MANÓCSKA: A Unified Verb Frame Database for Hungarian

Ágnes Kalivoda[1,2,3], Noémi Vadász[1,3], and Balázs Indig[1,2]

[1] Pázmány Péter Catholic University
[2] MTA–PPKE Hungarian Language Technology Research Group, Budapest, Hungary
[3] Research Institute for Linguistics, Hungarian Academy of Sciences
{kalivoda.agnes,vadasz.noemi,indig.balazs}@itk.ppke.hu

Abstract. This paper presents MANÓCSKA, a verb frame database for Hungarian. It is called *unified* as it was built by merging all available verb frame resources. To be able to merge these, we had to cope with their structural and conceptual differences. After that, we transformed them into two easy to use formats: a `TSV` and an `XML` file. MANÓCSKA is *open-access*, the whole resource and the scripts which were used to create it are available in a github repository. This makes MANÓCSKA reproducible and easy to access, version, fix and develop in the future. During the merging process, several errors came into sight. These were corrected as systematically as possible. Thus, by integrating and harmonizing the resources, we produced a Hungarian verb frame database of a higher quality.

Keywords: verb frame database, lexical resource, corpus linguistics, Hungarian

1 Introduction

Finding and connecting the arguments and adjuncts to the verb in a sentence is a trivial step for humans during sentence comprehension. For a parser, this task can only be solved using a verb frame database (in other terms, a valency dictionary). Because of their essential role in everyday NLP tasks, numerous lexical resources of this kind have been created, such as VerbNet [11] and FrameNets for several languages [1].

A couple of verb frame databases have been developed for Hungarian as well. However, each one has some weaknesses, first of all, they are not complete and precise enough. Our database, MANÓCSKA[4] is constructed using these already existing verb frame resources, aiming to harmonize them by merging them into a clearly structured, easy-to-use format.

To gain a better understanding of the issues presented in the following sections, let us sketch some important properties of the target language. Hungarian

[4] The resource and a detailed description of its structure can be found at `https://github.com/ppke-nlpg/manocska`.

is an agglutinative language, meaning that most of grammatical functions are marked with affixes (e.g. nouns can be declined with 18 case suffixes). In this way, Hungarian sentences have a relatively free word order. Furthermore, Hungarian is a pro-drop language: several components of a sentence can be omitted if they are grammatically or pragmatically inferable. This makes the corpus-driven analysis of verb valencies quite difficult. Finally, a considerable issue is raised by verbal particles (in other terms, preverbs). These are usually short words (like the ones in phrasal verbs of Germanic languages) which often change the meaning and the valency of their base verbs. By default, the verbal particle is written together with the verb as its prefix. In a lot of contexts, however, it can be detached from the verb and moved to a distant position. This can happen not only in the case of finite particle verbs, but also by infinitives and participles placed in the same clause. Thus, connecting the verbal particles to their base verbs during the parsing process is a task far from trivial.

During our work, we discovered several weaknesses of the original verb frame resources. Some of the errors could be corrected automatically, but most of them had to be corrected manually. This was done by writing the erroneous version and its correction into a separate file as a key–value pair. Thus, our manipulations did not affect the original resources and Manócska remained reproducible. Moreover, the merging process surfaced some theoretical controversies which are worth considering in the future.

The paper is structured as follows. After giving a brief overview about the resources, we discuss the main issues experienced during the merging process. This is followed by presenting the structure of Manócska. After that, we sketch the most important theoretical implications. Our conclusions close the paper.

2 Resources

Manócska contains six language resources, thus it covers all existing verb frame databases for Hungarian, even those which were previously not accessible freely in a database format. Five of them were built upon corpus data (see Table 1). It must be noted that there are considerable conceptual differences between the resources, e.g. regarding the set of verbal particles (see Section 3) or the distinction between arguments and adjuncts (which can be found only in MetaMorpho). We provide a short description about every resource in this section, recognizing their strengths and pointing out their weaknesses.

The name Mazsola refers to two versions of a verb frame database created by Bálint Sass as a part of his PhD dissertation about retrieving verb frames from corpus data [8]. The first version is a paper dictionary of the most frequent arguments and phrases occurring with the verb (Hungarian Verbal Structures) [10] which was produced automatically – using very simple heuristics to prefer the precision over recall –, based on the HNC corpus [12]. The content of the dictionary was manually corrected, but until now it was available only in paper format. The second version is larger, however, it is not reviewed. It is available online[5]

[5] http://corpus.nytud.hu/isz/

Table 1. Corpora used by the corpus-driven resources (third column) which are merged into Manócska. Their sizes are given in tokens, including punctuation marks.

Name of the corpus and its abbreviation	Size (tokens)	Resource using the corpus
Hungarian National Corpus (HNC)	187 600 000	Mazsola (2 versions)
Hungarian Webcorpus (Webcorpus)	589 000 000	Tádé
Hungarian Gigaword Corpus (HGC) v.2.0.3	978 000 000	Particle Verbs
Hungarian Gigaword Corpus (HGC) v.2.0.4	1 348 000 000	Infinitival Constructions

(after a free registration) and contains 28 million syntactically parsed sentences and half a million verbal structures [9]. Although several years have passed since the creation of these resources, no experiment was conducted to compare the two collections, neither in terms of usability nor of experimenting on other, larger corpora with the state-of-the-art tools and automating the correcting process.

The next resource, Tádé[6] is a frequency list of Hungarian verb frames created by spectral clustering [2], but in an unsupervised manner where the frames and their clustering are induced in the same pass [6]. The novelty of the approach lies in the sensitive thresholding technique which yields robust results and enables the inclusion of a broader class of frames which were not considered in the earlier works. The frames were extracted from the Webcorpus [3]. No language-centric tools were used during the creation of this resource, so it has many trivially correctable errors.

There are some notable differences between Mazsola and Tádé[7]. In the case of Mazsola, accuracy was in focus, in contrast with the pursuit to higher F-measure – and consequently higher recall – which can be seen by Tádé. Due to its higher precision, Mazsola is basically more suitable for everyday NLP tasks. It contains also the frequent lexical arguments of verbs which can not be found in Tádé. However, it must also be noted that Mazsola does not contain any infinitival arguments (neither versions), whereas Tádé does.

Beside Tádé, we used two frequency lists which were created by corpus-driven method. The first of them contains 27 091 particle verbs [5] extracted from HGC v2.0.3 [7]. It was checked and corrected manually, aiming for high precision. It does not contain any information about the verb frames, but it has a good coverage of the possible combinations of verbs and their particles including their joint frequency. The second list contains finite verbs which may have infinitives as their arguments[8]. It was extracted from HGC v2.0.4. It does not enumerate all infinitive arguments for each verb lexically (in contrast with

[6] `https://hlt.bme.hu/hu/resources/tade`

[7] Mazsola and Tádé are two puppets from a Hungarian puppet animated film which was popular in the early 1970s. The eponym of our database, Manócska is also a puppet from this film.

[8] `https://github.com/kagnes/infinitival_constructions`

TÁDÉ). Its only goal is to list verb and particle pairs that can have an infinitive as argument.

Last but not least, we included the verb frame database of METAMORPHO, a rule-based commercial machine translation system for Hungarian. This database was created by linguistic experts who aimed to describe Hungarian verb frame constructions in a granularity which was needed for the unambiguous translation to English. Thus, these rules have numerous lexical, syntactic and semantic constraints in order to explicitly isolate the verb senses. The creators used corpora to check their linguistic intuition, however, the database does not contain statistical frequencies. In this way, all rules appear as if they would have the same importance.

The aforementioned resources have different sizes and they are based on different sized corpora. The verb-related properties of the merged resource MANÓCSKA can be seen in Table 2. More than two-thirds of all verbs (33 937 out of 44 183) are present only in one or two used resources which makes the recall of MANÓCSKA really high.

Table 2. The number of frames, different verb lemmata and erroneous verbforms found in the resources. The size of MANÓCSKA is marked with **boldface**.

Resource	Frames	Verbs	Errors
MAZSOLA (dictionary)	6 203	2 185	47
MAZSOLA (database)	524 267	9 589	477
TÁDÉ	521 567	27 159	4 489
Particle Verbs	0	27 091	0
Infinitival Constructions	0	1 507	0
METAMORPHO	35 967	13 772	0
MANÓCSKA	**971 384**	**44 183**	**0**

3 Emerging Issues

In order to be able to merge the resources, we had to harmonize them. We assumed that the weaknesses of the databases will be corrected by the strengths of others. For instance, if a frame has high frequency in multiple independent databases, it can be safely considered a valid frame, while a frame which can be found only in one database with low frequency might be wrong or unimportant. By harmonization we also mean that the different structures and linguistic formalisms of the resources had to be converted into a standard format. During this process, several issues came to light.

Firstly, we had to cope with practical issues, e.g. the undocumented feature set used in METAMORPHO or the numerous verbal particle–verb mismatches (this is caused by the nature of Hungarian verbal particles, see Section 1). We could tackle these using ruled-based methods and manual corrections.

Secondly, we faced some more severe issues which have theoretical background. An interesting example is the fuzzy boundary between the verb modifiers and one of their subclasses, the verbal particles. In Manócska, the latter ones are separated from the verb with a pipe (because – by default – they are written together with the verb). The former ones are handled as lexical arguments, thus they have 'lemma with case marking' form. For example, in the case of *szörnyet|hal*, *szörnyet* (lit. 'monster.ACC') is defined as a verbal particle, while in *hal szörny[ACC]*, it is rather a lexical argument (both constructions mean 'to die on the spot').

Manócska contains 118 entries where a word is handled as verbal particle and as a lexical argument, respectively. There are altogether 33 words which are ambiguous from this point of view. In order to have a better understanding of these words' behaviour, we conducted a case study using HGC v.2.0.4. We looked for clauses 1) where the given word was in -1 position compared to the verb (immediately before it, but separated by a space) and 2) where it was in 0 position (written together with the verb). Orthography, of course, can not lead us to incontestable statements. However, it can show us the native speakers' intuition concerning these ambiguous words. If the word has -1 position, the writer of the clause handled it rather as a lexical argument, while 0 position indicates that it is handled as a verbal particle. Table 3 presents five cases where the orthographical uncertainty is remarkable.

Table 3. Five cases where there is no consensus regarding the category of the ambiguous word. The fourth column (-1) stands for the joint frequency of the given word and the verb, counting the cases when the given word is written separately from the verb. The fifth column (0) shows the number of cases when the given word is written together with the verb.

Ambiguous word	Verb	Meaning of the construction	-1	0
síkra 'plain.SUB'	*száll* 'to fly'	to come out in support of sy	423	320
nagyot 'big.ACC'	*hall* 'to hear'	to be hard of hearing	76	107
cserben 'tan_pickle.INE'	*hagy* 'to leave'	to let sy down	986	1 818
helyben 'place.INE'	*hagy* 'to leave'	to approve smth	986	2 132
véghez 'end.ALL'	*visz* 'to take'	to accomplish smth	1 260	3 054

4 The Structure of Manócska

Manócska is available in two formats: a `TSV` and an `XML` file. In the `TSV`, no distinction is made between arguments and adjuncts, as it does not contain all information that can be found in the MetaMorpho database, and the other five resources do not have this type of information. The `TSV` is easily parsable. Every row corresponds to one entry. The first column contains the verb lemma (the verbal particle is separated by a | character). The second column shows the verb frame which is represented by case-endings (e.g. 'with something' equals [INS], a word in instrumentalis). Columns 3–8 contain the frequencies of the verb frame in the six different resources. In the last column, a rank value can be seen which allows a cross-resource comparison of the given record's frequency.[9]

The XML-format (presented on Figure 1) contains all the six resources, including every fine-grained feature available in the MetaMorpho database (e.g. distinction between arguments and adjuncts – the latter marked with COMPL, information about the valencies' theta roles and semantic constraints like *animate* or *bodypart*). We handle the base verbs as the main elements. Each verb entry (Verb) is split into two optional subentries based on whether there is a verbal particle (Prev) or not (No Prev). Furthermore, each entry is subdivided depending on the possibility of an infinitival argument (Inf, No Inf). We chose these two as primary features, because recent research proved that these are essential features for real-life verb frame disambiguation in the case of Hungarian [4].

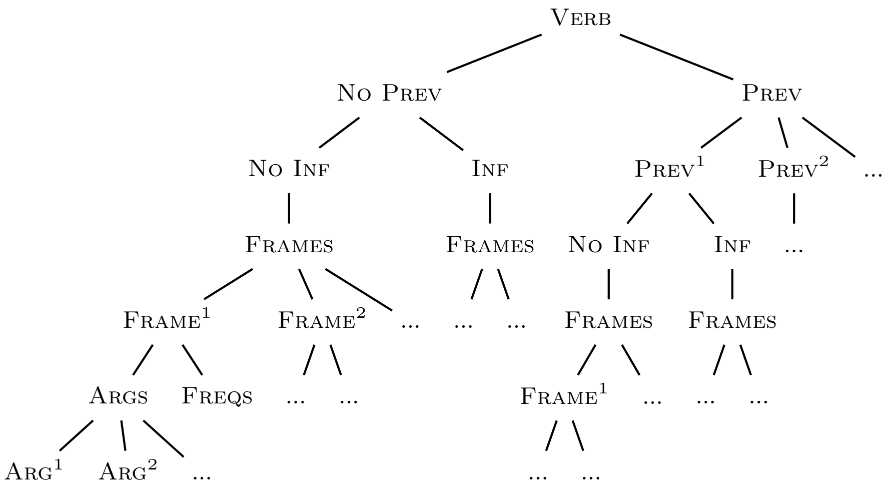

Fig. 1. The basic structure of the XML format.

The possible verb frames are collected within the Frames tag. Each frame can have meta attributes, e.g. a reference to its ID in the original resource. The frames are presented as lists of arguments (subject, object, obliquus) and adjuncts (both types within the Arg tag). Each of these must have a grammatical

[9] The rank value is computed by dividing the actual frame frequency of the given record and the summarized frame frequency for each resource, and finally by summarizing the divisions' results.

case or a postposition. Beside that, they may have extra constraints, e.g. some features which help to disambiguate the frames. We treat each feature as a key–value pair chosen from a predefined domain, presented as the attribute of the given Arg tag.

The frame frequencies coming from the different resources are attributes of the Freqs tag (as key–value pairs, with the key being the name of the resource). This formalism enables the user to easily add other resources in the future, including their own frequencies. The easily extendable, filterable, transformable form in conjunction with the GIT based public versioning and the availability of the production scripts[10] make Manócska a unique, open-access resource.

5 Theoretical Implications

To demonstrate the applicability of our resource, we created a custom naïve 'clustering' of the entries by different features, as we faced that no matter how we order the features in the XML-tree, there will always be many subtrees that are equivalent. We wanted to eliminate these duplicated subtrees and compress the database. This experiment revealed some nice patterns among the frames.

We eliminated all constraints from the arguments except their grammatical cases to achieve higher density. In this reduced "framebank", we looked for duplicate subtrees. Our search was not performed on the frame level, but rather on the level of the different verb–frame, verb–particle–frame combinations. We managed to gather many rather frequent groups of frames that can be paired with the verb or particle they occur with in any desired combination.

We argue that the essence of productivity can be revealed by recurring groups of frames. In a lot of cases, the verb itself can be substituted with several semantically related words, but interestingly, its frames can not vary so freely. This phenomenon becomes even more apparent if the verb has a particle which inherently encodes directionality and demands an argument which agrees with it in its grammatical case. In such structures, the verb seems to have very little syntactic, but rather semantic power in the predicate. For instance, the scheme '*be* (lit. in.ILL) + verb + smth.ACC smth.INS' mostly matches frames where the verb comes from a semantically related class of words having the core meaning 'to cover something with something' (e.g. *befed* 'to cover', *bearanyoz* 'to gild', *bedörzsöl* 'to rub in', *bepiszkít* 'to dirty', *besugároz* 'to irradiate', *beterít* 'to spread').

Another interesting phenomenon comes to light when we look at particle verbs having infinitival arguments. If we know that the particle has inherent directional meaning (e.g. *ki* 'out', *be* 'in', *el* 'away'), we can be almost certain that the verb is a verb of motion. There are only a few exceptions having abstract meaning: *el/felejt* 'to forget smth', *el/kezd* 'to begin smth', *ki/felejt* 'to leave out smth (by mistake)', *ki/próbál* 'to try out smth'. However, if we do not have any

[10] Due to licence reasons, the original resources could not be included but they can be asked for by the original copyright holders at the given addresses.

information about the particle, the chance that the given verb is semantically a verb of motion is only 38% (88 out of 232 verbs).

With the distributive inspection presented above, we can discover the real inner-workings of the verb frames including numerous examples which came from linguistic intuition and introspection along with the ones that maybe slipped our mind.

6 Conclusion

Manócska is a valuable, open-access database of Hungarian verb frames. Its `XML` format makes it possible to handle several built-in resources uniformly, but it is also possible to extract a single resource or a reduced feature set from the `XML`, if this is preferred for a specific task as demonstrated in Section 5.

This database is one step closer to be suitable for a lexical resource of a parser, helping it to connect the arguments to the verb in the right way. Beside everyday NLP tasks, it can be used for linguistic research as well. Due to its reproducibility, Manócska can be improved constantly by correcting previously unnoticed errors or by adding new resources.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1. pp. 86–90. ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998). https://doi.org/10.3115/980845.980860, `https://doi.org/10.3115/980845.980860`
2. Brew, C., Schulte im Walde, S.: Spectral clustering for german verbs. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. pp. 117–124. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002). https://doi.org/10.3115/1118693.1118709, `https://doi.org/10.3115/1118693.1118709`
3. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Calzolari, N. (ed.) Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pp. 203–210 (2004)
4. Indig, B., Vadász, N.: Windows in Human Parsing – How Far can a Preverb Go? In: Tadi, M., Bekavac, B. (eds.) Tenth International Conference on Natural Language Processing (HrTAL2016) 2016, Dubrovnik, Croatia, September 29-30, 2016, Proceedings. pp. (accepted, in press). Springer, Cham (2016)
5. Kalivoda, Á.: A magyar igei komplexumok vizsgálata [The Hungarian Verbal Complexes]. Master's thesis, PPKE-BTK (2016), `https://github.com/kagnes/hungarian_verbal_complex`
6. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting optional arguments of verbs. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA) (2016)

7. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA) (2014)
8. Sass, B.: Igei szerkezetek gyakorisági szótára – Egy automatikus lexikai kinyerő eljárás és alkalmazása [A Frequency Dictionary of Verbal Structures – An Automatic Lexical Extraction Procedure and its Application]. Ph.D. thesis, Pázmány Péter Katolikus Egyetem ITK (2011)
9. Sass, B.: 28 millió szintaktikailag elemzett mondat és 500 000 igei szerkezet [28 Million Syntactically Parsed Sentences and 500 000 Verbal Structures]. In: Tanács, A., Varga, V., Vincze, V. (eds.) XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015) [XI. Hungarian Conference on Computational Linguistics]. pp. 399–403. SZTE TTIK Informatikai Tanszékcsoport, Szeged (2015)
10. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára [Hungarian Verbal Structures – The Dictionary of the Most Frequent Arguments and Phrases]. Tinta Könyvkiadó, Budapest (2010)
11. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, University of Pennsylvania (2006), `http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf`
12. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002) European Language Resources Association, Paris. pp. 385–389 (2002)