# Extracting Idiomatic Hungarian Verb Frames

**Bálint Sass**

Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest
joker@nytud.hu

## Abstract

We describe a machine learning method for collecting idiomatic fixed stem verb frames. Firstly we collect frequent frame candidates from the output of a partial parser, secondly we apply a certain idiomaticity metric to the list to get the most idiomatic frames. The extracted frames will be translated to English and used as a resource in a Hungarian-to-English machine translation system.

## 1 Introduction

*Verb frame*: how many and what kind of NPs can or must appear together with a given sense of a given verb.

*Generalized* verb frame: frame without the verb.

*Position* within a frame is defined by . . .
- word order
- preposition / casemark / postposition

*Fixed-stem verb frame*: such frames where only one (or at most few) defined stem can appear at a given position of the frame.

For example in English frames to take stock of sg and to take sg into consideration, the fixed stem is stock in 'object' position and consideration in 'into' position respectively.

Hungarian is a highly inflectional free-word-order language. Verbs indicate explicit case markings for their complements. These case markings appear on the head of the NPs.

| | | | | | |
|---|---|---|---|---|---|
| János | szereti | Mari**t**. | John | loves | Mary. |
| S | V | O | S | V | O |
| János**t** | szereti | Mari. | Mary | loves | John. |
| O | V | S | S | V | O |

Figure 1: Word order vs. casemarks: the **-t** ending marks the object

There are many papers dealing with English frames [1, 2, 3, 4], but only few dealing with a free-word-order language [5, 6]. We build our system mainly on methods described in [5].

For translation purposes it is important to have those frequent fixed stem frames where where the meaning is not compositional or more importantly the translation is special.

What we need is *idiomatic fixed stem verb frames*.

## 2 Collecting Verb Frames

The corpus and parser we use are described in detail in [7] (in Hungarian).

The corpus is a special 10 million word subcorpus of the morphosyntactically annotated Hungarian National Corpus (HNC) [8], with short, hopefully one-frame sentences.

Our parser is a partial parser implementing a *cascaded regular grammar* engine. We use it with a simple grammar to dertermine NPs. After that we identify verb stems attaching separate verbal affixes to the stem. We cut off the most frequent deverbal verb suffix (-hat/het).

We generate a list of verb frame candidates with the so-called optionalization method.

In a sentence every NP is . . .
- either taken into account as bare case of the head;
- or as stem + case of the head;
- or omitted.

| | | |
|---|---|---|
| A polgármesteri hivatal**t** | bér**be** adták | a filmesek**nek**. |
| A polgármesteri hivatal**ACC** | bér**ILL** ad-ták | a filmesek**DAT**. |
| The town hall**ACC** | **into** payment give-PLUR3 | film-makers**DAT**. |
| The town hall | was let to | film-makers. |

Figure 2: An example sentence.

| | |
|---|---|
| ad | bérILL fi lmesDAT hivatalACC |
| ad | bérILL fi lmesDAT ACC |
| ad | bérILL fi lmesDAT |
| ad | fi lmesDAT hivatalACC ILL |
| ad | fi lmesDAT ACC ILL |
| ad | fi lmesDAT ILL |
| ad | fi lmesDAT hivatalACC |
| ad | fi lmesDAT ACC |
| ad | fi lmesDAT |
| ad | bérILL hivatalACC DAT |
| ad | bérILL ACC DAT |
| ad | bérILL DAT |
| ad | hivatalACC DAT ILL |
| ad | ACC DAT ILL |
| ad | DAT ILL |
| ad | hivatalACC DAT |
| ad | ACC DAT |
| ad | DAT |
| ad | bérILL hivatalACC |
| ad | bérILL ACC |
| ad | bérILL |
| ad | hivatalACC ILL |
| ad | ACC ILL |
| ad | ILL |
| ad | hivatalACC |
| ad | ACC |

Figure 3: Frame candidates of sentence on Fig. 2.

We collect all candidates and apply a simple frequency threshold (of 5) to the list.

Thank to optionalization . . .
- we get the same true frame from different sentences, where different adjuncts appear besides the same frame;
- we get rid of such adjuncts automatically, which appear in many different ways eg. as different cases.

Applying this method only the frame ad bérILL ACC DAT remains showing that this frame has one fixed and two free positions.

## 3 Considering Idiomaticity

Using the above method, there are times when we get a fixed stem frame just because the stem is frequent enough in a particular case, without having a special, idiomatic role.

To filter out such frames we apply an idiomaticity metric called *distributed frequency (DF)* based on [9] in a second step. In short, according to this metric a given frame is more idiomatic if its generalized frame is used with only few verbs, most idiomatic frames are used with only one verb.

More precisely, if a given generalized frame (**g**) appears with $n$ different verbs ($V_{1..n}$) more frequently than a threshold of 5 and the frequency of $(V_k, \mathbf{g})$ frames is $F_k$, then the formula for calculating *DF* for this generalized frame looks as follows:

$$DF(\mathbf{g}) = \sum_{k=1}^{n} \frac{F_k^a}{n^b}$$

The original paper deals with verb-object relation. We must apply the metric to verbs and generalized frames. We simply get the generalized frame as a string and then apply the method. The representation of the generalized frame of eg. add bérILL ACC DAT will be "bérILL ACC DAT" as a string.

In the paper in question nothing is said about how to assign a *DF*-value to different verb–frame pairs, they – because of the same generalized frame – seem to have the same *DF*-value. To prefer such pairs where the verb is more frequent, we multiply the *DF*-value with the relative frequency of verb within the generalized frame, so we define our eventual idiomaticity metric (so-called *DF-score*) as follows:

$$\textit{DF-score}(V_k, \mathbf{o}) = DF(\mathbf{o}) \cdot \frac{F_k}{\sum_{i=1}^{n} F_i}$$

Setting a certain threshold on *DF-score* we create a final list of 10000 verb frames. We can say that these are the most idiomatic fixed stem verb frames.

## 4 Case study

The generalized frame példaACC appears with many (namely 24) verbs, so these not very idiomatic frames are omitted: mond példaACC/to say example, vesz példaACC/to take example, említ példaACC/to mention example, mutat példaACC/to show example. Conversely, the generalized frame példaACC DAT appears only with one verb: mutat példaACC DAT[to show example for sy]/to set example for sy. As a result of idiomaticity-filtering mutat példaACC appearing 49 times is filtered out, and the idiomatic mutat példaACC DAT appearing only 13 times gets into the final verb frame list.

## 5 Pilot evaluation

Pilot-measurements on precision and recall of the idiomaticity-filtering step.

We manually annotated the frames, "which must be translated some special way" in some parts of the raw list generated by the collecting step. With this definition of goodness we get result seeming rather bad (precision ranges from 12 to 75% and recall ranges from 46 to 81%).

Comparison to an existing authoritative verb frame source.

We contrast manually a small sample of 17 frames from our final list with the Hungarian Concise Dictionary. There are 15 frames in the written dictionary, from which we found 5, so at first sight the recall to the dictionary is bad (5/15=33%). Conversely, it turns out that from the 17 frames found, 14 are true frames, so our method presents 9 new frames not appearing in the dictionary.

## Conclusion and future work

As a result of two level filtering described above we get a list of 10000 verb frames, which seems to be good enough to be the basis of a key lexical resource in a Hungarian-to-English MT system being prepared.

The parser itself needs improvement to be able to parse complex sentences, moreover a better grammar implementing a full-featured Hungarian NP-grammar should be used.

The binomial filtering method described in [1] can be tested for getting rid of frames which only occurs by error.

If we want to measure idiomaticity of free stem frames too, an other, more sensitive idiomacity metric should be worked out, possibly using az automatically acquired thesaurus like in [4].

## References

[1] Brent, M.: From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics **19**(2) (1993) 243–262

[2] Manning, C.D.: Automatic acquisition of a large subcategorization dictionary from corpora. In: Proceedings of the 31st Meeting of the Association for Computational Linguistics, Columbus, Ohio (1993) 235–242

[3] Briscoe, T., Carroll, J.: Automatic extraction of subcategorization from corpora. In: Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97), Washington, DC (1997)

[4] McCarthy, D., Keller, B., Carroll, J.: Detecting a continuum of compositionality in phrasal verbs. In: Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan (2003) 73–80

[5] Zeman, D., Sarkar, A.: Learning verb subcategorization from corpora: Counting frame subsets. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000), Athens, Greece (2000)

[6] Kis, B., Villada, B., Bouma, G., Ugray, G., Bíró, T., Pohl, G., Nerbonne, J.: A new approach to the corpus-based statistical investigation of hungarian multi-word lexemes. In: Proceedings of the 4rd International Conference on Language Resources and Evaluation (LREC2004). Volume V., Lisbon, Portugal (2004) 1677–1681

[7] Sass, B.: Vonzatkeretek a Magyar Nemzeti Szövegtárban [Verb frames in the Hungarian National Corpus]. In: Proceedings of the 3rd Magyar Számítógépes Nyelvészeti Konferencia [Hungarian Conference on Computational Linguistics] (MSZNY2005), Szeged, Hungary (2005) 257–264

[8] Váradi, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389

[9] Tapanainen, P., Piitulainen, J., Järvinen, T.: Idiomatic object usage and support verbs. In: Proceedings of the 17th COLING – 36th ACL, Montreal, Canada (1998) 1289–1293