Acta Linguistica Academica
DOI: 10.1556/2062.2022.00578

AKADÉMIAI KIADÓ

PrevDistro: An open-access dataset of Hungarian preverb constructions

ÁGNES KALIVODA*

Hungarian Research Centre for Linguistics, Hungary

Received: April 21, 2022 • Accepted: October 1, 2022

ABSTRACT

Hungarian has a prolific system of complex predicate formation combining a separable preverb and a verb. These combinations can enter a wide range of constructions, with the preverb preserving its separability to some extent, depending on the construction in question. The primary concern of this paper is to advance the investigation of these phenomena by presenting PrevDistro (Preverb Distributions), an open-access dataset containing more than 41.5 million corpus occurrences of 49 preverb construction types. The paper gives a detailed introduction to PrevDistro, including design considerations, methodology and the resulting dataset's main characteristics.

KEYWORDS

preverbs, constructions, corpus-driven, dataset, Hungarian grammar

1. INTRODUCTION

In recent years, advances in Natural Language Processing technology had an enormous impact on almost every field of linguistics. The number of studies opting for a large-scale quantitative approach in order to explore and analyze linguistic data is rapidly increasing, see Gries (2015), Brezina (2018) and Stefanowitsch (2020), just to mention some prime examples. Such an approach generally involves the use of computational methods, corpora and/or specialized lexical resources. Examples for the latter are inflectional datasets (see e.g. Beniamine et al. 2020) and derivational datasets (see e.g. Sánchez-Gutiérrez et al. 2018).

* Corresponding author. E-mail: kalivoda.agnes@nytud.hu

PrevDistro (Preverb Distributions)[1] is intended to enrich the range of these resources by providing high-quality, open-access data of 49 Hungarian preverb construction types. It was created by automatic data extraction based on the 1-billion-word Hungarian Gigaword Corpus (Oravecz et al. 2014). The dataset may serve as a starting point of subsequent linguistic research, including cross-linguistic studies. It is expected to reveal numerous trends in language use which would otherwise remain unnoticed or conjectural.

Section 2 introduces the phenomena which this dataset is focused on, as well as the core assumptions which had an effect on how PrevDistro was built. Section 3 presents the process of dataset creation. This is followed by a numeric overview of PrevDistro's main characteristics in Section 4. Finally, Section 5 summarizes the results and gives an outline of future work.

2. DESIGN CONSIDERATIONS

PrevDistro is primarily designed to account for three phenomena in Hungarian. These are preverbs – and preverb-like lexical items –, their position relative to the verb stem, and the construction types which may host these items.

2.1. Preverbs

The definition of preverbs is a notoriously problematic and debated topic in Hungarian linguistics. Issues arise from the fact that preverbs resemble other lexical items called bare nominal verb modifiers, both in terms of distribution and semantics. Kiefer (2007, 226–227) summarizes the similarities as follows. The syntactic behavior of the two groups is largely identical: (1) they are always placed before the finite verb in neutral sentences, (2) they may receive focal stress in this position, (3) if another component is focused in the sentence, they appear after the finite verb, (4) in answers given to yes/no questions, they can usually appear in themselves, their associate verb being ellipted. As for semantics, verb modifiers – including preverbs – can generally yield complex predicates with their associated verbs. They often change both the meaning and the valence of the verb.

Jakab (1976, 9–10) compares seven reference works, examining which lexical items are classified as preverbs in each one of these. He finds that they intersect in case of 28 items which is only 37.3% of the 75-item set he studies. These 28 lexical items are listed in Example (1). The glosses given here convey a "typical" sense of each preverb, but of course, preverbs forming specific complex predicates may obtain senses which are quite different from the ones presented here. The disagreement on the set of Hungarian preverbs did not change much since Jakab's study, see e.g. Komlósy (1992) and Kiefer (2007).

[1]PrevDistro was first introduced in my PhD dissertation (Kalivoda 2021) in Hungarian. Since then, it underwent some technical improvements and it was upgraded with important metadata: region, style and year of publication in most cases. The dataset was inspired by László Kálmán.

(1)

abba 'into'	*fenn* 'above'	*le* 'down'
agyon 'excessively'	*hátra* 'backward'	*meg* 'perfective'
alá '(to) under'	*helyre* 'to place'	*neki* 'to, against'
be 'into'	*hozzá* 'towards'	*oda* '(to) there'
bele 'into'	*ide* '(to) here'	*össze* 'together'
el 'away'	*keresztül* 'across'	*rá* 'onto'
fel 'up'	*ki* 'out'	*rajta* 'on'
félbe 'into half'	*körül* 'around'	*túl* 'over, beyond'
félre 'aside'	*közbe* 'in between'	*vissza* 'back'
felül 'above'		

Controversies are present in linguistic theory as well as practice. Lexical resources and corpora vary to a great extent in the words labeled as preverbs. The role of spelling is also worth mentioning here. For example, if certain verb modifiers and their associated verbs are written as one word in the Hungarian Gigaword Corpus – e.g. *békénhagy* 'leave sy alone' where *békén* lit. 'on peace' is a verb modifier –, the whole word will be labeled as UNKNOWN, while the resource Tádé (Kornai et al. 2016) presents these units as preverb-verb combinations.

In this study, preverbs are considered to be epiphenomenal. The sense of a 'preverb' category emerges of the similar distribution of a heterogeneous set of lexical items. If only a handful of them were accounted for in PrevDistro, this could considerably narrow down the possibilities of future research based on the dataset. It seemed to be a better option to take as wide cross-section of the potentially relevant lexical items as possible. This was achieved using PrevLex (Kalivoda 2019) as a starting point, a resource containing a total of 235 preverbs and preverb-like lexical items. This dataset contains all words labeled as preverbs which could be found in born-digital, Hungarian corpora and other lexical resources. Only those items were removed from it which are clearly parsing errors, e.g. *vízi* 'water-' which can never be a verb modifier but a first part of certain compounds.

It must be noted, however, that only the simplest morphological forms of these 235 items are included in the current version of PrevDistro. In what follows, a short enumeration of morphologically complex preverb types will be presented, as these are less known but quite interesting typologically.

Hungarian preverbs can be subjected to reduplication (2) which has an iterative function, expressing that the event or action happens repeatedly. Moreover, it typically entails that the event or action reoccurs at irregular time intervals. More on preverb reduplication can be found in Piñón (1991), Kiefer (1995) and Ackerman (2003).

(2)

***Bele~bele**-olvas-ott*	*a*	*vaskos*	*kötet-ek-be.*
into(PV)~into(PV)-read-3SG.PAST/DEF	the	massive	volume-PL-ILL

'He looked into the massive volumes from time to time.'

If a clause is imperative or progressive, the morpheme *-fele/felé* '-wards' may be suffixed to the preverb (Kerekes 2011). This phenomenon can primarily be observed in case of directional preverbs, but there is ample evidence showing that it can be attached to the preverb *meg*, a perfectivizer devoid of its original directional meaning 'to behind', see Example (3). It is usually associated with inverted clausal order which will be discussed in the next section.

(3) *Rohamosan* *tisztul* *a* *közélet* ***meg-fele***.
rapidly purify.3SG.PRES the public.life perfective-wards(PV)
'Public life is purifying rapidly.'

Certain preverbs, mainly the ones having adverbial origin, may host comparative suffixes. This is more typical for constructs with transparent meaning, but it can be attested in opaque preverb-verb combinations as well, see Example (4). To my knowledge, there is almost no research into this topic. The only source discussing the phenomenon in greater detail is Kálmán (2013).

(4) *ettől* *csak* ***ki-jjebb*** *és* ***ki-jjebb*** *ábrándul-tam.*
this.ABL just farther.out(PV) and farther.out(PV) go.off-1SG.PAST
'[…] this made me more and more disappointed.'

Finally, a peculiarity of some Hungarian preverbs and preverb-like items is that they can host inflectional markers, as shown in Example (5). The distribution of these forms is largely identical with the simple ones. They can even undergo reduplication, see Example (5). The inflectability of preverbs is a matter of lively debate, resulting in substantially different theoretical accounts – see É. Kiss (1998), Kálmán & Trón (2000), Surányi (2009), Rákosi & Laczkó (2011), Hegedűs (2013), Rákosi (2014) and Ackerman et al. (2022) among others.

(5) a. *Ver-jünk* *tábor-t,* ***rá-nk-fér*** *a* *pihenés!*
beat-1PL.SBJV/INDEF camp-ACC upon-1PL(PV)-get.3SG.PRES/INDEF the rest
'Let's set up camp, we need a rest.'

b. *A* *szél* ***nek-em~nek-em*** *csap,*
the wind to-1SG~to-1SG(PV) hit.3SG.PRES/INDEF
'The wind dashes against me from time to time […]'

In the future, the dataset will be extended with the corpus occurrences of complex forms listed above. However, in this initial phase, even the accurate detection of simple forms poses a great challenge which is mainly due to their rather free distribution relative to the verb stem. This will be addressed in the next subsection.

2.2. Clausal orders

A notable characteristic of Hungarian preverbs is their independent syntactic behavior in certain construction types. They may appear in three clausal orders relative to the verb stem.[2] Example (6a) presents direct order where the preverb – in this case *be* 'in(to)' – is prefixed to the verb stem. In (6b) the preverb precedes the verb, but they are separated by other elements, yielding discontinuous order. The third option shown by (6c) is inverted order, with the preverb following the verb, not necessarily immediately. The clausal distributions of preverbs are

[2]The terms used for clausal orders follow the terminology presented in Ackerman & LeSourd (1997). The preverb and its associated verb are marked with boldface.

discussed in a wide range of literature, see J. Soltész (1959), Ackerman & LeSourd (1997), É. Kiss (2021), Kalivoda (2021) among others.

(6) a. ***be-hoz-om***
in(PV)-bring-1SG.PRES/DEF
'I bring it in'

b. ***be*** *sem* ***hoz-om***
in(PV) not.even bring-1SG.PRES/DEF
'I do not even bring it in'

c. *nem* ***hoz-om*** ***be***
not bring-1SG.PRES/DEF in(PV)
'I do not bring it in'

PrevDistro is designed to account for clausal order information in great detail. Beside distinguishing the three ordering possibilities mentioned above, the aim was to determine the distance between the preverb and the verb stem in terms of tokens. In order to achieve this, each clause is mapped into a number line where the verb stem is defined as the origin, and the position of its associated preverb is counted relative to this. Prefixed preverbs – yielding direct order – get zero position, while preverbs in discontinuous order are found in an interval less than zero, and inverted order ones in an interval greater than zero, as shown in Figure 1.

2.3. Construction types

The aim of the project was to account for a range of preverb constructions that is as wide as technically possible, including finite and infinite verbs as well as derivates where a preverb and its associated verb can be detected. The exact amount of these construction types was not known in advance. It was outlined according to corpus data, in this case, the Hungarian Gigaword Corpus (see Section 3.1). In general, the names of construction types are created on the basis of

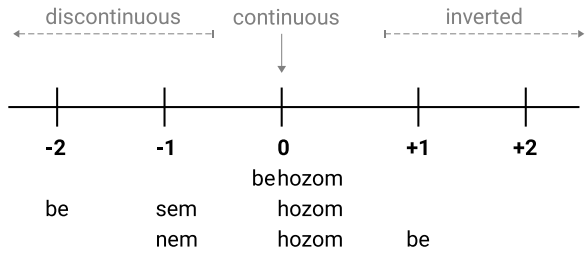

Figure 1. Calculating the preverb's position relative to the verb stem. Preverb-verb combinations of Example (6) are used as an illustration of the method

the last derivational suffix[3] identifiable after the verb stem. Thus, the construct in Example (7a) is labeled as -AndÓ and the one in Example (7b) as -Ú.[4]

(7) a. ***fel*** *nem* ***használ****-andó*
up(PV) not use-PTCP
'(something that) has not to be used up'

b. ***el-nevez****-és-ű*
away(PV)-name-NMLZ-ADJZ
'being named as'

There are four cases when this naming convention is abandoned. The first one is the case of finite verbs (*FIN*) where category preserving derivations can be detected but are not labeled explicitly. The second one occurs at superlative (8a) and absolute superlative (8b) constructions, both labeled as *leg-…-bb*. This is a case of circumfixing in Hungarian, having the peculiarity that the prefixed part – *leg* for superlative and *leges* for absolute superlative – can be repeated multiple times within a single construct. This is illustrated in Example (8b).

(8) a. *leg-****át-látsz****-ó-bb*
most-through(PV)-be.visible-PTCP-CMPR
'the most transparent'

b. *leges-leges-leg-****meg-old****-ott-abb*
extremely-extremely-most-perfective(PV)-solve-PTCP-CMPR
'(something that) is absolutely solved'

The third one labeled as *climbing* is an umbrella term denoting a group of constructions having the following characteristics: (1) The preverb and its associated verb have discontinuous order, and – seemingly, at least – they occur in separate clauses. (2) The preverb precedes a finite modal which is typically the verb *kell* 'must'. (3) If the verb forming a lexical unit with the preverb is finite, it is in subjunctive form. If it is a non-finite verbal complement, the subjunctive form is taken by the modal or the copula. (4) The presence of the complementizer *hogy* 'that' is always optional. Example (9) presents the simplest forms of this construction type. Searching for this construction in a corpus is far from trivial: the complementizer *hogy* is often abbreviated as *h* and the presence of a comma before it is also varying. More details on climbing can be found in Posgay (2002), Dóra (2006), É. Kiss (2009) and Kalivoda (2021, 75–78), among others.

[3]A more thorough insight into the Hungarian linguistic literature reveals controversial classification of several suffixes, especially when it comes to suffixes forming participles. Discussing the theoretical debates on the status of these suffixes is beyond the scope of this study, as they do not have direct significance for the dataset.

[4]Most suffixes in Hungarian show alternations due to regular vowel harmony. My notations for the alternating suffix vowels follow Kiefer (2000)'s conventions: *A* → *a* or *e*, *Á* → *á* or *é*, *O* → *o* or *ö*, *Ó* → *ó* or *ő*, *U* → *u* or *ü*, *Ú* → *ú* or *ű*, *V* → any vowel.

(9) a. ***össze*** *kell,* *hogy* ***áll-jon***
together(PV) must-3SG.PRES that stand-3SG.SBJV
'it must hang together'

b. ***össze*** *kell* ***áll-jon***
together(PV) must-3SG.PRES stand-3SG.SBJV
'it must hang together'

Finally, a group of constructions is labeled as *topicalization*. In this case, a preverb-verb combination is topicalized as an infinitive or as an adverbial participle, and it appears repeatedly as a finite verb, see Example (10a). If there is an auxiliary-like item in the clause – a modal or a copula – the associated verb is ellipted in the comment, as shown in Example (10b). However, this is a tendency rather than a rule. More on this can be found in Kalivoda (2021, 79–81).

(10) a. ***Ki-tel-ni*** ***ki-telik*** *tőle,* *az-t* *már* *ugye*
out(PV)-take-INF out(PV)-take.3SG.PRES he.ABL that-ACC already certainly
tud-om.
know-1SG.PRES/DEF
'As for being capable of it, he is capable of it, this is something I know for sure.'

b. ***Be-tilt-ani*** ***be*** *le-het,* *de* *at-tól* *még* *létezik.*
in(PV)-prohibit-INF in(PV) be-3SG.MOD/INDEF but that-ABL still exist-3SG.PRES
'As for banning, it can be banned, nevertheless it exists.'

Distinctions based on word class are intentionally avoided in the dataset. The first reason for this is practical: there are several cases when the automatic annotation of corpora can not really be trusted in this regard. For example, the preverb construct *fel-hajt-ó* can be tagged as noun (in the sense 'driveway'), adjective ('lifting') or adjectival participle ('sy driving st up'). The Hungarian Gigaword Corpus exhibits each of these possibilities, but the choice between them seems to be contingent. In PrevDistro, *fel-hajt-ó* is simply labeled as -Ó. The second reason for this decision is a conceptual one: grammatical word classes are not necessarily primitive concepts, but categories emerging from patterns of language use. A large body of literature can be consulted on this topic, e.g. Langacker (1991, 59–100), Croft (2001, 63–107), Kálmán (2016), Diessel (2019, 142–171).

3. DATASET CREATION

The section starts with an introduction to the Hungarian Gigaword Corpus which served as the text source of the dataset. Changes that were needed in order to enhance the quality of the corpus will also be discussed here, followed by a concise description of the dataset building process.

3.1. Text source

The Hungarian Gigaword Corpus – hereafter called HGC – is a 1-billion-word, automatically annotated general corpus. It is designed to represent a wide cross-section of Hungarian from

Table 1. The size of HGC 2.0.4 before and after filtering poems, sentences without any meaningful analysis and duplicates. Punctuation marks are counted as separate tokens

corpus	tokens	percent
original HGC	**1,348,000,000**	**100.00**
poems	5,661,000	0.42
UNKNOWN/SKIP	26,825,200	1.99
duplicates	271,217,600	20.12
filtered HGC	**1,044,296,200**	**77.47**

the later part of the 20th century and the start of the 21st century. The first dimension of the corpus is style. Within this dimension, each text is assigned to one of the following six categories: press, literature, popular science, official, personal or spoken. The other dimension is region, consisting of five categories: Hungary, Slovakia, Subcarpathia, Transylvania and Vojvodina. It must be emphasized that this is not a dialectal division, simply a geographical one. The published corpus does not contain the years of publication as metadata, but it is possible to extract this information for most texts, using the XML source files. In this study, version 2.0.4 of the corpus was used.[5] The corpus is tokenized, lemmatized and morphosyntactically tagged, but no higher level annotations – e.g. dependency relations – are available in it.

Three types of text were eliminated in order to make the corpus best suited for the purpose of this research. First, poems had to be left out since many of them exhibit clausal orders that are considerably – and deliberately – deviant from the naturally emerging patterns of language use. Second, it was necessary to filter non-Hungarian sentences as well as sentences where all diacritics were missing. All of these could have distorted the data if not removed beforehand. A simple heuristic was used for this task: a sentence was deleted if at least 80% of its tokens were annotated as UNKNOWN or SKIP, meaning they missed a proper morphosyntactic analysis in Hungarian. Finally, duplicates had to be removed. Deduplication was applied only to sentences longer than 8 tokens, assuming that short sentences – e.g. greetings – do in fact occur scores of times in exactly the same form. Although the method applied was cautious – favoring precision over recall –, the proportion of duplicates proved to be extremely high (20.12%). In the personal subcorpus, there were lengthy paragraphs being repeated more than a hundred times. The results of corpus cleaning are summarized in Table 1.

3.2. Workflow

The first step of data collection was to obtain all sentences from HGC which are likely to contain a preverb-verb combination in form of a finite verb, infinitive or verbal derivate. The original morphosyntactic annotation of the corpus was not used eventually, as it did not seem to be informative enough in several cases. Each word form was reanalyzed with the emMorph tool

[5]The newest version is HGC 2.0.5. There is no difference in the composition of the two versions, only small structural corrections were made. Further information can be found at: http://clara.nytud.hu/mnsz2-dev/hirek.html.

(Novák et al. 2016, 2017). This is an open-source, automatic morphological analyzer splitting word forms into a series of morphemes and morphological codes. Example (11) shows the difference between the original annotation of HGC and the output of emMorph. The original one uses a non-standard, Hungarian abbreviation system which in this case can be interpreted as 'adjective in superlative form with nominative singular ending', whereas emMorph's output reveals that this adjective is derived from a preverb-verb combination.

(11) *leg-**el-fogad**-ható-bb*
most-away(PV)-accept-able-CMPR
'the most acceptable'
→ HGC: FF.MN._FOK.NOM
→ emMorph: leg[/Supl] =leg+el[/Prev] =el+fogad[/V] =fogad+ható[_ModPtcp/Adj] =ható+bb[_Comp/Adj] =bb+[Nom] =

It was assumed that in the vast majority of preverb constructions, the preverb and its associated verb appear in the same clause. There are some exceptions to this – for example, the "climbing" construction described in Section 2.3 –, these were retrieved subsequently with specific queries. The potentially relevant sentences were assigned to subcorpora, each of them representing a possible construction type. In this way, a given sentence could be part of several subcorpora, and it could appear more than once in a single subcorpus, as shown in Example (12). Only a much later phase of data processing revealed which option is right – here it is the third one –, but at the current stage, it was necessary to assume that all combinations might be possible.

(12)

el	*kell*	*legyen*	***csesz**-ve*
away(PV)	must.3SG.PRES	be.3SG.SBJV	screw.up-PTCP

'it must be screwed up'
→ subcorpus of finite verbs: *el* + *kell* 'away(PV) + must.3SG.PRES'
→ subcorpus of finite verbs: *el* + *legyen* 'away(PV) + be.3SG.SBJV'
→ subcorpus of adverbial participles: *el* + *cseszve* 'away(PV) + screw.up-PTCP'

PrevLex (Kalivoda 2019) – a gold-standard resource containing 53,535 preverb-verb lemmas – was used to filter the potential preverb-verb combinations retrieved from the corpus. This led to the loss of some valid hits, simply because they are not listed in PrevLex, but the filtering of false, nonsense combinations was way more significant.

For each sentence, the preverb's position was calculated automatically, as described in Section 2.2. This was followed by the most time-consuming task: each preverb position in each subcorpus had to be studied and characterized by the formation of various filtering rules. These rules are typically regular expressions matching a series of morphosyntactic labels. For example, a frequently applied filtering rule removes hits if they contain the sequence 'finite verb + preverb + finite verb', since these are false positive hits, having a word erroneously annotated as preverb (e.g. *meg* which can be either a perfectivizing preverb or a conjunction meaning 'and'). These rules had to be tested one by one on the sentences relevant for the given position, monitoring whether they are indeed matching false hits and not matching a large number of true ones.

One might wonder if it would not be a better option to use dependency parsing for this task instead of the rule-based method described above, since separate preverbs can be paired with their associated verbs in a dependency parsed corpus by searching for compound:preverb relations. Experiments carried out by Pethő et al. (2022, 86–87) show that Stanza (Qi et al. 2020) – a well-known dependency parser – performs rather poor in this task. On a general sample, it achieves 87.40% accuracy, while on a difficult one – which contains more complex construction types – its accuracy is 81.91%. The emPreverb tool (Pethő et al. 2022), a rule-based algorithm inspired by the work presented here, outperforms Stanza both on the general sample (97.80%) and on the difficult one (93.62%).

Figure 2 gives an overview of the pieces of information that are to be extracted from a sentence. First, there are document-level data – metadata available in HGC – which can be accessed by the document identifier associated with the sentence. These are region, style and year of publication, see Section 3.1. Metadata are essential to track the geographical and temporal propagation or decline of constructions. Second, sentence-level data can be extracted: the given construction and its left and right contexts. Finally, construction-level data should be extracted in detail: the construction type, the preverb and its associated verb stem, the words interposed between these and the preverb's position which is calculated based on the intervening words.

4. BASIC STATISTICS OF THE DATASET

The resulting dataset consists of 41,547,495 records, each representing one hit in the corpus. It contains the occurrences of 235 preverbs or preverb-like lexical items in 49 construction types. Both preverbs and construction types have a positively skewed distribution which follows the macro-trends of Zipf's law (Zipf 1932), see Figure 3.

Turning to a numeric overview of the three clausal orders in Table 2, we find that continuous order is the most widespread and frequent one, shown by 72% of PrevDistro's records. It is followed by inverted order which has a relatively high token frequency (24.4%), but it can only be attested in 8 construction types.

Preverb-verb combinations may appear in inverted order if they are finite verbs, infinitives or participles functioning as the predicate of a non-neutral clause. According to PrevDistro, the only exception to this generalization is the comparative construction – marked with -bb – which can also be attested in inverted order sporadically, as shown in Example (13). It has to be noted,

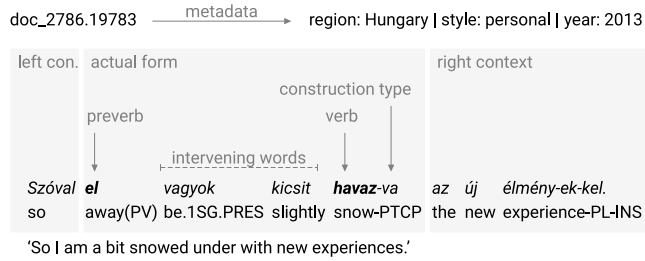

Figure 2. Pieces of information being extracted from a single sentence of the corpus

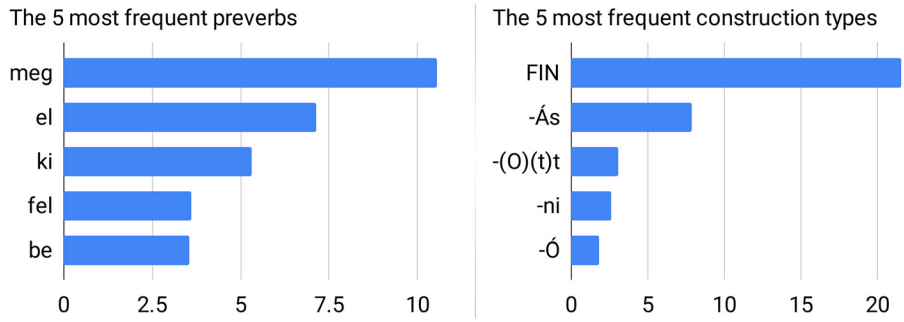

Figure 3. Token frequencies of the five most common preverbs and construction types, given in million tokens

Table 2. Clausal ordering possibilities of preverb-verb combinations. Column *tokens* shows the token frequency of each clausal order. Column *types* presents the number of construction types in which a given clausal order can be attested

clausal order	tokens	types
continuous	31,381,227	47
Inverted	8,670,459	8
discontinuous	1,495,809	34

however, that the input of comparative formation is always a -hAtÓ '-able' construction where this ordering possibility is quite common.

(13) a. *hol* ***játsz**-ható-bb* ***ki*** *leginkább a határőrség*
where play-able-CMPR out(PV) mostly the border.guards
'[…] where it is easiest to outwit the border-guards […]'

b. *Több pánt-tal feszes-ebb háló* ***ér**-hető-bb* ***el***.
more band-INS tight-CMPR net attain-able-CMPR away(PV)
'A tighter fishing net can be created by using more bands.'

The discontinuous order seems to be a mirror image of the inverted order: its token frequency is low (3.6% of the corpus data), but it can be attested in 34 construction types which is way more than previously assumed. Even deverbal nouns, adjectives and adverbs show this type of ordering, see Example (14). Mention must be made that only four clitic-like items can be placed between the preverb and the deverbal element in these derivates. These are: *nem* 'not' (99.35%), *sem* 'not even' (0.46%), *is* 'also' (0.15%), *se* 'not even' (0.03%).[6]

[6]The data are orthographically varied: usually, the whole unit is written as one single word, sometimes each piece is written separate and rarely they are connected with hyphens.

(14) a. ***el*** *is* ***vár****-ható-an*
away(PV) too expect-able-ly
'expectedly as well'

b. ***észre****-nem-****vevő****-sdi-t*
on.mind(PV)-not-taking-playing-ACC
'acting like one who did not notice anything'

c. *leg-****össze****-nem-****illő****-bb*
most-together(PV)-not-matching-CMPR
'as unmatched as possible'

These data are particularly interesting from a historical point of view, as they exhibit the ancient clausal order of negation. Even in the earliest surviving texts of Old Hungarian, negation shows an alternation of discontinuous ('preverb – negative particle – verb') and inverted ('verb – negative particle – preverb') orders, the former being the more frequent one (É. Kiss 2014). Over time, the situation changed to the contrary in the case of finite and non-finite verbs (Gugán 2015), but the data presented above show that the ancient clausal order of negation holds its ground in deverbal constructions.

Even this cursory glance at the data collected in PrevDistro shows that the variation of clausal orders was understated in several cases. Further, in-depth investigation of the data can certainly lead to novel findings.

5. CONCLUSION

This paper presented PrevDistro, a dataset of Hungarian preverb constructions. The resource consisting of 41.5 million records was extracted automatically from the Hungarian Gigaword Corpus. It is stored in TSV – tab separated values – format which is easily parsable for several programming tools, thus amenable to large-scale computational analysis. PrevDistro is open-access, licensed under GNU General Public License v3.0, archived in Zenodo under DOI 10.5281/zenodo.6349410.

In the near future, the dataset will be made browseable via an online search interface which does not require any technical knowledge. This would make it possible for a wider circle of linguists to carry out smaller-scale studies, e.g. a qualitative investigation of some specific constructions or preverbs. The corpus occurrences of morphologically complex preverbs – mentioned in Section 2.1 – are also planned to be included. A long-term project would be to augment PrevDistro by historical data, namely preverb constructions from the Old, Middle and New Hungarian periods.[7] This could open the way to quantitative diachronic studies which are currently cumbersome due to the differing structure and annotation formats of Hungarian historical corpora. Each upgrade of the dataset will be announced at https://zenodo.org/record/6349410.

[7]All historical periods can be covered by three corpora: the Old Hungarian Corpus (Simon 2014), the Old and Middle Hungarian corpus of informal language (Novák et al. 2018), and finally, the Hungarian Historical Corpus (Csengery 2006).

ACKOWLEDGEMENT

This research has been supported by the OTKA PD project No. 142317 funded by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the PD 22 funding scheme.

REFERENCES

Ackerman, Farrell. 2003. Lexeme derivation and multi-word predicates in Hungarian. Acta Linguistica Hungarica 50. 7–32.

Ackerman, Farrell, Ágnes Kalivoda and Robert Malouf. 2022. Paradigmatic organization as a solution to Zipfian distributions in Hungarian grammar. Manuscript in preparation.

Ackerman, Farrell and Philip LeSourd. 1997. Toward a lexical representation of phrasal predicates. In A. Alsina, J. Bresnan and P. Sells (eds.) Complex predicates. 67–106. Stanford, CA: CSLI Publications.

Beniamine, Sacha, Martin Maiden and Erich Round. 2020. Opening the romance verbal inflection dataset 2.0: A CLDF lexicon. Proceedings of the 12th Language Resources and Evaluation Conference. 3027–3035.

Brezina, Vaclav. 2018. Statistics in corpus linguistics: A practical guide. Cambridge: Cambridge University Press.

Croft, William. 2001. Radical construction grammar. Oxford: Oxford University Press.

Csengery, Kinga. 2006. Az elektronikus korpusz [The electronic corpus]. In: Ittzés, N. (Ed.), A magyar nyelv nagyszótára I: Segédletek [The Comprehensive Dictionary of Hungarian I: Manual]. Research Institute for Linguistics, HAS, Budapest, pp. 18–19.

Diessel, Holger. 2019. The grammar network: How linguistic structure is shaped by language use. Cambridge: Cambridge University Press.

Dóra, Zoltán. 2006. Kell mennem, kell menjek? [Kell mennem 'I have to go' or kell menjek 'I have to go'?]. Magyar nyelvőr 130(4). 413–421.

É. Kiss, Katalin. 1998. Verbal prefixes or postpositions? Postpositional aspectualizers in Hungarian. In C. de Groot and I. Kenesei (eds.) Approaches to Hungarian, Vol. 6. Szeged: JATE. 123–148.

É. Kiss, Katalin. 2009. Nekem el kell menni/el kell mennem/el kell, hogy menjek/el kell menjek/el kellek menni [Five ways to express 'I have to go']. In K. É. Kiss and A. Hegedűs (eds.) Nyelvelmélet és dialektológia. Piliscsaba: PPKE BTK. 213–227.

É. Kiss, Katalin. 2014. A tagadó és a kérdő mondatok változásai [Changes of negative and interrogative clauses]. In: Kiss, K.É. (Ed.), Magyar generatív történeti mondattan [The Diachronic Generative Syntax of Hungarian]. Akadémiai Kiadó, Budapest, pp. 34–49.

É. Kiss, Katalin. 2021. Predicative PPs. In K. É. Kiss and V. Hegedűs (eds.) Syntax of Hungarian: Postpositions and postpositional phrases. Amsterdam: Amsterdam University Press. 251–284.

Gries, Stefan Th. 2015. Quantitative linguistics. In J. Wright (ed.) International encyclopedia of the social and behavioral sciences, 2nd edn., Vol. 19. Amsterdam: Elsevier Ltd. 725–732.

Gugán, Katalin. 2015. És mégis: mozog? Tagadás és igemódosítók az ómagyarban és a középmagyarban [And yet it moves. Negation and verb modifiers in Old and Middle Hungarian]. Általános Nyelvészeti Tanulmányok 27. 153–178.

Hegedűs, Veronika. 2013. Non-verbal predicates and predicate movement. Utrecht: LOT.

J. Soltész, Katalin. 1959. Az ősi magyar igekötők: meg, el, ki, be, fel, le [The ancient Hungarian preverbs: meg, el, ki, be, fel, le]. Budapest: Akadémiai Kiadó. 263.

Jakab, István. 1976. A magyar igekötők állományi vizsgálata [Investigating the set of Hungarian preverbs] (Nyelvtudományi Értekezések 91). Budapest: Akadémiai Kiadó.

Kalivoda, Ágnes. 2019. Véges erőforrás végtelen sok igekötős igére [A finite resource for an infinity of preverb-verb combinations]. XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). 331–344.

Kalivoda, Ágnes. 2021. Igekötős szerkezetek a magyarban [Preverb constructions in Hungarian]. Doctoral dissertation. Pázmány Péter Catholic University, Budapest.

Kerekes, Judit. 2011. Az igekötők meghatározásának problémái [Issues concerning the definition of preverbs]. In: Gécseg, Z. (Ed.), LingDok10: Nyelvész-doktoranduszok dolgozatai [LingDok10: Papers written by PhD students in Linguistics]. Doctoral School in Linguistics, University of Szeged, Szeged, pp. 109–131.

Kiefer, Ferenc. 1995. Prefix reduplication in Hungarian. Acta Linguistica Hungarica 43(1/2). 175–194.

Kiefer, Ferenc (ed.). 2000. Strukturális Magyar Nyelvtan 3: Morfológia [Structural grammar of Hungarian 3: Morphology]. Budapest: Akadémiai Kiadó.

Kiefer, Ferenc. 2007. Jelentéselmélet [Theory of meaning], 2nd edn. Budapest: Corvina.

Kálmán, László. 2013. Egyre eljebb terjed [It is spreading more and more]. Published at the popular science portal Nyelv és Tudomány. https://www.nyest.hu/hirek/egyre-eljebb-terjed.

Kálmán, László. 2016. Bővítménykeretek mint konstrukciók [Argument frames as constructions]. In: Kas, B. (Ed.), "Szavad ne feledd!" Tanulmányok Bánréti Zoltán tiszteletére ["Hold that thought!" Papers in Zoltán Bánréti's honour]. Research Institute for Linguistics, HAS, Budapest, pp. 61–72.

Kálmán, László and Viktor Trón, 2000. A magyar igekötő egyeztetése [Agreement of the Hungarian preverb]. In: Büky, L., Maleczki, M. (Eds.), A mai magyar nyelv leírásának újabb módszerei IV [Recent methods of describing Present-day Hungarian IV]. University of Szeged, Szeged, pp. 203–211.

Komlósy, András. 1992. Régensek és vonzatok [Predicates and arguments]. In: Kiefer, F. (Ed.), Strukturális magyar nyelvtan 1: Mondattan [Structural grammar of Hungarian 1: Syntax]. Akadémiai Kiadó, Budapest, pp. 299–527.

Kornai, András, Dávid Márk Nemeskey and Gábor Recski. 2016. Detecting optional arguments of verbs. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2815–2818.

Langacker, Ronald W. 1991. Concept, image, and symbol: The cognitive basis of grammar. Berlin: Mouton de Gruyter.

Novák, Attila, Katalin Gugán, Mónika Varga and Adrienne Dömötör. 2018. Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. Language Resources and Evaluation 52. 1–28.

Novák, Attila, Péter Rebrus and Zsófia Ludányi. 2017. Az emMorph morfológiai elemző annotációs formalizmusa [Annotation format of the emMorph morphological analyzer]. XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017). 70–78.

Novák, Attila, Borbála Siklósi and Csaba Oravecz. 2016. A new integrated open-source morphological analyzer for Hungarian. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 1315–1322.

Oravecz, Csaba, Tamás Váradi and Bálint Sass. 2014. The Hungarian gigaword corpus. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). 1719–1723.

Pethő, Gergely, Bálint Sass, Ágnes Kalivoda, László Simon and Veronika Lipp. 2022. Igekötő-kapcsolás [Connecting preverbs and verbs]. XVIII. Magyar Számítógépes Nyelvészeti Konferencia. 77–91.

Piñón, Christopher. 1991. Falling in paradise: Verbs, preverbs and reduplication in Hungarian. Handout, Syntax Workshop, May 21, 1991. Stanford University, Stanford, CA. http://pinon.sdf-eu.org/covers/fp.html.

Posgay, Ildikó. 2002. Kell tanítsuk? [Kell tanítsuk? 'Should we teach it?']. In: Balázs, G., Adamikné Jászó, A., Koltói, G. (Eds.), Éltető anyanyelvünk: Írások Grétsy László 70. születésnapjára [Our vital native tongue: Papers dedicated to László Grétsy on the occasion of his 70th birthday]. Tinta Kiadó, Budapest, pp. 392–395.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 101–108.

Rákosi, György. 2014. A case of disagreement: On plural reduplicating particles in Hungarian. In A. Kertész and C. Rákosi (eds.) The evidential basis of linguistic argumentation (Studies in Language Companion Series 153). Amsterdam: John Benjamins Publishing Company. 179–198.

Rákosi, György and Tibor Laczkó. 2011. Inflecting spatial particles and shadows of the past in Hungarian. The Proceedings of the LFG11 Conference. 440–460.

Simon, Eszter. 2014. Corpus building from Old Hungarian codices. In K. É. Kiss (ed.) The evolution of functional left peripheries in Hungarian syntax (Oxford Studies in Diachronic and Historical Linguistics). Oxford: Oxford University Press. 224–236.

Sánchez-Gutiérrez, Claudia H., Hugo Mailhot, S. Hélène Deacon and Maximiliano A. Wilson. 2018. MorphoLex: A derivational morphological database for 70,000 English words. Behavior Research Methods 50(4). 1568–1580.

Stefanowitsch, Anatol. 2020. Corpus linguistics: A guide to the methodology (Textbooks in Language Sciences 7). Berlin: Language Science Press.

Surányi, Balázs. 2009. Adpositional preverbs, chain reduction, and phases. In M. den Dikken and R. M. Vago (eds.) Approaches to Hungarian, Vol. 11. Amsterdam: John Benjamins Publishing Company. 217–250.

Zipf, George Kingsley. 1932. Selected studies of the principle of relative frequency in language. Cambridge, MA: Harvard University Press.