ÁGNES KALIVODA

# PREVERB CONSTRUCTIONS IN HUNGARIAN

Doctoral (PhD) dissertation

## THESIS BOOKLET

Pázmány Péter Catholic University
Faculty of Humanities and Social Sciences
Doctoral School of Linguistics
Doctoral Programme in Language Technology

**Supervisor:**
**Prof. Gábor Prószéky**
Professor, Doctor of Sciences

Budapest
2021

# 1 Aims

The thesis investigates preverbs – in other terms: verbal particles, verbal prefixes – and preverb constructions in Hungarian. Its primary aim is to explore and describe these as completely as possible, using a corpus-driven approach. The three main topics discussed here are (1) defining a set of lexical items which can be regarded as preverbs, (2) describing the clausal orders of preverb constructions, and (3) exploring the productive preverb-verb patterns. A further aim is to create freely available resources which can serve as a starting point of subsequent linguistic research and which can also be used in language technology tasks.

# 2 Data and methods

The studies presented in this thesis are corpus-driven, i.e. they set out from the automatic analysis of extremely large bodies of text, aiming to detect phenomena that can not be explored by introspection.

**Corpora:** Most of my corpus analyses are based on the Hungarian Gigaword Corpus, version 2.0.4 (Oravecz et al. 2014) which was designed to represent a wide cross-section of Hungarian from the 20th and 21st centuries. In addition to this, I used three historical corpora in the course of a diachronic corpus study. The Old Hungarian period (896–1526) is represented by the Old Hungarian Corpus which contains all available Old Hungarian and some Middle Hungarian texts (Simon and Sass 2012). In order to investigate the Middle Hungarian period (1526–1772), I used the Old and Middle Hungarian corpus of informal language (Dömötör et al. 2017). This is focused on informal text types: private letters and court records of witch trials. The Modern Hungarian period (from 1772 to the present day) is represented by the Hungarian Historical Corpus (Ittzés 2009).

**Natural Language Processing Tools:** In several cases, it was unavoidable to improve the existing linguistic annotation of the corpora or to add new annotation layers. In order to achieve this, I used the emtsv (Indig et al. 2019; Váradi et al. 2018) and the magyarlanc 3.0 (Zsibrita et al. 2013) text processing systems. From a methodological point of view, the most challenging task was to explore the productive preverb-verb patterns. This required the identification of verb-forming suffixes, the extraction of argument frames, and the detection of semantically similar word groups. I used the emMorph morphological analyzer (Novák et al. 2016, 2017) for the first task, a method developed by Sass (2011) for the second one, and a word2vec embedding (Siklósi and Novák 2016) for the third one. Moreover, I developed new al-

gorithms to be able to investigate some specific linguistic phenomena, e.g. to determine the position of a – separate – preverb relative to the verb stem, or to identify diverse sound patterns.

## 3 The structure and the main theses of the dissertation

The dissertation begins with a short introduction, followed by a detailed description of the research methods. After that, I dedicate three large chapters to my main research questions which are as follows: (1) Which lexical items can be regarded as preverbs, and what are the grounds of their classification? (2) What kinds of clausal orders do preverb constructions show, when and to what extent can a preverb be separated from a finite/non-finite verb or a deverbal element? (3) How can we describe the productive preverb-verb patterns, and – based on this – what conclusions can be made about the semantics of preverbs? Finally, in a short but substantive chapter, I return to the evaluation of the approach introduced at the beginning of the dissertation. Having seen its flaws, I outline a different approach which takes the constructions as its starting point instead of the individual lexical items. I conclude by summarizing the most notable results and formulating my theses. Below I provide a more detailed description of the main chapters and the related theses.

**Chapter 2** discusses the notion of corpus-drivenness and presents two resources which I used in each of my corpus studies. One is a modified version of the Hungarian Gigaword Corpus – HGC – which is free of duplicate texts, poems and non-Hungarian sentences. The other is the PrevLex table which forms my first thesis:

1. Using the HGC corpus, I created PrevLex which is the largest manually checked, open-access table of preverb-verb combinations at the time of writing (consisting of 53 535 lexeme types). It contains hapaxes – words occurring only once in the data – as well as words annotated with UNKNOWN tags. Each lexeme is presented with its token frequency obtained from the HGC.

**Chapter 3** presents an attempt to define a set of lexical items that can be regarded as preverbs. I assume that there is a fuzzy boundary between preverbs and other bare nominal verb modifiers. The prototype-theory seems to be suited for the graded categorization needed here. With this in mind, I collect morphological and frequency-related features which might be useful in defining the set of preverbs. I measure the value of each feature in the case of 235

preverb-like lexical items. Using the results of my data collection, I create Preverb × Feature matrices which differ mainly in the way they represent the feature values. Based on the matrices, I measure the correlations between each feature-pair. Considering these correlations as well as the standpoints made in a range of relevant literature, and – undeniably – relying on my intuition, I assign *meg* – a perfectivizer with no literal meaning – as the prototype. I define the typical characteristics of preverbs based on the features of *meg*. Finally, I present three methods that may be suited for a feature-based classification of preverbs. After comparing these, I choose the method introduced by Smith et al. (1988). I set up a continuum ranging from the standard preverbs to the least preverb-like elements. In order to facilitate the discussion in the later chapters, I decide to split the continuum into four categories: prototypical (e.g. *meg* perfectivizer, *el* 'away'), central (e.g. *szét* 'apart', *vissza* 'back'), semi-peripheral (e.g. *agyon* 'to death', *félbe* 'into half') and peripheral (e.g. *szénné* 'to coal', *létre* 'into being') preverbs. The main results of this chapter are the following:

**2.** I defined and measured 10 features that can be used to characterize preverbs. I indicated, however, that not all features are equally relevant. The results of the corpus analysis can be accessed in the form of Preverb × Feature matrices.

**3.** I used the Preverb × Feature matrix containing absolute frequencies to compute the correlation of each feature-pair. Based on this, it was possible to show the process of grammaticalization in the case of preverbs by quantitative means. Productivity has a strong positive correlation with frequency, while the number of syllables and the morphological complexity show a negative correlation with these. Frequent and productive preverbs are typically short and monomorphic. I calculated the correlations on binary data as well, showing that the relations among the features under investigation do not change if the absolute token frequencies are omitted. I explained this by the fact that frequency is historically so closely related to other features – due to the grammaticalization process – that its effect is found in other features even if it is not considered as a feature on its own.

**4.** Based on the method introduced by Smith et al. (1988), I set up a continuum ranging from the prototypical preverbs to the bare nominal verb modifiers.

**Chapter 4** investigates the clausal orders of preverb constructions. I first perform a synchronic corpus study using the Hungarian Gigaword Corpus,

putting emphasis on construction types where the preverb is separable from the verb stem. I study the distribution of preverbs in the case of finite and non-finite verbs as well as deverbal elements. After that, I conduct a diachronic study which aims to quantify the changes of the prototypical preverbs' positions, from the Old Hungarian period to the present day. The main conclusions of this chapter can be summarized as follows:[1]

**5.** I have shown that prototypical preverbs tend to remain close to the finite verbs in terms of relative frequency, while more peripheral preverbs are found even in remote positions. In connection with this, I specified two factors that are likely to affect the distance of the preverb and the finite verb in inverted order constructions (V-PV). One is the formality of the text: in spontaneous – mostly spoken – language, the likelihood of an increased distance between the finite verb and its preverb is higher than in an edited, formal text. The other is the phonological weight of the constituents – including the preverb – following the verb. We can see a trend known as the Law of Increasing Terms or Behaghel's Law, which – according to É. Kiss (2007) – applies to the postverbal field of clauses in Hungarian: the shorter constituent precedes the longer one, unless it is blocked by a syntactic rule. This is also consistent with the observation that monosyllabic, prototypical preverbs are less likely to occur far from the finite verb than polysyllabic, peripheral preverbs.

**6.** I proved by corpus analysis that the preverb can take a distant preverbal position relative to its associated infinitive, but only if an auxiliary-like lexical item – mainly a finite verb – intervenes between them, and if the preverb can occupy the verb modifier position preceding this element (e.g. ***össze***$_{PV}$ *lehetne őket többé-kevésbé objektív módon is* ***mérni***$_{V}$ 'they could be compared in a more or less objective way'). This supports the hypothesis formulated by Kálmán C. et al. (1989): even if the preverb forms a lexical unit with the infinitive verb, it is more closely connected with the auxiliary-like item with respect to word order and prosody. I found a strikingly similar syntactic behavior in the case of passive constructions consisting of a copula and an adverbial participle. Here, the preverb associated with the participle can take a distant preverbal

[1] In the examples provided in Theses 5–12, the preverb (PV) and its associated verb (V) are marked with boldface, and the auxiliary-like finit verbs are underlined. One must also note that preverb-verb combinations display three ordering possibilities in Hungarian: (1) direct order – the preverb is prefixed to the verb stem, (2) discontinuous order – the preverb precedes the verb, but they are separated by other elements, (3) inverted order – the preverb follows the verb, often not immediately.

position if it occupies the verb modifier position preceding the copula (e.g. ***ki***$_{PV}$ *vannak ezek a marketinges dolgok* ***találva***$_{V}$ 'this marketing stuff is well-planned').

**7.** I found that in the case of adverbial participles, the inverted order is possible only if the participle functions as an adverb denoting a state or a manner (e.g. *ezzel* ***szoktatva***$_{V}$ ***át***$_{PV}$ 'by changing his/her habits in this way'). In passive constructions consisting of a copula and an adverbial participle, the preverb always precedes its associated verb, either in a direct or in a discontinuous order (e.g. ***el***$_{PV}$ *van* ***intézve***$_{V}$ 'it's arranged').

**8.** Regarding adjectival participles, I made the following observations: (1) If a participle having the suffix -hAtÓ '-able' functions predicatively – and there is no finite verb in the clause –, the -hAtÓ participle shows exactly the same behavior as finite verbs do (e.g. *az mindig akkor* ***vonható***$_{V}$ *már csak* ***le***$_{PV}$ 'it can be deducted only when ...'). Its associated preverb is separable in the same way as a finite verb's preverb, the distribution of the preverbs is clearly similar, and there is a parallel with finite constructions even when looking at the words that can be interposed between the preverb and the verb stem. All these facts indicate that -hAtÓ is not really a marker of adjective formation – as stated by Kiefer (2003) –, as the words suffixed with it show characteristics which are typical of verbs. (2) I found that adjectival participles suffixed with -AndÓ 'to be [verb]-ed' can function predicatively, and in this case, the inverted order is possible (e.g. *nem* ***tévesztendő***$_{V}$ ***össze***$_{PV}$ 'not to be confused with'), although this is undoubtedly rare (it can be attested in 1.85% of the cases, that is to say, 1 624 hits in the corpus).

**9.** The corpus analysis revealed the ubiquity of the discontinuous order: even deverbal nouns, adjectives and adverbs show this type of ordering. It must be noted, however, that only four clitic-like items can be placed between the preverb and the deverbal element in these derivates. These are: *nem* 'not', *sem* 'not even', *se* 'not even', *is* 'also' (e.g. ***el***$_{PV}$ *is* ***várhatóan***$_{V}$ *'expectedly as well', leg-****össze***$_{PV}$*-nem-****illőbb***$_{V}$ 'as unmatched as possible').

**10.** I studied a group of constructions in which the preverb associated with a verb in subjunctive form – or with a non-finite verbal complement of a verb in subjunctive form – precedes a finite modal which is typically the verb *kell* 'must' (e.g. ***el***$_{PV}$ *kell, hogy* ***menjek***$_{V}$ – ***el***$_{PV}$ *kell* ***menjek***$_{V}$ 'I must leave'). My main observations are as follows: (1) Variants with

and without the complementizer *hogy* 'that' are similar in terms of token frequency, regardless of whether the associated verb is a finite verb, an infinitive, or an adverbial participle. (2) Some short words can intervene between the finite modal and the complementizer. Moreover, constituents between the finite modal and the verb associated with the preverb are clearly similar to the ones which occur in infinitival constructions having discontinuous order.

**11.** I studied a group of constructions in which a preverb-verb combination is topicalized as an infinitive or as an adverbial participle, and it appears repeatedly as a finite verb (e.g. ***fel***$_{PV}$***jelenteni***$_{V}$ *azért* ***fel***$_{PV}$***jelentik***$_{V}$ 'as for pressing charges against him/her, they will do that'). I found that the preverb can be omitted in clauses having inverted order (e.g. ***be***$_{PV}$***tanulni***$_{V}$ *nem* ***tanultam***$_{V}$ *semmit* 'as for memorizing, I didn't memorize anything'). Within the range of topicalization constructions, I studied the characteristics of elliptical structures in which the repeatedly occurring preverb is followed by an auxiliary-like item (e.g. ***ki***$_{PV}$***írva***$_{V}$ ***ki***$_{PV}$ *van* 'as for being announced, it is announced').

**12.** The diachronic corpus study revealed an increase in the proportion of non-neutral sentences having inverted order (V-PV) from the Old Hungarian period to the present day. On the one hand, this trend can be explained by the fact that negative sentences having 'verb – negative particle – preverb' order made headway against the ones having 'preverb – negative particle – verb' order. On the other hand, an explanation might be that there is a continuous growth in the proportion of constructions where the use of structural focus became obligatory.

**Chapter 5** focuses on the exploration of productive preverb-verb patterns. I develop a method based on the corpus-driven study of 'preverb – derivational suffix – argument frame' triplets. I present the three most common ways of word formation in Hungarian: verb formation from nouns and verbs, and thirdly, verb formation using sound patterns. After that, I present the PREVCONS database containing 21 038 preverb-verb hapaxes. This resource makes it possible to explore the productive preverb-verb patterns by the accessibility of the triplets mentioned above. Finally, I present an attempt which aims to represent the different meanings associated with preverbs and the relationships between these meanings in a network-like structure based on PREVCONS, in the form of an ontology. My theses related to this chapter are the following:

**13.** I developed an algorithm to identify verbs which can be matched by sound patterns (e.g. *mormog*, 'mumble', *dörmög* 'grumble', *csemcseg*

'munch'), sorting the affected verbs into schema types. I have shown that even though the linguistic literature does not pay much attention to this way of verb formation, the proportion of verbs following a sound pattern is not negligible: these represent nearly one-tenth (9.4%) of the preverb-verb hapaxes.

**14.** I have shown that denominal verb formation plays the most significant role in the creation of new preverb-verb combinations. It can be detected in 35.2% of the hapaxes. At the same time, only 62 – by merging the alternating forms into one unit, only 56 – preverbs combine with denominal verbs.

**15.** I created the PrevCons database, an open-access resource for investigating preverb constructions. It contains 21 038 preverb-verb hapaxes along with information on their morphological structure, argument frame, semantics and context.

**16.** I created an open-source ontology which displays meanings associated with 56 preverbs, and the relationships between these meanings. The preverbs and the meanings are represented as entities, and three basic semantic relationships – synonymy, antonymy and hyperonymy – are considered as relations. The ontology is drawn as a plane graph.

In **Chapter 6**, I return to the concept which was my starting point, namely that the notion of preverb can be best captured by prototype-theoretical means. I check whether the original preverb continuum remains largely the same when considering distributional and semantic features of preverbs. The result shows that the two endpoints remain stable, while there is a considerable fluctuation in-between. The vagueness attested here leads to a viewpoint change from the study of lexical items to the study of constructions, largely based on László Kálmán's review on the first version of this thesis. The main contributions of this chapter are as follows:

**17.** I outlined an approach which sets out from the investigation of constructions. I pointed out two of its benefits over my former approach focusing on lexical items and using a prototype-theoretical framework: (1) There is no need to make arbitrary decisions which do not have a solid empirical basis. (2) By avoiding the categorization of lexical items in advance, the loss of information and the risk of over-generalization can be reduced considerably.

**18.** I created the PrevDistro dataset which contains the corpus occurrences of 49 preverb construction types, in each case indicating the preverb

and the verb lemma, the preverb's position relative to the verb stem, and other intervening words. In addition to this, the larger context of the construction – the whole sentence – can be accessed. The dataset consisting of 41.5 million records is open-source.

The dissertation contains several new scientific results, both from a theoretical and from a practical point of view. Its practical contribution to the field of linguistics and language technology is the publishing of PrevLex, PrevMatrix, PrevCons, PrevOnto and PrevDistro, all of them being valuable and freely available resources. Its theoretical contribution is twofold. On the one hand, it reveals numerous trends which would have remained unnoticed or conjectural in the absence of a corpus-driven method. On the other hand, it draws attention to some phenomena which are not uncommon, yet have so far been of interest to very few linguists. The methods and ideas presented here may be inspiring in the data-driven study of several linguistic phenomena.

## 4 Relevant publications

**Publications:**

Kalivoda, Ágnes 2021. Az igekötők produktív kapcsolódási mintái [Productive preverb-verb patterns in Hungarian]. *Argumentum* 17: 56–82. https://doi.org/10.34103/ARGUMENTUM/2021/4

Kalivoda, Ágnes 2019. Véges erőforrás végtelen sok igekötős igére [A finite resource for an infinity of particle verbs]. In: Berend, Gábor – Gosztolya, Gábor – Vincze, Veronika (eds.): *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)*. Szegedi Tudományegyetem, TTIK, Informatikai Intézet. Szeged. 331–344.

Kalivoda, Ágnes 2018a. Hungarian particle verbs in a corpus-driven approach. In: Gelbukh, Alexander (ed.): *Computational Linguistics and Intelligent Text Processing: 18th International Conference (CICLing 2017), Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I.* Springer International Publishing. Cham. 159–176.

Kalivoda, Ágnes 2018b. Az igekötős igék szintaxisa korpuszvezérelt megközelítésben [The syntax of Hungarian particle verbs in a corpus-driven approach]. In: Scheibl, György (ed.): *Lingdok 17.: Nyelvészdoktoranduszok dolgozatai.* Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola. Szeged. 159–176.

Kalivoda, Ágnes – Vadász, Noémi – Indig, Balázs 2018. MANÓCSKA: A Unified Verb Frame Database for Hungarian. In: Sojka, Petr – Horák, Aleš – Kopeček, Ivan – Pala, Karel (eds.): *Proceedings of the 21st International Conference on Text, Speech and Dialogue (TSD).* Springer-Verlag. Brno. 135–143.

Vadász, Noémi – Kalivoda, Ágnes – Indig, Balázs 2018. Egy egységesített magyar igei vonzatkerettár építése és felhasználása [Creation and application of a unified verb frame database for Hungarian]. In: Vincze, Veronika (ed.): *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018).* Szegedi Tudományegyetem, Informatikai Intézet. Szeged. 3–15.

Kalivoda, Ágnes 2017. Az igekötők gépi annotálásának problémái [Issues around the automatic annotation of Hungarian preverbs]. In: Ludányi, Zsófia (ed.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2017: XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia.* MTA Nyelvtudományi Intézet. Budapest. 100–109.

Vadász, Noémi – Indig, Balázs – Kalivoda, Ágnes 2017. Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével [Seeing clearly through the window – Argument frame disambiguation by means of preverbs and infinitival arguments]. In: Vincze, Veronika (ed.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017).* Szegedi Tudományegyetem, Informatikai Tanszékcsoport. Szeged. 3–12.

Indig, Balázs – Vadász, Noémi – Kalivoda, Ágnes 2016. Decreasing Entropy: How Wide to Open the Window? In: Martín-Vide, Carlos – Mizuki, Takaaki – Vega-Rodríguez, Miguel A. (eds.): *Theory and Practice of Natural Computing: 5th International Conference.* Springer International Publishing. Cham. 137–148.

**Conference presentations:**

Ackerman, Farrell – Kalivoda, Ágnes – Malouf, Robert 2021. A network analysis of Hungarian preverb constructions. *5th American International Morphology Meeting (AIMM5).* Hosted virtually at the Ohio State University, 26–29 August, 2021. (poster)

Kalivoda, Ágnes 2019. Véges erőforrás végtelen sok igekötős igére [A finite resource for an infinity of particle verbs]. *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019).* Szeged, 24–25 January, 2019. (talk)

Kalivoda, Ágnes – Vadász, Noémi – Indig, Balázs 2018. Manócska: A Unified Verb Frame Database for Hungarian. *21st International Conference on Text, Speech and Dialogue (TSD 2018).* Brno, 11–14 September, 2018. (poster)

Vadász, Noémi – Kalivoda, Ágnes – Indig, Balázs 2018. Egy egységesített magyar igei vonzatkerettár építése és felhasználása [Creation and application of a unified verb frame database for Hungarian]. *XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018).* Szeged, 18–19 January, 2018. (talk)

Kalivoda, Ágnes 2017a. Hungarian verbal particles in a corpus-driven approach. *13th International Conference on the Structure of Hungarian (ICSH13).* Budapest, 29–30 June, 2017. (poster)

Kalivoda, Ágnes 2017b. Hungarian particle verbs in a corpus-driven approach. *Computational Linguistics and Intelligent Text Processing – 18th International Conference (CICLing 2017).* Budapest, 17–23 April, 2017. (poster)

Kalivoda, Ágnes 2017c. Az igekötők gépi annotálásának problémái [Issues around the automatic annotation of Hungarian preverbs]. *XI. Alkalmazott Nyelvészeti Doktoranduszkonferencia (Alknyelvdok 2017).* Budapest, 3 February, 2017. (talk)

Vadász, Noémi – Indig, Balázs – Kalivoda, Ágnes 2017. Ablak által világosan – Vonzatkeret-egyértelműsítés az igekötők és az infinitívuszi vonzatok segítségével [Seeing clearly through the window – Argument frame disambiguation by means of preverbs and infinitival arguments]. *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017).* Szeged, 26–27 January, 2017. (talk)

Indig, Balázs – Vadász, Noémi – Kalivoda, Ágnes 2016. Decreasing Entropy: How Wide to Open the Window? *5th International Conference on the Theory and Practice of Natural Computing (TPNC 2016).* Sendai, 12–13 December, 2016. (talk)

Kalivoda, Ágnes 2016. Az igekötős igék szintaxisa korpuszvezérelt megközelítésben [The syntax of Hungarian particle verbs in a corpus-driven approach]. *Nyelvészdoktoranduszok 20. Országos Konferenciája (LingDok 17).* Szeged, 30 November – 1 December, 2016. (talk)

## References


Dömötör, Adrienne – Gugán, Katalin – Novák, Attila – Varga, Mónika 2017. Kiútkeresés a morfológiai labirintusból – korpuszépítés ó- és középmagyar kori magánéleti szövegekből [Finding the way out of the morphological maze: Building a corpus of Old and Middle Hungarian informal texts.] *Nyelvtudományi Közlemények* 113: 85–110.

É. Kiss, Katalin 2007. Az ige utáni szabad szórend magyarázata [An explanation of the free word order after the verb]. *Nyelvtudományi Közlemények* 104: 124–152.

Indig, Balázs – Sass, Bálint – Simon, Eszter – Mittelholcz, Iván – Kundráth, Péter – Vadász, Noémi 2019. emtsv – Egy formátum mind felett [emtsv – One format to rule them all]. In: Berend, Gábor – Gosztolya, Gábor – Vincze, Veronika (eds.): *XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019)* Szegedi Tudományegyetem, TTIK, Informatikai Intézet. Szeged. 235–247.

Ittzés, Nóra 2009. A magyar nyelv nagyszótára [The Hungarian Explanatory Dictionary]. In: Fábián, Zsuzsanna (ed.): *Szótárírás és szótárírók*. Lexikográfiai füzetek 4. Akadémiai Kiadó. Budapest. 65–80.

Kálmán C., György – Kálmán, László – Nádasdy, Ádám – Prószéky, Gábor 1989. A magyar segédigék rendszere [The system of auxiliaries in Hungarian]. *Általános Nyelvészeti Tanulmányok* XVII: 49–103.

Kiefer, Ferenc 2003. Alaktan [Morphology]. In: É. Kiss, Katalin – Siptár, Péter – Kiefer, Ferenc (eds.): *Új magyar nyelvtan.* Osiris. Budapest. 127–199.

Novák, Attila – Rebrus, Péter – Ludányi, Zsófia 2017. Az emMorph morfológiai elemző annotációs formalizmusa [The annotation formalism of the emMorph morphological analyzer]. In: Vincze, Veronika (ed.): *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2017).* Szegedi Tudományegyetem Informatikai Intézet. Szeged. 70–78.

Novák, Attila – Siklósi, Borbála – Oravecz, Csaba 2016. A New Integrated Open-source Morphological Analyzer for Hungarian. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* European Language Resources Association (ELRA). Portorož. 1315–1322.

Oravecz, Csaba – Váradi, Tamás – Sass, Bálint 2014. The Hungarian Gigaword Corpus. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).* European Language Resources Association (ELRA). Reykjavík. 1719–1723.

Sass, Bálint 2011. *Igei szerkezetek gyakorisági szótára – egy automatikus lexikai kinyerő eljárás és alkalmazása* [Frequency dictionary of verbal structures – an automatic lexical extraction method and its application]. Doctoral (PhD) dissertation. Pázmány Péter Katolikus Egyetem. Budapest.

Siklósi, Borbála – Novák, Attila 2016. Beágyazási modellek alkalmazása lexikai kategorizációs feladatokra [Using embedding models for lexical categorization]. In: Tanács, Attila – Varga, Viktor – Vincze, Veronika (eds.): *XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2016).* Szegedi Tudományegyetem, TTIK, Informatikai Intézet. Szeged. 3–14.

Simon, Eszter – Sass, Bálint 2012. Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből [Language technology and cultural heritage – Corpus building from Old Hungarian codices]. *Általános Nyelvészeti Tanulmányok* 24: 243–264.

Smith, Edward E. – Osherson, Daniel N. – Rips, Lance J. – Keane, Margaret 1988. Combining prototypes: A selective modification model. *Cognitive Science* 12: 485–527.

Váradi, Tamás – Simon, Eszter – Sass, Bálint – Mittelholcz, Iván – Novák, Attila – Indig, Balázs 2018. E-magyar – A Digital Language Processing System. In: Calzolari, Nicoletta et al. (eds.): *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA). Miyazaki. 1307–1312.

Zsibrita, János – Vincze, Veronika – Farkas, Richárd 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Angelova, Galia – Bontcheva, Kalina – Mitkov, Ruslan (eds.): *Recent Advances in Natural Language Processing (RANLP 2013).* INCOMA Ltd. Sumen. 763–771.