Hungarian verbal particles in a corpus-driven approach

Ágnes Kalivoda
Pázmány Péter Catholic University, Faculty of Humanities and Social Sciences

Introduction

Verbal particles in Hungarian can occupy various positions in the sentence: they can be preverbal, immediately preverbal (written together with the verb) or postverbal. The state and syntax of verbal particles are discussed in a wide range of theoretical literature (e.g. Komlósy 1992, Kiefer – Ladányi 2000, É. Kiss 2004, Surányi 2009, among others). This paper presents a corpus-driven approach, focusing on the distribution patterns of verbal particles in more than 21.5 million sentences, which is a new method in this topic.

The aim of this paper is to answer the following questions: **1)** How far can a particle be placed from its verb in Hungarian? **2)** Which factors determine whether a particle should stay close to its verb or can be moved to a remote position? **3)** Is it possible to specify the set of verbal particles with the help of computational linguistics?

The main goal of this research is not to support or disprove any theoretical statements. Instead, it aims to show the usefulness of corpus-driven methods and the importance of relatively big and reliable data which could turn predictions into facts.

Methodology

The statements of this paper reflect the evidence provided by the HUNGARIAN GIGAWORD CORPUS (HGC), version 2.0.4 (Oravecz–Váradi–Sass, 2014). The distribution of verbal particles is measured by defining the finite verb as 0 position and calculating the other positions compared to this. Preverbal particles are found in an interval less than zero and postverbal ones in an interval greater than zero. See Kalivoda (2016) for technical details.

Short answers to the research questions raised

1) The data extracted from HGC show that the dominant position among preverbal particles is the -1 position. This covers more than 99% of the preverbal cases. The left periphery of the finite verb has a strict syntax, the verbal particle can not move far away. The maximal left-position turned out to be -4, meaning that there were no more than three intervening words between the particle and the verb. As for postverbal particles, the measurements showed that they can be located in a wider scope compared to the preverbal ones, however, they can be found in the +1/2 positions in 99.9% of the postverbal cases. It is also worth mentioning that more than 35% of verbal particles are postverbal in Hungarian, meaning that sentences containing focused elements are quite frequent.

2) In the present state of research, two factors seem to determine the possible distance of postverbal particles:[1] (1) the opposition of written and spoken – edited and unedited – text, (2) phonological constraints. In the first case, I used the metadata of HGC in order to decide whether a given sentence was originally said or written. At the +1 position, the proportion of spoken sentences is 16.5%, however, it keeps growing as we look to more distant positions. At the +7 position, 62% of the sentences come from spoken text. Regarding the phonological constraints, relatively long verbal particles (consisting of two or more syllables) can stand in

1 I did not consider every possible factor, e.g. semantics, because it is very far from trivial how to measure semantical features on large data, automatically.

remote positions, while short ones are not placed too far from the verb. I measured the length of the first three words behind the finite verb in the postverbal corpus, and counted the average of length values in these three positions, separately.[2] The obtained data show a tendency predicted by the Law of Increasing Terms, a cross-language principle presented by Otto Behaghel (1932). According to this principle, the shorter constituent prefers to precede the longer one, if there is no syntactic rule that could prevent it. This principle holds in the case of the right periphery – phrases placed behind the finite verb – in Hungarian (É. Kiss 2007), and it was possible to quantify it.

3) The category of verbal particles is constantly disputed, it has no clear boundaries (see Komlósy 1992: 495–497). Instead of trying to define this category in terms of necessary and sufficient features, Forgács (2005) proposes a new approach. According to him, it would be advisable to apply the centre-periphery model in the categorization of verbal particles. He applies this model based on the origin and development of these words (e.g. *meg* is in the centre as it is the oldest and most grammaticalized verbal particle in Hungarian.) Kerekes (2011) follows this concept, using the term prototype theory. My findings provide a new aspect to these theories: there is a group of verbal particles that prefer to stay close to the finite verb (e.g. *ki* 'out', *fel* 'up'), while other can stand in remote positions (e.g. *oda* 'there (as a direction)', *vissza* 'back'). The former ones could be specified as a prototypical set of verbal particles, and the latter ones as a less prototypical set – as verbal particles basically tend to stay close to the verb, according to the measurements on HGC.

To sum up, the corpus-driven research of Hungarian particle verbs has proven to be useful for theoretical linguistics. There is evidence of postverbal particles moving away in spontaneous speech more often than in edited texts. By examining the phonological effects, Behaghel's Law of Increasing Terms became quantifiable. Finally, a new aspect is introduced in the prototype theory of verbal particles.

References

Behaghel, O.: *Deutsche Syntax IV.* Carl Winters, Heidelberg (1932)

Forgács, T.: Grammatikalizálódás az igekötők körében. In: Oszkó, B. – Sipos, M. (eds.) *Uráli grammatizáló,* Budapesti Uráli Műhely 4, Budapest, MTA Nyelvtudományi Intézet, 88–116 (2005)

É. Kiss, K.: Egy igekötőelmélet vázlata. *Magyar Nyelv 100,* 15–43 (2004)

É. Kiss, K.: Az ige utáni szabad szórend magyarázata. *Nyelvtudományi Közlemények 104,* 124–152 (2007)

Kalivoda, Á.: *A magyar igei komplexumok vizsgálata,* MA thesis. Available at: https://github.com/kagnes/hungarian_verbal_complex (2016)

Kerekes, J.: Az igekötők meghatározásának problémái. In: Gécseg Zsuzsanna (ed.) *LingDok 10. Nyelvészdoktoranduszok dolgozatai,* Szeged, JATEPress, 109–130 (2011)

Kiefer, F. – Ladányi, M.: Az igekötők. In: Kiefer Ferenc (ed.) *Strukturális magyar nyelvtan. III. Morfológia.* Budapest, Akadémiai Kiadó, 453–518 (2000)

Komlósy, A.: Régensek és vonzatok. In: Kiefer Ferenc (ed.) *Strukturális magyar nyelvtan. I. Mondattan.* Budapest, Akadémiai Kiadó, 299–527 (1992)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., et al. (eds.): *Proceedings of the 9th International Conference on Language Resources and Evaluation,* May 26-31, 2014, Reykjavik, Iceland, ELRA, 1719–1723 (2014)

Surányi, B.: Verbal particles inside and outside vP. *Acta Linguistica Hungarica 56,* 201–249 (2009)

2 Behagel's law is about phrases, but HGC does not contain phrase-level annotation, meaning that the only thing that could be measured confidently was the word length.