MANÓCSKA: A UNIFIED VERB FRAME DATABASE FOR HUNGARIAN

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

[1]Pázmány Péter Catholic University,
[2]MTA–PPKE Hungarian Language Technology Research Group,
[3]Research Institute for Linguistics, Hungarian Academy of Sciences

1. Introduction

- MANÓCSKA is a **unified** verb frame database, built by merging all available verb frame resources for Hungarian. 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.
- 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.

Find MANÓCSKA!

`https://github.com/ppke-nlpg/manocska`

2. Resources

MANÓCSKA uses five resources built upon corpus data. Three of these (**MAZSOLA dictionary** [7], **Particle Verbs** [1] and **Infinitival Constructions** [2]) are manually corrected as well.
The **MAZSOLA database** [6] contains 28 million syntactically parsed sentences and half a million verbal structures. **TÁDÉ** [3] is a frequency list of Hungarian verb frames created in an unsupervised manner.
METAMORPHO [5] was created manually, by linguistic experts. Its frames contain numerous lexical, syntactic and semantic constraints in order to explicitly isolate the verb senses.

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**

MANÓCSKA's coverage is really high, while only a few verbs can be found in the intersection of the resources.

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

3. Emerging Issues

Practical issue No. 1: the undocumented feature set used in METAMORPHO.
Practical issue No. 2: the numerous verbal particle–verb mismatches.
These could be tackled using ruled-based methods and manual corrections.

An important theoretical issue is the fuzzy boundary between the verb modifiers and one of their subclasses, the verbal particles (also known as preverbs).

In MANÓCSKA, verb modifiers are stored in the ARG entries, while verbal particles are placed in PREVs. The table below shows five cases where there is no consensus regarding the category of the ambiguous word. The case study was conducted using the Hungarian Gigaword Corpus, v.2.0.4 [4].

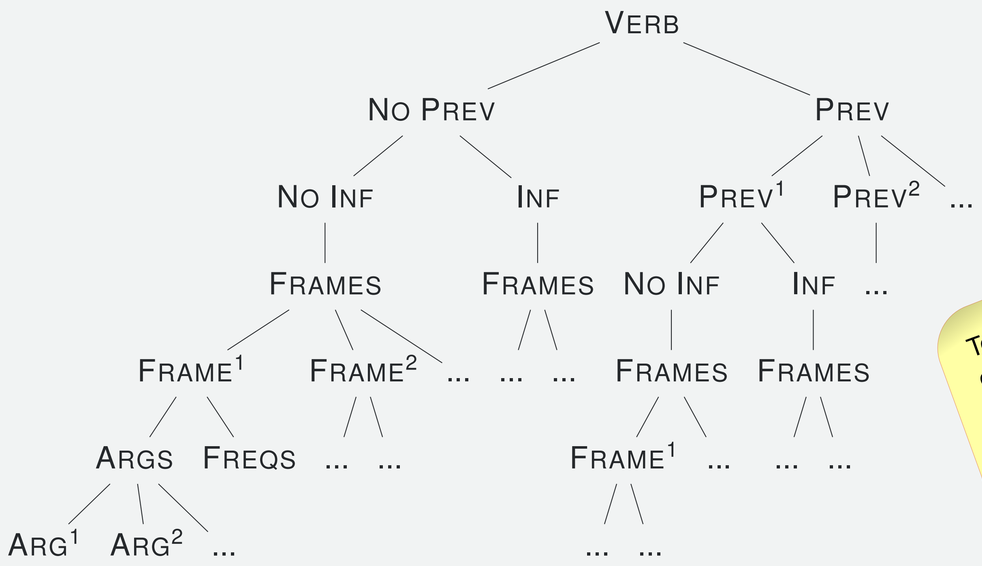
Ambiguous word	Verb	Meaning of the construction	-1	0
síkra 'plain.SUB'	*száll* 'to fly'	to come out in support of sy	423	320
nagyot 'big.ACC'	*hall* 'to hear'	to be hard of hearing	76	107
cserben 'tan_pickle.INE'	*hagy* 'to leave'	to let sy down	986	1 818
helyben 'place.INE'	*hagy* 'to leave'	to approve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

4th column (-1): frequency of the given word and the verb written as distinct words
5th column (0): frequency of the given word and the verb written as one word

4. The `xml` format of MANÓCSKA

The `XML`-format contains all the six resources, including every feature available in the METAMORPHO database (e.g. distinction between arguments and adjuncts, information about the valencies' theta roles and semantic constraints like *animate* or *bodypart*).

- VERB: the base verb (the main element)
- PREV: holds together particle verbs formed from the base verb
- NO PREV: needed if the base verb stands without verbal particle
- INF and NO INF: these show whether an infinitival argument is possible
- FRAMES: contains the possible verb frames
- each FRAME is presented as a list of arguments and adjuncts (both types stand within the ARG tag)
- FREQS: frame frequencies coming from the different resources



To compress the emerging equivalent subtrees we created a naïve clustering on them, see below.

5. Theoretical Implications

We created a custom naïve clustering of the entries in the following way:

1. we eliminated all resource dependent constraints from the arguments except their grammatical cases to achieve higher density with less but more standard frames
2. in this reduced "framebank", we looked for "duplicate subtrees"
3. we focused on the different verb–frame, verb–particle–frame combinations to uncover their common features like frequent patterns and semantic productiveness

As a result, the experiment revealed some interesting patterns among the frames.

6. Example for a pattern found by the clustering

The scheme '*be* (in.ILL) + verb + smth/sy.ACC smth.INS' mostly matches frames where the verb comes from a semantically related class of words having the core meaning 'to cover something/somebody with something'.

be|fed 'to cover'	*be|piszkít* 'to dirty'
be|aranyoz 'to gild'	*be|sugároz* 'to irradiate'
be|dörzsöl 'to rub in'	*be|terít* 'to spread'

In such structures, the verb seems to have very little syntactic, but rather semantic power in the predicate.

[1] Ágnes Kalivoda. *A magyar igei komplexumok vizsgálata*. `https://github.com/kagnes/hungarian_verbal_complex`. 2016.
[2] Ágnes Kalivoda. *Infinitival Constructions in Hungarian*. `https://github.com/kagnes/infinitival_constructions`. 2017.
[3] András Kornai, Dávid Márk Nemeskey and Gábor Recski. 'Detecting Optional Arguments of Verbs'. In: *LREC 2016*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. ISBN: 978-2-9517408-9-1.
[4] Csaba Oravecz, Tamás Váradi and Bálint Sass. 'The Hungarian Gigaword Corpus'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2014.
[5] Gábor Prószéky, László Tihanyi and Gábor Ugray. 'Moose: a robust high-performance parser and generator'. In: *Proceedings of the 9th EAMT Conference. La Valletta: Foundation for International Studies*. Ed. by J. Hutchins. 2004, pp. 138–142.
[6] Bálint Sass. '28 millió szintaktikailag elemzett mondat és 500 000 igei szerkezet'. In: *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015) [XI. Hungarian Conference on Computational Linguistics]*. Ed. by Attila Tanács, Viktor Varga and Veronika Vincze. Szeged: SZTE TTIK Informatikai Tanszékcsoport, 2015, pp. 399–403.
[7] Bálint Sass et al. *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára*. Budapest: Tinta Könyvkiadó, 2010.

This research has been supported by the European Union, co-financed by the European Social Fund EFOP-3.6.3-VEKOP-16-2017-00002.