidates: in Stevenson et al. (2004), a statistical method is applied to measure the acceptability of possible semi-compositional constructions, which correlates reasonably well with human judgments.

In the following, methods developed for identifying semi-compositional constructions are summarized shortly.

Dias (2003) presents a hybrid method for identifying multiword expressions. His system is based on word statistics and information from POS-tagging and syntactic parsing.

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.

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 con-

struction 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, which achieves 72% in classifying verb and noun combinations as literal or idiomatic. As far as it is suggested in their paper, combinations that are called semi-compositional constructions in this thesis are classified as either literal or idiomatic, depending on their individual characteristics (recall their being situated on a scale of productivity and idiomacity discussed in Chapter 4).

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.

Samardžić and Merlo (2010) analyze English and German semi-compositional constructions in parallel corpora: they pay special attention to their manual and automatic alignment. They found that linguistic features (i.e. the degree of compositionality) and the frequency of the construction both have an effect on aligning the constructions.

Gurrutxaga and Alegria (2011) extract idiomatic and semi-compositional noun + verb combinations 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.

9.3 Experiments

Earlier studies on the detection of semi-compositional constructions generally take syntactic information as a starting point (e.g. Cook et al. (2007), Bannard (2007), Tan et al. (2006), Tu and Roth (2011)), that is, their goal is to classify verb + object constructions selected on the basis of syntactic patterns as literal or idiomatic. However, we do not aim at classifying FX candidates filtered by syntactic patterns but at identifying them in running text without assuming that syntactic information is necessarily available. Thus, in our investigations[1],

[1]Thanks are due to István Nagy T. and Gábor Berend for their help in conducting the experiments described here.

we will pay distinctive attention to the added value of syntactic features on the system's performance.

9.3.1 Rule-based methods

We used several rule-based methods to detect English semi-compositional constructions in running texts. Texts were tokenized, POS-tagged (Toutanova and Manning, 2000), stemmed and dependency relations were assigned to sentences in a pre-processing step (using Stanford parser (Klein and Manning, 2003)).

The POS-rule method meant that each n-gram for which the pre-defined patterns (e.g. `VB.? (NN|NNS)`) could be applied was accepted as semi-compositional construction. Since the methods to follow rely on morphological information (i.e. it is required to know which element is a noun), matching the POS-rules is a prerequisite to apply those methods for identifying MWEs.

The 'Suffix' method exploited the fact that many nominal components in semi-compositional constructions are derived from verbs. Thus, in this case only constructions that matched our POS-rules and contained nouns ending in certain derivational suffixes were allowed.

The 'Most frequent verb' (MFV) method relied on the fact that the most common verbs function typically as light verbs (e.g. *do*, *make*, *take*, *have* etc.) Thus, the 12 most frequent verbs typical of semi-compositional constructions were collected and constructions that matched our POS-rules and where the stem of the verbal component was among those of the most frequent ones were accepted.

The 'Stem' method pays attention to the stem of the noun. In the case of semi-compositional constructions, the nominal component is typically one that is derived from a verbal stem (*make a decision*) or coincides with a verb (*have a walk*). In this case, we accepted only candidates that had a nominal component whose stem was of verbal nature, i.e. coincided with a stem of a verb.

Syntactic information can also be exploited in identifying MWEs. Typically, the syntactic relation between the verb and the nominal component in a semi-compositional construction is `dobj` or `partmod`. If it is a prepositional semi-compositional construction, the relation between the verb and the preposition is `prep`. The 'Syntax' method accepts candidates among whose members the above syntactic relations hold.

We also combined the above methods to identify semi-compositional constructions in our databases (the union of candidates yielded by the methods is denoted by $\cup$ while the intersection of the candidates is denoted by $\cap$ in the respective tables). Rule-based methods were evaluated first on our Wikipedia database (see Chapter 3), then methods were adapted to the SzegedParalellFX and results are presented in 9.3.2.

9.3.2 Results of rule-based methods

Results on the rule-based identification of semi-compositional constructions can be seen in Table 9.1. Our methods were first implemented for the source domain and later rules were modified by experts according to the characteristics of the target domain (rule-based domain adaptation). The corpus that is smaller in size and contains simpler annotation was selected as the source domain, i.e. the Wiki50 corpus is the source domain whereas the SzegedParalellFX is the target domain. For adaptation, characteristics of the corpora must be considered: in our case, differences in annotation principles and the topics of texts determined the modifications in our methods.

In our investigations, the performance of the POS-rules method is considered to be the baseline as opposed to earlier studies focusing on the idiomacity of verb + object pairs, where labeling all verb + object pairs as idiomatic was used as the baseline method. Since we did not restrict ourselves to identify semi-compositional constructions where the noun is the object of the verb, besides, our definition of semi-compositional constructions includes only part-of-speech information ("verb + noun combinations"), we voted for a baseline where only POS-rules are applied.

In the case of the source domain, the recall of the baseline (POS-rules) is high, however, its precision is low (i.e. not all of the candidates defined by the POS patterns are semi-compositional constructions). The 'Most frequent verb' (MFV) feature proves to be the most useful: the verbal component of the semi-compositional construction is lexically much more restricted than the noun, which is exploited by this feature. The other two features put some constraints on the nominal component, which is typically of verbal nature in semi-compositional constructions: 'Suffix' simply requires the noun to end in a given n-gram (without exploiting further grammatical information) whereas 'Stem' allows nouns derived from verbs. When combining a verbal and a nominal feature, union results in high recall (the combinations `typical verb + non-deverbal noun` or `atypical verb + deverbal noun`

Method	**SOURCE**			**T w/o ADAPT**			**T+ADAPT**		
POS-rules	7.02	76.63	**12.86**	5.2	81.47	**9.78**	5.07	79.4	9.52
Suffix	9.62	16.3	12.1	9.7	15.84	12.03	10.5	15.24	12.43
MFV	33.83	55.16	**41.94**	20.59	64.16	31.18	28.81	54.64	37.73
Stem	8.56	50.54	14.64	7.43	62.01	13.26	7.66	61.55	13.62
Suffix ∩ MFV	44.05	10.05	16.37	32.13	10.74	16.1	**48.31**	10.24	16.9
Suffix ∪ MFV	19.82	61.41	29.97	15.69	69.26	25.59	19.02	59.64	28.84
Suffix ∩ Stem	10.35	11.14	11.1	10.27	11.41	10.8	11.14	11.07	11.1
Suffix ∪ Stem	8.87	57.61	15.37	7.49	66.44	13.46	7.74	65.71	13.84
MFV ∩ Stem	39.53	36.96	38.2	27.96	49.4	**35.71**	38.87	43.45	**41.03**
MFV ∪ Stem	10.42	68.75	18.09	7.92	76.78	14.35	8.25	72.74	14.82
Suffix ∩ MFV ∩ Stem	**47.37**	7.34	12.7	**35.09**	8.05	13.1	47.41	7.62	13.13
Suffix ∪ MFV ∪ Stem	10.16	**72.28**	17.82	7.76	**78.52**	14.13	8.05	**74.29**	14.53

Table 9.1: Results of rule-based methods for semi-compositional constructions in terms of precision, recall and F-measure

are also found) while intersection yields high precision (`typical verb + deverbal noun` combinations are found only).

We also evaluated the performance of the 'Syntax' method without directly exploiting POS-rules. Results are shown in Table 9.2. It is revealed that the feature `dobj` is much more effective in identifying semi-compositional constructions than the feature `prep`, on the other hand, `dobj` itself outperforms POS-rules. If we combine the `dobj` feature with the best feature (namely, MFV), we can achieve an F-measure of 26.46%. The feature `dobj` can achieve a recall of 59.51%, which suggests that about 40% of the nominal components in our database are not objects of the light verb. Thus, approaches that focus on only verb-object pairs (e.g. Cook et al. (2007), Bannard (2007), Tan et al. (2006), Tu and Roth (2011)) fail to identify a considerable part of semi-compositional constructions found in texts.

Method	**P**	**R**	**F**
Dobj	10.39	59.51	17.69
Prep	0.46	7.34	0.86
Dobj ∪ Prep	2.09	38.36	3.97
Dobj ∩ MFV	31.46	22.83	**26.46**
Prep ∩ MFV	3.24	5.12	4.06
Dobj ∪ Prep ∩ MFV	8.78	19.02	12.02

Table 9.2: Results of syntactic methods for semi-compositional constructions in terms of precision, recall and F-measure

Methods developed for the source domain were also evaluated on the target domain with-

out any modification (T w/o ADAPT column in Table 9.1). Overall results are lower than in the case of the source domain, which is especially true for the 'MFV' method: while it performed best on the source domain (41.94%), it considerably declines on the target domain, reaching only 31.18%. The intersection of a verbal and a nominal feature, namely, 'MFV' and 'Stem' yields the best result on the target domain. It is interesting to see, however, that the intersection of all three features can achieve a better result on the target domain than on the source domain, which suggests that there are more instances of `typical verb` + `deverbal noun` combinations in the target domain.

Techniques for identifying semi-compositional constructions were also adapted to the other domain. The SzegedParalellFX corpus contained annotation for nominal and participial occurrences of semi-compositional constructions. However, the number of nominal occurrences was negligible (58 out of 1100) hence we aimed at identifying only verbal and participial occurrences in the corpus. For this reason, POS-rules and syntactic rules were extended to treat postmodifiers as well (participial instances of semi-compositional constructions typically occurred as postmodifiers, e.g. *photos taken*).

Since the best method on the Wiki50 corpus (i.e. 'MFV') could not reach such an outstanding result on the parallel corpus, we conducted an analysis of data on the unannotated parts of SzegedParalell. It was revealed that *have* and *go* mostly occurred in non light verb senses in these types of texts. *Have* usually denotes possession as in *have a son* vs. *have a walk* while *go* typically refers to physical movement instead of an abstract change of state (*go home* vs. *go on strike*). The reason for this might be that it is primarily everyday topics that can be found in magazines or novels rather than official or scientific topics, where it is less probable that possession or movement is described. Thus, a new list of typical light verbs was created which did not contain *have* and *go* but included *pay* and *catch* as they seemed to occur quite often in the unannotated parts of the corpus and in this way, an equal number of light verb candidates was used in the different scenarios.

The T+ADAPT column of Table 9.1 shows the results of rule-based domain adaptation. As for the individual features, 'MFV' proves to be the most successful on its own, thus, the changes in the verb list are beneficial. Although the features 'Suffix' and 'Stem' were not modified, they perform better after adaptation, which yields that there might be more deverbal nominal components in the PART class of the target domain, which could not be identified without extended POS-rules. In the light of this, it is perhaps not surprising that their combination with 'MFV' also reaches better results than on the source domain. The

intersection of 'MFV' and 'Stem' performs best after adaptation as well. Adaptation techniques add 1.5% to the F-measure on average, however, this value is 6.55% in the case of 'MFV' and 7.3% if syntax is also exploited (see Table 9.3).

The added value of syntax was also investigated in both the source and the target domains after adaptation. As represented in Table 9.3, syntax clearly helps in identifying semi-compositional constructions: on average, it adds 2.58% and 2.37% to the F-measure on the source and the target domains, respectively. The best result on the source domain, again, is yielded by the 'MFV' method, which is about 30% above the baseline. On the target domain, it is still the intersection of 'MFV' and 'Stem' that performs best, however, 'MFV' also achieves a good result. The fact that semi-compositional constructions do form a syntactic phrase can explain why syntactic information improves the system.

Method	**SOURCE+SYNT**			**T w/o ADAPT+SYNT**			**T+ADAPT+SYNT**		
POS-rules	9.35	72.55	16.56	6.8	73.57	12.44	6.89	72.97	12.59
Suffix	11.52	15.22	13.11	12.56	14.52	13.47	12.81	14.52	13.61
MFV	40.21	51.9	**45.31**	24.28	57.5	34.15	34.82	51.19	41.45
Stem	11.07	47.55	17.96	10.07	56.67	17.1	10.16	56.19	17.2
Suffix ∩ MFV	11.42	54.35	18.88	37.55	9.88	15.65	**55.03**	9.76	16.58
Suffix ∪ MFV	23.99	57.88	33.92	19.06	62.14	29.17	23.06	55.95	32.66
Suffix ∩ Stem	12.28	11.14	11.68	13.73	10.59	11.96	14.02	10.59	12.07
Suffix ∪ Stem	11.46	54.35	18.93	10.07	60.6	17.28	10.18	60.12	17.4
MFV ∩ Stem	46.55	34.78	39.81	33.1	44.52	**37.97**	44.04	40.48	**42.18**
MFV ∪ Stem	13.36	64.67	22.15	10.47	69.64	18.2	10.99	66.9	18.88
Suffix ∩ MFV ∩ Stem	**50.0**	6.79	11.96	**41.06**	7.38	12.51	53.98	7.26	12.8
Suffix ∪ MFV ∪ Stem	13.04	**68.2**	21.89	10.22	**71.07**	17.87	10.64	**68.33**	18.49

Table 9.3: Results of rule-based methods enhanced by syntactic features for semi-compositional constructions in terms of precision, recall and F-measure

9.3.3 Machine learning based methods

In addition to the above-described rule-based approach, we defined a machine learning method for automatically identifying semi-compositional constructions. For this, the Conditional Random Fields (CRF) classifier[2] was used (MALLET implementations (McCallum, 2002)). The basic feature set includes the following categories (Szarvas et al., 2006) enhanced with some FX-specific features such as MFV, stem and suffix features:

[2]A conditional random field is an undirected probabilistic graphical model, mostly used for labeling sequential data (Lafferty et al., 2001).

- **orthographical features:** capitalization, word length, bit information about the word form (contains a digit or not, has uppercase character inside the word, etc.), character level bi/trigrams; suffix;
- **dictionaries** of first names, company types, denominators of locations, the most frequent light verbs and stems of nouns; noun compounds collected from the English Wikipedia;
- **frequency information:** frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token which was derived from the Gigaword dataset[3];
- **shallow linguistic information:** Part of Speech; dependency relations;
- **contextual information:** sentence position, trigger words (the most frequent and unambiguous tokens in a window around the word under investigation) from the train text, the word between quotes, etc.

For conducting machine learning experiments, we randomly separated the target domain (i.e. SzegedParalellFX) into 70% as training set and 30% as test set. As the target domain contained several different topics, we separated all documents into training and test parts. We evaluated our various models in this resulting test set. In order to be able to compare rule based and machine learning based methods, we reevaluated our rule-based methods on this test set as well. Results can be seen in Tables 9.4 and 9.5.[4]

9.3.4 Results of machine learning based methods

To identify semi-compositional constructions we used Wiki50 and SzegedParalellFX to train CRF classification models (they were evaluated in a leave-one-document-out cross validation scheme[5]). Results are shown in Table 9.6. As the numbers indicate, the automatic detection of semi-compositional constructions in natural language texts is not an easy task, probably

[3]Linguistic Data Consortium (LDC), catalogId: LDC2003T05

[4]If results are compared to those achieved on the whole SzegedParalellFX, it is revealed that methods perfom considerably better on the whole dataset. This might be explained by the fact that the data were automatically separated into training and test sets and the distribution of different types of semi-compositional constructions might not be balanced, for instance, applying domain-specific POS-rules does not change the results, which suggests that there are no postmodifiers in the test dataset.

[5]In this setting, multiple evaluations are carried out on the dataset, which is separated into documents: each time one document being the test set and all the other documents the training set.

Method	**SOURCE**			**T w/o ADAPT**			**T+ADAPT**		
POS-rules	7.02	76.63	12.86	4.28	73.33	8.09	4.28	73.33	8.09
Suffix	9.62	16.3	12.1	9.83	14.58	11.74	9.92	15.42	12.07
MFV	33.83	55.16	**41.94**	16.25	51.25	24.67	22.48	49.17	30.85
Stem	8.56	50.54	14.64	6.55	57.08	11.75	6.55	57.08	11.75
Suffix ∩ MFV	44.05	10.05	16.37	32.35	9.17	14.28	**48.94**	9.58	16.03
Suffix ∪ MFV	19.82	61.41	29.97	13.01	56.67	21.17	15.51	55.0	24.2
Suffix ∩ Stem	10.35	11.14	11.1	11.59	11.25	11.42	11.6	12.08	11.84
Suffix ∪ Stem	8.87	57.61	15.37	6.55	60.42	11.82	6.55	60.42	11.82
MFV ∩ Stem	39.53	36.96	38.2	24.08	40.83	**30.29**	30.72	39.17	**34.43**
MFV ∪ Stem	10.42	68.75	18.09	6.64	67.5	12.09	6.97	67.08	12.63
Suffix ∩ MFV ∩ Stem	**47.37**	7.34	12.7	**40.0**	7.5	12.63	51.35	7.92	13.72
Suffix ∪ MFV ∪ Stem	10.16	**72.28**	17.82	6.53	**69.17**	11.94	6.8	**68.75**	12.39

Table 9.4: Results of rule-based methods for semi-compositional constructions in terms of precision, recall and F-measure evaluated on the 30% of SzegedParalellFX.

Method	**SOURCE+SYNT**			**T w/o ADAPT + SYNT**			**T+ADAPT+SYNT**		
POS-rules	9.35	72.55	16.56	5.93	69.17	10.92	5.93	69.17	10.92
Suffix	11.52	15.22	13.11	12.1	14.17	13.05	12.1	14.17	13.05
MFV	40.21	51.9	**45.31**	19.54	49.17	27.96	28.28	45.83	34.97
Stem	11.07	47.55	17.96	9.0	54.17	15.43	9.0	54.17	15.43
Suffix ∩ MFV	11.42	54.35	18.88	34.92	9.17	14.52	52.5	8.75	15.0
Suffix ∪ MFV	23.99	57.88	33.92	15.82	54.17	24.48	19.52	51.25	28.27
Suffix ∩ Stem	12.28	11.14	11.68	15.17	11.25	12.92	15.17	11.25	12.92
Suffix ∪ Stem	11.46	54.35	18.93	8.85	57.08	15.32	8.85	57.08	15.32
MFV ∩ Stem	46.55	34.78	39.81	27.81	39.17	**32.53**	37.18	36.25	**36.7**
MFV ∪ Stem	13.36	64.67	22.15	9.0	64.17	15.79	9.56	63.75	16.63
Suffix ∩ MFV ∩ Stem	**50.0**	6.79	11.96	**45.0**	7.5	12.86	**56.67**	7.08	12.59
Suffix ∪ MFV ∪ Stem	13.04	**68.2**	21.89	8.77	**65.42**	15.46	9.21	**65.0**	16.14

Table 9.5: Results of rule-based methods enhanced by syntactic features for semi-compositional constructions in terms of precision, recall and F-measure evaluated on the 30% of SzegedParalellFX.

due to the sparsity of data (there are many semi-compositional constructions that occurred only once or twice in the corpora) and the highly semantic nature of the task (e.g. the sequence *make decisions* may or may not be a semi-compositional construction, depending on context: it is definitely a semi-compositional construction in *the government will make decisions on foreign policy issues* whereas it is not in *they will make decisions taken by the government publicly available*).

However, as in the case of the rule-based approach, FX-specific features were adapted to the target corpus. In this way, for instance, the MFV dictionary did not contain *have* and

Corpus	Precision	Recall	F-measure
Wiki50	60.40	41.85	49.44
SzegedParalellFX	63.60	39.52	48.75

Table 9.6: Results of leave-one-out approaches in terms of precision, recall and F-measure.

go but *pay* and *catch* instead. In the case of the 'Stem' feature, we used domain specific dictionaries. Furthermore, when we trained on the parallel corpus, we extended the syntax feature rules with `partmod`. If we compare the results of rule-based approaches with the leave-one-out method, it is revealed that in both of the two corpora the CRF based approach can achieve better results.

In NLP applications, it is often the case that existing solutions must be applied to domains where only a limited amount of annotated resources can be found. Thus, regarding the two corpora as two different domains, we also conducted some experiments on machine learning based domain adaptation in order to model a scenario where a lot of outdomain (source) data are available and only a small amount of annotated indomain (target) data can be found. We applied the popular domain adaptation model described in Daumé III (2007). This model exploits the existence of source-, target-specific and general information, which means that each feature is represented three times in the model (Wiki50-based, SzegedParalellFX-based and general values are mapped to each feature).

As Wiki50 was the source domain, we used it as the training set with the above presented extended features, and we added to this training set some randomly selected sentences from the training set of the target domain. We extended the source training set with 10%, 20%, 25%, 33% and 50% of the target domain training sentences in a step-by-step fashion. The size of the target domain training set was about 10,000 sentences, hence the source dataset was extended with 1000, 2000, 2500, 3300 and 5000 annotated sentences from the target domain, respectively. As Table 9.7 shows, we evaluated the model trained with the source domain specific feature set (BASE) and the domain adapted trained model (ADAPT) too.

As the results show, the addition of even a little amount of target data has beneficial effects on performance in both the BASE and the ADAPT settings. Obviously, the more target data are available, the better results are achieved. Interestingly, the addition of target data affects precision in a positive way (adding only 10% of parallel data improves precision by about 11%) and recall in a negative way, however, its general effect is that the F-measure im-

proves. Results can be enhanced by applying the domain adapted model. Compared to the base settings, with this feature representation, the F-measure improves 1.515% on average, again primarily due to the higher precision, which clearly indicates that the domain adaptation techniques applied are optimized for precision in the case of this particular setting and datasets. The advantage of applying both domain-adapted features and adding some target data to the training dataset can be further emphasized if we compare the results achieved without any target data and with the basic feature set (34.88% F-score) and with the 50% of target data added and the adapted feature set (44.65%), thus, an improvement of almost 10% can be observed.

Training set	**BASE**			**ADAPT**		
Wiki50	29.79	42.08	34.88	31.04	43.33	36.18
Wiki50 + 10% paralell training set	40.44	37.91	39.14	42.72	37.91	40.18
Wiki50 + 20% paralell training set	40.09	38.75	39.40	43.60	38.33	40.79
Wiki50 + 25% paralell training set	41.96	39.16	40.51	47.37	37.5	41.86
Wiki50 + 33% paralell training set	45.78	38.41	41.79	46.44	40.83	43.46
Wiki50 + 50% paralell training set	47.89	37.91	42.32	49.24	40.83	44.65

Table 9.7: Results of machine learning approach for semi-compositional constructions in terms of precision, recall and F-measure, evaluated on the 30% of SzegedParalellFX.

9.3.5 Discussion of results

Our adapted methods achieved better results on the target domains than the original ones. However, there is not much difference between the performance on the source and the target domains, which might be related to differences in the distribution of (a)typical light verb constructions. It seems that the target domain contains more instances of `typical verb` + `deverbal noun` combinations than the source domain. However, 'MFV' proves to be the most important feature for both domains, which suggests that with a well-designed domain-specific list of POS-rules and light verb candidates, competitive results can be achieved on any domain, especially if enhanced with syntactic features.

The adaptation of light verb candidates was carried out on the basis of semantic considerations, that is, verbs that typically occurred in one of their non light verb senses were omitted. On the other hand, this feature proved to be the most crucial one in identifying semi-compositional constructions and its added value was the highest in rule-based domain adaptation (more than 4 times as much as the average value). It is suggested that adaptation rules and techniques related to semantic information are the most beneficial in domain

adaptation, which is also demonstrated by applying domain-specific semantic rules in named entity recognition (Chiticariu et al., 2010).

The importance of domain-specific annotated data is also underlined by our machine learning experiments. Simple cross-training (i.e. training on Wiki50 and testing on Szeged-ParalellFX) yields relatively poor results but adding some parallel data to the training dataset efficiently improves results (especially precision).

If rule-based methods and machine learning approaches are contrasted, it can be seen that machine learning settings almost always outperform rule-based methods, the only exception being when there are no parallel data used in training. This suggests that if no annotated target data are available, it might be slightly more fruitful to apply rule-based methods, however, if there are annotated target data and a larger corpus from another domain, domain adaptation techniques and machine learning may be successfully applied. In our settings, even about 1000 annotated sentences from the target domain can considerably improve performance if large outdomain data are also exploited.

Obviously, our methods can be further improved. First, stemming of the nominal components of semi-compositional constructions can be enhanced by e.g. wordnet features in order to eliminate false negative matches originating from the stemming principles of the Porter stemmer (e.g. the stems of *decision* and *decide* do not coincide). Second, the lists of possible light verb candidates can be extended as well. Finally, investigations on other domains and corpora would also be beneficial, which we would like to carry out as future work.

9.4 The role of detecting multiword expressions in a processing toolchain

The identification of multiword expressions also evokes the question of when to identify them in the processing toolchain. For a natural language text to be automatically processed, it first must be segmented into paragraphs, sentences and words. This is followed by morphological analysis and POS-tagging, then syntactic parsing takes place, which is followed by semantic parsing (if any). Higher level applications such as machine translation or information retrieval can then exploit linguistic information provided in these preprocessing steps.

Traditionally, the identification of multiword expressions is based on syntactic informa-

tion, that is, it follows syntactic parsing, e.g. Martens and Vandeghinste (2010) make use of dependency trees when identifying syntactically motivated multiword expressions. However, the answer to this question depends on the information exploited when identifying multiword expressions on the one hand and on the specific type of the multiword expressions to be identified on the other hand. Needless to say, if syntactic information is used in identifying a multiword expression (as it is in e.g. Sass (2008; 2009)), the identification process cannot precede syntactic parsing. On the other hand, in most cases multiword expressions require special treatment at early levels of parsing as well – for instance, it is advisable for the syntactic parser to treat certain types of multiword expressions as one unit (compounds, semi-compositional constructions, multiword prepositions and conjunctions etc.).

In this spirit, 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.

As our results show, morphological analysis and POS-tagging seem to be necessary for the identification of semi-compositional constructions. This means that they cannot be efficiently detected before POS-tagging information is available, however, syntactic information also contributes to their identification. On the other hand, we also argued in Chapter 6 that treating semi-compositional constructions as one syntactic unit has beneficial effects on certain applications, e.g. information extraction (see also Chapter 11). This may be reached if in a post-processing step after parsing, semi-compositional constructions are identified and the syntactic relation between their nominal and verbal component is relabeled in order to assure that they have a special relationship and belong together. From an applicational point of view, the advantages of treating compound verbs (and named entities) as one unit in phrase-based statistical machine translation are emphasized in Pal et al. (2010).

9.5 How to identify Hungarian semi-compositional constructions?

The recognition of semi-compositional constructions cannot be solely based on syntactic patterns for other (productive or idiomatic) combinations may exhibit the same verb + noun scheme (see Chapter 4). However, in agglutinative languages such as Hungarian, nouns can have several grammatical cases, some of which typically occur in a semi-compositional 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 semi-compositional constructions being right before or after the verb[6]), it is most likely a semi-compositional construction. To illustrate this, we offer some examples:

(9.1) *vizet* *hoz*
water-ACC brings
'to bring some water'

(9.2) *zavarba* *hoz*
trouble-ILL brings
'to embarrass'

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

(9.3) *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 semi-compositional 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* 'bring' is under investigation. Thus, morphosyntactic information can enhance the identification of semi-compositional constructions. Similarly, Das

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

et al. (2010) propose a method for extracting Bengali complex predicates on the basis of shallow morphological information and seed lists of light verbs.

Most semi-compositional constructions have a verbal counterpart derived from the same stem as the noun, which entails that it is mostly deverbal nouns that occur in semi-compositional constructions (as in *make/take a* ***decision*** compared to ***decide*** or ***döntést*** *hoz* vs. ***dönt*** in Hungarian, see also Test 12 in Chapter 4). The identification of such nouns is possible with the help of a morphosyntactic parser that is capable of treating 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 semi-compositional construction. Thus, an algorithm that makes use of morphosyntactic and derivational information and previously given lists can be constructed to identify Hungarian semi-compositional constructions in texts.

As our earlier results on Hungarian data indicate (Zsibrita et al., 2010), identifying multiword named entities before morphological parsing leads to better results in POS-tagging than identifying them only after morphological parsing. The identification algorithm used very simple features such as orthographical form, lemmas, upper/lowercase characters, lists etc. This might suggest that for other types of multiword expressions, simple methods are worth examining, which can later serve as a basis for implementing more complex systems. For instance, multiword conjunctions and prepositions (i.e. non-productive multiword expressions that do not exhibit syntactic flexibility) could be identified with the help of orthographical information and lists (thus, there is no need for morphological or syntactic information). Nominal compounds can also be identified by using orthographical information, lemmas and POS-tags (thus, morphological analysis and POS-tagging are necessary in this case) and lists can also enhance the performance of the system.

With regard to the identification of (especially Hungarian) semi-compositional constructions and the implementation of an FX-detector, the following features may prove to be essential besides co-occurrence statistics (cf. Muischnek and Kaalep (2010) on Estonian data):

- orthography;
- POS-tags (we are looking for verb and noun combinations);
- lemma (for the verb because of its morphosyntactic flexibility);
- morphological analysis (the suffix of the noun can have an indicative role in agglutina-

tive languages and the nominal component with a verbal stem can also be suggestive);

- windows of proper size (because of split semi-compositional constructions and determiners and modifiers that might intervene between the two basic components of the construction).

In the future, we plan to develop an algorithm based on the above principles to identify semi-compositional constructions in Hungarian texts.

9.6 Summary of results

In this chapter, questions related to the automatic identification of semi-compositional constructions were discussed. Statistical, rule-based and hybrid models for identifying multiword expressions were presented and the questions of how and when to identify multiword expressions in general and semi-compositional constructions in particular were raised. As for semi-compositional constructions, we presented our rule-based and CRF-based methods that are capable of identifying English constructions:

- their identification can be based on information provided by a morphological parser, POS-tagger and syntactic parser:
  - lemmas
  - stems of the verb and the noun
  - POS-tags
  - suffixes
  - dependency relations;
- 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;
- their identification can take place in a post-processing step after syntactic parsing;
- in this way, the output of the syntactic parser can be effectively exploited by higher-level applications.

The following chapters will present the NLP treatment of semi-compositional constructions in several applications where it is crucial to know which text spans are to be seen as one unit. In other words, at the level of those applications it is taken for granted that semi-compositional constructions are known to the system.

Chapter 10

Semi-compositional constructions and word sense disambiguation

10.1 Introduction

Word sense disambiguation (WSD) aims at resolving ambiguities (homonymy, polysemy) in texts. This problem has been present in natural language processing since the beginnings, and it is an important intermediate task for most NLP applications (e.g. text comprehension, human-machine interaction, machine translation and information retrieval and extraction).

In this chapter, the task of word sense disambiguation is analyzed from the viewpoint of semi-compositional constructions. After a brief description of the task and existing approaches, a database developed for Hungarian is presented, which serves as a base for our investigations. It is further discussed how lexical features of semi-compositional constructions can be exploited in WSD and a case study involving two Hungarian verbs is also carried out. The chapter concludes with a summary of results.

10.2 The task of word sense disambiguation

The main goal of word sense disambiguation is to select the appropriate sense for the word under investigation in context from a pre-defined sense inventory. Senses are usually determined on the basis of a dictionary or ontology. For instance, the Hungarian word *nap* can be used in the sense of 'sun' or 'day' as well. The result of word sense disambiguation can

be exploited in machine translation (the proper translational equivalent can be selected) and information extraction and retrieval systems as well (only documents containing the word in the relevant sense are retrieved to the user).

Word sense disambiguation applications can be divided into two major groups on the basis of the limits of their applicability and the degree of granularity. With regard to scope, distinction can be made between "all-words" (applied to overall vocabulary) and "lexical sample" (applied to selected word forms only) methods of WSD. In other words, the two methods differ in that all or only some of the words in a corpus are disambiguated. As for granularity, fine-grained (at least 6-7 senses) and coarse-grained (at most 4-5 senses) levels can be distinguished.

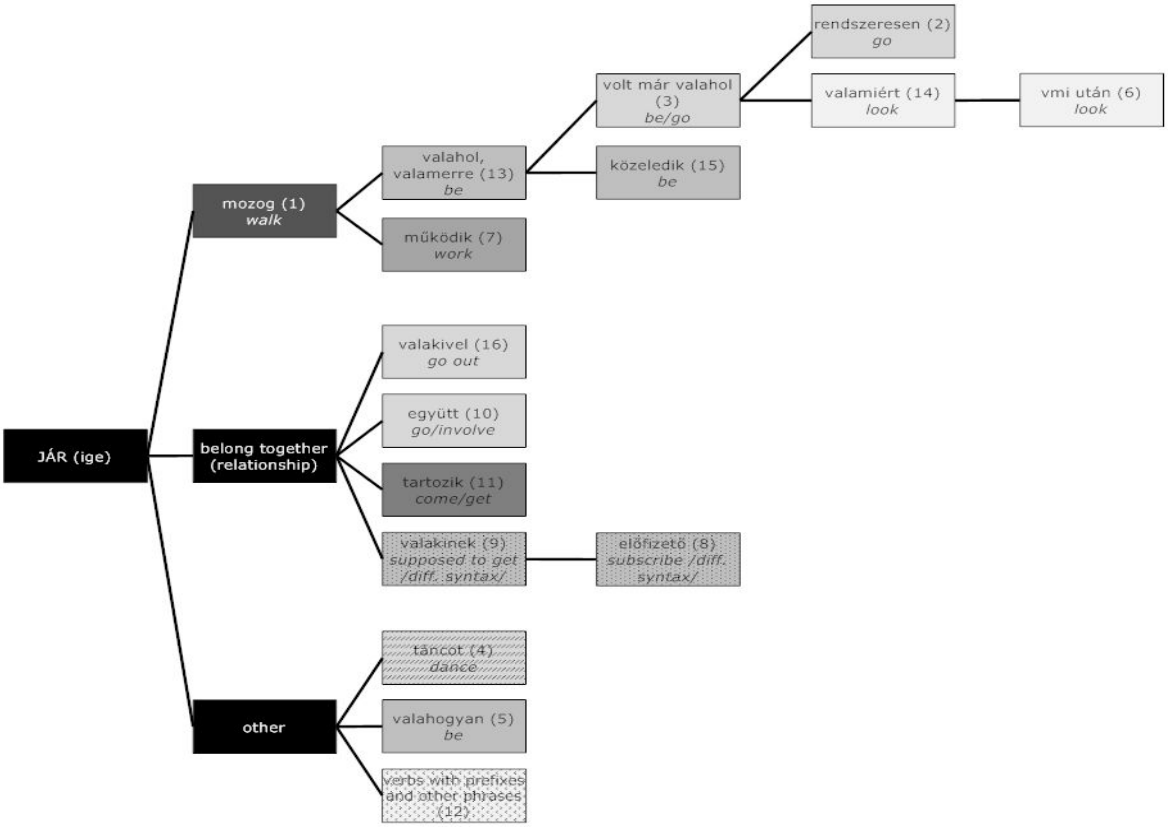


Figure 10.1: Senses of *jár* 'go'

Figure 10.1 represents different levels of granularity in the case of the verb *jár* 'go'. In the Hungarian word sense disambiguated corpus (Vincze et al., 2008), 16 different senses were selected for annotation. Every box in the chart represents an individual sense denoted by numbers and an approximate English equivalent is also provided. However, the senses can be unified into 3 major groups: *movement*, *relationship* and *other*. Hence, the verb *jár* 'go' has 16 fine-grained but only 3 coarse-grained senses.

Word sense disambiguation research concerning, first, English and, later, other languages as well was related in a greater part to SensEval (Kilgariff, 2001; Mihalcea and Edmonds, 2004) workshops organized by ACL-SIGLex. The book Word Sense Disambiguation (Agirre and Edmonds, 2006) and two publications of the SemEval workshops (Agirre et al., 2007; Erk and Strapparava, 2010) as the next step in the SensEval series provide a detailed overview of results up till then.

In relation to the development of Hungarian-English and English-Hungarian machine translation systems, word sense disambiguation tasks in Hungarian have been carried out for a long time (Miháltz, 2005; Miháltz and Pohl, 2005).

To create the first Hungarian WSD corpus (Vincze et al., 2008), 39 suitable word samples (21 nouns, 12 verbs and 6 adjectives) were selected for the purpose of word sense disambiguation. These word forms are considered to be frequent in Hungarian language usage and they have more than one well-defined sense. The Hungarian National Corpus (Váradi, 2002) and its Heti Világgazdaság (HVG) subcorpus provided the basis for corpus text selection. This corpus is a fine-grained lexical sample corpus. When building the corpus, the XML format designed for corpora prepared for WSD tasks of the SensEval/SemEval international conference workshops was followed.

In the first phase of the work, possible senses of the selected 39 words were defined. In this process, we relied on paper and electronic forms of The Concise Dictionary of the Hungarian Language (Pusztai, 2003) on the one hand and our linguistic intuition on the other hand. Senses that could be definitely distinguished on the basis of their dictionary definition were considered separate senses.

Following international standards, annotation of corpus samples were carried out by two independent linguists, which means that they were not allowed to cooperate. Double annotation made it possible to measure the consistency level of the database and to correct incidental annotation errors after comparing the two. Finally, a third, independent annotator checked the cases where annotations were dissimilar and finalized the tags of these samples.

When planning the corpus, 300-500 samples of each word form were to be annotated. This makes it possible that the size of subcorpora prepared for the individual word forms can be compared to data available for other languages. However, the finalized database also contains unannotated samples and samples with single annotation, which were annotated only by one of the linguists.

To perform a WSD task, well-defined and clearly distinguishable senses are of primary

importance. Thus, the role of the lexicon and the lexical representation of words is essential in WSD as well (see Chapter 8). Besides, the number of senses also plays a crucial role in successful annotation: the more senses are supposed, the less precise the annotation will be, especially when senses are vaguely defined or on the contrary, definitions for different senses are too similar. On the basis of an earlier study (Vincze et al., 2008), 3-5 possible senses for each word to be disambiguated seem to yield the highest level of precision for both human annotators and the computer and this number of possible senses seems to be ideal for NLP applications as well.

10.3 Semi-compositional constructions in word sense disambiguation

Multiword expressions pose a special problem for WSD because in many of the cases it is possible to know which sense the given word form assumes within the phrase. E.g. in the following Hungarian proverb the sense of *víz* 'water' can be identified precisely: 'a mass of water covering an area of the Earth's surface':

(10.1) *sok* *víz* *lefolyik* *a* *Dunán* *addig*
much water flows the Danube-SUP till.that.time
'it'll be a long time'

However, in other examples the selection of the sense is not so straightforward, thus, it is still a question whether parts of a fixed expression should be tagged on the basis of their literal meaning or a special sense (e.g. 'other' or 'collocation') should be reserved for these cases: in the case of decomposable idioms, the sense attributed to their parts within the MWE may be also given (see Chapter 2).

As for semi-compositional constructions, based on Alonso Ramos's (2004) classification, Sanromán Vilas (2009) proposes to classify verbs found in Spanish semi-compositional constructions into the following classes:

- pure light verbs,
- light verbs that share some semantic content with their general senses,
- light verbs used in the same way as in the general sense,

- light verbs with only light verb senses.

In line with this, for verbal components in the last two groups it suffices to define only one sense: e.g. *eszközöl* 'to carry out' as having only a light verb sense or *mond* 'to say / tell' used in its general sense as in:

(10.2) *köszönetet* *mond*
thank-ACC says
'to say thank you'

However, the possibility of reserving a special light verb sense for verbal components is still open for the other two groups. This option can prove to be useful when there is no need for fine-grained WSD. However, there are cases when a verbal component occurs together with several semantic groups of nouns, in each case instantiating a (slightly) different sense (Apresjan et al., 2007). In such cases, it is possible to assume several well-distinguishable light verb senses. An illustrious example from Hungarian is the verb *hoz* 'bring', which seems to occur in two different grammatical constructions with nouns belonging to four semantic classes (see also Chapter 8):

- a noun denoting profit or loss: [~t] hoz (e.g. *nyereséget hoz* 'to produce profit')
- a noun denoting a verbal act: [~t] hoz (e.g. *intézkedést hoz* 'to make arrangements')
- a noun denoting a turn: [~t] hoz (e.g. *fordulatot hoz* 'to change')
- a noun denoting a state: [~ba / ~ra] hoz (e.g. *zavarba hoz* 'to confuse')

With a noun belonging to the first semantic class, the English equivalent of *hoz* could be 'produce' while with that belonging to the second one it means 'to create something' or 'to make something come into being'. Finally, with nouns denoting turns or states, *hoz* means 'to cause that something starts to be in another state' and 'to cause that something starts to be in a specific state', respectively. The first two senses might be unified as 'to create'. This reduction of senses is supported by the shared grammatical structure as well: in both types of constructions, the nominal component bears an accusative case. The latter two senses can also be subsumed as 'to cause that something starts to be in another state', however, they differ in the grammatical case of the nominal component. In this way, three "general" and two light verb senses of *hoz* can be assumed:[1]

[1]The reader may notice that one of the general senses coincides with one of the light verb senses, namely, 'to create' occurs in both groups. Their unification, however, is dubious since *hoz* used in the general 'to produce'

- general senses:
  - to carry sg to a nearer place
  - to go together with sg
  - to produce (about plants)
- light verb senses:
  - to create
  - to cause that something starts to be in another state

As for the nominal component in *nyilvánosságra hoz* (publicity-SUB brings) 'to publish', it can also have two basic senses, namely 'publicity' and 'public' (i.e. people as a whole). Thus, when determining the senses of the nominal component and the verb in the construction *nyilvánosságra hoz*, there are ten (i.e. two times five) variations for the whole construction. To choose the correct version, the following steps may prove useful (provided that proper sense definitions are available for both the noun and the verb). First, it is necessary to recognize that the noun and the verb form a semi-compositional construction, which can be carried out by using a properly designed algorithm (see Chapter 9). In this way, the three "general" senses of the verb are eliminated. Then, the morphosyntactic analysis of the construction reveals that the nominal component is in the sublative case, which excludes the possibility of having the verbal sense 'to create' in the construction. Thus, *hoz* has the sense 'to cause that something starts to be in another state', which sense is paired with nouns denoting a state. Finally, of the two possible senses of *nyilvánosság*, it is 'publicity' that denotes a state, thus, this is the appropriate sense for the nominal component.

If no light verb senses of the verbal component are assumed, *hoz* can have the following merged senses:

1. to cause that something starts to be in another state/place (generally with a nominal component in the illative or sublative case);
2. to cause, create, yield (generally with a nominal component in the accusative case).

sense does not always appear in a semi-compositional construction as in: *A kukorica tavaly szépen hozott* (the corn last.year nicely produce-past3sg) 'The corn produced well last year'.

Again, the morphosyntactic analysis of the nominal component can enhance the choice of the appropriate sense. On the other hand, the senses are determined on the basis of dictionary entries, thus, the lexical representation of semi-compositional constructions (Chapter 8) also plays an important role in word sense disambiguation.

10.4 A case study: *kerül* 'get' and *tart* 'hold/keep/last'

In order to perform some experiments on learning models for word sense disambiguation, we selected two verbs, *kerül* 'get' and *tart* 'hold/keep/last' from the Hungarian WSD corpus – these verbs also occurred frequently in our database of semi-compositional constructions – and redefined senses for them, with respect to their possible usage as light verbs. Texts of the original WSD corpus were annotated for the newly defined senses, which were the following:

Kerül:

1. cost sg (e.g. *5000 dollárba kerül* 'it costs $5000')
2. move, get somewhere (e.g. *gyermekkorában Angliába került* 'he got to England as a child')
3. avoid (e.g. *kerüli a dohányzást* 'he avoids smoking')
4. happen, take place (e.g. *aláírásra került* 'it was signed')
5. other

Tart:

1. move, get somewhere (e.g. *a hegy felé tart* 'he is going to the mountain')
2. be in a state (e.g. *kapcsolatot tart* 'he keeps in touch with sy')
3. fear sy/sg (e.g. *tart a férjétől* 'she is afraid of her husband')
4. take some time (e.g. *három hétig tart* 'it takes three weeks')
5. perform, carry out (e.g. *beszédet tart* 'deliver a speech')
6. think, recognize (e.g. *hülyének tart* 'take him for an idiot')

7. other

Among the senses defined above, there are some FX-specific ones, that is, senses that are often associated with semi-compositional constructions, namely, the fourth one of *kerül* and the fifth one of *tart*: they are similar to the definition given for the lexical function **Oper** (Apresjan, 2004). Originally, there were 9 senses of *tart* and 8 senses of *kerül*, thus, it can be analyzed what the effect of sense reduction will be on the accuracy of manual annotation and the performance of WSD algorithms.

10.4.1 Analysis of corpus data

Two annotators worked on the texts and annotated 310 instances of *kerül* and 311 instances of *tart*. The agreement rate between the annotations can be seen in Table 10.1. Differences in annotation were later resolved by a third annotator, yielding the gold standard annotation. As the data reveal, one of the annotators achieved almost the same accuracy[2] as the gold standard annotation while the other one also had somewhat lower but still good results. The overall accuracy of annotations was 94.47%, which is 2.2% better than the accuracy with the original senses (92.27%).

	kerül	*tart*
Annotator 1 vs. Annotator 2	94.52	89.07
Annotator 1 vs. gold standard	95.16	90.68
Annotator 2 vs. gold standard	99.35	98.07

Table 10.1: Agreement rates between annotations

Differences in annotations were related to the question of annotating parts of multiword expressions on the basis of their literal meaning or a special sense should be given to them (cf. 10.3). Annotator 1 usually chose the meaning ‘be in a state’ for constructions such as:

(10.3) *sakkban* *tart*
chess-INE holds
‘to hold in check’

tiszteletben *tart*
respect-INE holds
‘to respect’

[2]Accuracy is the ratio of the number of properly classified instances over the number of instances that have been evaluated.

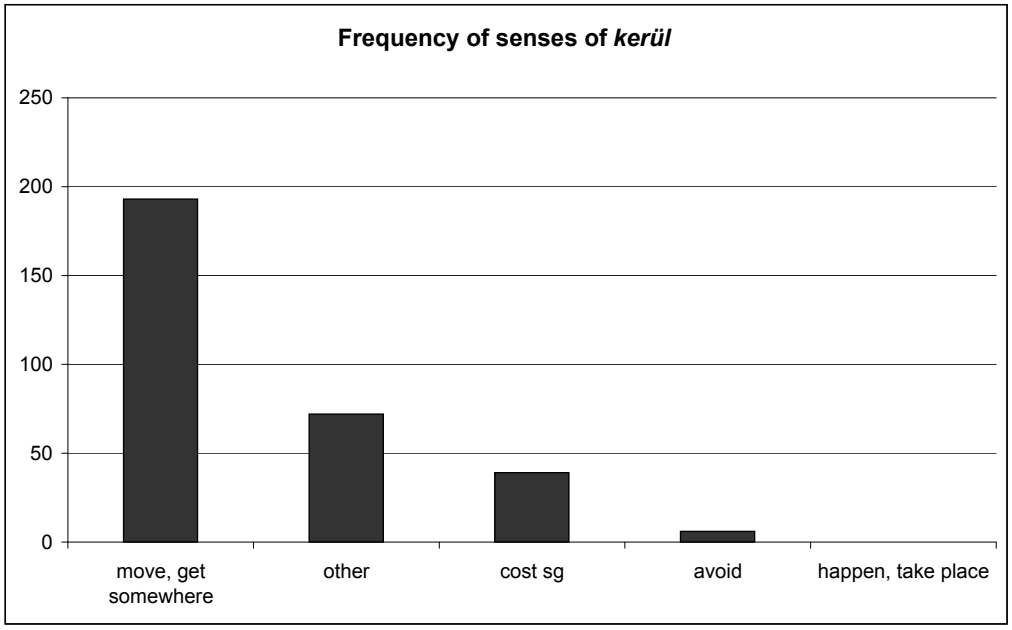
whereas Annotator 2 and the gold standard annotation chose the 'other' class for them. However, the fine-tuning of the annotation principles would eliminate this difference in annotation. In all, these results suggest that fewer senses and more precise definitions can enhance the quality of manual annotation.

The frequency of each sense of the two verbs are shown in Figure 10.2. The most frequent sense was 'to move, to get somewhere' for *kerül* (193 instances) and 'to think, to recognize' for *tart* (136 instances). As for the FX-specific senses of the two verbs, the percentage rate of 'to perform, to carry out' was about 9.3%, however, 'to happen, to take place' did not occur at all. The latter can be attributed to the domain of the texts. The WSD corpus contains texts from Heti Világgazdaság (HVG), a newspaper with economy-related topics. If the semi-compositional constructions containing *kerül* found in the HVG subcorpus of the Szeged Treebank are examined in detail, it is revealed that only one of them exhibits the sense 'to perform, to carry out', thus, it is suggested that this is a characteristics of the language use and style of this newspaper. This may be also traced back to the fact that semi-compositional constructions with the verb *kerül* are considered less acceptable than other constructions (see Chapter 5) and journalists try to avoid them.

10.4.2 Experiments

When evaluating word sense disambiguation procedures, the rate of the most frequent sense (MFS) is usually considered as baseline accuracy since this is the precision that can be obtained trivially. A system output (disambiguated occurrences of word forms) can be considered evaluable if it assigns proper senses to word forms in a proportion higher than the rate of the MFS.

As the above verbs have only one FX-specific sense each, if semi-compositional constructions are identified before word sense disambiguation (as suggested in Chapter 9), the proper sense of the verb can be easily selected given that the FX-specific sense is distinctively marked in the sense inventory. On the other hand, morphosyntactic information might also prove useful: the case suffix of the noun can be indicative of the sense of the verb, e.g. the sublative case next to the verb *kerül* is likely to co-occur with a light verb instance of the verb. However, in our experiments, no semi-compositional constructions were previously known to the system, that is, light verb senses were not treated differently from the other senses defined for the two verbs.



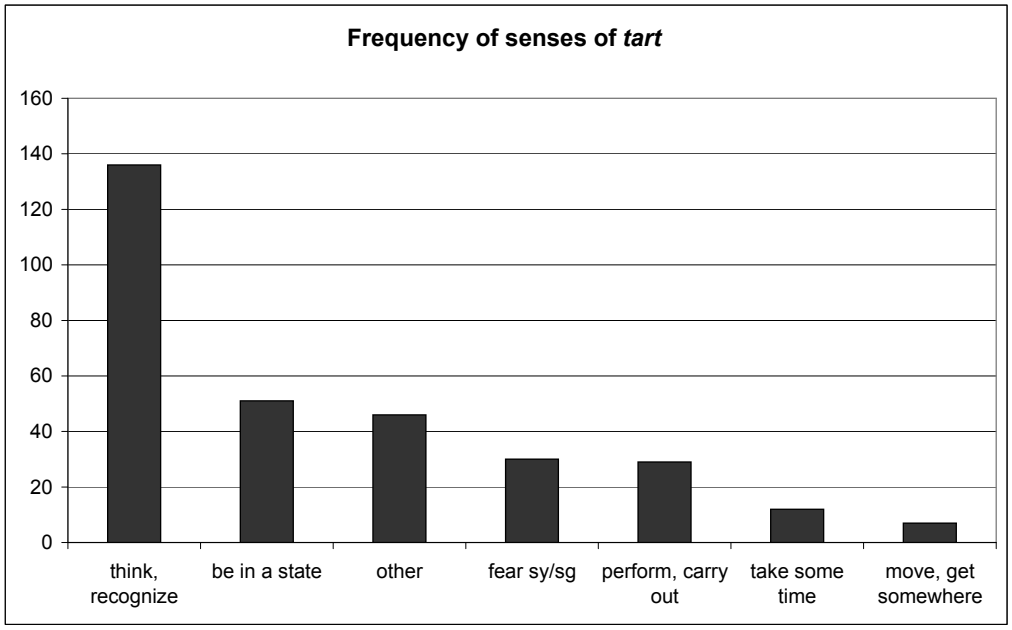


Figure 10.2: The frequency of senses of *kerül* and *tart*

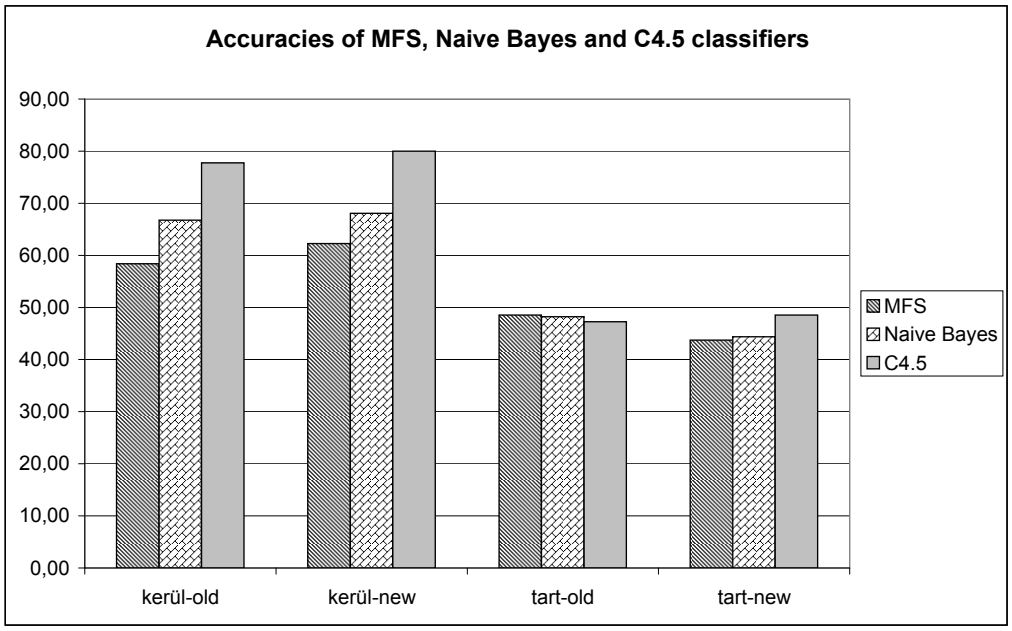


Figure 10.3: Accuracy of classifiers

To build supervised learning models it is necessary to convert the examples of the task to a representation that can be easily handled by the algorithm. In our experiments we applied a token unigram Vector Space Model[3] as feature representation and local contextual features in a window of size +/-3. We considered only nouns, verbs, adjectives and adverbs as contextual features and used lemmatized word forms. In the case of word sense disambiguation, this representation is obviously too simple, as morphosyntactic, topic and other features are not considered here. Thus, the results described here are intended as a comparative baseline.

We performed a leave-one-out evaluation of Naïve Bayes[4] and C4.5[5] classifiers with default parameters, as of the Weka package (Hall et al., 2009) and 5 instances per leaf. We carried out the experiments for both the old and the new datasets (i.e. texts annotated with the old and the new senses). The accuracy achieved by the classifiers for each setting is presented in Figure 10.3.

[3]Vector space model is an algebraic model for representing documents as vectors of identifiers.

[4]A Naïve Bayes classifier is a simple probabilistic classifier, which assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

[5]C4.5 is an algorithm that generates decision trees on the basis of which classification can be carried out.

Our results show that statistical models outperform the most frequent sense (MFS) heuristic currently used by applications that perform no WSD with the exception of the old *tart* dataset, where neither C4.5 nor Naïve Bayes can beat the baseline (48.55% in this case). On the other hand, both classifiers achieve a better accuracy than the MFS in the case of the new datasets.

In the case of *kerül*, there is not much difference in accuracy concerning the old and new datasets: the statistical models outperform the MFS heuristic to approximately the same degree, however, the new dataset yields slightly better results than the old one (the best model (C4.5) achieves 80%). This can be attributed to the fact that no matter the original 8 senses of *kerül* were reduced to 5, the number of senses present in the corpus did not change significantly: 5 old and 4 new senses occurred in the corpus. Thus, the difficulty of sense classification was similar in both scenarios.

As for *tart*, the datasets contain annotated samples for all the senses defined (i.e. 9 senses in the old and 7 in the new datasets). This makes the WSD process more complicated since there are more classes where an instance of the verb to be disambiguated can belong to. It is also interesting to see that after the sense reduction, the frequency of the MFS – thus the baseline – decreased (from 48.55% to 43.73%), which reflects that the reduction of senses was carried out by redefining senses from another perspective (e.g. instead of distinguishing abstract and concrete senses of *tart*, senses referring to motion and states were introduced) and not by merging the already existing ones. However, sense reduction has a positive effect on the performance of the statistical methods: in the case of the new dataset, both Naïve Bayes and C4.5 achieve higher accuracies than the baseline.

The benefits of the new sense inventories are more perceivable if the F-measures for similarly defined senses are contrasted. Figure 10.4 illustrates the precision, recall and F-scores of the C4.5 algorithm for two senses of *tart* (one of which is the FX-specific sense) and one of *kerül* which were similarly defined in the old and new sense inventories. In the case of *kerül*, the precision and F-measure are slightly improved while recall remains the same in the new dataset. Concerning the senses of *tart*, there are considerable differences especially in the precision values, moreover, recall values are also higher, which yields substantially higher F-measures for the newly defined senses (1.85 and 1.57 times higher for the two senses, respectively).

All in all, the results of our experiments show that the new sense inventories lead to higher results in both manual and automatic word sense disambiguation. A higher level of

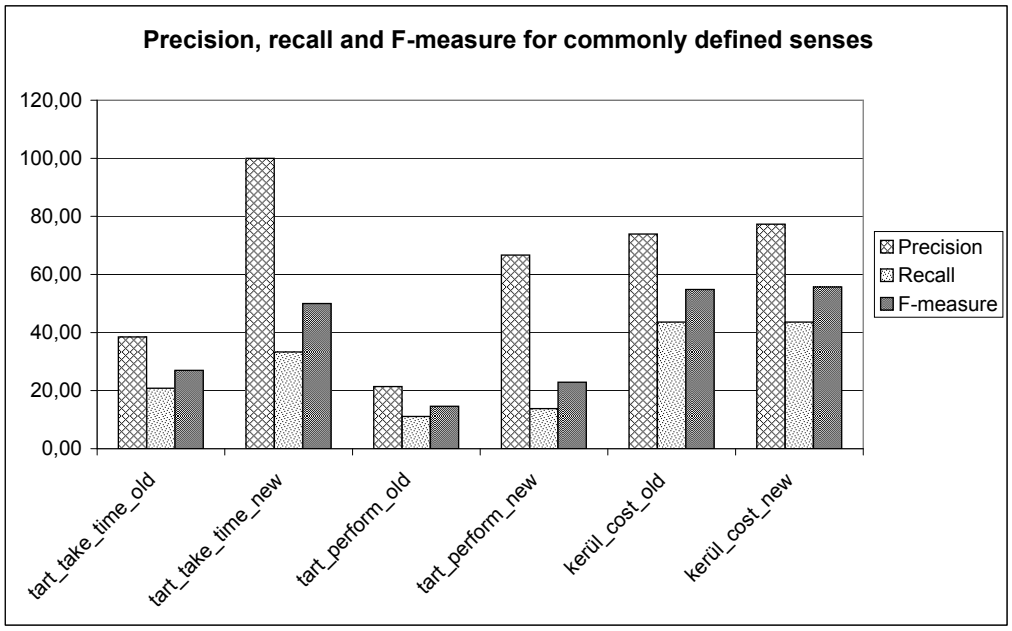


Figure 10.4: Precision, recall and F-measure for commonly defined senses

agreement and accuracy can be reached for the verb *kerül* than *tart*, however, the difference is considerable only between the performance of the statistical models. This can be explained by the greater number of senses to be considered: as for *tart*, there were 7 senses while *kerül* had only 4 senses that occurred in the datasets. Comparing the results of our experiments on the old and new sense inventories defined for the two verbs, it is revealed that the reduction of senses apparently leads to higher performance in the case of both verbs on the one hand, and better results can be achieved for a verb with fewer senses (i.e. for *kerül*) on the other hand. Focusing on the FX-specific sense of *tart*, the positive effects of the new sense inventories were also recognizable in that the F-measure of the new 'perform' class was 1.57 times higher than the old one. It is expected that more sophisticated models (exploiting e.g. morphosyntactic information) can achieve even better results, which hopefully will be soon implemented.

10.5 Summary of results

In this chapter, semi-compositional constructions were analyzed from the aspect of word sense disambiguation. Emphasis was put on the following points:

- well-defined senses are necessary for both the nominal component and the verb;
- light verb senses of the verb can be distinguished from other senses on the basis of theoretical considerations;
- detecting multiword expressions – thus, semi-compositional constructions too – prior to WSD can have a positive effect on the performance of the system;
- morphosyntactic information can also enhance WSD;
- the reduction of senses and more precise sense definitions seem to yield a higher accuracy in both manual and automatic word sense disambiguation.

In order to perform WSD tasks more efficiently, precise sense definitions for light verbs and a competent algorithm for identifying semi-compositional constructions are highly needed. These require theoretical and empirical work as well, which hopefully will be carried out in the future.

Chapter 11

Semi-compositional constructions in information extraction and information retrieval

11.1 Introduction

Information extraction (IE) seeks to process large amounts of unstructured text, 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.

Information retrieval (IR) aims at selecting the appropriate documents from a set that match the user's query. It differs from information extraction in several aspects:

- IR provides relevant (parts of) documents to the user while IE extracts structured information from unstructured text;
- IR solutions are typically easier and quicker than IE systems;
- IR requires less expert knowledge, thus it is easier to expand existing techniques to other domains;
- the processing of the results of an IR application requires more human effort than that of an IE application.

IR techniques are widely used in several NLP fields, e.g. question answering and summarization (Jurafsky and Martin, 2008). They are based on keyword searching, that is, keywords within a user query are looked for in documents and those that contain the relevant words are given to the user.

In this chapter, the role of semi-compositional constructions are analyzed in three subfields of information extraction and in information retrieval in general. It is shown how semantic frame mapping, semantic role labeling and modality detection can profit from the proper treatment of semi-compositional constructions and their usability in information retrieval is also discussed.

11.2 Information extraction

In this section, the treatment of semi-compositional constructions is discussed in three different fields of information extraction, namely, semantic frame mapping, semantic role labeling and modality detection.

11.2.1 Semantic frame mapping

Semantic frames represent lexical units as a pair of words and a sense and they also include information on the syntactic and semantic features of the participants of the event the frame describes. A semantic frame can contain multiple lexical units that express the same sense, that is, they are synonyms. One of the biggest databases that contain semantic frames is FrameNet (Baker et al., 1998).

Within the project FrameTagger, a domain-specific system for information extraction was to be developed which requires a well-defined, well-structured grammar that can serve as a linguistic preprocessor for identifying semantic frames (Prószéky, 2003; Alexin et al., 2004; Farkas et al., 2004). Texts to be analyzed come from the domain of business news – they constitute a subcorpus of the business texts included in the Szeged Treebank (Csendes et al., 2005). The pieces of news are annotated with semantic frames: abstract event patterns are mapped to sentences and semantic roles are also assigned to the participants of the events.

The most important linguistic category to identify frame elements is that of verb (in general it is possible to identify the topic of the news by relying simply on verbs); nevertheless, nouns play a significant role as well (there are frames with a noun as target word). However,

the syntactic analysis of the sentences proved insufficient since in the case of frame mapping, there is need for the semantic analysis of the events. For this, lists of nouns, adjectives and verbs that occur in the texts to be analyzed were compiled where semantic features of them (and their complements) are also provided. The semantic frames contain only the relevant semantic features of the complements that can co-occur with the central element of the frame (generally, with the verb). In this way, it can be assured when mapping semantic frames to the text that within a given frame, the central element (the target word) is matched only with elements having the semantic features specified in the frame description.

Within the FrameTagger project, ten newsgroups were selected for a more detailed investigation from the business news subcorpus of the Szeged Treebank. The topics of these newsgroups were as listed below:

- one year report;
- change of owner;
- profile;
- contract;
- planned income;
- income;
- litigation;
- opening of a new plant;
- privatization;
- midterm report.

These groups of news altogether contain 1557 pieces of news – within them, 437 different semantic frames can be identified. An example for a semantic frame is offered here:

(11.1)
```
<event schema="litigation.litigation.19">
<rv role="order">msd=N lemma=döntés|lemma=ítélet case=nom</rv>
<rv role="_1">msd=V lemma=születik</rv>
<rv role="_2">msd=N case=ine possessed_by=company</rv>
<rv role="company">msd=N human|company</rv>
</event>
```

As the above example shows, semantic frames include three types of information about their parts. Thus, semantic frames are defined by:

- lemmas of possible words occurring in the frame

  In the above example, the lemmas *döntés* 'decision' and *ítélet* 'judgment' are listed as possible subjects of the verb *születik* 'be born'.

- semantic features

  Semantic features determine a class of words that exhibit the given feature and thus can occur within the frame. In (11.1), the role *company* can be fulfilled by a noun which denotes either a human being or a company.

- POS-tags, grammatical cases

  Finally, grammatical information is also encoded in semantic frames: the grammatical case and the part of speech are also offered for each participant of the event.

Semantic frames and semi-compositional constructions

Among the 437 semantic frames there are 32 that contain a semi-compositional construction. However, this number does not equal to the one of semi-compositional constructions found in the database since there are frames that consist of the combination of several nominal components with a given verb, for instance:

(11.2)
```
<event schema="litigation.litigation.10">
<rv role="office">msd=N office case=nom</rv>
<rv role="_1">msd=V lemma=hoz</rv>
<rv role="order">msd=N lemma=határozat|lemma=ítélet|
lemma=rendelet|lemma=szabály direct_object</rv>
<rv role="_2">msd=N case=del</rv>
</event>
```

This semantic frame incorporates the semi-compositional constructions *határozatot hoz* 'to make a verdict', *ítéletet hoz* 'to make a verdict', *rendeletet hoz* 'to make a decree' and *szabályt hoz* 'to make a rule'. In this way, it generalizes over these four cases and integrates them into one semantic frame, which results in reducing the number of semantic frames.

However, providing a list of lemmas that co-occur with the given verb in a given frame proves to be insufficient for covering all the possible cases since there are two main problems with this approach. First, the semantic frame defined in this way is highly representative of the texts they are extracted from, which might result in the fact that the tagger can successfully map the correct semantic frame for a sentence containing the word *rendelet* 'order', however, it may fail when the sentence contains the word *kormányrendelet* 'order issued by the government'. Thus, semantic frames seem to be domain-specific (in the case of FrameTagger, they are characteristic of the business domain), which might hinder the adaptation of the recognition system to other domains. Second, a simple list of lemmas does not offer the possibility to make generalizations since this way of representation does not include any restrictions on what can be an item of a list and what cannot. On the other hand, recall that semantic features used in the representation of semantic frames are able to express generalizations over nouns that can fulfill the given roles. However, if we chose to express all semantic relations through semantic features, their number should be multiplied, which would lead to a greater deal of complexity on the one hand, and many unique semantic feature would also appear (i.e. those valid for only one group of words) on the other hand. It entails that the size of the database of semantic frames together with the semantic features of nouns would considerably grow, which is undesirable since the semantic parsing of texts would be more time-consuming.

With these in mind, it can be proposed that instead of lists, other ways of generalizations should be incorporated into the representation of semantic frames. In Chapter 7, nouns co-

occurring with the verb *hoz* were classified into semantic groups and in Chapter 8, lexical entries of the verbs *hoz* 'bring' and *köt* 'bind' were presented. It was argued in Chapter 8 that instead of listing all possible nominal components that can co-occur with a verb, it proves to be sufficient to give only their hypernym in the dictionary entry of the verb. This approach can be also applied to semantic frames as well.

Two case studies: *köt* 'bind' and *hoz* 'bring'

In the following, the approach described above will be illustrated with the examples *köt* 'bind' and *hoz* 'bring'.

Semantic frames containing the semi-compositional construction *szerződést köt* 'make a contract' are provided below (in their original form):

(11.3)
```
<event schema="contract.contract.2">
    <rv role="company1">msd=N case=nom company|human</rv>
    <rv role="company2">msd=N case=nom company|human</rv>
    <rv role="_1">msd=V lemma=köt</rv>
    <rv role="topic">msd=N lemma=szerződés |lemma=keretszerződés |
    lemma=hitelszerződés | lemma=megállapodás direct_object</rv>
</event>
```

(11.4)
```
<event schema="contract.contract.6">
    <rv role="company1">msd=N case=nom company|human</rv>
    <rv role="_1">msd=V lemma=köt</rv>
    <rv role="topic">msd=N lemma=szerződés | lemma=keretszerződés
    | lemma=hitelszerződés | lemma=megállapodás direct_object</rv>
    <rv role="company2">msd=N case=ins company|human</rv>
</event>
```

(11.5)
```
<event schema="contract.contract.4">
        <rv role="company1">msd=N case=nom company|human</rv>
        <rv role="company2">msd=N case=nom company|human</rv>
        <rv role="_1">msd=V lemma=köt</rv>
        <rv role="topic" modified_by_adj="_2">msd=N lemma=szerződés
        direct_object</rv>
        <rv role="_2">msd=A lemma=együttműködési</rv>
    </event>
```

(11.6)
```
<event schema="contract.contract.5">
        <rv role="company1">msd=N case=nom company|human</rv>
        <rv role="_1">msd=V lemma=köt</rv>
        <rv role="topic" modified_by_adj="_2">msd=N lemma=szerződés
        direct_object</rv>
        <rv role="_2">msd=A lemma=együttműködési</rv>
        <rv role="company2">msd=N case=ins company|human</rv>
    </event>
```

Thus, there are four such semantic frames which comprise the noun *szerződés* 'contract', one of its synonyms (*megállapodás* 'agreement') and three hyponyms (*együttműködési szerződés* 'cooperation agreement', *keretszerződés* 'framework agreement' and *hitelszerződés* 'credit agreement').[1] However, in principle, all subtypes (i.e. hyponyms) of *szerződés* 'contract' might occur in business texts hence with reference to the corresponding synset in the Hungarian WordNet (Miháltz et al., 2008) (compare the lexical representations discussed in Chapter 8), the following representations are suggested instead of (11.3), (11.4) and of (11.5), (11.6), respectively:

(11.7)
```
<event schema="contract.contract.6">
        <rv role="company1">msd=N case=nom company|human</rv>
        <rv role="_1">msd=V lemma=köt</rv>
        <rv role="topic">msd=N lemma={szerződés:1} direct_object</rv>
        <rv role="company2">msd=N case=ins company|human</rv>
    </event>
```

[1]There are two pairs of frames, which differ in their syntactic structure but the roles are the same in each case.

(11.8) <event schema="contract.contract.2">
<rv role="company1">msd=N case=nom company|human</rv>
<rv role="company2">msd=N case=nom company|human</rv>
<rv role="_1">msd=V lemma=köt</rv>
<rv role="topic">msd=N lemma={szerződés:1} direct_object</rv>
</event>

In the above semantic frames, synsets are denoted by curly brackets and their sense number. An extra rule should be added to the frame tagger architecture in order to assure that hyponyms of the synsets be also mapped to the same semantic frame. Since in the Hungarian WordNet, the synset *szerződés:1* has approximately 50 direct hyponyms, this way of representation requires less semantic frames, what is more, all possible kinds of contracts (or at least those included in HuWN) can be covered with one single semantic frame. In order to unify the examples (11.7) and (11.8), first, the generalized notion of the **Conv** (conversion) lexical function will be shortly presented.

Coyne and Rambow (2009) discuss the relationship between Meaning–Text Theory and FrameNet. They propose an extended notion of the **Conv** lexical function: now it can relate verbs with different numbers of actants if they describe the same situation. In this particular case, the coordinated subject in (11.8) corresponds to a subject and a noun in the instrumental case in (11.7). The two frames can be unified in the following way:

(11.9) <event schema="contract.contract.6">
<rv role="company1">msd=N case=nom company|human</rv>
<rv role="_1">msd=V lemma=köt</rv>
<rv role="topic">msd=N lemma={szerződés:1} direct_object</rv>
<rv role="company2">msd=N case=ins company|human</rv>
$\mathbf{Conv}_{1\ \rightarrow\ 1+3\,2\ \rightarrow\ 2\,3\ \rightarrow\ 1+3}(köt)$ = köt
<event>

Thus, actants 1 and 3 (i.e. the two companies) are coordinated in the subject position whereas actant 2 (szerződés:1) remains unchanged: it still occurs in the object position. In this way, more compact semantic frames can be introduced, which yields that fewer semantic frames are required to capture the same phenomena (in this specific case, the original 4 frames are merged into 1).

Similarly, (11.2) can be substituted by (11.10):

(11.10)
```
<event schema="litigation.litigation.10">
<rv role="office">msd=N office case=nom</rv>
<rv role="_1">msd=V lemma=hoz</rv>
<rv role="order">msd=N lemma={jogszabály:1} direct_object</rv>
<rv role="_2">msd=N case=del</rv>
</event>
```

The synset *jogszabály:1* 'act' contains as hyponyms all kinds of rules, laws, decrees etc. that might possibly occur in business texts and with the help of this approach, they can be surely mapped to the proper semantic frame.

11.2.2 Semantic role labeling

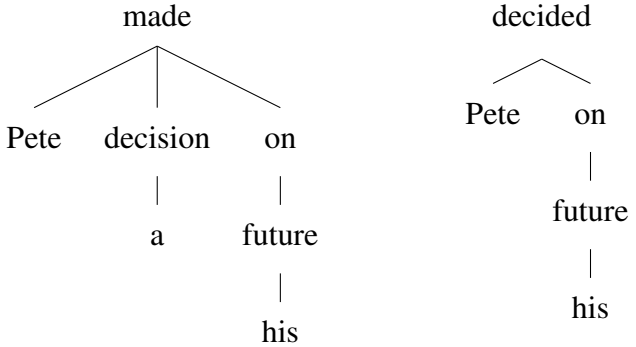
For several IE applications it is essential to identify phrases in a clause and to determine their grammatical role (subject, object, verb) as well. This can be carried out by a syntactic parser and is a relatively simple task. However, the identification of the syntactic status of the nominal component is more complex in the case of semi-compositional constructions for it is a quasi-argument of the verb not to be confused with other arguments (see the alternative proposal for the syntactic structure of semi-compositional constructions in Chapter 6). 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 semi-compositional construction while the other one a verbal counterpart of the construction:

(11.11) Pete **made a decision** on his future.

Pete **decided** on his future.

Their syntactic structure can be represented by the following dependency trees:

(11.12)



In the sentence with the verbal counterpart, the event of deciding involves two arguments: *he* and *his future*. In the sentence with the semi-compositional construction, the same arguments can be found. If a precise syntactic analysis is needed, it is crucial to know which argument belongs to which governor. In Chapter 6 we argued that 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. As for the automatic identification of semi-compositional constructions, we emphasized in Chapter 9 that in a post-processing step after syntactic parsing, the special relation of the nominal and the verbal component should be marked. 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** as it would be assumed if *decision* was not marked as a quasi-argument of the verb). To represent this differently:

(11.13) EVENT: decision-making
ARGUMENT$_1$: Pete
ARGUMENT$_2$: his future

Instead of:

(11.14) *EVENT: making
ARGUMENT$_1$: Pete
ARGUMENT$_2$: decision
ARGUMENT$_3$: his future

Again, the precise identification of semi-compositional constructions can highly improve the performance of parsers in recognizing semantic relations between the complex verb and its arguments.

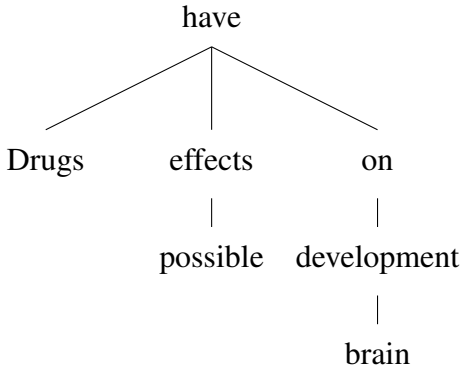
11.2.3 Modality detection

In information extraction, many applications seek to extract factual information from text. That is why it is of high importance to distinguish uncertain and/or negated propositions from factual information. In most cases, what the user needs is factual information, thus, uncertain or negated propositions should be treated in a special way. Depending on the exact task, the

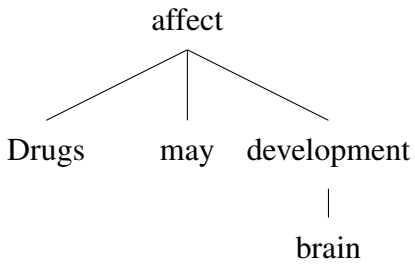
system should either neglect such propositions or separate them from factual information (later, the user can decide whether s/he needs them). In the following, the treatment of semi-compositional constructions in modality detection is presented and it is argued that seeing the verbal and the nominal component as one complex predicate has beneficial effects on the modality detection of events.

Let us consider the following examples:

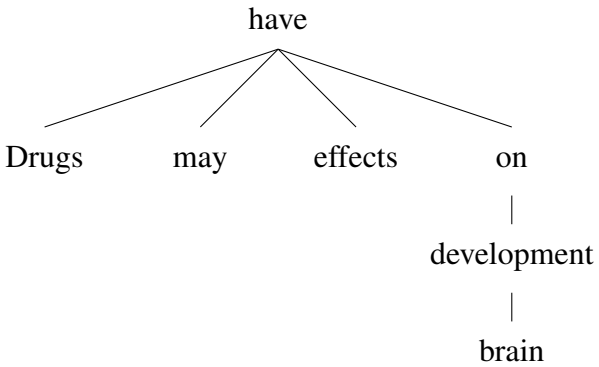
(11.15) Drugs have possible effects on brain development.



(11.16) Drugs may affect brain development.



(11.17) Drugs may have effects on brain development.

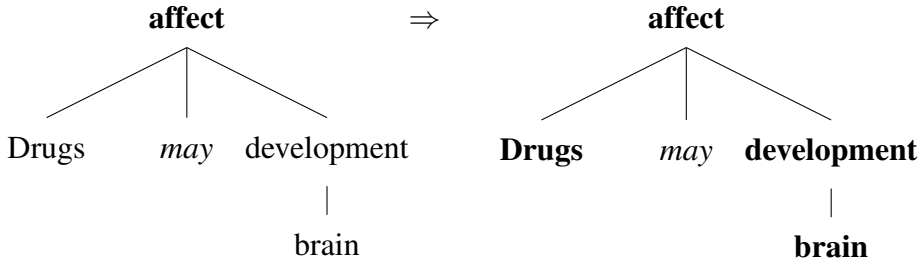


The semantic content of the propositions are the same, however, they differ in the way the proposition is expressed. In (11.15) and (11.17), there is a semi-compositional construction (*have effects*) whereas in (11.16), we can find the verbal counterpart of the construction (*affect*). In (11.15), uncertainty is signaled by the adjective *possible* while in the other two

examples, the auxiliary *may* is present. However, it is essential to assure that all the three sentences are analyzed similarly since they mean the same.

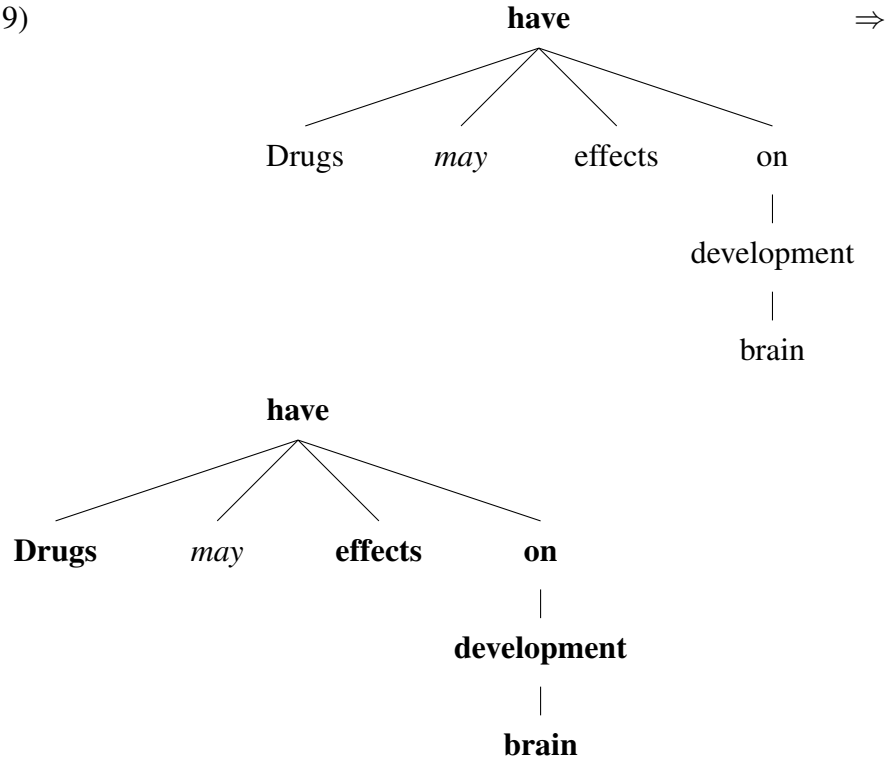
Current modality detectors mostly use syntactic features in order to determine what is in the scope of negation or speculation (Kilicoglu and Bergler, 2009; Farkas et al., 2010). If an element is negated/speculated, all its dependents are negated/speculated (as it is usually defined in negation/uncertainty detection systems). It entails that in (11.18) 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). In this way, it is highly important to make sure that speculation also extends to the whole proposition in the sentences with the semi-compositional constructions as well.

(11.18)



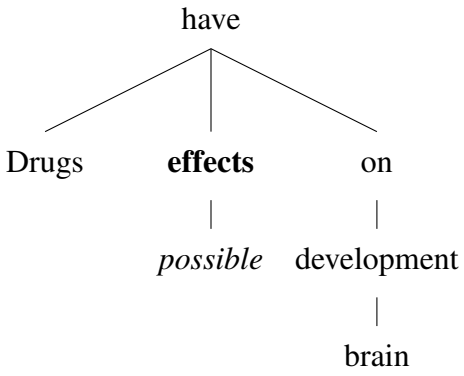
In (11.19), it is the main verb of the sentence (*have*) that is modified by the speculative element *may*, thus, all of its dependents are under speculative scope.

(11.19)



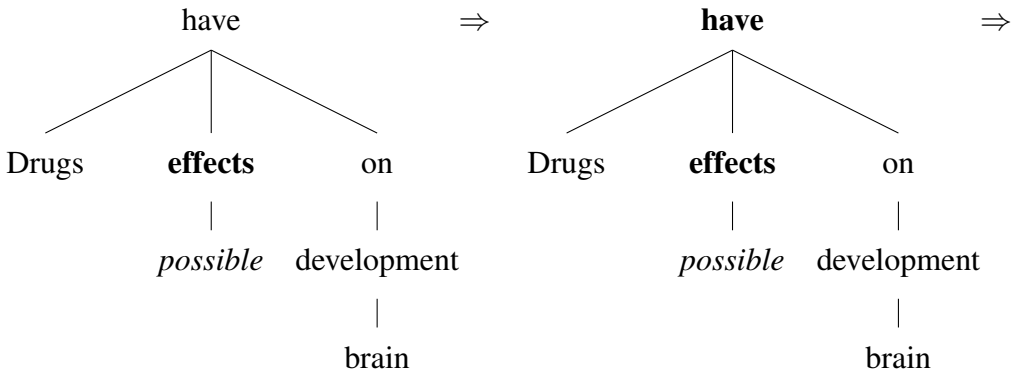
However, in (11.15) it is only the nominal component *effects* that is modified by the adjective. If the rules for detecting the scope of speculation are observed, speculation cannot be extended to the verb hence to the whole proposition since the verb is not an argument of *effects* (i.e. the modified element) as it is represented in (11.20).

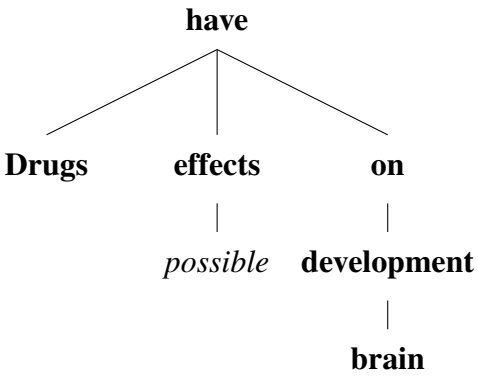
(11.20)



Apparently, the speculation scopes of (11.18) and (11.19) on the one hand and (11.20) on the other hand 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 (see Chapter 6). 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:

(11.21)





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 semi-compositional 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.

11.3 Information retrieval

IR systems usually make use of thesauri, wordnets and synonym lists in order to be more effective: lexical semantic relations such as synonymy, hypernymy or holonymy might be exploited when mapping the query to the words occurring in the text. For instance, documents containing the word *bike* also conform to the query *bicycle*, which is a synonym of the former.

With regard to semi-compositional constructions, recall that most of the constructions have a verbal counterpart with the same meaning and, what is more, some of the constructions are also synonymous with other constructions (see Chapters 7 and 5 for details). We developed a bilingual list of semi-compositional constructions on the basis of the data gained from the SzegedParallelFX corpus, which were manually converted into a database of pairs of English–Hungarian semi-compositional constructions, that is, translational equivalents where both the source and the target language consist of a semi-compositional construction are provided in a list. Moreover, the verbal counterparts of the semi-compositional constructions are also included in the list (wherever applicable). The database contains 343 pairs of semi-compositional constructions (see Appendix D).

This list can be integrated into information retrieval applications as a small thesaurus with synonyms, providing the possibility of finding a document that does not contain the verb in

the query but the semi-compositional construction with the same meaning, e.g. for the query *participate* (or its nominal form, *participation* – if it is stemmed) documents containing *take part* can also be retrieved.

Moreover, due to the bilingual nature of this list, it can be also applied in cross-language information retrieval: e.g. the queries are in English and the documents retrieved are in Hungarian since the queries in one language can be automatically matched to queries in the other language with the help of this list and it also makes it possible to enter a verbal query in one language and to retrieve a document in the other language that contains a synonym construction (e.g. *esküszik* and *make an oath*). Search for nominal queries can also enhanced by this list: the fact that semi-compositional constructions most typically contain a deverbal noun (cf. Chapter 5) can be exploited and in a broad sense, the nominal component can be seen as a synonym of the verbal counterpart. In this way, the Hungarian query *nyomozás* 'investigation' can yield English documents including *investigate* since it is the nominal component of the semi-compositional construction *nyomozást folytat* 'to make an investigation', which is matched to *investigate* in the bilingual list.

The list can also be used for extending wordnets. As it was discussed in Chapter 8, only the most frequent semi-compositional constructions were included in the Hungarian WordNet. However, in the Princeton Wordnet, the typical tendency is that the synset contains the verbal counterpart as a literal, which is defined by a semi-compositional 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.

To conclude, the integration of this list into an IR system can result in enhancing the recall of the system (i.e. the number of relevant documents increases) and the performance of wordnet-based systems can also improve due to the extension of wordnets.

11.4 Summary of results

In this chapter, the treatment of semi-compositional constructions was discussed in the fields of information extraction and retrieval. The following issues were highlighted:

- we proposed a more dense and more compact way of representing semantic frames containing semi-compositional constructions;

- the role and effect of the precise identification of semi-compositional constructions on semantic role labeling were emphasized;
- considering semi-compositional constructions as one unit has also beneficial effects on modality detection;
- a bilingual list of semi-compositional constructions and their verbal counterparts can enhance the recall of (cross-language) information retrieval systems.

As it was argued, these results related to semi-compositional constructions can improve the performance of the above-mentioned applications, however, it would be useful to investigate other fields of IE or IR applications as well. This will be left as future work.

Chapter 12

Semi-compositional constructions and machine translation

12.1 Introduction

The aim of this chapter is to overview the machine translatability of semi-compositional constructions. The lack of compositionality leads to problems concerning the machine translatability of multiword expressions in general and semi-compositional constructions in particular. Some of the solutions offered for these problems, which are related to the semi-productivity of such constructions, are shown and the way lexical functions can help the machine translatability of semi-compositional constructions is demonstrated along with a proposal for the dictionary encoding of such constructions. Current methods for machine translation are presented with special emphasis on the treatment of semi-compositional constructions along with several examples. The chapter concludes with a summary of results.

12.2 On the machine translatability of multiword expressions

In machine translation, one of the most challenging tasks is the proper treatment of multiword expressions. Some Hungarian examples are listed here together with their English equivalents:

(12.1) *gyáva* *nyúl*
coward rabbit
'chicken'

hatos *lottó*
six lottery
'6/45 lottery'

kreol *bőrű*
creole skinned
'dark-skinned'

jóban-rosszban
good-INE-bad-INE
'for better for worse'

The translation of multiword expressions is a hard task for both a human translator and a machine translation program, since their meaning is not totally compositional, that is, it cannot be computed on the basis of the meanings of the parts of the multiword expression and the way they are related to each other. Thus, the result of translating the parts of the multiword expression can hardly be considered as the proper translation of the original expression.

12.2.1 Problems concerning the machine translation of multiword expressions

When translating multiword expressions, translation programs must face two main problems. On the one hand, parts of the multiword expression do not always occur next to each other in the sentence (split MWEs). In this case, the computer must first recognize that the parts of the multiword expression 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 semi-compositional constructions is required for successful translation (Dura and Gawrońska, 2005).

In the following, some examples containing semi-compositional constructions are shown to illustrate the problems mentioned. The English sentences were translated by MetaMorpho[1], an English-Hungarian machine translation program (Novák et al., 2008). The results,

[1]The program is freely available at `www.webforditas.hu`.

that is, the Hungarian versions are given together with the precise English equivalent gained by retranslating the results into English. Finally, the correct Hungarian version of the original sentence is also presented.

(12.2) All the trees have already come into bloom. (original English sentence)
Minden, ami a fáknak már van, bejön virágba. (MetaMorpho)
Everything, which there is already for the trees, comes into a flower. (retranslation)
Már minden fa virágba borult. (Hungarian equivalent)

(12.3) The decision that he made was useless. (original English sentence)
Az a döntés, amit csinált, haszontalan volt. (MetaMorpho)
The decision that he did was useless. (retranslation)
A döntés, amit hozott, haszontalan volt. (Hungarian equivalent)

(12.4) He made a useless decision. (original English sentence)
Egy haszontalan döntést hozott. (MetaMorpho)

In (12.2) there is a semi-compositional construction (*come into bloom*) the parts of which occur next to each other, however, the program does not know this expression, that is why its parts are translated separately (word-by-word translation: *come* as *bejön* and *into bloom* as *virágba* bloom-ILL). Since the expression is not compositional, the result is ungrammatical and meaningless. In (12.3) a split collocation can be found: parts of the multiword expression *make a decision* are separated (other divergences from the word forms given by dictionaries are due to grammar). The program does not treat the multiword expression as a whole that is why no acceptable translation is provided. This is confirmed by (12.4): when the nominal component and the verb are not syntactically split, the program can provide a perfectly sound translation.

12.2.2 A possible solution

Váradi (2006) offers three different treatments of multiword expressions in computational linguistics. First, totally fixed expressions must be listed in the dictionary: the meanings of words in the English expression *French fries* do not equal to the ones of the words in the Hungarian equivalent *sült krumpli* (fried potato), thus, this expression must have a separate lexical entry in the dictionary. Second, productive expressions can be translated totally freely:

in the case of *French wines*, the translation of the parts of the expression provides the correct result (*francia borok*), thus, *French wines* does not form a separate lexical entry. Third, semi-fixed expressions are not worth listing in the dictionary because they are productive in the case of certain (semantic) groups of words. The scheme of the expression *French-speaking population* can be used for creating new expressions such as *Spanish-speaking population*, *Chinese-speaking population* etc. Local grammars have a leading role in the treatment of semi-fixed expressions in machine translation.

On the basis of the tests given in Chapter 4, bare common noun + verb constructions can be divided into three groups, which can be compared to the above-mentioned groups of multiword expressions. For convenience, basic characteristics of the groups are briefly summarized here.

First, most of the tests give grammatical results for productive constructions. Constructions belonging to this group mostly describe conventionalized actions. Their structure is semantically transparent, their meaning can easily be calculated from the meaning of the verb, the noun and the suffix of the noun (that is, their compositionality is of high degree), therefore they are highly productive – this is why they are called productive constructions.

Second, tests give ungrammatical results for idioms. Constructions of this type are not semantically transparent, the meaning of the complex construction cannot be computed from the meanings of the parts of the expression, therefore their productivity is very low.

Third, there is a group of expressions for which some tests give grammatical, while other tests give ungrammatical results. This group of bare common noun + verb constructions is called semi-compositional constructions because they are situated in between the compositional productive constructions and the non-compositional idioms.

A parallel can be drawn between the three treatments of multiword expressions and the three subgroups of bare common noun + verb constructions. Productive constructions can be translated by using the word-by-word method, that is, they do not have to be listed in the dictionary, whereas idioms must be treated similarly to totally fixed expressions, thus, they must be listed in the dictionary. Nevertheless, semi-compositional constructions are too compositional for being listed in the dictionary since the relation between the parts of the constructions is constant. This relation can be formalized with the help of lexical functions (see Chapter 7).

12.3 Lexical functions in machine translation

The results of Chapter 7 suggest that on the one hand, the semantic type of the noun can predict what verb will be the value of a given lexical function, being the noun its argument, and, that on the other hand, certain verbs are typical values of a given lexical function. Furthermore, it has been shown that the groups of semi-compositional constructions tend to correlate with the groups of lexical functions. Thus, in Hungarian, it can be predicted to a certain degree what verb will co-occur with a given noun in the case of a given syntactic relation. Here is an example: if the accusative form of the noun *tájékoztatás* 'informing', that is, *tájékoztatást* needs a light verb, then two facts must be considered. First, the accusative form of the noun refers to the predicate-object syntactic relation, which is described by **Oper**, and verbs such as *ad* 'give', *tesz* 'do', *hoz* 'bring', *vesz* 'take', *kap* 'get' are the most common values of this lexical function in Hungarian. Second, *tájékoztatás* 'informing' is a noun denoting a verbal act, and nouns denoting verbally performed actions or speech acts are usually paired with the verbs *tesz* 'do', *ad* 'give' and *hoz* 'bring'. Thus, the constructions *tájékoztatást ad* and *tájékoztatást tesz* can be predicted (actually, both constructions can be found in the database used in this research, see also Chapter 7).

These results can be fruitfully applied in machine translation as well. However, for a successful translation, these generalizations must be made for both the source language and the target language. Invaluable sources of these generalizations are explanatory combinatorial dictionaries containing different semantic and syntactic relations between lexemes coded by the means of lexical functions. 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 semi-compositional constructions within the entry of the nominal component.

Lexical functions have been already used in several NLP fields (Diachenko, 2006; Reuther, 2006; Apresjan et al., 2002; Apresjan et al., 2007) and the applicability of lexical functions in machine translation is emphasized in Apresjan and Tsinman (2002). The two languages they focus on are English and Russian. If a list containing all the values of lexical functions applied to a lexeme is available for both languages, machine translation becomes much easier

and more precise, since it is only the two lists that must be compared and the corresponding lexeme can be easily chosen (as an illustration, the Hungarian equivalent of this construction is provided as well):

(12.5) **Oper**$_1$ (*nadežda*) = [*~u*] *pitat'*

(12.6) **Oper**$_1$ (*hope*) = *cherish*

(12.7) **Oper**$_1$ (*remény*) = [*~t*] *táplál*

Thus, a comprehensive list of semi-compositional 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 are possible sources of such lists. Since in foreign language equivalents of semi-compositional constructions, the nominal components are usually literal translations of each other (Vincze, 2009d), by collating the corresponding noun entries in these lists or in explanatory-combinatorial dictionaries, the foreign language variant of the given semi-compositional construction can easily be found.[2] On the other hand, since one noun can occur in several semi-compositional constructions, it is necessary that either the semantic relation is encoded between the verb and the noun (in the form of lexical functions)[3] or semi-compositional constructions of the two languages are matched. When building our English-Hungarian database based on data from the SzegedParalellFX, we voted for the second option, that is, we constructed a list of pairs of English-Hungarian semi-compositional constructions (see Appendix D). However, this list can be easily transformed into a formalized database of lexical functions and their values – as future work, we would like to address this task as well.

12.4 Machine translation methods and semi-compositional constructions

In this section, basic methods of machine translation are presented based on Jurafsky and Martin (2008). We will discuss how these methods can treat multiword expressions in gen-

[2]Thus, it is essential that the entry of the nominal component also include a reference to the semi-compositional constructions it may appear in, cf. Chapter 8.

[3]While encoding the relations between the nominal component and the verb in the form of lexical functions might, however, augment the size of the dictionary, this is in fact secondary in the case of electronic dictionaries (cf. Prószéky (2004)).

eral and semi-compositional constructions in particular.

12.4.1 Direct translation

In direct translation, each word is translated to the target language while processing the text, which basically yields a word-by-word translation. Some simple morphological information and reordering rules may be exploited during translation and a large vocabulary is needed.

In the example offered by Jurafsky and Martin (2008), a semi-compositional construction can be found, which is translated from English to Spanish:

(12.8) *Mary didn't slap the green witch*

Maria	*no*	*dió*	*una*	*bofetada*	*a*	*la*	*bruja*	*verde*
Mary	not	gave	a	slap	to	the	witch	green

In order to successfully translate this English sentence to Spanish, the following steps are necessary:

Input	Mary didn't slap the green witch
Morphology	Mary DO-PAST not slap the green witch
Lexical transfer	Maria PAST no dar una bofetada a la verde bruja
Local reordering	Maria no dar PAST una bofetada a la bruja verde
Morphology	Maria no dió una bofetada a la bruja verde

Table 12.1: Translating with a direct system

First, the sentence undergoes shallow morphological analysis (*didn't* is decomposed into *DO-PAST not*), then individual words are translated into Spanish. Note that for this step, it is essential that the dictionary used contain the semi-compositional construction *dar una bofetada* as the equivalent of *slap*, possibly in the entries of both the noun and the verb (see Chapter 8), otherwise no acceptable translation can be provided. In the next step, words are reordered according to the rules of Spanish word order and some morphological processes yield the final past tense form of the verb.

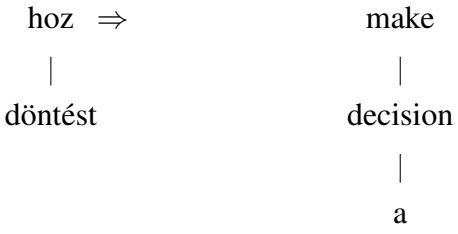
12.4.2 Transfer-based translation

Transfer-based methods make use of syntactic parsing: the sentence in the source language is parsed, then transformation rules produce the parse of the sentence in the target language (i.e. the two parse trees are mapped to each other by using transfer rules), and finally, from

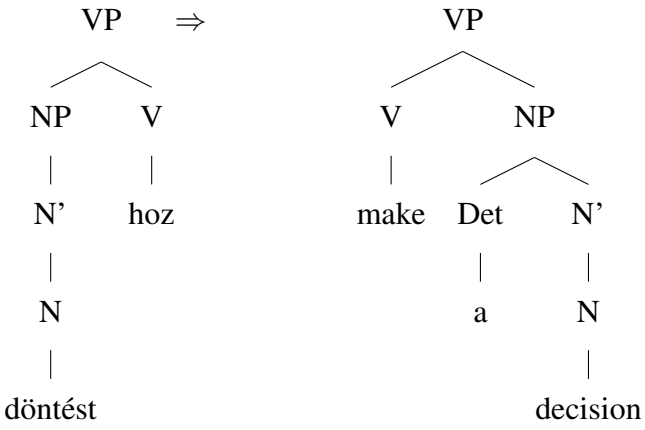
the parse tree in the target language a sentence is generated. Besides, lexical transfer rules based on large dictionaries are also necessary for successful translation.

In the following, two examples of translating Hungarian semi-compositional constructions into English will be provided together with their syntactic trees and transformation rules. Let our first example be *döntést hoz – make a decision*. The dependency and constituency trees of the two constructions are given here:

(12.9)

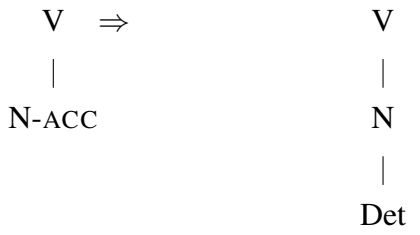


(12.10)

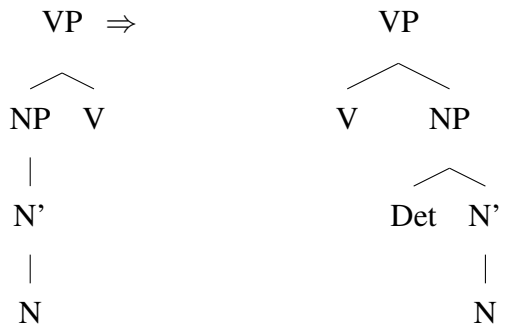


If the parses are compared, it is revealed that a determiner should be inserted into the English phrase as compared to Hungarian (naturally, if the direction of the translation is reversed, the insertion rule would change into a deletion rule). On the other hand – since the canonical order of the construction differs in Hungarian and in English –, the nominal and the verbal component should be swapped on the surface level. Transformation rules are given in a generalized form:

(12.11)

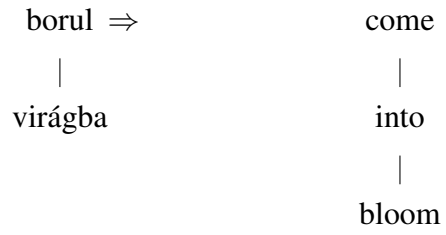


(12.12)

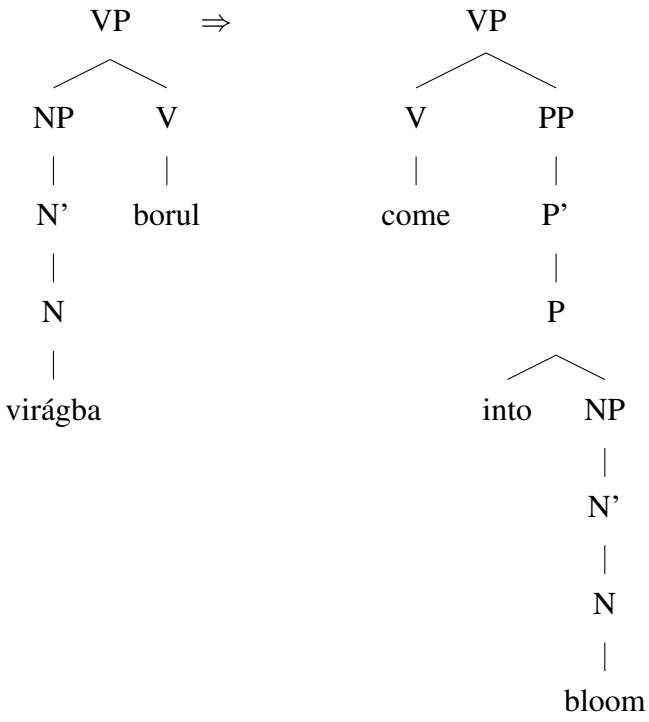


In the previous example, a semi-compositional construction with an accusative nominal component was translated into another construction with a nominal component in the object position. As for a more complex example, *virágba borul – come into bloom* is examined in detail, where the internal structure of the NP also changes in a more chiseled way. Again, the dependency and the constituency parses of the trees are provided:

(12.13)



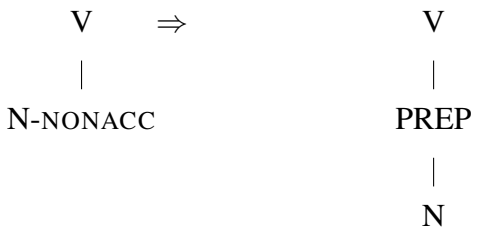
(12.14)



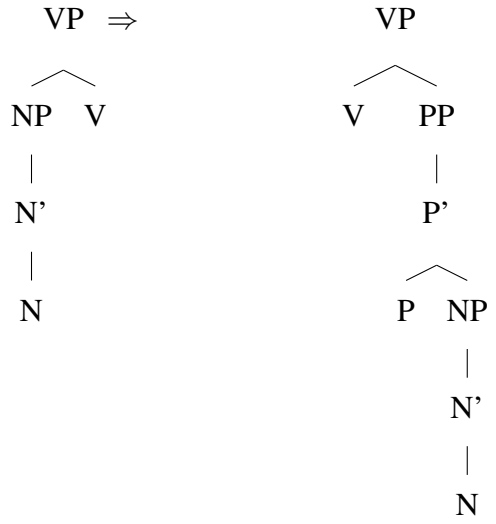
As can be seen from the examples, the dependency-based translation process involves the insertion of a preposition – the exact lexical unit should be selected as determined by the case suffix of the nominal component in Hungarian (thus, morphological information is also

exploited and in the case of not totally compositional phrases, the lexicon is also evoked so as to select the proper preposition). In the constituency tree, the Hungarian NP is substituted by a PP in English (which, on the other hand, also contains an NP). Reordering rules are also necessary as discussed above. The generalized versions of the above transformation rules can be seen here (N-NONACC stands for a noun that is not in the accusative case:

(12.15)



(12.16)



The insertion of the lexical items into the nodes is determined by the dictionary, which should include semi-compositional constructions as well in order to provide an adequate translation.

Thus, for translating semi-compositional constructions with the transfer method, morphological parsing, syntactic parsing, reordering rules, transformation rules and lexical rules are of high importance.

12.4.3 Interlingua

The interlingua method requires that the semantic content of the text in the source language be extracted and formalized in an interlingua (i.e. meaning representation), which is then expressed in the target language. With this method, it cannot be guaranteed that semi-compositional constructions are translated as semi-compositional constructions since it is

on the level of semantics that translation takes place, instead of the level of words or syntactic phrases. As it has already been emphasized in Chapter 5, most semi-compositional constructions have a verbal counterpart with (approximately) the same meaning hence it is possible that the verbal counterpart will be used in the target language.

The sentence *Stan did not give a kiss to Wendy* can have the following meaning representation:
(12.17)

$$\begin{bmatrix} EVENT & kissing \\ AGENT & Stan \\ TENSE & past \\ POLARITY & negative \\ THEME & Wendy \end{bmatrix}$$

However, from this representation it is not obvious whether the event of kissing was originally expressed by a semi-compositional construction (*give a kiss*) or its verbal counterpart (*kiss*). Thus, when expressing this meaning in the target language (e.g. Hungarian), both *csókot ad* and *megcsókol* might be used.

For applying the interlingua method, the way of representing meaning should be thoroughly elaborated and natural language words and phrases should be aligned to meaning representations (i.e. words and phrases that have the same meaning should be grouped together along with their meaning). Natural language words and phrases are grouped in synsets in wordnets and wordnets of different languages use an International Language Index (ILI), which allows synsets belonging to the same concept to be readily accessible in every database. Thus, multilingual wordnets can be used as large dictionaries where a set of words in one language corresponds to another set of words in the other language. This feature can be exploited in machine translation. On the other hand, our bilingual list of semi-compositional constructions and their verbal counterparts (Appendix D) can also be a useful resource similar to wordnets in creating a database of concepts and meaning representations.

12.4.4 Statistical machine translation

Statistical machine translation systems make use of probabilities in order to provide translations that are faithful to the original utterance and sound natural in the target language.

Since one utterance can have several possible translations, the best translation candidate is selected on the basis of statistical measures. In order to get the best translation possible, the probability for a given sentence to occur in the target language (language model) and the probability of the source language sequence being a translation of the target language sequence (translation model) are calculated. Statistical machine translation systems exploit huge parallel corpora in collecting probable translations for utterances. An example of a statistical translator is Google Translate, available at `http://translate.google.com/`.

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) and Liu et al. (2010) and one-to-many alignment can be exploited in identifying previously unknown multiword expressions (Caseli et al., 2010; Caseli et al., 2009; Zarrieß and Kuhn, 2009; Tsvetkov and Wintner, 2010). The Szeged-ParalellFX corpus, which is manually aligned on the sentence level, is annotated for semi-compositional constructions (see Chapter 3), thus, the constructions can be used as anchors for automatic word alignment. These features can be exploited in statistical machine translation systems and the annotation of semi-compositional constructions would most probably have a beneficial effect on translating semi-compositional 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 semi-compositional constructions without any language resource (such as bilingual lists or manually aligned corpora). For instance, *make a decision* and *take a decision* are both perfectly sound constructions in English (in our database, *make a decision* is the second most frequent semi-compositional construction whereas *take a decision* is the fifth one (see Table 3.8)). 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' (*döntést hoz* is the tenth most frequent Hungarian semi-compositional construction in the database) hence *döntést hoz* would be probably judged by the system as the best candidate for translating these expressions into Hungarian.

12.5 Summary of results

In this chapter, some possible ways of translating semi-compositional constructions with the help of a computer were presented. The following points were discussed:

- the lack of total compositionality results in difficulties of translating multiword expressions in general and semi-compositional constructions in particular;
- lexical functions can formalize the relation between the verbal and the nominal component of semi-compositional constructions, which can be exploited in machine translation applications and dictionary building;
- since each method needs a large dictionary containing entries on multiword expressions, our bilingual database of semi-compositional constructions and their verbal counterpart can be utilized in machine translation;
- statistical machine translation can make use of the SzegedParalellFX corpus annotated for semi-compositional constructions.

Although there is still a lot to do in the field of the automatic translation of multiword expressions, we hope that the language resources developed in this research will contribute to the improvement of the state-of-the-art methods.

Chapter 13

Summary

The main aim of this thesis was to investigate semi-compositional common noun + verb constructions from both theoretical and computational linguistic aspects. Emphasis was put primarily on Hungarian data, however, constructions from other languages such as English, Russian or Spanish were also taken into account.

For methodological purposes, we constructed three corpora of semi-compositional constructions. The Szeged Treebank (Vincze and Csirik, 2010), the SzegedParalell corpus (Vincze et al., 2010a) and the Wiki50 corpus (Vincze et al., 2011b) were annotated for semi-compositional constructions, thus yielding a database that consists of 1524 Hungarian and 827 English constructions. Annotation guidelines were based on earlier theoretical results, however, the data collected from the corpora served as a basis for drawing further theoretical conclusions on the behavior of semi-compositional constructions. Thus, theory and practice nicely intertwined in the data collecting methodology of this thesis.

Research questions answered in this thesis can be divided into two categories: theoretical and computational linguistic issues. A crucial goal of the thesis was to apply theoretical results to the greatest extent possible in natural language processing, thus making a direct connection from theory to practice. With this point in mind, thesis results are summarized in this chapter.

13.1 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.

13.1.1 Semi-compositional constructions as a subtype of multiword expressions

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

13.1.2 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). 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.

13.1.3 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 or in aspect and Aktionsart**. As opposed to productive constructions, 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.

13.1.4 The syntax of semi-compositional constructions

The syntactic analyses of semi-compositional constructions were provided in the frameworks of constituency and dependency grammars. 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.

13.1.5 The semantics of semi-compositional constructions

The semantics of semi-compositional constructions was analyzed within the framework of Meaning–Text Theory. The syntactic-semantic relation between the nominal component and the verb can be formalized by using lexical functions. For the analysis, 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**. 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 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. These results can be applied in word sense disambiguation and in machine translation.

13.1.6 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. 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.

13.2 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 adaptation of theoretical results. In the following, thesis results related to computational linguistics are shortly summarized.

13.2.1 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. Our results suggest that **shallow morphological information** (such as lemmas, stems, POS-tags and suffixes) is sufficient to detect semi-compositional

constructions in text 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.

13.2.2 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** lead to a higher accuracy in both manual and automatic word sense disambiguation.

13.2.3 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 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 (in Appendix D) can enhance the recall of (cross-language) **information retrieval** systems.

13.2.4 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. 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.

13.3 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 Chapter 4 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** (see Chapter 6) and **semantic correlations between the type of the noun and the verb** also exist in several languages as illustrated with Russian and Hungarian examples (see Chapter 7). 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**.

13.4 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;
- 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 13.1 also visualizes the interrelationships among the theoretical linguistic and computational linguistic chapters of the thesis.

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

Table 13.1: Connections between topics 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 variativity with verbal counterparts can be exploited;
- they are regarded as one unit in statistical machine translation.

Table 13.2 offers a summary of these points.

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

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

Table 13.2: The dual nature of semi-compositional constructions and fields of the thesis

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

Agirre, Eneko; Edmonds, Philip. 2006. *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Agirre, Eneko; Màrquez, Lluís; Wicentowski, Richard (eds.). 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June.

Al-Haj, Hassan; Wintner, Shuly. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 10–18, Beijing, China, August. Coling 2010 Organizing Committee.

Alexin, Zoltán; Csirik, János; Gyimóthy, Tibor. 2004. Programcsomag információkinyerési kutatások támogatására [Toolchain for enhancing information extraction research]. In Alexin, Zoltán; Csendes, Dóra (eds.), *II. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 41–48, Szeged, Hungary, December. University of Szeged.

Alexin, Zoltán. 2007. A frázisstrukturált Szeged Treebank átalakítása függőségi fa formátumra [Converting the phrase structure based Szeged Treebank to dependency format]. In Tanács, Attila; Csendes, Dóra (eds.), *V. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 263–266, Szeged. Szegedi Tudományegyetem.

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

Alonso Ramos, Margarita. 2004. *Las construcciones con verbo de apoyo*. Visor Libros, Madrid.

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

Apresjan, Jurij D.; Tsinman, Leonid L. 2002. Formal'naja model' perifrazirovanija predloženij dlja sistem pererabotki tekstkov na estestvennyx jazykax. *Russkij jazyk v naučnom osveščenii*, 2(4):102–146.

Apresjan, Jurij D.; Boguslavsky, Igor M.; Iomdin, Leonid I.; Tsinman, Leonid L. 2002. Lexical Functions in NLP: Possible Uses. In Klenner, Manfred; Visser, Henriëtte (eds.), *Computational Linguistics for the New Millennium: Divergence or Synergy? Festschrift in Honour of Peter Hellwig on the Occasion of his 60th Birthday*, pp. 55–72, Frankfurt am Main / New York. Peter Lang.

Apresjan, Jurij D.; Boguslavsky, Igor M.; Iomdin, Leonid I.; Tsinman, Leonid L. 2007. Lexical Functions in Actual NLP-Application. In Wanner, Leo (ed.), *Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk*, pp. 203–233, Amsterdam / Philadelphia. Benjamins.

Apresjan, Jurij D. 2004. O semantičeskoj nepustote i motivirovannosti glagol'nyx leksičeskix funkcij. *Voprosy jazykoznanija*, (4):3–18.

Apresjan, Jurij D. 2005. O Moskovskoj semantičeskoj škole. *Voprosy jazykoznanija*, (1):3–30.

Apresjan, Juri. 2009. The theory of lexical functions: An update. In Beck, David; Gerdes, Kim; Milićević, Jasmina; Polguère, Alain (eds.), *Proceedings of the Fourth International Conference on Meaning-Text Theory – MTT'09*, pp. 1–14, Montreal, Canada. Université de Montréal.

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

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

B. Kovács, Mária. 2000. A funkcióigék kérdéséhez [On the question of function verbs]. *Magyar Nyelvőr*, 124:370–372.

Baker, Collin F.; Fillmore, Charles J.; Lowe, John B. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pp. 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Baker, Mark C. 1988. *Incorporation: A Theory of Grammatical Function Changing*. The University of Chicago Press, Chicago.

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.

Beavers, John Travis. 2006. *Argument/Oblique Alternations and the Structure of Lexical Meaning)*. Ph.D. thesis, Stanford University.

Bejcek, Eduard; Stranák, Pavel. 2010. Annotation of multiword expressions in the Prague Dependency Treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Bolshakov, Igor A.; Gelbukh, Alexander. 1998. Lexical Functions in Spanish. In *Proceedings of the Simposium Internacional de Computación - CIC'98*, pp. 383–395, Mexico D.F.

Bouma, Gerlof. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, pp. 109–114, Uppsala, Sweden, July. Association for Computational Linguistics.

Bu, Fan; Zhu, Xiaoyan; Li, Ming. 2010. Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 116–124, Beijing, China, August. Coling 2010 Organizing Committee.

Butnariu, Cristina; Kim, Su Nam; Nakov, Preslav; Ó Séaghdha, Diarmuid; Szpakowicz, Stan; Veale, Tony. 2010. Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 39–44, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Caseli, Helena de Medeiros; Villavicencio, Aline; Machado, André; Finatto, Maria José. 2009. Statistically-driven alignment-based multiword expression identification for technical domains. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pp. 1–8, Singapore, August. Association for Computational Linguistics.

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

Chiticariu, Laura; Krishnamurthy, Rajasekar; Li, Yunyao; Reiss, Frederick; Vaithyanathan, Shivakumar. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of EMNLP 2010*, pp. 1002–1012, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Chomsky, Noam. 1995. *The Minimalist Program*. Current studies in linguistics. MIT Press, Cambridge, MA.

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

Comrie, Bernard. 1976. The syntax of causative constructions: Cross-language similarities and divergences. In Shibatani, M. (ed.), *The grammar of causative constructions: Syntax and semantics, Vol. 6.*, pp. 261–312, New York. Academic Press.

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.

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

Coyne, Bob; Rambow, Owen. 2009. Meaning-Text-Theory and Lexical Frames. In Beck, David; Gerdes, Kim; Milićević, Jasmina; Polguère, Alain (eds.), *Proceedings of the Fourth International Conference on Meaning-Text Theory – MTT'09*, pp. 119–128, Montreal, Canada. Université de Montréal.

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

Das, Dipankar; Pal, Santanu; Mondal, Tapabrata; Chakraborty, Tanmoy; Bandyopadhyay, Sivaji. 2010. Automatic Extraction of Complex Predicates in Bengali. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pp. 37–45, Beijing, China, August. Coling 2010 Organizing Committee.

Daumé III, Hal. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Diachenko, Pavel. 2006. Lexical Functions in Learning the Lexicon. In Méndez-Vilas, Antonio; Solano Martín, Aurora; Mesa González, José Antonio; Mesa González, Julian (eds.), *Current Developments in Technology-Assisted Education*, pp. 538–542, Bajadoz. FORMATEX.

Dias, Gaël. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment – Volume 18*, 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.

É. Kiss, Katalin. 2002. *The Syntax of Hungarian*. Cambridge University Press, Cambridge.

Eckhardt, Sándor (ed.). 1992a. *Francia–magyar kéziszótár [French–Hungarian dictionary]*. Akadémiai Kiadó, Budapest.

Eckhardt, Sándor (ed.). 1992b. *Magyar–francia kéziszótár [Hungarian–French dictionary]*. Akadémiai Kiadó, Budapest.

Erk, Katrin; Strapparava, Carlo (eds.). 2010. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Uppsala, Sweden, July.

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.

Farkas, Richárd; Vincze, Veronika; Móra, György; Csirik, János; Szarvas, György. 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*, pp. 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

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

Fillmore, Charles J. 1968. The Case for Case. In Bach, Emmon; Harms, Robert T. (eds.), *Universals in linguistic theory*, pp. 1–90, New York. Holt.

Fillmore, Charles J. 1977. The Case for Case Reopened. In Cole, Peter; Sadock, Jerry (eds.), *Syntax and Semantics 8: Grammatical relations*, pp. 59–82, New York. Academic Press.

Freed, Alice F. 1979. *The Semantics of English Aspectual Complementation*. Reidel, Dordrecht.

Gábor, Kata; Héja, Enikő. 2006. Predikátumok és szabad határozók [Predicates and adjuncts]. In Kálmán, László (ed.), *KB 120 - A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósy András tiszteletére*, pp. 135–152, Budapest. Tinta Kiadó.

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

Gracia i Sole, Lluisa. 1986. *La teoria tematica*. Ph.D. thesis, Universitat Autonoma de Barcelona, Barcelona.

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*, pp. 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

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

Grétsy, László; Kemény, Gábor. 1996. *Nyelvművelő kéziszótár [Purists' dictionary]*. Auktor Könyvkiadó, Budapest.

Grimshaw, Jane; Mester, Armin. 1988. Light Verbs and Theta-Marking. *Linguistic Inquiry*, 19:205–232.

Guenthner, Franz; Blanco, Xavier. 2004. Multi-Lexemic Expressions: An Overview. In Leclère, Christian; Laporte, Éric; Piot, Mireille; Silberztein, Max (eds.), *Lexique, Syntaxe et Lexique-Grammaire / Syntax, Lexis & Lexicon-Grammar. Papers in honour of Maurice Gross. Lingvisticae Investigationes Supplementa 24.*, pp. 239–252, Grenoble. LADL, CNRS.

Gurrutxaga, Antton; Alegria, Iñaki. 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*, pp. 2–7, Portland, Oregon, USA, June. Association for Computational Linguistics.

Halácsy, Péter; Kornai, András; Németh, László; Sass, Bálint; Varga, Dániel; Váradi, Tamás; Vonyó, Attila. 2005. A hunglish korpusz és szótár [The hunglish corpus and dictionary]. In Alexin, Zoltán; Csendes, Dóra (eds.), *MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 134–142, Szeged, Hungary, December. University of Szeged.

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.

Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Haugereid, Petter; Bond, Francis. 2011. Extracting Transfer Rules for Multiword Expressions from Parallel Corpora. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 92–100, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heltai, Pál; Gósy, Mária. 2005. A terpeszkedő szerkezetek hatása a feldolgozásra [The effect of sprawling constructions on processing]. *Magyar Nyelvőr*, 129:470–487.

Hendrickx, Iris; Mendes, Amália; Pereira, Sílvia; Gonçalves, Anabela; Duarte, Inês. 2010. Complex Predicates Annotation in a Corpus of Portuguese. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pp. 100–108, Uppsala, Sweden, July. Association for Computational Linguistics.

Hornby, Albert S.; Wehmeier, Sally (eds.). 2002. *Oxford Advanced Learners' Dictionary*. Oxford University Press, Oxford.

Iordanskaja, Lidija; Paperno, Slava. 1996. *A Russian–English Collocational Dictionary of the Human Body*. Slavica Publishers, Columbus, Ohio.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Jackendoff, Ray. 2010. *Meaning and the Lexicon: The Parallel Architecture 1975–2010*. Oxford University Press, Oxford.

Jurafsky, Daniel; Martin, James H. 2008. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2nd edition.

Kaalep, Heiki-Jaan; Muischnek, Kadri. 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*, pp. 57–64, Trento, Italy, April. Association for Computational Linguistics.

Kaalep, Heiki-Jaan; Muischnek, Kadri. 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)*, pp. 23–26, Marrakech, Morocco, June.

Kálmán, László. 2006. Miért nem vonzanak a régensek? [Why do governors not govern?]. In Kálmán, László (ed.), *KB 120 - A titkos kötet. Nyelvészeti tanulmányok Bánréti Zoltán és Komlósy András tiszteletére*, pp. 229–246, Budapest. Tinta Kiadó.

Kearns, Kate. 1998. Extraction from *make the claim* constructions. *Journal of Linguistics*, 34:53–72.

Kearns, Kate. 2002. *Light verbs in English*. Manuscript.

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.

Keszler, Borbála. 1994. *Magyar leíró nyelvtani segédkönyv [A guidebook to descriptive Hungarian grammar]*. Nemzeti Tankönyvkiadó, Budapest.

Keszler, Borbála. 2000. A szintagmák [Syntagmas]. In Keszler, Borbála (ed.), *Magyar grammatika*, pp. 349–366, Budapest. Nemzeti Tankönyvkiadó.

Kiefer, Ferenc; Ladányi, Mária. 2000. Az igekötők [Preverbs]. In Kiefer, Ferenc (ed.), *Strukturális magyar nyelvtan. 3. Alaktan*, pp. 453–518, Budapest. Akadémiai Kiadó.

Kiefer, Ferenc. 1990–1991. Noun Incorporation in Hungarian. *Acta Linguistica Hungarica*, 40(1–2):149–177.

Kiefer, Ferenc. 2003. A kétféle igemódosítóról [On the two types of verb modifiers]. *Nyelvtudományi Közlemények*, 100:177–186.

Kiefer, Ferenc. 2006. *Aspektus és akcióminőség különös tekintettel a magyar nyelvre [Aspect and Aktionsart with special emphasis on Hungarian]*. Akadémiai Kiadó, Budapest.

Kiefer, Ferenc. 2007. *Jelentéselmélet [Semantics]*. Corvina, Budapest.

Kilgariff, Adam (ed.). 2001. *Proceedings of Senseval 2: Second International Workshop on the Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, Toulouse.

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

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

Király, Rudolf (ed.). 1993a. *Magyar–portugál kéziszótár [Hungarian–Portuguese dictionary]*. Akadémiai Kiadó, Budapest.

Király, Rudolf (ed.). 1993b. *Portugál–magyar kéziszótár [Portuguese–Hungarian dictionary]*. Akadémiai Kiadó, Budapest.

Klein, Dan; Manning, Christopher D. 2003. Accurate unlexicalized parsing. In *Annual Meeting of the ACL*, volume 41, pp. 423–430.

Komlósy, András. 1992. Régensek és vonzatok [Governors and arguments]. In Kiefer, Ferenc (ed.), *Strukturális magyar nyelvtan. I. Mondattan*, pp. 299–527, Budapest. Akadémiai Kiadó.

Koutny, Ilona; Wacha, Balázs. 1991. Magyar nyelvtan függőségi alapon [Dependency-based Hungarian grammar]. *Magyar Nyelv*, 87(4):393–404.

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)*, pp. 7–10, Marrakech, Morocco, June.

Lafferty, John D.; McCallum, Andrew; Pereira, Fernando C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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

Laporte, Éric; Ranchhod, Elisabete; Yannacopoulou, Anastasia. 2008. Syntactic variation of support verb constructions. *Lingvisticae Investigationes*, 31(2):173–185. DOI: 10.1075/li.31.2.04lap.

Larson, Richard. 1988. On the double object construction. *Linguistic Inquiry*, 19:335–391.

Lengyel, Klára. 1999. A segédigék kérdéséhez. Válasz Uzonyi Kiss Judit és Tuba Márta cikkére [On the question of auxiliaries. A reply to the paper by Judit Uzonyi Kiss and Márta Tuba]. *Magyar Nyelvőr*, 123:116–129.

Lengyel, Klára. 2000. A segédigék és származékaik [Auxiliaries and their derivatives]. In Keszler, Borbála (ed.), *Magyar grammatika*, pp. 252–256, Budapest. Nemzeti Tankönyvkiadó.

Levin, Beth. 1993. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago.

L'Homme et al., Marie-Claude. 2007. DiCoInfo: Dictionnaire fondamental de l'informatique et de l'Internet. http://olst.ling.umontreal.ca/dicoinfo/.

Lőrincz, Julianna. 2004. A funkcióigés szerkezetek és a főnévi taggal azonos tövű igék varianciája [The variation of function verb constructions and verbs derived from the same root as the noun]. In Kurtán, Zsuzsa; Zimányi, Árpád (eds.), *A nyelvek vonzásában. Köszöntő kötet Budai László 70. születésnapjára*, pp. 104–109, Eger - Veszprém. Veszprémi Egyetemi Kiadó.

Liu, Zhanyi; Wang, Haifeng; Wu, Hua; Li, Sheng. 2010. Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 825–833, Uppsala, Sweden, July. Association for Computational Linguistics.

Magay, Tamás; Országh, László (eds.). 2001a. *Angol–magyar kéziszótár [English–Hungarian dictionary]*. Akadémiai Kiadó, Budapest.

Magay, Tamás; Országh, László (eds.). 2001b. *Magyar–angol kéziszótár [Hungarian–English dictionary]*. Akadémiai Kiadó, Budapest.

Maleczki, Márta. 2008. Határozatlan argumentumok [Indefinite arguments]. In Kiefer, Ferenc (ed.), *Strukturális magyar nyelvtan. IV. A szótár szerkezete*, pp. 129–184, Budapest. Akadémiai Kiadó.

Manning, Christopher D.; Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Mansoory, Niloofar; Bijankhan, Mahmood. 2008. The possible effects of Persian light verb constructions on Persian WordNet. In Tanács, Attila; Csendes, Dóra; Vincze, Veronika; Fellbaum, Christiane; Vossen, Piek (eds.), *Proceedings of the Fourth Global WordNet Conference*, pp. 297–303, Szeged, Hungary, January. University of Szeged.

Martens, Scott; Vandeghinste, Vincent. 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*, pp. 85–88, Beijing, China, August. Coling 2010 Organizing Committee.

McCallum, Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

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. 1974. Esquisse d'un modèle linguistique du type "Sens<->Texte". In *Problèmes actuels en psycholinguistique. Colloques inter. du CNRS, no. 206*, pp. 291–317, Paris. CNRS.

Mel'čuk, Igor. 1988. *Dependency Syntax: theory and practice*. State University of New York Press, Albany, NY.

Mel'čuk, Igor. 1989. Semantic Primitives from the Viewpoint of the Meaning-Text Linguistic Theory. *Quaderni di Semantica*, 10(1):65–102.

Mel'čuk, Igor. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In Wanner, Leo (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, pp. 37–102, Amsterdam/Philadelphia. Benjamins.

Mel'čuk, Igor. 1998. Collocations and Lexical Functions. In Cowie, A. P. (ed.), *Phraseology. Theory, Analysis, and Applications*, pp. 23–53, Oxford. Clarendon Press.

Mel'čuk, Igor. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. In Agel, V.; Eichinnger, L.; Eroms, H.-W.; Hellwig, P.; Herringer, H. J.; Lobin, H. (eds.), *Dependency and Valency. An International Handbook of Contemporary Research, vol. 1.*, pp. 188–229, Berlin-New York. W. de Gruyter.

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.

Mével, Jean-Pierre (ed.). 2004. *Dictionnaire HACHETTE*. Hachette Livre, Paris.

Meyers, Adam; Reeves, Ruth; Macleod, Catherine. 2004a. 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.

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

Mihalcea, Rada; Edmonds, Phil (eds.). 2004. *SensEval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics, Barcelona, Spain, July.

Miháltz, Márton; Pohl, Gábor. 2005. Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból [A proposal for the automatic construction of semantically annotated multilingual training corpora from parallel corpora for word sense disambiguation]. In Alexin, Zoltán; Csendes, Dóra (eds.), *MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 418–419, Szeged, Hungary, December. University of Szeged.

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

Mihăltz, Márton. 2005. Towards A Hybrid Approach To Word-Sense Disambiguation In Machine Translation. In *Proceedings of Modern Approaches in Translation Technologies Workshop at RANLP-2005*, Borovets.

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

Muischnek, Kadri; Kaalep, Heiki Jaan. 2010. The variability of multi-word verbal expressions in Estonian. *Language Resources and Evaluation*, 44(1-2):115–135.

Nagy T., István; Vincze, Veronika; Berend, Gábor. 2011. Domain-dependent identification of multiword expressions. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Newson, Mark; Hordós, Marianna; Pap, Dániel; Szécsényi, Krisztina; Tóth, Gabriella; Vincze, Veronika. 2006. *Basic English Syntax With Exercises*. Bölcsész Konzorcium, Budapest.

Nicholson, Jeremy; Baldwin, Timothy. 2008. Interpreting Compound Nominalisations. In *LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 43–45, Marrakech, Morocco.

Novák, Attila; Tihanyi, László; Prószéky, Gábor. 2008. The MetaMorpho translation system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 111–114, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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

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

Oravecz, Csaba; Nagy, Viktor; Varasdi, Károly. 2005. Lexical idiosyncrasy in MWE extraction. In *Proceedings from the Corpus Linguistics Conference Series*, pp. 134–142, Birmingham.

Pal, Santanu; Naskar, Sudip Kumar; Pecina, Pavel; Bandyopadhyay, Sivaji; Way, Andy. 2010. Handling named entities and compound verbs in phrase-based statistical machine

translation. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pp. 46–54, Beijing, China, August. Coling 2010 Organizing Committee.

Pearce, Darren. 2001. Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, CMU.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Piao, Scott S. L.; Rayson, Paul; Archer, Dawn; Wilson, Andrew; McEnery, Tony. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment – Volume 18*, pp. 49–56, Morristown, NJ, USA. Association for Computational Linguistics.

Prószéky, Gábor; Koutny, Ilona; Wacha, Balázs. 1989. Dependency Syntax of Hungarian. In Maxwell, Dan; Schubert, Klaus (eds.), *Metataxis in Practice (Dependency Syntax for Multilingual Machine Translation)*, pp. 151–181, Dordrecht. Foris.

Prószéky, Gábor. 2003. NewsPro: automatikus információszerzés gazdasági rövidhírekből [NewsPro: automatic information extraction from short business news]. In Alexin, Zoltán; Csendes, Dóra (eds.), *Magyar Számítógépes Nyelvészeti Konferencia*, pp. 161–166, Szeged, Hungary, December. University of Szeged.

Prószéky, Gábor. 2004. Az elektronikus papírszótártól az "igazi" elektronikus szótárak felé [From the electronic paper-based dictionary towards to "real" electronic dictionaries]. In Fóris, Ágota; Pálfy, Miklós (eds.), *A lexikográfia Magyarországon*, pp. 81–87, Budapest, Hungary. Tinta Könyvkiadó.

Pusztai, Ferenc (ed.). 2003. *Magyar értelmező kéziszótár [The Concise Dictionary of the Hungarian Language]*. Akadémiai Kiadó, Budapest.

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

Ramisch, Carlos; Villavicencio, Aline; Boitet, Christian. 2010b. mwetoolkit: a framework for multiword expression identification. In Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan; Piperidis, Stelios; Tapias, Daniel (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Ramisch, Carlos; Villavicencio, Aline; Boitet, Christian. 2010c. Web-based and combined language models: a case study on noun compound identification. In *Coling 2010: Posters*, pp. 1041–1049, Beijing, China, August. Coling 2010 Organizing Committee.

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

Répási, Györgyné; Székely, Gábor. 1998. Lexikográfiai előtanulmány a fokozó értelmű szavak és szókapcsolatok szótárához [A lexicographic pilot study for the dictionary of intensifying words and phrases]. *Modern Nyelvoktatás*, 4(2–3):89–95.

Reuther, Tilmann. 1996. On Dictionary Entries for Support Verbs: The Cases of Russian VESTI, PROVODIT' and PROIZVODIT'. In Wanner, Leo (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, pp. 181–208, Amsterdam/Philadelphia. Benjamins.

Reuther, Tilmann. 2006. Beginnen - Enden - Weitergehen: Die "Werkbank Kollokationen" als Instrument inner- und zwischensprachlicher Phraseologieforschung. In Binder, Eva; Stadler, Wolfgang; Weinberger, Helmut (eds.), *ZEIT - ORT - ERINNERUNG. Slawistische Erkundungen aus sprach-, literatur-und kulturwissenschaftlicher Perspektive. Festschrift für Ingeborg Ohnheiser und Christine Engel zum 60. Geburtstag*, pp. 467–485, Innsbruck. Institut für Sprachen und Literaturen der Universität Innsbruck, Abteilung Sprachwissenschaft.

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.

Samardžić, Tanja; Merlo, Paola. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pp. 52–60, Uppsala, Sweden, July. Association for Computational Linguistics.

Sanches Duran, Magali; Ramisch, Carlos; Aluísio, Sandra Maria; Villavicencio, Aline. 2011. Identifying and Analyzing Brazilian Portuguese Complex Predicates. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 74–82, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sanromán Vilas, Begoña. 2009. Towards a semantically oriented selection of the values of Oper_1. The case of *golpe* 'blow' in Spanish. In Beck, David; Gerdes, Kim; Milićević, Jasmina; Polguère, Alain (eds.), *Proceedings of the Fourth International Conference on Meaning-Text Theory – MTT'09*, pp. 327–337, Montreal, Canada. Université de Montréal.

Sass, Bálint; Váradi, Tamás; Pajzs, Júlia; Kiss, Margit. 2010. *Magyar igei szerkezetek: A leggyakoribb vonzatok és szókapcsolatok szótára [Verbal constructions in Hungarian: a dictionary of the most frequent arguments and collocations]*. Tinta Könyvkiadó, Budapest.

Sass, Bálint. 2007. First attempt to automatically generate Hungarian semantic verb classes. In Davies, M.; Rayson, P.; Hunston, S.; Danielsson, P. (eds.), *Proceedings of the 4th Corpus Linguistics Conference*, Birmingham.

Sass, Bálint. 2008. The Verb Argument Browser. In Horák, Aleš; Kopeček, Ivan; Pala, Karel; Sojka, Petr (eds.), *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, pp. 187–192, Berlin, Heidelberg. Springer Verlag.

Sass, Bálint. 2009. "Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára ["Raisin" – a tool for analyzing the argument structure of Hungarian verbs]. In Váradi, Tamás (ed.), *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból / Selected Papers from the First Applied Linguistics PhD Conference*, pp. 117–129, Budapest. MTA Nyelvtudományi Intézet.

Sass, Bálint. 2010. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból [Extracting parallel multiword verbs from parallel corpora]. In Tanács, Attila; Vincze, Veronika (eds.), *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 102–110, Szeged. Szegedi Tudományegyetem.

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.

Sinha, R. Mahesh K. 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*, pp. 40–46, Singapore, August. Association for Computational Linguistics.

Sinha, Rai Mahesh. 2011. Stepwise Mining of Multi-Word Expressions in Hindi. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 110–115, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

Szarvas, György; Farkas, Richárd; Kocsor, András. 2006. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In *Discovery Science*, pp. 267–278.

Székely, Gábor. 2003. *A fokozó értelmű szókapcsolatok magyar és német szótára [Hungarian-German dictionary of intensifying phrases]*. Tinta Könyvkiadó, Budapest.

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.

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

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

Tóth, Csilla. 2007. Kollektivitás és disztributivitás vizsgálata a magyar statikus helyhatározóragok és névutók körében [Analysing collectivity and distributionality of Hungarian stative locative suffixes and postpositions]. *Nyelvtudományi Közlemények*, 104:222–242.

Toutanova, Kristina; Manning, Christopher D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP 2000*, pp. 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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

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.

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

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

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

Verkuyl, Henk J. 1972. *On the Compositional Nature of the Aspects*. Reidel, Dordrecht.

Villavicencio, Aline; Kordoni, Valia; Zhang, Yi; Idiart, Marco; Ramisch, Carlos. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1034–1043, Prague, Czech Republic, 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; Szarvas, György; Almási, Attila; Szauter, Dóra; Ormándi, Róbert; Farkas, Richárd; Hatvani, Csaba; Csirik, János. 2008. Hungarian word-sense disambiguated corpus. In Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph;

Odjik, Jan; Piperidis, Stelios; Tapias, Daniel (eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

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

Vincze, Veronika; Nagy T., István; Berend, Gábor. 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*, pp. 116–121, Portland, Oregon, USA, June. ACL.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011b. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

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*, pp. 207–218, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2008a. A puszta köznév + ige komplexumok státusáról [On the status of bare common noun + verb constructions]. In Sinkovics, Balázs (ed.), *LingDok 7. Nyelvész-doktoranduszok dolgozatai*, pp. 279–297, Szeged, Hungary. University of Szeged.

Vincze, Veronika. 2008b. *A főnév + ige szerkezetek lexikai reprezentációjáról [On the lexical representation of noun + verb constructions]*. Presented at the 12th National Conference of PhD Students of Linguistics. Szeged, Hungary, 2–3 December 2008.

Vincze, Veronika. 2009a. 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*, pp. 112–122, Budapest. MTA Nyelvtudományi Intézet.

Vincze, Veronika. 2009b. *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*, pp. 265–278, Szeged. Szegedi Tudományegyetem.

Vincze, Veronika. 2009c. 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*, pp. 390–393, Szeged. Szegedi Tudományegyetem.

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

Vincze, Veronika. 2009e. On the Machine Translatability of Semi-Compositional Constructions. In Váradi, Tamás (ed.), *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból / Selected Papers from the First Applied Linguistics PhD Conference*, pp. 166–178, Budapest. MTA Nyelvtudományi Intézet.

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*, pp. 327–332, Budapest. Tinta Könyvkiadó.

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. Nyelvészdoktoranduszok dolgozatai*, pp. 225–243, Szeged, Hungary. University of Szeged.

Vintar, Špela; Fišer, Darja. 2008. Harvesting multi-word expressions from parallel corpora. In Calzolari, Nicoletta; Choukri, Khalid; Maegaard, Bente; Mariani, Joseph; Odjik, Jan; Piperidis, Stelios; Tapias, Daniel (eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Viszket, Anita. 2004. *Argumentumstruktúra és lexikon: A puszta NP grammatikai sajátosságai a magyarban és ezek következményei a predikátumok lexikonbeli argumentumstruktúrájára [Argument structure and lexicon: Grammatical features of bare NPs in Hungarian and their consequences on the argument structure of predicates within the lexicon]*. Ph.D. thesis, Eötvös Loránd University, Budapest, Hungary.

Wanner, Leo (ed.). 1997. *Recent Trends in Meaning-Text Theory*. Benjamins, Amsterdam/Philadelphia.

Wanner, Leo (ed.). 2007. *Selected Lexical and Grammatical Issues in the Meaning-Text Theory. In Honour of Igor Mel'čuk*. Benjamins, Amsterdam / Philadelphia.

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

Wierzbicka, Anna. 1982. Why can you *have a drink* when you can't **have an eat*? *Language*, 58:753–799.

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

Zipf, George K. 1949. *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, MA.

Zsibrita, János; Vincze, Veronika; Farkas, Richárd. 2010. Ismeretlen kifejezések és a szófaji egyértelműsítés [Unknown expressions and POS-tagging]. In Tanács, Attila; Vincze, Veronika (eds.), *MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia*, pp. 275–283, Szeged, Hungary, December. University of Szeged.

Appendices

Appendix A

List of the most frequent Hungarian semi-compositional constructions

Semi-compositional constructions that occurred at least 3 times in the corpora are listed below. Their status is also marked: p refers to constructions being more similar to productive noun + verb combinations and i refers to those being more similar to idioms.

irányt ad	p	608	ajánlatot tesz	p	108
részt vesz	i	278	igénybe vesz	i	99
szerződést köt	p	217	nyilvántartásba vesz	i	95
forgalomba hoz	i	205	lehetőséget ad	p	84
nyilvánosságra hoz	i	197	rendelkezésre áll	i	80
eszébe jut	i	191	engedélyt ad	p	77
sor kerül	i	182	tanácsot ad	p	76
figyelembe vesz	i	167	bérbe ad	p	68
határozatot hoz	p	153	eleget tesz	i	64
döntést hoz	p	136	véget ér	i	63
hatályba lép	i	111	üzletet köt	p	60

megállapodást köt	p	52
időt tölt el	p	48
hírül ad	i	45
szolgáltatást nyújt	p	42
lépést tesz	p	38
szert tesz	i	38
életbe lép	i	34
választ ad	p	34
szerepet játszik	p	33
tevékenységet folytat	p	33
tudomásul vesz	i	33
munkát végez	p	32
tárgyalást folytat	p	32
kérdést tesz fel	p	28
igényt tart	i	27
előadást tart	p	26
haszonkölcsönbe ad	i	26
nyomon követ	i	26
nyereséget realizál	i	25
nyilatkozatot tesz	p	25
tevékenységet végez	p	25
helyet foglal	p	23
hozzáférést ad	p	23
megrendezésre kerül	i	23
különbséget tesz	p	22
segítséget nyújt	p	22
ülést tart	p	21
helyet kap	i	20
tájékoztatást ad	p	20
forgalomba kerül	i	18
használatba vesz	i	18
pénzt keres	i	18
számon tart	i	18
tiszteletben tart	i	17
állást foglal	i	16
helyt áll	i	16
lehetőség nyílik	p	16
őrizetbe vesz	i	16
szolgáltatást kínál	p	16
törvényt hoz	p	16
családot alapít	p	15
feleségül vesz	i	15
kezdetét veszi	i	15
számot ad	i	15
támogatást kap	p	15
alapul vesz	i	14
hibát követ el	p	14
kilátás nyílik	p	14
okot ad	p	14
szemügyre vesz	i	14
szerepet kap	p	14
véget vet	i	14
véghez visz	i	14
csődbe megy	i	13
felelősséget vállal	p	13
férjhez megy	i	13
jogot ad	p	13
lehetőséget nyújt	p	13
másolatot készít	p	13
otthont ad	p	13
összhangba hoz	i	13
piacra dob	i	13
pillantást vet	p	13
előkészületet tesz	p	12
hitelt nyújt	p	12
hozzájárulását adja	p	12

kísérletet tesz	p	12
megbízást ad	p	12
megjegyzést tesz	p	12
szóba jön	i	12
életet él	i	11
feljelentést tesz	p	11
figyelemmel kísér	i	11
hatást gyakorol	p	11
információt ad	p	11
ítéletet hoz	p	11
kárt tesz	p	11
kezelést végez	p	11
kézhez vesz	i	11
támogatást nyújt	p	11
választ kap	p	11
helyet ad	i	10
látogatást tesz	p	10
parancsot ad	p	10
támogatást ad	p	10
teljesítményt nyújt	p	10
tudomást szerez	p	10
bérbe vesz	i	9
erőfeszítést tesz	p	9
fontolóra vesz	i	9
hangot ad	p	9
hasznát veszi	p	9
hatalmába kerít	p	9
intézkedést hoz	p	9
kapcsolatban áll	i	9
koncertet ad	p	9
következtetésre jut	i	9
következtetést von le	p	9
lehetőséget kínál	p	9
lépést tart	p	9
nyilvánosságra kerül	i	9
partra száll	i	9
piacra kerül	i	9
sikerrel jár	i	9
sikert arat	p	9
számításba vesz	i	9
tekintetbe vesz	i	9
tudomására hoz	i	9
utasítást ad	p	9
üzembe helyez	i	9
áldását adja	p	8
alkalmat ad	p	8
biztosítékot ad	p	8
célba vesz	i	8
célt ér	i	8
életre hív	i	8
hasznot hoz	p	8
hírt ad	p	8
ígéretet tesz	p	8
intézkedést tesz	p	8
javaslatot tesz	p	8
késedelembe esik	i	8
meglepetést okoz	p	8
módot ad	p	8
rendbe hoz	i	8
részét képezi	p	8
számba vesz	i	8
szolgálatot tesz	p	8
tárgyalás folyik	p	8
alkalmazásra kerül	i	7
bejelentést tesz	p	7
éjszakát tölt	p	7

eljárást folytat	p	7
élményt nyújt	p	7
felvilágosítást ad	p	7
gondot fordít	i	7
haladékot ad	p	7
hangsúlyt fektet	p	7
interjút ad	p	7
kapcsolatot tart	p	7
kapcsolatot vesz fel	p	7
kibocsátásra kerül	i	7
kivételt képez	i	7
látványt nyújt	p	7
lehetőséget kap	p	7
lélegzetet vesz	p	7
mozdulatot tesz	p	7
nyilvántartást vezet	p	7
nyomást gyakorol	p	7
örömmel fogad	i	7
példát mutat	p	7
pénzt költ	p	7
perben áll	i	7
szívességet tesz	p	7
útra kel	i	7
ügyletet köt	p	7
üzenetet küld	p	7
vállat von	i	7
alapot ad	p	6
bajba jut	i	6
birtokba vesz	i	6
biztosítékot nyújt	p	6
biztosra vesz	i	6
célt tűz ki	p	6
elsőbbséget élvez	p	6
emelkedést mutat	p	6
erőt vesz	p	6
felhatalmazást ad	p	6
felhívja a figyelmet	p	6
figyelmet fordít	p	6
gyűlést tart	p	6
háborút visel	p	6
helyt ad	i	6
kapcsolatot hoz létre	p	6
kockázatot vállal	p	6
meglepetés ér	p	6
nyereséget mutat	p	6
postára ad	i	6
segítséget ad	p	6
szándéknyilatkozatot tesz	p	6
szóba áll	i	6
szolgáltatást ad	p	6
szünetet tart	p	6
túlzásba visz	i	6
utat tesz meg	p	6
vádat emel	p	6
változást hoz	p	6
védelmet nyújt	p	6
áldozatul esik	i	5
bizalmat vet	p	5
bosszút áll	p	5
ellenőrzést végez	p	5
említést tesz	p	5
eredményt hoz	p	5
értekezletet tart	p	5
esküt tesz	p	5
fedezetet nyújt	p	5
feledésbe merül	i	5

felmentést ad	p	5
fordulatot hoz	p	5
garanciát ad	p	5
gondot okoz	p	5
halálra ítél	i	5
használatba ad	i	5
hatást tesz	p	5
helyezést elér	p	5
hiányt szenved	i	5
igazat ad	i	5
irányt vesz	p	5
írásba foglal	i	5
kapcsolatba lép	i	5
kétségbe von	i	5
kézben tart	i	5
kivételt tesz	p	5
konferenciát tart	p	5
korlátot szab	p	5
közgyűlést tart	p	5
kudarcot vall	i	5
lökést ad	p	5
magyarázatot ad	p	5
megbeszélést folytat	p	5
műsorra tűz	i	5
nevet ad	p	5
nyugovóra tér	i	5
összefüggésbe hoz	i	5
sorra kerül	i	5
szabályozást ad	p	5
szerepet vállal	p	5
szóba kerül	i	5
szövetséget köt	p	5
tájékoztatót tart	p	5
tanácskozást tart	p	5
tapasztalatot szerez	p	5
titokban tart	i	5
tudomást vesz	p	5
veszteséget mutat	p	5
zavarba jön	i	5
ajánlatot kap	p	4
álomba merül	i	4
benyomást gyakorol	p	4
biztosítást köt	p	4
búcsút mond	p	4
csalódás ér	p	4
csalódást okoz	p	4
csapást mér	i	4
csődbe jut	i	4
életbe léptet	i	4
életre kel	i	4
eligazítást ad	p	4
eredményre vezet	i	4
értelembe vesz	i	4
felelősségre von	i	4
felhasználásra kerül	i	4
felhatalmazást kap	p	4
figyelmet felhív	p	4
győzelmet arat	p	4
hangsúlyt helyez	p	4
hátat fordít	i	4
háttérbe szorul	i	4
házasságot köt	p	4
hitet vet	p	4
kárt okoz	p	4
kézbe vesz	i	4
kijelentést tesz	p	4

kirándulást tesz	p	4
lehetőséggel él	i	4
lendületet ad	p	4
megállapításra jut	i	4
megbízást kap	p	4
meglepetéssel szolgál	i	4
mód nyílik	p	4
nyilatkozatot ad	p	4
nyilvántartást végez	p	4
összejövetelt tart	p	4
ötlete támad	i	4
panaszt tesz	p	4
példát követ	p	4
próbára tesz	i	4
rendet rak	p	4
részt vállal	i	4
sétát tesz	p	4
szakvéleményt ad	p	4
szerepet tölt be	p	4
szerephez jut	i	4
színpadra lép	i	4
szóba hoz	i	4
szót ejt	i	4
szövetségre lép	i	4
sztrájkba lép	i	4
teret nyer	p	4
tetten ér	i	4
tudomására jut	i	4
utalás történik	p	4
útmutatást ad	p	4
üzletpolitikát folytat	p	4
valóra vált	i	4
végére ér	i	4
vendégül lát	i	4
veszély fenyeget	i	4
vizsgálatot végez	p	4
vizsgát tesz	p	4
zavarba hoz	i	4
ajánlatot ad	p	3
árulást követ el	p	3
baja esik	i	3
bajba kerül	i	3
barátságot köt	p	3
befolyást gyakorol	p	3
békét köt	p	3
beleegyezést ad	p	3
bemutatásra kerül	i	3
bemutatót tart	p	3
benyomást kelt	p	3
benyomást tesz	p	3
beszámolót tart	p	3
bevezetésre kerül	i	3
célul tűz ki	i	3
csalódást kelt	p	3
egyezményt köt	p	3
egyezséget köt	p	3
egyezségre jut	i	3
elejét veszi	p	3
élen jár	i	3
elhatározásra jut	i	3
ellenőrzést tart	p	3
előnyben részesít	i	3
engedélyt kér	p	3
érdeklődést felkelt	p	3
érettségit tesz	p	3
értéket felmér	p	3

érvénybe lép	i	3
érvényre jut	i	3
eszébe ötlik	i	3
féken tart	i	3
felmérést készít	p	3
felszámolás alatt áll	i	3
felveszi a versenyt	p	3
felvételt nyer	p	3
foglalkozást űz	p	3
fordulatot vesz	p	3
formát ölt	p	3
főszerepet játszik	p	3
gátat szab	p	3
gyakorlatot végez	p	3
gyanút kelt	p	3
gyengülést mutat	p	3
háborúban áll	i	3
harcot vív	p	3
hasonlóságot mutat	p	3
határidőt szab	p	3
helyet foglal el	p	3
hiányt mutat	p	3
hivatalba lép	i	3
igényt támaszt	p	3
ígéretet kap	p	3
iránymutatást ad	p	3
irányt mutat	p	3
jogot gyakorol	p	3
jogot kap	p	3
jót tesz	p	3
jóváhagyását adja	p	3
jutalmat kap	p	3
kapcsolatba kerül	i	3
kapcsolatot létesít	p	3
kapcsolatot teremt	p	3
kedvet kap	i	3
kérdést intéz	p	3
kérdést vet fel	p	3
készen áll	i	3
kezet fog	p	3
kezet ráz	p	3
kézhez kap	i	3
kiadásra kerül	i	3
kiállítást rendez	p	3
kifizetésre kerül	i	3
lehetőséget biztosít	p	3
lehetővé tesz	i	3
leírást ad	p	3
letétbe helyez	i	3
megállapodásra jut	i	3
meghatározásra kerül	i	3
meglepetésként ér	i	3
megoldást talál	p	3
méretet ölt	i	3
nehezére esik	i	3
növekedést mutat	p	3
nyomás alá kerül	i	3
nyomot hagy	i	3
nyomozást folytat	p	3
örömet szerez	p	3
papírra vet	i	3
példát hoz	p	3
pénzt hoz	p	3
perbe fog	i	3
pofon üt	i	3
rendelkezésre bocsát	i	3

sajtótájékoztatót tart	p	3
segítséget kér	p	3
sorban áll	i	3
sorrendbe állít	i	3
sort kerít	p	3
sportot űz	p	3
szabályt ad	p	3
szemmel tart	i	3
szóhoz jut	i	3
szolgáltatást végez	p	3
szót szól	i	3
szót vált	p	3
szóvá tesz	i	3
találkozót tart	p	3
támogatást élvez	p	3
tanúságot tesz	p	3
tervbe vesz	i	3
testet ölt	i	3
tudtul ad	i	3
túlzásba esik	i	3
tüzet nyit	i	3
ügyet vet	i	3
üldözőbe vesz	i	3
vallomást tesz	p	3
védelembe vesz	i	3
véleményt ad	p	3
vereséget szenved	p	3
vizsgálatot folytat	p	3

Appendix B

List of the most frequent English semi-compositional constructions

Semi-compositional constructions that occurred at least 3 times in the corpora are listed below.

take place	36	commit suicide	8
make a decision	29	do service	8
take part	26	fall in love	8
play a role	25	give advice	8
take care	18	hold an event	8
take a decision	16	take a step	8
make a remark	14	take advantage	8
take a look	13	take the role	8
give an order	11	make a shift	7
make a mistake	11	pay attention	7
take a seat	11	take action	7
make a sign	10	come into force	6
give a concert	9	have access	6
meet a requirement	9	make a noise	6

make use	6	take a measure	4
provide service	6	take into consideration	4
put an end	6	take on board	4
take a photo	6	tell a lie	4
bring into line	5	burst into tears	3
give an account	5	do business	3
have effect	5	find a solution	3
have impact	5	give air	3
hold a meeting	5	give an opinion	3
make a promise	6	give a right	3
make an effort	5	give a ruling	3
make his debut	5	have a good time	3
make progress	5	have a rest	3
meet a criterion	5	have an accident	3
pay a visit	5	have an impact	3
take a leave	5	have flu	3
take control	5	have influence	3
take into account	5	have sex	3
catch sight	4	hold a celebration	3
do work	4	make a contribution	3
give access	4	make a journey	3
give an insight	4	make a report	3
give notice	4	make a trip	3
have dinner	4	make a voyage	3
hold a competition	4	make an appearance	3
hold a contest	4	make an attempt	3
make fun	4	make confession	3
make a speech	4	make his escape	3
meet a need	4	make his way	3
play a (key) role	4	make lace	3
spend time	4	make preparation	3
take a job	4	offer a service	3

provide assistance	3
reach agreement	3
render a service	3
set a trap	3
take a shower	3
take a walk	3
take an action	3
take a bath	3
take effect	3
take his place	3

Appendix C

Lists of semi-compositional constructions with the verbs *ad* ‘give’, *vesz* ‘take’, *hoz* ‘bring’ or *tesz* ‘do’

C.1 Semi-compositional constructions with the verb *ad* ‘give’

adótanácsot ad
ajánlást ad
ajánlatot ad
alapot ad
áldását adja
alkalmat ad
árnyékot ad
áttekintést ad
bátorságot ad
becslést ad
beleegyezést ad
beleszólást ad
bérbe ad
beszámolót ad
betekintést ad
bevezetést ad
bizonyságot ad
biztosítékot ad
definíciót ad
elégtételt ad
elfogatóparancsot ad ki
eligazítást ad
előírást ad
előrejelzést ad
elsőbbséget ad
engedélyt ad

engedményt ad
eredményt ad
erőt ad
értelmet ad
értelmezést ad
értésére ad
esélyt ad
estet ad
extraszolgáltatást ad
fedezetet ad
feleletet ad
felhatalmazást ad
felhívást ad
felmentést ad
felvilágosítást ad
férjhez ad
garanciát ad
haladékot ad
hálát ad
hangot ad
használatba ad
haszonkölcsönbe ad
határidőt ad
hatáskört ad
hátteret ad
helyet ad
helyt ad
hírt ad
hírül ad
hitelt ad
hozzáférést ad
hozzájárulását adja
időt ad
igazat ad
igazolást ad
információt ad
interjút ad
iránymutatást ad
irányt ad
írásba ad
ismeretet ad át
ismertetést ad
jelet ad
jelét adja
jelleget ad
jelt ad
jelzést ad
jogosítást ad
jogot ad
jóváhagyását adja
kártalanítást ad
kedvezményt ad
kegyelmet ad
képet ad
keretet ad
kézbe ad
kezet ad
kiegészítést ad
kifejezést ad
kivételt ad
koncertet ad
kölcsönt ad
közleményt ad (ki)
lehetőséget ad
leírást ad
lemondást ad be

lendületet ad
létjogosultságot ad
lökést ad
magánórát ad
magyarázatot ad
megbízást ad
megoldást ad
megrendelést ad
mérleget ad
módot ad
nevelést ad
nevet ad
nyilatkozatot ad
nyomatékot ad
okot ad
osztalékelsőbbséget ad
osztályozást ad
otthont ad
öblítést ad
összefoglalót ad
ösztönzést ad
ötletet ad
parancsot ad
példát ad
poflevest ad
postára ad
próbaórát ad
publikációt ad
rangot ad
reményt ad
rendelést lead
részletezést ad
riasztást ad
segélyt ad
segítséget ad
súlyt ad
szabályozást ad
szabályt ad
szakvéleményt ad
számot ad
szárnyakat ad
szavát adja
szavazatát adja
szavazatelsőbbséget ad
szavazatokat ad
szerepet ad
szeretetet ad
szolgáltatást ad
szólókoncertet ad
tájékoztatást ad
tájékoztatót ad
támogatást ad
támpontot ad
tanácsot ad
tanújelét adja
teret ad
termést ad
továbbképzést ad
tudtára ad
tudtul ad
tulajdonba ad
utasításba ad
utasítást ad
útbaigazítást ad
útmutatást ad
üzenetet ad át

választ ad
válókeresetet ad be
védelmet ad
védőoltást ad
véleményt ad
zajt ad

C.2 Semi-compositional constructions with the verb *vesz* 'take'

alapul vesz
alkalmazásba vesz
bérbe vesz
birtokba vesz
biztosra vesz
búcsút vesz
célba vesz
elejét veszi
erőt vesz
értelembe vesz
feleségül vesz
felveszi a kapcsolatot
felveszi a versenyt
figyelembe vesz
fontolóra vesz
fordulatot vesz
használatba vesz
hasznát veszi
homályba vesz
igénybe vesz
irányt vesz
kézbe vesz
kezdetét veszi
kézhez vesz
lélegzetet vesz
leltárba vesz
lendületet vesz
levegőt vesz
nyilvántartásba vesz
őrizetbe vesz
örömmel vesz
pártfogásba vesz
példának vesz
példát vesz
részt vesz
startot vesz
számba vesz
számításba vesz
szemügyre vesz
táncórát vesz
tekintetbe vesz
tervbe vesz
tudomást vesz
tudomásul vesz
tulajdonba vesz
utat vesz
üldözőbe vesz
védelembe vesz
vizsgálat alá vesz
zuhanyt vesz

C.3 Semi-compositional constructions with the verb *hoz* 'bring'

áldozatot hoz
állapotba hoz
állásfoglalást hoz
áttörést hoz
bajt hoz
bevételt hoz
divatba hoz
döntést hoz
egyensúlyba hoz
előírást hoz
előnyt hoz
enyhülést hoz
eredményt hoz
fordulatot hoz
forgalomba hoz
hasznot hoz
határozatot hoz
helyzetbe hoz
hírbe hoz
hírül hoz
intézkedést hoz
iránymutatást hoz
ítéletet hoz
jogszabályt hoz
lázba hoz
megoldást hoz
mentségére felhoz
mozgásba hoz
nyereséget hoz
nyilvánosságra hoz
összefüggésbe hoz
összhangba hoz
példát hoz
pénzt hoz
profitot hoz
rendbe hoz
rendeletet hoz
rendelkezést hoz
sikert hoz
szabályt hoz
szakvéleményt hoz
szavazatot hoz
szégyenbe hoz
szégyent hoz
szóba hoz
törvényt hoz
tudomására hoz
változást hoz
végzést hoz
vészt hoz
zavarba hoz

C.4 Semi-compositional constructions with the verb *tesz* 'do'

ajánlást tesz
ajánlatot tesz
ártalmatlanná tesz
bejelentést tesz

benyomást tesz
célzást tesz
eleget tesz
ellenajánlatot tesz
előkészületet tesz
említést tesz
engedményt tesz
érettségit tesz
erőfeszítést tesz
esküt tesz
észrevételt tesz
felfedezést tesz
feljelentést tesz
felszólítást tesz
fogadalmat tesz
fogadást tesz
folyamatba tesz
fordulatot tesz
hatást tesz
hitet tesz
ígéretet tesz
indítványt tesz
intézkedést tesz
javaslatot tesz
jelentést tesz
jót tesz
kárt tesz
kérdést tesz fel
kijelentést tesz
kirándulást tesz
kísérletet tesz
kivételt tesz
kört tesz
közszemlére tesz
különbséget tesz
látogatást tesz
lehetővé tesz
lépést tesz
megállapítást tesz
megjegyzést tesz
megkülönböztetést tesz
mérlegre tesz
módosítást tesz
mozdulatot tesz
nyilatkozatot tesz
panaszt tesz
pénzzé tesz
próbára tesz
próbát tesz
rendbe tesz
rendet tesz
sétát tesz
szándéknyilatkozatot tesz
szemrehányást tesz
szert tesz
szívességet tesz
szolgálatot tesz
szóvá tesz
tanúságot tesz
tanúvallomást tesz
utat tesz
vallomást tesz
vizsgát tesz

Appendix D

List of English-Hungarian semi-compositional constructions and their verbal counterparts

	applause breaks out	tapsvihar tör ki	
ask	ask a question	kérdést tesz fel	kérdez
	attach weight	súlyt ad	
fruit	bear fruit	gyümölcsöt hoz	gyümölcsözik
	become prey	áldozatul esik	
	bid farewell	búcsút int	búcsúzik
complain	bring a complaint	panaszt tesz	panaszol
	burst into applause	tapsban tör ki	
	burst into flame	lángba borul	
	burst into tears	könnyekben tör ki	
	call attention	figyelmet felhív	
glance	cast a glance	pillantást vet	pillant
doubt	cast doubt	kétségbe von	
	come into circulation	forgalomba kerül	
	come into effect	érvénybe lép	
	come into force	hatályba lép	

	come into use	használatba kerül	
	come to life	életre kel	éled
	come to terms	megállapodásra jut	megállapodik
conclude	come to the conclusion	következtetésre jut	következtet
	come to understanding	megállapodásra jut	megállapodik
	commit a mistake	hibát vét	hibázik
sin	commit a sin	bűnt követ el	bűnöz
	conduct a research	kutatást folytat	kutat
lecture	deliver a lecture	előadást tart	előad
	deliver experience	tapasztalatot nyújt	
	deliver an opinion	véleményt ad	véleményez
service	deliver service	szolgáltatást nyújt	szolgáltat
	do a good office	jó üzletet csinál	
	do business	üzletet köt	
harm	do harm	kárt tesz	(meg)károsít
honour	do honour	tiszteletet ad	
	do kindness	szívességet tesz	
research	do research	kutatást folytat	kutat
service	do service	szolgálatot tesz	szolgál(tat)
conclude	draw a conclusion	következtetést levon	következtet
	earn money	pénzt keres	
succeed	enjoy success	sikert arat	
	enter into force	hatályba lép	
control	exercise control	hatalmat gyakorol	
	fall in love	szerelembe esik	
	fall into pieces	darabokra esik	
solve	find a solution	megoldást talál	megold
	follow practice	gyakorlatot követ	
	form a part	részét képezi	
	gain admittance	bebocsátást nyer	
	gain ground	teret nyer	
	gain strength	erőt nyer	
	get a role	szerepet kap	
	get advice	tanácsot kap	

	get an offer	ajánlatot kap	
	get odds	fogadást tesz	fogad
command	give a command	parancsot ad	parancsol
	give a concert	koncertet ad	
describe	give a description	leírást ad	leír
	give a good upbringing	jó nevelést ad	
	give a hand	segédkezet nyújt	
hint	give a hint	célzást tesz	céloz
lesson	give a lesson	leckét ad	
license	give a license	engedélyt ad	engedélyez
mark	give a mark	jelet ad	jelez
picture	give a picture	képet ad	
	give a relation	beszámolót nyújt/ad	beszámol
reply	give a reply	választ ad	válaszol
	give a say	beleszólást ad	
	give a speech	beszédet tart	
state	give a statement	állítást tesz	állít
	give a token	jelét/zálogát adja	
treat	give a treatment	kezelést ad	kezel
try	give a try	próbát tesz	megpróbál
wound	give a wound	sebet ejt/üt	(meg)sebez
	give access	hozzáférést ad	
	give account	számot ad	beszámol
advise	give advice	tanácsot ad	tanácsol
	give an account	számot ad	beszámol
answer	give an answer	választ ad	válaszol
assign	give an assignment	megbízást ad	megbíz
impress	give an impression	benyomást kelt	
	give an insight	betekintést ad	
	give an opinion	véleményt ad	véleményez
order	give an order	parancsot ad	parancsol
assent	give assent	hozzájárulását adja	hozzájárul
	give audience	audienciát tart	
bear	give birth	életet ad	
consider	give consideration	megfontolás tárgyává teszi	megfontolja

delight	give delight	örömet szerez/élvezetet nyújt	
disturb	give disturbance	zavart kelt	
help	give help	segítséget nyújt	segít
	give home	otthont ad	
inform	give information	információt ad	
liberate	give liberty	szabaddá tesz	kiszabadít
name	give name	nevet ad	elnevez
	give notice	figyelmeztetést ad	figyelmeztet
	give opportunity	lehetőséget ad	
empower	give power	erőt ad	
protect	give protection	védelmet ad	véd
	give a relation	beszámolót nyújt/ad	beszámol
relieve	give relief	enyhülést nyújt/ad	enyhít
	give responsibility	felelősséget ad	
	give right	jogot ad	(fel)jogosít
raise	give rise	alkalmat ad	
rule	give ruling	döntést hoz	dönt
	give space	teret ad	
vaccinate	give vaccination	oltást ad	(be)olt
	give way	elsőbbséget ad	
	give weight	súlyt ad	
	go on sale	forgalomba kerül	
	grant concession	engedményt tesz/ad	engedményez
	have a conversation	beszélgetést folytat	beszélget
	have a few words	szót vált	
	have a meeting	találkozót tart	
	have a nervous break-down	idegösszeomlást kap	
	have a party	partit rendez	
shower	have a shower	zuhanyt vesz	zuhanyozik
walk	have a walk	sétát tesz	sétál
	have an accident	balesetet szenved	
	have an affair	viszonyt folytat	
agree	have an agreement	megegyezésre/egyezségre jut	megegyezik / -

intend	have intent	szándékában áll	szándékozik
war	have war	háborút visel	háborúzik
	hold a ball	bált tart	
	hold a camp	tábort tart	
celebrate	hold a celebration	ünnepséget tart	
	hold a championship	bajnokságot rendez	
	hold a competition	versenyt rendez	
	hold a conference	konferenciát tart	
	hold a conversation	társalgást/beszélgetést folytat	társalog/beszélget
	hold a consultation	konzultációt tart	
	hold a contest	versenyt rendez	
demonstrate	hold a demonstration	demonstrációt tart	demonstrál
	hold a festival	fesztivált rendez	
	hold a meeting	találkozót tart	
	hold a party	partit rendez	
perform	hold a performance	előadást tart	előad
demonstrate	hold a demonstration	demonstrációt tart	demonstrál
	hold a position	betölt tisztséget/hivatalt/állást	
	hold a race	versenyt rendez/tart	
	hold a record	rekordot tart	
	hold a service	misét/istentiszteletet tart	misézik / -
	hold a session	ülést/összejövetelt tart	ülésezik/ -
	hold a summit	csúcstalálkozót tart	
	hold a talk	tárgyalást tart	tárgyal
	hold an auction	árverést tart	árverez
	hold an election	választást tart	
	hold an event	eseményt rendez	
penalize	impose penalty	büntetést ad	(meg)büntet
	keep a secret	titkot tart	
	keep a steady eye	rajta tartja a szemét	
	keep an eye	szemmel tart	
	keep contact	kapcsolatot tart	
	keep in (good) condition	valamilyen (jó) állapotban tart	
	keep in mind	észben tart	
	keep it secret	titokban tart	titkol

	keep order	rendet tart	
	keep silence	csöndben marad	
tally	keep tally	számon tart	
	keep to the rules	tartja magát a szabályokhoz	
counterattack	launch a counterattack	ellentámadást indít	
found	lay a foundation	lerakja az alapjait	megalapoz/megalapít
plan	lay a plan	tervet lefektet	megtervez
claim	lay claim	igényt tart	igényel
	lead into temptation	kísértésbe visz	(meg)kísért
	leave an impression	benyomást kelt	
	leave room	teret enged	
	leave to fate	sorsára hagy	
weightlift	lift weight	súlyt emel	
	light a fire	tüzet gyújt	
	live a life	életet él	
	maintain interest	fenntartja az érdeklődést	
use	make (full) use	legjobb hasznát veszi	
bet	make a bet	fogadást köt	fogad
claim	make a claim	igényt támaszt	
comment	make a comment	megjegyzést tesz	megjegyez
compare	make a comparison	összehasonlítást tesz	összehasonlít
complain	make a complaint	panaszt tesz	panaszol
contract	make a contract	szerződést köt	
contribute	make a contribution	hozzájárulást tesz	hozzájárul
	make a countenance	képet vág	
	make a counterargument	ellenérvet felhoz	
decide	make a decision	döntést hoz	dönt
declare	make a declaration	kijelentést tesz	kijelent
deduce	make a deduction	következtetésre jut	következtet
discover	make a discovery	felfedezést tesz	felfedez
	make a film	filmet készít	
journey	make a journey	utazást tesz	utazik
leap	make a leap	előrelépést tesz	előrelép
	make a mistake	hibát vét	hibázik

move	make a move	lépést tesz	lép
	make a movie	filmet készít	
promise	make a promise	ígéretet tesz	ígér
question	make a question	kérdést feltesz	kérdez
	make a reflection	észrevételt tesz	észrevételez
regulate	make a regulation	szabályt hoz	szabályoz
remark	make a remark	megjegyzést tesz	megjegyez
report	make a report	jelentést tesz	jelent
sacrifice	make a sacrifice	áldozatot hoz	
	make a shift	módot talál	kimódol
sign	make a sign	jelet ad	jelez
	make a speech	beszédet tart	
state	make a statement	állítást tesz	állít
step	make a step	lépést tesz	lép
stir	make a stir	feltűnést kelt	
suggest	make a suggestion	javaslatot tesz	javasol
survey	make a survey	felmérést végez	felmér
trip	make a trip	kirándulást tesz	kirándul
visit	make a visit	látogatást tesz	(meg)látogat
voyage	make a voyage	utazást tesz	utazik
	make advance	megteszi az első lépést	
	make allowance	figyelembe/számításba vesz	
	make alterations	módosítást/változtatást eszközöl	módosít/változtat
allude	make an allusion	célzást tesz	céloz
apologize	make an apology	bocsánatot kér	
	make an effort	erőfeszítést tesz	
err	make an error	hibát vét	hibázik
exaggerate	make an exaggeration	túlzásba esik	túloz
	make an exception	kivételt tesz	kivételez
impress	make an impression	benyomást kelt	
observe	make an observation	megfigyelést tesz	megfigyel
	make an oath	esküt tesz	esküszik/esküdik
offer	make an offer	ajánlatot tesz	
	make career	karriert csinál	
confess	make confession	vallomást tesz	vall
contact	make contact	kapcsolatba lép/kapcsolatot felvesz	

fun	make fun	tréfát űz	tréfál
	make gain	hasznot húz	
	make his name	hírnevet szerez	
	make his way	utat tör	
investigate	make investigation	vizsgálódás tárgyává teszi	megvizsgál
	make lace	csipkét ver	
	make law	törvényt hoz	
mention	make mention	említést tesz	említ
	make money	pénzt hoz	
noise	make noise	zajt csap	zajong
order	make order	parancsot ad	parancsol
	make peace	békét köt	
prepare	make preparation	előkészületet tesz	előkészül
proclaim	make proclamation	közhírré tesz	
profit	make profit	hasznot hoz	
progress	make progress	haladást tesz	halad
publish	make publication	közhírré tesz	
recommend	make recommendation	ajánlást tesz/ javaslatot tesz	ajánl/javasol
use	make use	hasznát veszi	hasznosít
	meet a condition	eleget tesz a feltételeknek	
	meet a requirement	eleget tesz a követelménynek	
bid	offer a bid	ajánlatot tesz	
service	offer a service	szolgáltatást nyújt	szolgáltat
solve	offer a solution	megoldást ad	megold
	offer an opportunity	lehetőséget ad	
	offer chance	esélyt ad	
visit	pay a visit	látogatást tesz	meglátogat
attend	pay attendance	látogatást tesz	meglátogat
	pay attention	figyelmet szentel	odafigyel
	pay respect	tiszteletét teszi	
investigate	perform investigation	nyomozást folytat	nyomoz
stunt	perform stunt	mutatványokat végez	
confide	place confidence	bizalmat vet	megbízik
emphasize	place emphasis	hangsúlyt tesz	hangsúlyoz

	play a (key) role	(kulcs)szerepet játszik	
	play a part	szerepet játszik	
	play a role	szerepet játszik	
	provide access	hozzáférést nyújt	
	provide an opportunity	lehetőséget ad	
assist	provide assistance	segítséget nyújt	segít
service	provide service	szolgáltatást nyújt	szolgáltat
support	provide support	támogatást ad	támogat
	pursue a policy	politikát folytat	
question	put a question	kérdést tesz fel	kérdez
end	put an end	véget vet	végez
risk	put at risk	kockára tesz	kockáztat
confide	put confidence	bizalmat helyez	(meg)bízik
	put energy into	energiát fektet	
bid	put in bid	árajánlatot tesz	
	put in place	helyére tesz	
	put into circulation	forgalomba hoz	
	put into effect	hatályba léptet	
stage	put on stage	színpadra visz	
pressure	put pressure	nyomást gyakorol	
	raise a question	kérdést vet fel	
decide	reach a decision	döntésre jut	dönt
	reach agreement	megállapodásra jut	megállapodik
	receive a call	hívást fogad	
	receive an answer	választ kap	
	receive assistance	segítséget kap	
service	render a service	szolgálatot tesz	szolgál(tat)
	represent a sight	látványt nyújt	
	risk threatens	veszély fenyeget	
trap	set a trap	csapdát állít	
fire	set fire	tüzet gyújt	
	shake hands	kezet ráz	kezel
	shed blood	vért ont	
	shoot a film	filmet forgat	
sign	show a sign	jelet mutat	jelez

trend	show a trend	tendenciát mutat	tendál
improve	show improvement	fejlődést mutat	fejlődik
	show interest	érdeklődést mutat	érdeklődik
respect	show respect	tiszteletet mutat	
	spend money	pénzt költ	
	spend time	időt tölt	időzik
	step on stage	színpadra lép	
awe	strike awe	félelmet ébreszt	megfélemlít
	suffer heart attack	szívrohamot kap	
bathe	take a bath	fürdőt vesz	fürdik
	take a break	szünetet tart	
breathe	take a breath	levegőt vesz	lélegzik
decide	take a decision	döntést hoz	dönt
	take a leave	búcsút vesz	búcsúzik
	take a position	állást felvesz	
	take a role	szerepet vállal	
seat	take a seat	helyet foglal	
shower	take a shower	zuhanyt vesz	zuhanyozik
step	take a step	lépést tesz	lép
seat	take a seat	helyet foglal	elhelyezkedik
turn	take a turn	fordulatot vesz	megfordul
video	take a video	videóra vesz	(le)videóz
voyage	take a voyage	utazást tesz	utazik
walk	take a walk	sétát tesz	sétál
account	take account	számításba vesz	(be)számít
	take action	akcióba lép/cselekvésbe fog	
	take an action	akcióba lép/cselekvésbe fog	
	take an exam	vizsgát tesz	vizsgázik
inventory	take an inventory	leltárt felvesz	leltározik
care	take care	gondot visel	gondoz
defend	take defense	védelmébe vesz	véd
	take effect	életbe lép	
	take effort	erőfeszítést tesz	
	take examination	vizsgát tesz	vizsgázik
	take his place	elfoglalja a helyét	

consider	take in consideration	figyelembe vesz	
	take instructions	útmutatást kap	
account	take into account	számításba vesz	számol
consider	take into consideration	figyelembe vesz	
	take into custody	őrizetbe vesz	
	take into possession	birtokba vesz	
	take on a look	kifejezést ölt	
participate	take part	részt vesz	részesül
pleasure	take pleasure	örömet lel	örül
	take possession	birtokba vesz	
refuge	take refuge	menedéket vesz	
risk	take risk	kockázatot vállal	kockáztat
shape	take shape	alakot ölt	
	watch a film	filmet néz	
	win an opportunity	lehetőséget kap	
	work his way	utat tör	