Conjugated Infinitives in the Hungarian National Corpus

Gergely Bottyán[1,2] and Bálint Sass[1]

[1] Department of Corpus Linguistics, Research Institute for Linguistics, Hungarian Academy of Sciences
[2] English Linguistics PhD Program, Doctoral School in Linguistic Sciences, ELTE
{bottyang,joker}@nytud.hu

1 Introduction

The infinitive is one of those linguistic forms with which nonfiniteness, i.e. the verbal feature meaning lack of tense, number and person markers, is usually associated. This is a direct consequence of the fact that we only find nonfinite infinitives in Slavic languages and in most Germanic and Romance languages. However, in languages as diverse as Hungarian, Portuguese and Welsh, for example, there are both nonfinite infinitives and conjugated infinitives, i.e. infinitives that are inflected for number and person [1].

The two types of Hungarian infinitive are exemplified in Table 1.

Table 1. The two types of Hungarian infinitive

I. Reggel fel kell kel*ni*.	
morning up must wake-*INF*	
One has to wake up in the morning.	
II. Ír*ni***a** kell.	
Read-*INF*-[**3rd sing**] must	
(S)he must write.	

Hungarian infinitives of the conjugated type (II. in Table 1) have recently attracted considerable attention from generativist syntacticians. Much effort has been made to specify the sentential contexts in which conjugated infinitives occur and the structural representation of phrases formed with conjugated infinitives [2–4]. As is generally the case with syntactic research done in the Chomskyan paradigm, the authors of these studies relied on their own linguistic intuition and no systematic data collection procedure was followed. Nevertheless, the specification of contexts provided in [2] is said to be exhaustive and based on empirical material.

The present paper reports on the investigation that we have performed on the basis of the 153.7 million word lemmatized, morphosyntactically tagged and

disambiguated Hungarian National Corpus [5]. Our principal aim was to check the validity of the claim that all linguistic items (hereafter called licensors) that take conjugated infinitival complements are identified in [2] by making a list of such items in the corpus data. A further aim was to specify which licensors occur with which conjugated infinitives in the extracted sentences. As is common in language technology, the tasks were carried out partly manually, partly automatically, with an iterative method.

2 The Procedure

First, all the sentences that contained conjugated infinitives were automatically extracted from the corpus, on the basis of the morphosyntactic annotation. Our working hypothesis was that licensors were to be found in the clause that contained the conjugated infinitive. Since the corpus is not tagged for clauses, candidates for clauses were identified with our own approximation. This approximation was based on clause-final punctuation marks and clause-initial conjunctions [6]. From the resulting set of clause candidates those members were filtered out that contained a licensor identified in [7], and the conjugated infinitive was recorded along with its licensor. In the remaining set of clause candidates new licensors were looked for manually, and the procedure was applied all over again.

3 Results Obtained

After three iterations, the following measures were obtained. The lemmatized list of licensors has 197 members. Clauses containing these licensors cover 223140 (98%) of the 228367 conjugated infinitive tokens that occur in the Hungarian National Corpus. The number of lemmatized licensor – conjugated infinitive pairs identified is 17874.

An extract from the resulting data collection can be seen in Table 2.

4 Conclusions and Further Research

On the basis of the results of our investigation, the following conclusions can be drawn. (i) A number of sentential contexts in which Hungarian conjugated infinitives occur are missing from the list in [2], thus it is not exhaustive. (ii) If at least one feature of a grammatical construction (in this case, the inflectional suffix of the conjugated infinitive) is tractable in the Hungarian National Corpus or any other richly annotated corpus of its size, it is worth the effort extracting data from the language resource partly automatically, partly manually before jumping into hasty conclusions.

Further research based on our data collection should establish whether there are semantic restrictions on the range of licensors that take conjugated infinitival complements in Hungarian. Similarly, checking the grammaticality of the

Table 2. Extract from the lemmatized licensor – conjugated infinitive pairs data collection. The headword is the lemma of the licensor, which is followed by its number of occurrences in the corpus. Then come the lemmas of the licensed conjugated infinitives in decreasing order of frequency

köteles [18 db]
3x: alávet
2x: ad megakadályoz tart
1x: átvesz biztosít gondol gondoskodik igazol marad megad megtesz visszafizet

kötelesség [138 db]
5x: biztosít
4x: lesz vesz
3x: ad hoz megjelenik megtesz szól vállal
2x: elhatárolód ellát ellenőriz elvisel értesít foglal gondoskodik ismer megvéd tájékoztat tesz visel
1x: áll átír beavatkozik beküld beszámol betart bocsát csinál eljön elmegy elvégez emel emlékezik|emlékez épít felajánl felfegyverez felismer fellép felvilágosít figyelmeztet fizet folytat fordul fölnevel gazdálkodik gyarapít hajt házasodik huny hurcol indít iszik kardoskodik kér kijelöl kikényszerít kiszab kiüresít kivesz kizeng köszön küld küzd marad megakadályoz megemlékezik|megemlékez meghallgat megismer megkérdez megkeres megőriz megszavaz megtanul megtárgyal megtart megválaszt megvív meggyőződik|meggyőződ meztelenít mozgósít néz összegyűjt politizál sorol szerez takarít támad támogat tart teljesít tisztáz törekedik tud tudat túljut tűz ügyel véd védekezik végez verekedik vet virraszt

nonfinite counterparts of the example sentences belonging to the extracted licensor – conjugated infinitive pairs in the corpus could help us specify the overlaps between the distribution of nonfinite and conjugated Hungarian infinitives. Both directions of research require a tool that enables the analyst to retrieve those sentences in the corpus that belong to a given licensor – conjugated infinitive pair in the collection. The authors of this study are planning to develop such a tool and make both the data collection and the tool available to the research community.

References

1. Miller, D. G.: Where do conjugated infinitives come from? Diachronica, 20, 1. (2003) 45–81
2. Tóth, I.: Inflected infinitives in Hungarian. Ph. D. dissertation. Tilburg: University of Tilburg. (2000)
3. É. Kiss, K.: Agreeing infinitives with a case-marked subject. In The syntax of Hungarian. Cambridge: Cambridge University Press. (2002) 210–221
4. Tóth, I.: Can the Hungarian infinitive be possessed? In Kenesei, I. and Siptár, P. (eds.), Proceedings of the conference "Approaches to Hungarian". Budapest: Akadémiai Kiadó. (2002) 135–160
5. Váradi, T.: On Developing the Hungarian National Corpus. In Vintar, Š. (ed.), Proceedings of the Workshop "Language Technologies – Multilingual Aspects". Ljubljana: University of Ljubljana. (1999)

6. Gábor, K., Héja, E. and Mészáros, Á.: Kötőszók korpusz-alapú vizsgálata [A corpus-based investigation of conjunctions]. In Alexin, Z. and Csendes, D. (eds.), MSZNY 2003 - I. Magyar Számítógépes Nyelvészeti Konferencia [Abstracts of the 1st Hungarian Conference on Computational Linguistics]. Szeged: University of Szeged. (2003) 305–306
7. É. Kiss, K.: A személyragos alaptagú főnévi igeneves kifejezés [Verb phrases with a conjugated infinitival head]. In É. Kiss, K., Kiefer, F. and Siptár, P., Új magyar nyelvtan [A new Hungarian grammar]. Budapest: Osiris. (1999) 118–121