Distinguishing Compositional and Non-Compositional Verbal Complexes: A Corpus-Based Approach

Tibor Szécsényi[a] and Lívia Gyulai[b]

[a,b]University of Szeged, Szeged, Hungary

[a]szecsenyi.tibor@szte.hu; [b]liviagyulai95@gmail.com

Abstract: In this study we present research focusing on the argument structure of non-compositional verbs modified by verbal modifiers, using data obtained from a corpus and an asemantic version of the principle of compositionality. We examine the compositionality of the meanings of complex expressions from a syntactic perspective, focusing on their formal features and formal behavior instead of their semantic properties. Our main question is whether it is possible to predict, solely using syntactic methods, whether the meaning of these expressions is compositional or the verb modified by a verbal modifier has a non-compositional reading. We propose that systematic changes in the argument structure of the verb in the presence of a verbal modifier result in a compositional reading, while non-systematic ones give a non-compositional reading. In this paper, we present a method that makes this information automatically retrievable from the corpus without any need for semantics.

Keywords: compositionality; corpus-based; argument structure, verbal modifier; Hungarian

1. Compositionality and Idiomacity: Asemantic Compositionality

The Fregean Compositionality Principle can be formulated in various ways; differences in formulations are discussed in detail in works such as Zoltán Gendler Szabó (2000; 2012; 2017). For now, let's start with Partee's simple formulation (Partee 2004, 153):

(1) The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined.

The Compositionality Principle ensures that speakers of natural languages can understand each other, despite the fact that these languages contain an infinite number of linguistic expressions and sentences. Therefore, in the interest of mutual intelligibility, we consider the Compositionality Principle to hold for every multi-word expression in every natural language, with no exception: expressions in natural languages are compositional. We highlight two components of the Compositionality Principle here: 1. "the meaning of the parts" and 2. "the way of combination".

However, in natural language, the meaning of some expressions is not compositional but rather idiomatic.

(2) a. hot dog
b. the apple of someone's eye
c. someone kicks the bucket

In the examples in (2), even though we understand the meanings of the individual words and how they are combined, we cannot determine the meaning of the entire expression. However, since we assume the universality and exceptionlessness of the Compositionality Principle, we must conclude that in these expressions, the way the parts are combined is unique, or that the parts themselves have unique meanings within these expressions.

1.1 Idiomaticity as Lexical Ambiguity

If we consider the way the components are combined in the examples in (2) to be unique, implying that idiomatic expressions have unique structures, we can attribute a unique meaning to these unique structures. However, we cannot fully explain why idiomatic expressions do not differ more from compositional expressions. Idiomatic expressions almost always have a literal meaning as well, and, even under the idiomatic reading, some of the idioms have their own internal syntax: some words, such as *someone* in examples (2b–c) are modifiable. Other components of idioms are fixed and can only be specific words. If the structure is unique, why is there a need for the components to be fixed?

Another way to define the meaning of idioms according to the Compositionality Principle is to consider the meaning of (some of) the words appearing in them as unique. Since the meanings of words are always unique and rely solely on convention, we can say that the words in idiomatic expressions appear as separate lexical items in the lexicon. In the example in (2c), the verb *kick* has two distinct lexical representations: kick_1 provides the "conventional", compositional meaning, and kick_2 represents the idiomatic meaning found in the example:

(3) kick_1:

$$\begin{bmatrix} \text{form} & kick \\ \text{category} & \text{V} \\ \text{argument structure} & \langle \text{NP}[nom]_{\text{i}},\ \text{NP}[acc]_{\text{j}} \rangle \\ \text{meaning} & '(\text{i})\ strike\ or\ propel\ forcibly\ (\text{j})\ with\ the\ foot' \end{bmatrix}$$

(4) kick_2:

$$\begin{bmatrix} \text{form} & kick \\ \text{category} & \text{V} \\ \text{argument structure} & \langle \text{NP}[nom]_{\text{i}},\ \text{NP}[\text{form: } the\ bucket]_{\text{j}} \rangle \\ \text{meaning} & '(\text{i})\ die' \end{bmatrix}$$

In the lexical entry in (3), where the word *kick* has its compositional meaning, the meaning description references the roles of the subject (i) and the object (j) in the situation described by the verb. However, in (4), the lexical entry provides the entire idiomatic meaning of the word, with only the subject (i) indicated as the variable participant. The two lexical descriptions differ not only in how they describe meaning but also in detailing argument structure. Both the compositional and idiomatic uses of the verb have two arguments. However, while the compositional *kick* imposes only syntactic restrictions on its object, the idiomatic usage also imposes constraints on the phonological form, i.e., the lexical content of the object. Specifically, only the object *the bucket* can co-occur with it in the sentence. Another distinction lies in the fact that the compositional verb utilizes the meaning descriptions of both its arguments to define the verb's meaning, whereas the idiomatic verb completely disregards the compositional meaning of the object.

For idiomatic expressions containing lexical elements, we do not calculate with the meaning of all the lexical expressions when determining the meaning of the entire expression (as it would be the case in accordance with the Compositionality Principle). Therefore, we can consider expressions idiomatic

if they contain a component whose meaning is disregarded but whose presence is obligatory, with the meaning of the whole expression being lexically determined.

1.2 Functional Categories in Idioms

Idiomatic expressions can also contain functional elements such as the verbal modifiers under investigation in the present paper. In this case, it becomes even more challenging to determine what makes the expression idiomatic. At times functional expressions are difficult to assign basic or compositional meanings.

Example (5) is a good illustration of the compositional behavior of a verbal modifier in Hungarian. One of the most common uses of the Hungarian verbal modifier *be* ('into') occurs with verbs that describe the movement of something. The *be* ('into') verbal modifier is used when we also specify the endpoint of this movement in the sentence, and this endpoint is located inside some space. The endpoint of the movement is expressed as an illative (*-ba*) phrase, and the *be* verbal modifier functions as a kind of clitic.

(5) a. Péter rúgott egy labdá-t.
Peter kicked a ball-ACC
'Peter kicked a ball.'

b. Péter be rúgott egy labdá-t a kapu-ba.
Peter into kicked a vall-ACC the goal-ILL
'Peter kicked the ball into the goal.'

Sentence (5b) is entirely compositional, and we can replace the verb, the object, and the illative complement with other expressions.

However, the verbal modifier *be* ('into') and the verb *rúg* 'kick' can be associated not only with the meaning seen in (5b) but also with another, idiomatic meaning:

(6) Péter be rúgott a sör-től.
Peter into kicked the beer-ABL
'Peter got drunk from the beer.'

In sentence (6), neither *be* 'into' nor *rúg* 'kick' are used with their original, compositional meanings ('move into a place' and 'strike or propel something with the foot forcibly', respectively). The *be* + *rúg* construction is an idiomatic expression here, where the verbal modifier *be* can only be interpreted in conjunction with the *rúg* verb as 'get drunk', and conversely, the verb *rúg* only conveys this meaning in the presence of this verbal modifier.

Two important differences can be observed between the idiomatic sentence in (6) and the idiomatic expressions in (2). One is that the idioms in (2), the idiomaticity of which relies exclusively on lexical expressions, can be interpreted not only idiomatically but also literally. In the case of sentence (6), however, there is no literal reading; they are purely idiomatic.

The other difference is that in sentence (6) the expressions within the idiomatic construction not only change in meaning when used idiomatically compared with their compositional usage but also in other properties. As seen in the examples in (5), the verb *rúg* 'kick' takes an accusative complement, and the verbal modifier *be* ('into') takes an illative one. However, in the idiomatic construction in sentence (6), neither of these is present, but optionally there can be a modifier in ablative case. Idiomaticity, therefore, entails not only uniqueness of meaning but also changes in other properties such as argument structure, as seen in the lexical descriptions in (3)–(4).

1.3 Asemantic Compositionality

In identifying idioms, we have relied on native language intuition so far: an expression is considered idiomatic when its meaning is not identical to the literal meaning. However, this presupposes that expressions have a literal meaning, precisely one, and any other meanings deviating from that are idiomatic. Some expressions, though, lack a literal meaning (as seen in (6)), while others have multiple meanings that we do not want to identify as idiomatic. The challenge therefore is determining the meanings of expressions, whether they are literal or idiomatic.

Identifying idiomatic expressions may prove highly subjective. Our aim is to present a method for identifying a subset of idioms without relying on native speaker intuition or the meanings of expressions. Primarily, we demonstrate a method for the identification of idiomatic verbal modifier + verb constructions which does not involve a qualitative analysis of individual sentences or constructions but is solely based on quantitative analysis. We obtain the necessary amount of linguistic data from a pre-parsed linguistic corpus with morphological and syntactic features, and, rather than relying on the meaning of words and expressions, we consider the way they are used.

In the process of idiom identification, we rely on the previously mentioned observation that idiomatic expressions differ from compositional ones not only in meaning but also in other properties. In our case, we focus on the observation that verbal modifiers can change the argument structure of the verb. We infer the idiomaticity of the verbal modifier + verb construction from the uniqueness of the argument structure changing ability of verbal modifiers. Naturally, this allows us to identify only those idiomatic expressions in which the verbal modifier triggers a unique change in argument structure and not those where, in addition to the idiomatic reading, there is also a literal reading.

Since this way of differentiating between compositional and idiomatic constructions does not take into account the meaning of the expressions, we are essentially using an alternative, extended formulation of the Compositionality Principle. If we replace *meaning* with *properties*, meaning essentially becomes just one of the properties falling within the scope of the principle. This way, we obtain a Compositionality Principle that can be applied without reference to meaning, creating an Asemantic Compositionality Principle:

(7) The properties of an expression are a function of the properties of its parts and of the way they are syntactically combined.

Before demonstrating the applicability of asemantic compositionality to verbal modifier + verb constructions in Section 1.5, we review the basic properties of verbal modifiers in Section 1.4 below.

1.4 About Verbal Modifiers

Most of the verbal modifiers form a closed class with their primary function being indicating direction or creating a new, idiomatic meaning: *be* ‘into’, *ki* ‘out’, *le* ‘down’, *fel* ‘up’, *el* ‘away’, át ‘over’, *rá* ‘on’, *ide* ‘here’, *oda* ‘there’, *szét* ‘apart’, össze ‘together’, *vissza* ‘back’. One of the most commonly used verbal modifiers, *meg*, however, does not have such a directional meaning; instead, it changes the aspect of the verb, making it telic.

With the basic verbal modifiers, we can observe that the verbal modifier duplicates and appears on the argument of the verb as a case suffix, too, as we have seen in example (5b), as well as in sentence (8) below.

(8) Péter **rá** rakja a könyv-et az asztal-**ra**
Peter on put the book-ACC the table-SUBL
‘Peter puts the book on the table.’

As seen above, the verbal modifier *rá* 'on' appears before the verb but is also linked to the argument of the verb with the help of the sublative suffix.

Directional markings are considered the primary function of verbal modifiers; in these cases, they form a compositional structure with the verb. There are additional functions of the verbal modifiers in Hungarian like marking telicity.

1.4.1 Telicity in Hungarian

Verbal modifiers generally mark telicity, and not the outer aspect of perfectivity, nor specific Aktionsarten. As stated in Kardos' (2016) study, many elements in the Hungarian language indicate telicity, including the verb, its arguments, and potentially the pragmatic context. In this article, Kardos compares Hungarian with the English language: "Hungarian verbal particles and resultative/locative expressions mark telicity by directly placing bounds on events by virtue of serving an event maximalizing function, whereas the English counterparts of these elements do not have such direct event-bounding effects." (Kardos 2016, 1). Her study highlights cross-linguistic differences in connection with the aspectual role of verbal particles and resultative/locative expressions and the referential properties of telic verbal predicates. In her article, she uses the Hungarian variant of in/for X time unit test to probe for telicity. As discussed in the study of É. Kiss (2008, 1), telic and resultative VMs are the ones with telic function: "Resultative particles mark telic sentences describing an inherently delimited change of state, by denoting the resultant state of the individual undergoing the change. Terminative particles mark telic sentences describing an inherently delimited change of location, by denoting the end location of the moving individual."

1.4.2 Word-forming Function

The other additional function of verbal modifiers is the word-forming function. They can attach to verbs and create a new word. As a result, we either get a semantically compositional new meaning that also expresses direction, or we get a new, idiomatic meaning that is non-compositional. In the examples in Table 1, we can observe the second scenario, with new, idiomatic meanings.

VM	VERB	NEW VERB	MEANING
be 'into'	*jár* 'to walk'	*bejár*	to attend
el 'away'	*jár* 'to walk'	*eljár*	to proceed
meg 'PERF'	*jár* 'to walk'	*megjár*	he/she/it was bad for him/her; it's acceptable
össze 'together'	*jár* 'to walk'	összejár	to gather, to meet habitually

Table 1. The word-forming functions of verbal modifiers.

1.5 The Compositionality of Verbal Modifiers

Applying the concept of the asemantic compositionality presented in the previous section, the compositionality of verbs modified by a verbal modifier can be studied as follows: if the behavior of the verb modified by the verbal modifier can be predicted based on the independent, unique properties of the verb (e.g. in terms of the case of their arguments), then the verb can be considered compositional. However, in case the typical properties of the verb and the verbal modifier cannot explain the presence or absence of a complement, it can be considered idiomatic. We cannot forget about the so called half-compositional construction in Hungarian, such as *fel* + *nő* 'grow up'. Despite the other, semantically non-compositional meaning of the verbal complex (*fel* + *nő* + allative case 'grow up to the task'), we stress that semantic and syntactic compositionality are different. When we talk about *fel* + *nő* 'grow up', we consider the VM+V complex syntactically compositional because the VM *fel* 'up' causes telicity.

As already discussed above, in this study we apply the so-called asemantic compositionality principle in (7). Therefore, we consider idiomacity as a change in the argument structure of the verb, not as a change in the meaning of the verbal complex, like the original Fregean compositionality principle. With this modified compositionality principle our aim is to characterize the behavior of verbal modifiers using purely syntactic methods.

One of the many functions of verbal modifiers is the earlier mentioned word-forming function. However, this way not only the meaning of the new verbal complex can be observed, but the argument structure of the verb may also change. According to Gyulai (2019), the verbal modifier can change the verb's argument structure in one or more of five ways. The appearance of a verbal modifier can:

- increase the frequency of a certain argument type,
- decrease the frequency of a certain argument type,
- cause the appearance of a new argument type,
- forbid the appearance of a certain argument type, or
- require the appearance of a certain argument type.

These changes can appear alone but also mixed, for example, the appearance of the verbal modifier increases the frequency of one argument type that was already in the argument structure and decreases the one that was frequent before the verbal modifier appeared.

In this study it was also shown that the different argument structure changing effects of verbal modifiers work as follows: there are changes which can be observed by a group of verbs in the corpus, but there are also unique changes that affect only one verb. In our theoretical framework, this contrast is exploited to distinguish compositional verbal complexes from non-compositional ones: systematic changes affecting larger groups of verbs are compositional, while the verbal complex is non-compositional if the change is unique to it.

In what follows, we demonstrate the method used to extract the argument structure properties of verbs and verbal modifier + verb constructions from the corpus in Chapter 2. In Chapter 3, we outline the details of the method by which we distinguish compositional and idiomatic verbal modifier + verb constructions from each other.

2. Argument Structure Representation with Vectors of Scalar Frequency Values

Our aim is to determine the asemantic compositionality or idiomacity of verbal modifier + verb structures based on the observations presented at the end of the previous chapter. To do this, we need to examine the changes in their argument structure. For this, we need to know the argument structure of verbs with and without a verbal modifier so that we can identify their argument structure changing effects. If the same effect can be observed with multiple verbs for a particular verbal modifier, we say that the verbal modifier is in a compositional relation with the verb. If it only occurs with one or a few verbs, we consider it idiomatic.

To characterize the argument structure-changing ability of the verbal modifier, it is not sufficient to specify the argument structure of the verb by listing its arguments, as the verbal modifier often changes only the frequency of occurrence of the arguments. However, determining the frequency of occurrence of individual arguments can only be achieved through quantitative methods, by examining corpus data. A potential difficulty for now is that determining which constituent is an argument of the verb and which is an adjunct requires manual annotation.

Sentence (9) lacks the subject (due to pro-drop), but it includes an inessive case complement (*Péter-ben*) and an adjunct in the same case form (*január-ban*). The position of the constituents cannot be used as

a diagnostic for distinguishing complements from adjuncts because Hungarian is a free word order language, and *Péterben még nem bíztam januárban* (with the same meaning) is also a grammatical sentence.

(9) Január-ban még nem bíztam Péter-ben.
January-INE yet not trusted.1SG Peter-INE
'In January, I didn't trust Péter yet.'

Following the methodology described detailed (Szécsényi 2019) or this volume (Szécsényi and Szécsényi 2024), we use an argument structure representation in our analysis where we do not distinguish between arguments and adjuncts, but treat them uniformly: we refer to both arguments and adjuncts collectively as *supplements*. Furthermore, we characterize the supplement types of the verb based on the frequency values determined from corpus data: we specify how frequently each type of supplement appears in the corpus in the same clause as the verb. For each verb, we specify the argument structure with the frequency values associated with all possible supplement types: if one type of supplement never appears with the verb in a clause, its frequency is 0.0; if another appears in 80% of cases, its frequency is 0.8, and so on. Thus, we characterize each verb with the same number of probability values, that is, a probability vector.

We represented the constituents found in the clauses with one of their characteristic morpho-syntactic features. For the constituents whose morphological case was specified (nouns, adjectives, pronouns, etc.), we used their case form (one of 21 different types). For constituents with a verbal head, we differentiated between *that*-clauses, indicative, imperative, and conditional clauses, and infinitival clauses (5 types). Furthermore, we utilized the types of adpositions, adverbials, and verbal modifiers, which sums up to a total of 29 different types, therefore, the argument structure vectors of verbs are 29-dimensional.

From the manually annotated Szeged Dependency Treebank reference corpus (Vincze et al. 2010) containing 1,200,000 tokens, we determined the argument structure vectors of all verbs (13,556), a part of which is shown in Figure 1. Verbs were considered together with their verbal modifiers, as the verbal modifier can influence the argument structure of the verb. (Gyulai 2019; 2021).

A	B	C	D	E	F	G	N	O	P	Q	R	S	T	U	V	W
pvform	lemma	pvv	moo	cau	pot	n	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL
0	korlátoz	0+korlátoz	ind	-	-	41	0,68	0,93	0,00	0,17	0,07	0,15	0,02	0,12	0,00	0,00
0	terhel	0+terhel	ind	-	-	41	0,88	0,88	0,00	0,17	0,07	0,07	0,10	0,00	0,00	0,02
0	távozik	0+távozik	ind	-	-	41	0,59	0,00	0,00	0,10	0,27	0,05	0,17	0,15	0,00	0,29
0	ugrik	0+ugrik	ind	-	-	41	0,63	0,10	0,05	0,05	0,07	0,39	0,10	0,05	0,17	0,10
0	vágyik	0+vágyik	ind	-	-	41	0,27	0,00	0,00	0,02	0,02	0,85	0,00	0,02	0,05	0,00
be	ül	be+ül	ind	-	-	41	0,32	0,00	0,00	0,02	0,02	0,02	0,10	0,05	0,88	0,00
el	felejt	el+felejt	inf	-	-	41	0,00	0,22	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00
el	jár	el+jár	ind	-	-	41	0,63	0,00	0,00	0,39	0,12	0,07	0,07	0,27	0,00	0,02
haza	megy	haza+megy	inf	-	-	41	0,02	0,00	0,02	0,00	0,00	0,05	0,00	0,00	0,00	0,00
meg	indul	meg+indul	ind	-	-	41	0,85	0,00	0,02	0,12	0,07	0,12	0,15	0,05	0,00	0,07
0	alakít	0+alakít	ind	-	-	40	0,60	0,93	0,00	0,18	0,28	0,08	0,08	0,10	0,00	0,00
0	bízik	0+bízik	ind	-	-	40	0,45	0,00	0,00	0,88	0,00	0,00	0,00	0,08	0,05	0,00
0	ellenőriz	0+ellenőriz	ind	-	-	40	0,55	0,75	0,03	0,18	0,08	0,00	0,05	0,08	0,00	0,00
0	emlékezik	0+emlékez	ind	-	-	40	0,38	0,00	0,00	0,03	0,00	0,95	0,00	0,15	0,00	0,00
0	növekedik	0+növeked	ind	-	-	40	0,88	0,03	0,00	0,30	0,18	0,15	0,33	0,18	0,00	0,05
0	válik	0+válik	ind	-	+	40	0,80	0,00	0,03	0,20	0,05	0,08	0,18	0,10	0,00	0,00
be	állít	be+állít	inf	-	-	40	0,00	0,50	0,00	0,05	0,05	0,03	0,03	0,03	0,00	0,00
ki	lép	ki+lép	ind	-	-	40	0,35	0,00	0,00	0,08	0,28	0,15	0,08	0,08	0,03	0,00
meg	fogalmaz	meg+fogal	ind	-	-	40	0,83	0,85	0,00	0,18	0,05	0,05	0,05	0,13	0,00	0,08
meg	követel	meg+köve	ind	-	-	40	0,75	0,70	0,00	0,10	0,05	0,03	0,00	0,20	0,00	0,00

Figure 1. A partial view of the argument structure vectors for verbs.

Columns A–F display the characteristics of the verbs: Verbs were considered separately for different morphosyntactic categories (conditional, imperative, indicative, and infinitive) and various derived forms. Column G shows the frequency of the verb's occurrence, and the following 29 columns present the probabilities of different supplement types occurring alongside that verb (only 10 are visible in the figure). For example, the verb *bízik* ('trust') without a particle occurred 40 times in the corpus, with a 0.45 probability of having a nominative supplement, 0.0 probability of an accusative or dative supplement, and 0.88 probability of an inessive supplement, and so on. (Apart from the nominative, accusative, and dative supplements, the supplement types are indicated by the actual case markers.) For easier interpretation, probability values have been color-coded: the darker the red in a cell, the higher the probability of that supplement type.

Earlier studies have also examined the semantic change of words from a historical perspective using word vector representations (Hamilton, Leskovec, and Jurafsky 2016; Giulianelli, Kutuzov, and Pivovarova 2022). However, in those works, the vector representations assigned to words were based on word embeddings (Mikolov et al. 2013), which were primarily used to describe word meanings directly. The vector representation we use, on the other hand, is purely morphological and syntactic in nature.

3. Identifying the Effect of Verbal Modifiers on Argument Structure

The research we present in this study is about the automatic determination of the argument structure changing effects of verbal modifiers based on corpus data. The initial hypothesis is that there is a syntactic relationship between the verbs with and without verbal modifiers: the verbal modifier causes the same change in the argument structure of many verbs, this is what we call asemantic compositionality. The theoretical framework we are using is the earlier mentioned argument structure representation with scalar vectors (Szécsényi 2019). Our hypothesis is that we can observe the changes caused by the verbal modifier in the appearance of different supplement types if we compare the argument structures (vectors) of a certain verb with and without a verbal modifier.

3.1 The Corpus

The corpus used for this research is the Hungarian National Corpus 2 (Oravecz, Váradi, and Sass 2014) and we examine the verbal modifiers: *el* 'away' and *ki* 'out'. We studied the 100 verbs the most frequently associated with these verbal modifiers based on the Hungarian National Corpus 2.

We had to exclude auxiliaries and verbs that never occurred, or did so only very infrequently, without a verbal modifier, therefore we have 93 verbs with *el* 'away' and 87 with *ki* 'out'.

3.2 Data Processing

The data processing was carried out as follows: The data was retrieved using CQL from the Hungarian National Corpus 2 (Oravecz, Váradi, and Sass 2014). We searched for sentences containing verbal complexes written together, verbal modifiers separated from the verb by a maximum of +/- 5 tokens, and verbs without verbal modifiers. For each verbal modifier, a random sample of up to 10,000 sentences was retrieved. The sentences were saved in .txt files, with a subsequent analysis by the Hungarian NLP tool 'magyarlanc' (Zsibrita, Vincze, and Farkas 2013). This NLP tool performs a dependency analysis of the sentences. This was followed by sentence segmentation into clauses and maximal XPs and the aggregation of argument structure vectors for each verb.

3.3 Method

The output of data processing is 3 tables. The first table contains all the verbs under investigation, that is, the ones connected to *el* 'away' and the ones connected to *ki* 'out', this time without verbal modifiers.

pvform	lemma	pvv	mood	cau	pot	n	HKM	inf	nom	acc	dat	BAN
0	ad	0+ad	ind	-	+	6604	0,12	0,00	0,51	0,89	0,26	0,09
0	ad	0+ad	ind	-	-	6832	0,11	0,02	0,51	0,85	0,22	0,09
0	akadályoz	0+akadály	ind	-	-	3846	0,13	0,00	0,58	0,77	0,01	0,14
0	alakít	0+alakít	ind	-	+	1632	0,02	0,00	0,53	0,88	0,00	0,16
0	alakít	0+alakít	ind	-	-	7627	0,03	0,00	0,64	0,83	0,00	0,18
0	alakul	0+alakul	ind	-	+	1869	0,04	0,00	0,85	0,01	0,00	0,20
0	alakul	0+alakul	ind	-	-	4928	0,02	0,00	0,86	0,00	0,01	0,15
0	állapít	0+állapít	ind	-	-	18	0,06	0,00	0,67	0,50	0,06	0,28
0	állít	0+állít	ind	-	-	7785	0,34	0,01	0,61	0,22	0,04	0,08

Figure 2. Result of data processing. Verbs without verbal modifiers.

In Figure 2 ‘pvform’ zero means that there is no verbal modifier attached to the word in the sentences. The columns contain the frequency of different supplement types in the corpus. The rows, as mentioned earlier, define the argument structure vector of the particular verbs. The highlighted row above means that with the verb *ad* ‘to give’ without a verbal modifier we have 6,832 sentences, and the unified argument structure vector is as follows: the frequency of the nominative supplement type beside this verb in the analyzed sentences is 0.51, the frequency of the accusative supplement type is 0.85, etc.

pvform	lemma	pvv	mood	cau	pot	n	HKM	inf	nom	acc	dat	BAN
el	ad	el+ad	ind	-	-	5973	0,03	0,00	0,34	0,73	0,19	0,17
el	áll	el+áll	ind	-	-	5119	0,06	0,01	0,61	0,14	0,02	0,07
el	árul	el+árul	ind	-	-	6350	0,38	0,00	0,53	0,29	0,06	0,06
el	bírál	el+bírál	ind	-	-	1927	0,04	0,00	0,43	0,74	0,01	0,13
el	bír	el+bír	ind	-	-	3035	0,06	0,01	0,64	0,54	0,00	0,03
el	bocsát	el+bocsát	ind	-	-	5960	0,03	0,00	0,24	0,66	0,00	0,13
el	dől	el+dől	ind	-	-	5991	0,31	0,00	0,49	0,01	0,00	0,21
el	dönt	el+dönt	ind	-	+	3990	0,51	0,01	0,54	0,18	0,00	0,11

Figure 3. Result of data processing. Verbs with the verbal modifier *el* ‘away’.

In Figure 3 we see the same information when the verb appears with the verbal modifier *el* ‘away’. The question is how we can identify the effect of verbal modifiers on argument structure based on these tables. How is it possible to identify the changes caused by the presence of verbal modifiers in the argument structure compared to occurrences without verbal modifiers? We claim that by calculating the difference between the argument structure vectors, we can determine how much the frequency of a certain supplement type has increased or decreased with the appearance of the verbal modifier compared to the argument structure of verbs without a verbal modifier.

3.4 Systematic and Unsystematic Changes

We assume that there are systematic changes in the argument structure which occur in many verbs only when a certain verbal modifier appears. Some unsystematic, unique changes affect only one verb. The systematic and unsystematic changes can be used as a diagnostic based on the notion of asemantic compositionality introduced in 1.3. A [verbal modifier + verb] complex expression is asemantically compositional if the change effected by the verbal modifier is systematic. But it is asemantically non-compositional if the change is unique, unsystematic. It is also important to remark that verbal modifiers have various compositional effects, with their function and behavior differing depending on the verbs they are connected to (Gyulai 2021).

3.5 DBSCAN Algorithm

To highlight unique changes, the non-compositional verbal complexes, we used DBSCAN. DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density. It groups ‘densely grouped’ data points into a single

cluster. Our hypothesis was that if we group our verbs with DBSCAN, the verbs that are clearly far from a certain group will be the ones in which the change caused by the verbal modifier is unique. This way these can be identified, leaving the rest of the verbs where the change effected by the verbal modifier is systematic. Figure 4 illustrates how DBSCAN works.

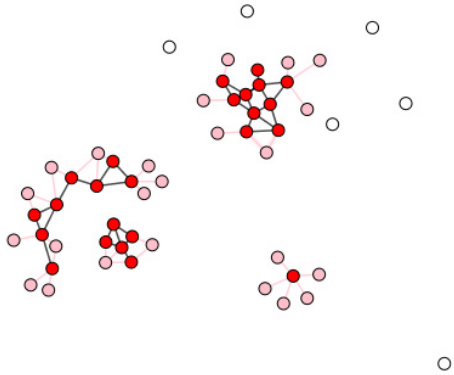

Figure 4. Example of the result of DBSCAN algorithm.

This algorithm finds those points that have enough neighbors within a range, the red points in Figure 4. These are the central points of the groups. Those points that have at least one central neighbor are the pink ones. These belong to the same group as their central neighbors. The rest of the points are considered "outliers", the white ones. These outliers can be identified as non-compositional verbs. As outliers we can list e.g. *tölt* 'fill, load'– *el*+*tölt* 'spend time', árul 'sell' – *el*+árul 'betray', *tűnik* 'it seems' – *ki*+*tűnik* 'stands out', *merül* 'sink, dive' – *ki*+*merül* 'be exhausted'. These verbal complexes are clearly both syntactically and semantically non-compositional because the change in their argument structure caused by the VM is unique and their meaning becomes idiomatic. Once we exclude the outliers, we can classify the rest of our verbs, looking for similarities between them. In the next chapter, we introduce the clustering method we used for this process.

3.6 K-means Clustering

After identifying all the verbal complexes modified by a verbal modifier in which the argument structure changing effect of the verbal modifier is unique, we used the K-means clustering algorithm for the rest of the verbal complexes. The K-means cluster analysis groups multi-dimensional vectors into K clusters in such a way that each element is assigned to the cluster whose center (mean) is the closest to it, based on Euclidean distance. This way verbs that exhibit similar changes when the verbal modifier is added could automatically be grouped together.

The input for the program was provided by the table that contains the difference of the verb's scalar vectors. In the case of K-means clustering, we need to specify how many clusters we want to create. In this experiment, the number of clusters ranged from 2 to 20: since we didn't know how many different argument structure changing effects the phrasal particle could have, we needed to try various possibilities. The output of the program is a table that combines all the clusters (2 groups, 3 groups, 4 groups, and so on) attached to the input file. After transferring the data to an Excel file, we can manually analyze the clusters suggested by the clustering tool.

Our hypothesis regarding K-means clustering is that after excluding verbs where the VM caused a unique change, the remaining ones can be classified with the following output: clear classes with easily detectable changes and homogeneous classes without clear changes. In the next section, we will take a closer look at some outputs of the clustering.

3.7 Results

3.7.1 Asemantically Compositional

In this section, we analyze the results of the K-means clustering. Since the cluster analysis works as described above, we must decide which grouping is the most ideal. Based on intuition and the manual analysis of the data we decided to go with clustering the data into seven groups.

Some of the resulting tables can be seen in Figures 5–9. All the tables and the software used are available in the https://github.com/szecsenyi/Szecsenyi-Gyulai-2024 GitHub repository. In the tables red cells indicate an increase in the frequency of supplement type, and blue cells mark a decrease. The darker red the cell, the bigger the increase, with darker blue cells indicating a higher level of decrease.

According to our hypothesis, there are "clear" classes where the argument structure changing effect is obvious. First, we show two classes of the table of *ki* 'out'. Verb group 0 shown in Figure 5, has 4 elements, and for all the four of them we can see an increase in the column BÓL 'from'. It is the expected, compositional behavior of this verbal modifier.

pvfor	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	ADP	ADV	FRON	IN	TO	7
ki	lóg	ki+lóg	0,02	0,00	0,00	0,03	-0,01	-0,19	-0,31	0,00	-0,01	-0,01	-0,04	-0,04	0,00	0,53	0,01	0,01	0,54	-0,52	-0,05	0
ki	marad	ki+marad	0,00	0,00	0,13	0,00	-0,01	-0,23	-0,14	-0,01	-0,03	0,09	0,00	0,01	0,00	0,51	-0,15	-0,16	0,52	-0,44	-0,12	0
ki	marad	ki+marad	-0,01	-0,01	-0,06	0,00	-0,02	-0,14	-0,06	-0,01	-0,02	0,00	0,00	0,01	0,00	0,49	-0,05	-0,13	0,49	-0,28	-0,06	0
ki	olvas	ki+olvas	0,08	0,00	0,05	-0,04	0,00	-0,08	-0,03	0,00	-0,01	-0,03	0,00	-0,02	0,00	0,49	0,02	-0,12	0,47	-0,10	0,00	0

Figure 5. K-means clustering, *ki* 'out' group 0.

Group 1 in Figure 6 has 6 elements, and for all 6 of them, we can see an increase in the accusative form. The explanation of this effect is that in case the verbal modifier appears with these verbs, the result of the action they express also appears, in the form of an accusative. This is due to the telic function of the verbal modifier *ki* 'out'. Of course, this does not mean that the verbal modifier serves a telic function in all the verbal modifier + verb constructions listed here. Among them, there may be semantically idiomatic constructions that share the argument structure-changing capability with telic verbal modifiers.

pvfor	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	ADP	ADV	FRON	IN	TO	7
ki	állít	ki+állít	-0,31	-0,01	-0,22	0,52	-0,01	0,19	0,14	0,00	0,02	0,00	-0,04	0,04	0,00	0,02	0,05	0,08	0,08	0,33	-0,08	1
ki	gondol	ki+gondol	0,34	0,01	0,12	0,32	0,00	0,04	0,04	0,04	0,02	0,02	0,00	0,01	0,00	0,00	0,05	0,13	0,01	0,09	0,02	1
ki	hallgat	ki+hallgat	-0,03	0,00	0,02	0,32	0,00	0,14	0,14	-0,13	-0,01	-0,06	0,00	-0,04	0,00	-0,01	0,16	-0,04	-0,05	0,27	-0,11	1
ki	kér	ki+kér	-0,13	-0,01	0,10	0,18	0,00	0,11	0,01	-0,01	-0,01	0,01	-0,01	0,05	0,00	0,00	0,05	0,08	0,01	0,15	-0,01	1
ki	mond	ki+mond	0,04	0,00	0,10	0,26	0,00	0,09	0,03	0,00	0,01	0,05	0,00	-0,01	0,00	0,03	-0,05	-0,01	0,12	0,00	1
ki	sorsol	ki+sorsol	0,01	0,00	0,05	0,27	0,01	0,00	-0,04	0,03	0,00	-0,02	-0,12	0,00	-0,03	0,06	0,31	-0,19	0,05	0,21	-0,14	1

Figure 6. K-means clustering, *ki* 'out' group 1.

We also assumed that there would be groups in which there would be no change, the homogenous group. With these verbs the appearance of the verbal modifier did not change much, this is also due to a compositional effect of the verbal modifier.

pvfor	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	ADP	ADV	FRON	IN	TO	7
ki	ad	ki+ad	-0,07	-0,01	0,02	-0,10	-0,13	0,09	0,07	-0,04	0,02	-0,01	-0,03	-0,01	-0,02	0,01	0,06	0,02	-0,01	0,15	-0,08	4
ki	alakít	ki+alakít	0,04	0,00	-0,31	0,07	0,01	0,03	0,10	0,03	0,06	0,00	0,00	0,02	0,00	0,03	0,08	0,02	0,05	0,18	0,03	4
ki	alakul	ki+alakul	0,05	0,00	0,05	0,00	0,01	0,07	0,07	0,02	0,01	-0,06	0,00	0,01	0,00	0,02	-0,25	-0,18	0,04	-0,05	0,02	4
ki	bontakozik	ki+bontak	0,00	-0,01	0,04	0,00	0,01	0,05	0,07	0,00	-0,01	-0,08	-0,02	0,03	0,00	0,04	0,17	-0,02	0,07	0,19	-0,05	4
ki	egészít	ki+egészít	0,00	0,00	-0,09	0,10	-0,03	0,07	-0,01	0,00	0,11	0,03	0,00	0,01	0,00	-0,05	0,02	0,06	-0,07	0,04	0,01	4
ki	épít	ki+épít	0,00	0,00	-0,02	0,17	-0,01	0,07	0,02	-0,17	0,06	0,02	-0,02	0,00	0,00	-0,02	0,04	0,00	-0,02	0,14	-0,18	4
ki	fejlődik	ki+fejlődik	0,00	0,00	0,10	0,00	0,01	0,04	0,04	0,01	-0,01	-0,20	-0,01	-0,01	0,00	0,18	0,07	-0,05	0,18	0,15	-0,01	4
ki	fejt	ki+fejt	0,19	0,00	0,14	0,04	0,01	0,12	0,03	0,02	-0,02	0,08	-0,07	0,02	-0,01	0,00	0,04	-0,06	0,02	0,17	-0,06	4
ki	fizet	ki+fizet	-0,01	0,00	-0,08	0,17	0,04	0,04	0,01	0,00	-0,02	-0,01	-0,01	0,00	0,00	0,03	0,01	-0,03	0,02	0,02	0,01	4
ki	hirdet	ki+hirdet	-0,03	0,00	-0,09	0,09	0,01	0,07	0,12	-0,03	-0,03	0,00	0,00	0,00	0,00	-0,01	0,01	0,02	-0,02	0,18	-0,04	4
ki	jelöl	ki+jelöl	0,02	0,00	-0,11	0,07	-0,03	0,01	0,05	0,10	-0,08	0,03	-0,03	0,00	0,00	0,00	0,08	-0,03	0,00	0,06	0,09	4
ki	jön	ki+jön	0,02	0,00	-0,12	0,01	0,00	0,01	0,04	0,02	0,04	0,03	-0,04	0,00	0,00	0,16	0,01	0,03	0,16	0,04	-0,01	4
ki	kíván	ki+kíván	-0,02	0,07	-0,17	-0,25	-0,02	0,16	-0,02	-0,04	-0,04	0,00	-0,01	-0,02	-0,02	-0,01	-0,03	-0,06	0,00	0,13	-0,08	4
ki	küld	ki+küld	0,04	0,01	-0,04	-0,08	0,01	0,03	0,03	-0,03	0,00	-0,02	-0,09	-0,02	0,00	0,04	0,00	0,02	0,02	0,05	-0,12	4
ki	lát	ki+lát	-0,22	-0,02	-0,09	-0,28	-0,05	-0,09	0,08	0,11	-0,01	0,00	0,04	0,00	0,00	0,18	0,08	-0,04	0,22	-0,01	0,17	4
ki	mutat	ki+mutat	0,06	0,00	-0,10	0,01	-0,02	0,17	0,05	-0,02	0,00	-0,01	-0,02	0,01	0,00	0,01	-0,02	-0,07	0,02	0,24	-0,04	4
ki	nevez	ki+nevez	-0,02	0,00	0,15	0,18	-0,61	0,07	0,06	0,20	0,03	0,00	0,03	0,00	0,01	0,00	0,07	-0,07	0,00	0,16	0,24	4
ki	oszt	ki+oszt	-0,06	-0,01	-0,11	0,12	0,07	0,11	0,12	-0,12	0,02	0,00	-0,02	-0,01	0,00	0,01	0,12	0,07	0,00	0,26	-0,12	4
ki	rajzolódik	ki+rajzoló	0,06	-0,01	-0,03	0,01	0,00	0,01	0,10	-0,22	-0,05	0,10	-0,03	0,01	0,00	0,09	0,04	0,01	0,07	0,17	-0,31	4
ki	szab	ki+szab	-0,06	0,00	-0,02	0,07	-0,35	0,06	0,07	-0,14	-0,06	-0,08	-0,01	0,00	-0,03	-0,01	0,07	0,04	0,04	0,13	-0,13	4
ki	szab	ki+szab	-0,09	0,00	-0,09	0,05	-0,34	0,13	0,09	0,21	0,01	-0,05	0,00	0,00	-0,06	0,00	0,16	0,00	0,00	0,21	0,17	4
ki	szállít	ki+szállít	0,01	0,00	-0,19	-0,07	-0,02	0,04	-0,01	-0,01	-0,04	-0,01	-0,04	-0,01	0,00	0,09	-0,04	0,02	0,06	0,01	-0,05	4
ki	szolgál	ki+szolgál	-0,09	0,00	-0,12	0,26	-0,03	0,03	0,04	-0,15	0,00	-0,01	0,00	0,00	-0,02	0,01	0,02	0,05	0,01	0,09	-0,15	4
ki	választ	ki+választ	0,02	-0,01	0,05	-0,05	-0,02	0,01	0,09	0,00	0,03	-0,03	0,00	0,00	0,00	0,06	0,07	-0,08	0,06	0,10	0,00	4
ki	vet	ki+vet	-0,10	0,00	0,05	0,06	-0,21	0,03	0,03	0,03	-0,03	0,00	-0,09	0,00	0,00	0,03	0,00	-0,01	0,04	0,07	-0,09	4

Figure 7. K-means clustering, *ki* 'out' group 4.

Moving to the tables with the verbal modifier *el* 'away', we see an increase in the column TÓL 'away from', which is due to the compositional behavior of *el* 'away'.

pvfori	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	TÓL	ADP	ADV	FRON	IN	TO	7	
el	áll	el+áll	-0,01	-0,02	-0,05	0,11	-0,01	-0,23	-0,13	-0,09	-0,07	-0,04	-0,01	0,00	-0,03	-0,09	0,41	-0,12	0,02	0,33	-0,44	-0,10	0	
el	marad	el+marad	-0,03	-0,01	0,08	0,00	-0,02	-0,06	-0,04	-0,03	0,10	0,00	0,00	0,00	0,00	0,00	0,31	0,01	-0,07	0,31	-0,17	-0,08	0	
el	vár	el+vár	0,31	0,00	-0,10	-0,29	0,00	-0,07	-0,04	-0,12	0,07	-0,02	0,01	-0,02	0,00	-0,03	0,20	-0,02	0,01	0,14	-0,11	-0,14	0	
el	vár	el+vár	0,27	0,00	-0,11	-0,25	0,00	-0,05	-0,07	-0,19	-0,05	-0,02	0,00	0,00	0,00	-0,01	0,19	-0,03	-0,04	0,18	-0,13	-0,21	0	
el	vesz	el+vesz	0,00	0,00	0,21	0,24	0,02	0,21	0,21	0,03	0,03	0,05	0,35	0,00	0,01	0,01	0,16	0,06	0,05	0,17	0,41	0,38	0	
el	vesz	el+vesz	0,01	0,00	-0,30	-0,21	0,00	-0,30	-0,24	0,00	-0,01	-0,02	-0,14	-0,01	0,00	0,03	0,18	-0,03	0,00	0,21	-0,49	-0,15	0	

Figure 8. K-means clustering, *el* 'away' group 0.

In group 1 all the elements show an increase in the accusative supplement type, which shows the telic function of *el* 'away'.

pvfori	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	TÓL	ADP	ADV	FRON	IN	TO	7	
el	dönt	el+dönt	0,25	0,00	-0,15	0,15	0,00	-0,08	-0,07	0,00	-0,01	-0,04	-0,02	-0,37	0,00	0,00	0,00	-0,09	-0,12	-0,37	-0,17	-0,01	1	
el	dönt	el+dönt	0,23	0,00	-0,07	0,27	0,00	-0,05	-0,08	0,00	0,01	-0,04	-0,01	-0,37	0,00	0,00	0,00	-0,07	-0,18	-0,37	-0,17	-0,01	1	
el	ér	el+ér	-0,01	0,00	-0,25	0,26	-0,02	0,01	0,00	0,02	0,06	0,00	-0,05	0,00	-0,01	0,00	0,00	0,04	-0,01	0,01	0,03	-0,05	1	
el	ír	el+ír	-0,09	0,00	-0,30	0,27	-0,01	-0,04	0,01	-0,02	-0,01	0,02	-0,01	-0,07	-0,01	0,00	0,00	0,00	0,01	-0,06	-0,02	-0,04	1	
el	követ	el+követ	0,07	0,00	-0,30	0,23	0,00	0,06	-0,02	0,01	-0,04	-0,04	0,00	-0,01	0,00	-0,01	-0,01	0,12	-0,05	-0,03	0,07	0,00	1	
el	mesél	el+mesél	0,15	0,00	-0,11	0,28	0,03	0,02	0,00	0,00	0,00	0,00	0,00	-0,24	0,00	0,01	0,00	0,01	0,01	-0,23	0,03	0,01	1	
el	olvas	el+olvas	-0,01	0,00	0,07	0,28	0,00	-0,16	0,07	0,00	0,00	-0,19	0,00	-0,33	0,00	0,00	0,00	-0,01	-0,08	-0,33	-0,19	0,01	1	
el	veszít	el+veszít	0,00	0,00	0,11	0,47	0,00	0,03	-0,03	0,02	0,02	-0,04	0,00	0,00	0,00	-0,25	0,00	0,02	-0,02	0,25	0,02	0,03	1	

Figure 9. K-means clustering, *el* 'away' group 1.

According to our hypothesis, we got three kinds of groups:

1. the primary function of the verbal modifier appears
2. clear argument structure changing effect of the verbal modifier – compositional behavior of the verbal modifier, additional functions such as telicity also identified
3. no clear change, homogenous groups – compositional behavior of the verbal modifier, not always compositional meaning.

3.7.2 Asemantically Non-Compositional

Apart from the results presented above, we would also like to report on the argument structure of non-compositional VM+V constructions selected by DBSCAN. Through some examples, we demonstrate how the asemantic non-compositionality manifests within the theoretical framework we employ.

pvfori	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	TÓL	NÁL	VÁ	ADP	ADV	FRON	IN	TO
ki	kap	ki+kap	-0,04	0,00	0,11	-0,84	0,00	0,25	0,20	0,26	0,06	0,00	-0,02	-0,01	-0,02	-0,01	0,40	-0,01	0,00	0,00	-0,01	0,38	0,35	0,20

Figure 10. Argument structure of *kikap* 'is scolded'.

As previously detailed, the examples presented above are compositionally asemantic, meaning that the appearance of the VM leads to similar changes in the verbal argument structure for multiple verbs. In contrast, in asemantic non-compositional VM+V constructions, such as *ki+kap* 'is scolded' in Figure 10, the change caused by the VM is unpredictable and unique. Examining the example of *ki+kap* 'is scolded', we can see that prior to the appearance of the VM, the accusative supplement type had a high frequency of occurrence. However, with the appearance of the VM *ki* 'out', this supplement type disappears from the argument structure, and most commonly, the supplement type TÓL 'away from stg' appears, referring to the person causing the scolding.

pvfori	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	TÓL	NÁL	VÁ	ADP	ADV	FRON	IN	TO
ki	köt	ki+köt	0,13	0,00	-0,07	-0,60	0,00	0,30	0,15	-0,05	-0,14	-0,01	-0,02	-0,01	-0,31	0,00	0,00	0,10	0,00	0,05	0,17	-0,01	0,53	-0,39

Figure 11. Argument structure of *ki+köt* 'sb. ends up in a certain situation'.

The example in Figure 11 is also a combination with the VM *ki* 'out'. The basic meaning of the verb *köt* is 'to tie sth.'. The accusative supplement type decreases significantly with the appearance of the VM, and we observe the emergence of a new supplement type, BAN 'in'. Nevertheless, the appearance of

this supplement type brings along a new meaning, as now the verb *ki+köt* carries the meaning of 'sb. ends up in a certain situation' or requires the appearance of a subordinative clause with the meaning 'insist on doing sth.'. Thus, the unique argument structure changing effect of the VM is not only non-compositional in an asemantic sense but also semantically.

pvfori	lemm	pvv	HKM	inf	nom	acc	dat	BAN	ON	RA	VAL	UL	BA	RÓL	HOZ	BÓL	TÓL	NÁL	ADP	ADV	FRON	IN	TO
el	választ	el+választ	0,02	-0,01	0,35	-0,26	-0,03	-0,02	-0,03	-0,06	0,01	-0,06	-0,01	0,00	-0,01	-0,02	0,48	0,00	-0,09	-0,12	0,45	-0,04	-0,07

Figure 12. Argument structure of *el+választ* 'to separate'.

In Fig. 12, the argument structure of the verb *el+választ* 'to separate' can be seen. The verb *választ* 'to choose' without the VM generally means 'to choose' and usually appears with an accusative supplement type in a sentence, as evidenced by the decrease in the occurrence frequency of the accusative supplement type. In contrast, the meaning of the verb *el+választ* means 'to separate', and the complement type TÓL 'away from sth.' comes to the forefront, expressing what or whom we are separating the subject of the sentence from.

4. Summary

In this study, we presented a new method for the identification of non-compositional [verbal modifier + verb] complexes in Hungarian. In the first section, we introduced a modified version of the Fregean compositionality principle, which we call asemantic compositionality. According to this principle, the properties of an expression are a function of the properties of its parts and of the way they are syntactically combined. In the second section, we discussed the closed class of verbal modifiers in Hungarian and their primary and additional functions. We discussed the word-forming function of verbal modifiers and the non-compositional verbal complexes through the example of *rúg* 'kick' combined with the verbal modifier *be* 'into'. Then we presented how to represent argument types with scalar vectors. Instead of the binary argument-adjunct distinction we proposed a scalar characterization of argumenthood: each expression that can occur with the verb is characterized by a probability value based on the frequency of their occurrence in the corpus. The argument structure representation is a list of scalar values, which is a scalar vector. To distinguish compositional and non-compositional verbal complexes in Hungarian we studied [verbal modifier + verb] complexes. Our aim was to identify these complexes with the help of automatic methods and computational linguistic tools. Our hypothesis was that if we compare the argument structures of a certain verb with and without a verbal modifier, we can observe the changes caused by the verbal modifier in the appearance of different types of supplements. The corpus we used is the Hungarian National Corpus 2 (Oravecz, Váradi, and Sass 2014). After analyzing the data, we used two clustering algorithms: DBSCAN and K-means clustering. DBSCAN was used to exclude the outliers, the idiomatic verbal complexes and the rest of the words were clustered by the K-means clustering algorithm, which shows the systematic changes the addition of a verbal modifier triggers.

In the final chapter, we not only discussed asemantically compositional constructs but also showed the argument structure of non-compositional VM+V constructs, demonstrating the changes caused by the appearance of verb particles in the verbal argument structure.

The vector-based representation of argument structure facilitates the differentiation of various (compositional) uses of verbal modifiers, as well as some non-compositional uses.

Works Cited

É. Kiss, Katalin. 2008. "The Function and the Syntax of the Verbal Particle". In *Event Structure And The Left Periphery: Studies on Hungarian*, edited by Katalin É. Kiss, 17–55. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-4755-8_2.

Gendler Szabó, Zoltán. 2000. "Compositionality as Supervenience". *Linguistics and Philosophy* 23 (5): 475–505. https://doi.org/10.1023/A:1005657817893.

Gendler Szabó, Zoltán.. 2012. “The Case for Compositionality.” In *The Oxford Handbook of Compositionality*, edited by Markus Werning, Wolfram Hinzen, and Edouard Machery, 64–80. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199541072.013.0003.

Gendler Szabó, Zoltán. 2017. “Compositionality.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Metaphysics Research Lab, Stanford University.

Giulianelli, Mario, Andrey Kutuzov, and Lidia Pivovarova. 2022. “Do Not Fire the Linguist: Grammatical Profiles Help Language Models Detect Semantic Change.” In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 54–67. Dublin: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.6.

Gyulai, Lívia. 2019. “Nem kompozicionális igekötős igék argumentumszerkezetének korpuszalapú vizsgálata.” In *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2019. XIII. Alkalmazott Nyelvészeti Doktoranduszkonferencia*, edited by Zsófia Ludányi and Tekla Etelka Gráczi, 44–58. Budapest: MTA Nyelvtudományi Intézet. https://doi.org/10.18135/Alknyelvdok.2019.13.4.

Gyulai, Lívia. 2021. “Az igekötők legjellemzőbb argumentumszerkezetváltoztató hatásainak korpuszalapú vizsgálata.” In *Alknyelvdok15. Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből*, edited by Tekla Etelka Gráczi and Zsófia Ludányi, 176–98. Budapest: Nyelvtudományi Kutatóközpont.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. Berlin: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1141.

Kardos, Éva. 2016. “Telicity Marking in Hungarian.” *Glossa: A Journal of General Linguistics* 1 (1). https://doi.org/10.5334/gjgl.52.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *arXiv:1310.4546 [Cs, Stat]*, October. http://arxiv.org/abs/1310.4546.

Oravecz, Csaba, Tamás Váradi, and Bálint Sass. 2014. “The Hungarian Gigaword Corpus.” In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 1719–23. Reykjavik: European Language Resources Association (ELRA).

Partee, Barbara Hall. 2004. *Compositionality in Formal Semantics: Selected Papers*. Exploration in Semantics 1. Blackwell Publishing.

Szécsényi, Krisztina, and Tibor Szécsényi. 2024. “Determining Argument Structure Variants by Numerical Optimization.” In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistic Colloquium 2023*, edited by Markéta Janebová, Michaela Čakányová, and Joseph Emonds. Olomouc: Palacký University.

Szécsényi Tibor. 2019. “Argumentumszerkezet-variánsok korpusz alapú meghatározása.” In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, edited by Gábor Berend, Gábor Gosztolya, and Veronika Vincze, 315–29. Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Vincze, Veronika, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. “Hungarian Dependency Treebank.” In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 1855–62. Valletta, Málta: European Language Resources Association.

Zsibrita, János, Veronika Vincze, and Richárd Farkas. 2013. “Magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian.” In *Proceedings of RANLP 2013*, 763–71.