FDVC – Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions for Hungarian

Sass Bálint
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
sass.balint@nytud.hu
Pajzs Júlia
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
pajzs@nytud.hu

We present a method for creating a special dictionary which ...
— is a corpus-driven [1], frequency dictionary;
— has *verb phrase constructions* (VPCs) as entries;
— is a *meaningless* dictionary in the sense of [2];
— is for Hungarian but the core methodology is *language independent*;
— is created in a *mostly automatic* way with less manual lexicographic work;
— can be created with a low budget;
— is hoped to be useful in language teaching and NLP both.

We follow the sinclairian approach of corpus-driven lexicography. We take a corpus and "jettison ruthlessly" [3] all verbs and constructions which have zero or low frequency in our corpus. In other words, we take the data from the corpus as is, and we do not allow the lexicographer to add any "missing" constructions. Knowing that "authenticity alone is not enough: evidence of conventionality is also needed" [3] we take the most frequent VPCs into account and record and display their frequency in the dictionary. We focus on frequent patterns and do not "seek to cover all possible meanings and all possible uses" [4].

The target is all Hungarian VPCs (consisting of a verb plus some NP or PP dependents) from verb subcategorization frames through light verb constructions to completely rigid figures of speech. The so called *multiword verbs* (e.g. *to take sg into account* or *to get rid of sg*) are at the heart of our approach. Having *fixed* (like the object *rid* above) and *free* (like the *of* position above) positions both they are borderline cases between verb subcategorization frames and real multiword expressions. Contrary to common intuition, they are expressly frequent, they can not be treated marginally. If we take the fixed position as a part of the multiword verb itself, we can treat simple and multiword verbs the same way: both can have some free positions beside. Entries in FDVC are VPCs, the microstructure apparently integrates phraseology as the a basic units are phrases. We arrange the VPCs around a verb in a subsequent step to form more traditional dictionary entries.

On the one hand, the FDVC can be called a "meaningless dictionary". It does not contain explicit definitions, just enumerates the frequent VPCs together with corpus frequencies. Most dictionary users are looking up only basic information, for these tasks meaningless dictionaries are efficient [2]. On the other hand, it contains also a corpus sentence examplifying the meaning. Furthermore, this meaning is fairly concrete, as VPCs (being collocations) usually have one and only one meaning [5]. In fact, most VPCs are real constructions, "form and meaning pairings" [6], as they cannot be broken down into smaller units without loss of meaning. Each VPC represent a pattern of use, and can be paired with one sense of its main (simple or multiword) verb.

The dictionary creation process is mostly automatic: starting from the morphosyntactically tagged and disambiguated Hungarian National Corpus (HNC) [7] we obtain raw dictionary entries using some NLP tools; only the final editing step is manual lexicographic work. We proceed the following way:

HNC →

1. chunking to have verbs and NP/PP dependents;

   → corpus (with richer annotation) →

2. an algorithm based on cumulative frequency of corpus patterns to collect frequent VPCs, with appropriate treatment of fixed and free positions (details and evaluation can be found in [8]); and another algorithm to collect suitable examples for VPCs;

   → frequent VPCs →

3. manual lexicographic work: error correcting and example selecting

   → dictionary

It should be emphasized that step 2 supersedes a substantial amount of manual lexicographic work. As a result of this step VPCs (arranged around verbs) are presented in XML form, so the lexicographer can edit the entries in an XML editor. In step 3 he/she basically has to check if the patterns suggested by the program are correct, and to choose among the example sentences the most appropriate ones. When doing this, the suggestions made in [9] are taken into account (choosing full-sentence examples, or at least clauses with full predicate structure, avoid personal names etc.). Sometimes none of the example sentences are correct or appropriate for illustration, in this cases other ones are retrieved from the HNC by a special corpus query system [10]. In this form, the task of the lexicographer is considerable easier and the result needs much less corrections than before.

It should be noted, that the algorithms in step 2 are language independent, so the methodology can be applied to any language, if we have a POS tagger and a suitable chunker. This methodology allows creating smaller budget dictionaries as the programming and support costs (step 1 and 2) are estimated to 1 man-year, and the lexicographic work (step 3) is also about 1 man-year for a dictionary containing about 3000 verbs and 8000 VPCs altogether.

Beside the traditional (alphabetically ordered by verb) presentation we plan to have several indexes. All of them can be generated automatically:

- aggregated list of all VPCs sorted by frequency – in fact this is the true FDVC;
- an index by general patterns (i.e. VPCs without the verb);
- an index by number of fixed/free positions;
- a frequency list of verb stems;
- an index by lemmas in fixed positions.

Here is an example entry for the verb *elver* (*to beat*) in XML form. It is in the stage after step 2 (amended by manually choosing one corpus example from the auto-generated ten for each VPC).

```
<entry>
<verb lemma="elver" freq="744"/>
<pattern freq="284">
<frame><p c="-t"/></frame>
<type str="1:01" len="1" fixed="0" free="1"/>
<cits>
  <cit>hogy minap elvertelek azért,</cit>
</cits>
  <pattern freq="36">
  <frame><p c="" l="jég"/><p c="-t"/></frame>
  <type str="3:11" len="3" fixed="1" free="1"/>
  <cits>
    <cit type="sentence">Már ahol a jég nem verte el a termést!</cit>
  </cits>
  </pattern>
</pattern>
<pattern freq="95">
<frame><p c="-n"/><p c="-t" l="por"/></frame>
<type str="3:11" len="3" fixed="1" free="1"/>
<cits>
  <cit type="sentence">vagy hogy egy pár túlbuzgó
                       helyi tanácselnökön verjék el a port.</cit>
</cits>
</pattern>
</entry>
```

The corresponding dictionary entry showing the most important three verb phrases constructions of this verb is:

> *elver* `[744]`
> *elver* -t `[284]` hogy minap elvertelek azért, ...
>   *elver* **jég** -t `[36]` Már ahol a jég nem verte el a termést!
> *elver* -n **por**-t `[95]` vagy hogy egy pár túlbuzgó helyi tanácselnökön verjék el a port.

English translation of the entry:

> *beat* `[744]`
> *beat* OBJECT `[284]` that I beat you yesterday, because ...
>   *beat* **ice** OBJECT `[36]` Just where the hail did not destroy the crop!
> *beat* ON **dust**-OBJECT `[95]` or to blame some overzealous local mayors.

Verb phrase constructions are translated word by word while example sentences have overall translations, so it can be seen that when *hail destroys* something Hungarians say *the ice beats* it; and *to blame sy* is put in Hungarian something like *to beat the dust on sy*.

We described the creation of a Corpus-driven Frequency Dictionary of Verb Phrase Constructions (FDVC) for the Hungarian language. We collected automatically all VPCs from corpus, and presented them to the lexicographer in a convenient XML form, significantly reducing the manual lexicographical work this way. Core algorithms are language independent. Using this methodology we can obtain a lexical database, which is at first a good learner dictionary which lists all frequent VPCs and “helps students to write and speak idiomatically” [3]. Beyond that, it is a rich lexical resource from which many NLP tasks could benefit from information retrieval to machine translation.

References

[1] Tognini-Bonelli, E.: Corpus Linguistics at Work. John Benjamins (2001)

[2] Maarten, J.: Meaningless dictionaries. In: Proceedings of the XIII EURALEX International Congress, Institut Universitari de Linguistica Aplicada, Universitet Pompeu Fabra, Barcelona (2008) 409–420

[3] Hanks, P.: The lexicographical legacy of John Sinclair. International Journal of Lexicography **21**(3) (2008) 219–229

[4] Hanks, P.: The probable and the possible: Lexicography in the age of the internet. In: Proceedings of AsiaLex 2001, Yonsei University, Seoul, Korea (2001)

[5] Yarowsky, D.: One sense per collocation. In: Proceedings of the workshop on Human Language Technology, Princeton, New Jersey (1993) 266–271

[6] Goldberg, A.E.: Constructions at Work. Oxford University Press (2006)

[7] 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

[8] Sass, B.: A unified method for extracting simple and multiword verbs with valence information and application for Hungarian. In: Proceedings of RANLP 2009, Borovets, Bulgaria (2009) (to appear)

[9] Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychly, P.: GDEX: Automatically finding good dictionary examples. In: Proceedings of the XIII EURALEX International Congress, Institut Universitari de Linguistica Aplicada, Universitet Pompeu Fabra, Barcelona (2008) 425–432

[10] Sass, B.: The Verb Argument Browser. In: Sojka P. et al. (eds.): 11th International Conference on Text, Speech and Dialogue. LNCS, Vol. 5246., Brno, Czech Republic (2008) 187–192