

Frequency Dictionary of Verb Phrase Constructions

An automatic lexical acquisition method
and its applications

theses of the Ph.D. dissertation

Bálint Sass

supervisor:
Gábor Prószéky, D.Sc.

Pázmány Péter Catholic University,
Faculty of Information Technology,
Multidisciplinary
Technical Sciences
Doctoral School



Budapest, 2011.

Introduction

'*Részt vesz vmiben.*' (take part in sg) '*Górcső alá vesz vmit.*' (take sg under '*górcső*' = examine sg) Although verb subcategorization frames and multiword expressions are a separate field of research both in natural language processing and lexicography, there are such complicated constructions in several languages which are verb subcategorization frames and collocations *at the same time*. These constructions consist of (at least) two content units – normally a verb and a nominal (with casemark/postposition/preposition) –, and additionally one (or more) valences are also inherent part of the construction. In addition to the above Hungarian examples similar constructions can be found in several languages indeed: '*get rid of*' (English), '*få lov til*' (Danish; get permission to sg), '*imati pravo na*' (szerb; have the right to sg), '*houden rekening met*' (holland; take sg into consideration), '*zijn van toepassing op*' (holland; concern sg), '*avoir effet sur*' (francia; have effect to sg).

In the examples mentioned above, there are always two dependents: one of them is filled by a concrete, fixed word constituting a collocation with the verb, while in case of the other dependent, only its place is located by a casemark or a preposition. It can be seen that the dependents are usually connected to the verb by the same linguistic tools (casemarks, postpositions, prepositions or word order constraints); regardless of the fact that the dependent word is a fixed collocate or just an accidental word filling in a valence slot. In the '*részt vesz vmiben*' construction the object is a collocate (marked by the '*-t*' casemark),

while in the *'górcső alá vesz vmit'* the object is a valence slot. This alternation occurs also among the constructions of the same verb. The *'pillantást vet vkire'* (cast a glance at sy) and the *'szemére vet vmit'* (upbraid sy with sg) constructions consist of an object and a dependent with *'-ra/-re'* casemark equally, but in the first case the object is the collocate and the dependent with *'-ra/-re'* casemark is the valence slot, and in the other case just vice versa.

Such constructions – although our intuition as a native speaker tells us the contrary often – are expressly frequent, they constitute an important segment of the constructions of a language, they cannot be treated marginally. They have non-compositional, idiomatic meaning often. Accordingly, they must be included in dictionaries and in language resources of automatic natural language processing tools both. In most cases, it is worth storing their translations as a separate unit, because these translations often contain not predictable elements.

There is a need for a data driven computerized method which makes order in the overlapping system of relation markers, and separates dependents containing a fixed word from dependents which can be filled freely. A method that discovers which word is an integral part of a given verb phrase construction as a collocate, and what necessary valence slots are connected to the construction besides. In one word, a method that is able to extract typical verb phrase constructions from corpus. Main result of the dissertation is this method (section 3.3 of the dissertation), and the monolingual Hungarian verb phrase construction dictionary (section 4.2 of the dissertation) which is prepared using this method.

The dictionary – which is based on the simplest model of verb phrase constructions – makes the usefulness of the extraction method tangible. But what gives the real significance of the method is it can be extended in several directions. Firstly, because the model is language independent, after appropriate language specific preprocessing the extraction method can be applied to several languages without modification, therefore, similar dictionaries can be prepared for various languages. Secondly, the method is able to handle more complicated

constructions (see for example '*gyenge lábakon áll*' (stand on weak legs = weak) which contains an additional adjectival collocate compared to the above constructions), and also noun-centered, adjective-centered constructions and so on. Thirdly, applying the model in a special way, the same mentioned lexical acquisition method can be made to handle parallel verb phrase constructions, namely verb phrase constructions and their translations. Such a way, the method can discover parallel construction pairs which are asymmetric, that means the two parts corresponds each other but formally totally different.

Methodology

One of the central issues in present-day computational lexicography is the question how much of the traditional manual work can be taken over by computers, how far lexicography can get with purely *automatic* tools. The *corpus* is the resource from which the material of a dictionary can be collected automatically. In my research, I follow the strictly *corpus-driven* approach. I use the corpus not only as an aid, but taking the corpus authentic and representative, I derive the full linguistic information about verb phrase constructions solely from corpus data. During the corpus-driven dictionary material collection the typical verb phrase constructions are determined automatically, and – based on corpus frequency – a part of them are chosen automatically to include into the dictionary. Present-day huge corpora provide a solid basis for characterizing rarer phenomena too.

In recent decades, results of corpus-driven lexicography revolutionized the preparation of the dictionaries in many ways. One important result is that the relevance of *multiword lexical units* – collocations, phrasemes, idiomatic or institutionalized expressions – is recognised, and such expressions gain more and more pronounced presence in new dictionaries. As Sinclair said “many, if not most, meanings require the presence of more than one word for their normal realisation.” In my research, I treat the formally different constructions in a unified framework, whether they are single-word or multiword units, verbs or verb phrase constructions. During the dictionary creation process I put the multiword verb phrase constructions as full-fledged

lexemes in the focus of my approach as the examples mentioned in the introduction shows. My *type independent* approach allows to represent the complete verb phrase constructions always, that is no unit is left out which is relevant in terms of the construction. Completeness of constructions is also an important requirement during the evaluation.

The Hungarian language has free word order, at least in the sense that the verb and the dependents can occur almost in arbitrary order in the sentence, with possible intermittent units. In other words: verb phrase constructions can be continuous or non-continuous, they can occur in any order variant. Dealing with Hungarian, we can handle word order variability efficiently if we choose a linguistic framework which fits the nature of the language well, namely *dependency grammar*. In dependency grammar analysis basic units are usually words. In contrast, in my research I have chosen the *morpheme* as basic unit to be able to interpret bound morphemes expressing the relation between the verb and the dependent (namely casemarks) as independent units in addition to words. During collecting the typical verb phrase constructions, I do not follow the usual approach which pays only attention whether two words are next to each other or not, but in our case elements of a verb phrase construction are always in a particular *dependency relationship* with each other. These dependency relationships themselves become full-fledged elements of verb phrase constructions, thereby the mentioned unified framework extends also to verb phrase constructions without a collocate, including verb subcategorization frames.

New Scientific Results

The topic of the dissertation is extracting typical verb phrase constructions from corpus. We focus primarily on constructions which are multiword units and subcategorization frames at the same time, namely complex verb with valences. Such constructions are for example *'hasznot húz vmiből'* (pull benefit from sg = benefit from sg), *'igényt tart vmire'* (lay claim to sg) or *'lehetővé tesz vmit'* (make sg possible). These constructions contain a lexically free dependent (LFD) (*'vmiből'* (from sg), *'vmire'* (to sg), *'vmit'* (sg)), and a lexically fixed dependent (LXD) (*'hasznot'* (benefit), *'igényt'* (claim), *'lehetővé'* (possible)) too.

First task was to develop a model for Hungarian which can represent all types of verb phrase constructions including the above mentioned type. The solution is a special graph representation based on dependency analysis.

Shaping this model is covered in section 2.1 in the dissertation, new results can be summarized as follows:

Thesis 1

I developed a model for the Hungarian language which is able to uniformly represent clauses and also formally very different verb phrase constructions inherent in clauses. Basic unit of representation is the clause that is a central verb and its dependents together. Dependents are represented by their most important content unit (the head of

the phrase in case of nominal phrase dependents) and the relation marker connecting the dependent and the verb (a casemark or a postposition in case of nominal phrase dependents). To sum up:

clause = verb + set of dependents

dependent = relation marker + content unit

Publications related to the thesis:

(Sass, 2009c), (Sass, 2009a), (Sass, 2008), (Sass, 2005)

The model can be depicted graphically by a 1-level deep dependency tree at best. The root is the verb, edges are the relation markers, and vertices are the content units. The general dependency tree corresponding to the model can be seen in Fig. 1, and also the concrete representation of one of the above constructions.

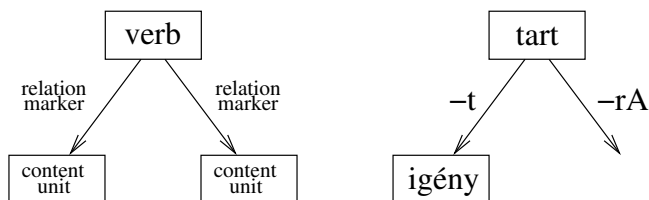


Figure 1: Visualization of the model by means of dependency tree. The general dependency tree corresponding to the model can be seen on the left side with relation markers and content units. In turn, a concrete construction can be seen on the right side, namely the *'igényt tart vmire'* (lay claim to sg). The arbitrary content unit which occurs at the *'-rA'* LFD is not part of this construction.

— • —

The next question is if we take a corpus, how its representation according to the above model can be worked out. Naturally, this representation can be derived from a dependency treebank, the other possibility

is to run a dependency parser on a POS tagged corpus. There is no available dependency treebank for Hungarian of appropriate size, and also no dependency parser developed yet. My dissertation does not cover the development of a Hungarian dependency parser (it could be the topic of another dissertation), but for my research I need a large corpus equipped with good quality representation according to my model.

I have chosen the 187 million word Hungarian National Corpus as a representative Hungarian corpus, and investigated whether the suitable representation can be produced using a simple rule based approximate method. It turned out that the clause boundary detection and the partial shallow syntactic parsing of clauses (essentially identification of verbs and nominal dependents) can be done in appropriate quality in a rule based way.

The processing of the corpus is discussed in section 2.2 in the dissertation, the moral of this section is uttered in the following thesis:

Thesis 2

I showed that starting from POS tagged and disambiguated corpus a reliable model based representation can be produced, using rule based clause boundary detection and rule based shallow syntactic parsing with a relatively simple set of rules.

Publications related to the thesis:
(Sass, 2006b), (Sass, 2005)

Although, in the future the quality of the representation can be improved using a complete dependency parser, it is good enough in its current state to be the starting point for further research.



The resulting representation is in itself a valuable resource. As a special corpus it opens the door to different queries which are unusual

in corpus query systems: we can prescind from the word order, and investigate verb phrase constructions uniformly, independently from their actual word order. Therefore, I developed the Verb Argument Browser corpus query system which is suitable for investigating typical dependents occurring along verbs and verb-noun collocations. This tool displays the characteristic words occurring as a given dependent, together with the appropriate corpus examples.

Basically, the Verb Argument Browser provides two kinds of typical dependents. On the one hand, frequent words with „literal meaning” which often constitute a semantically coherent class; such as the different kinds of food appearing as direct objects of ‘to eat’. On the other hand, frequent words which is part of an idiomatic, complex verb or locution; such as ‘*kása*’ (mush) as the object of ‘*eszik*’ (eat) which is not here because it is a typical food nowadays, but due to the fact that it constitute a saying with the verb: ‘*nem eszik olyan forrón a kását*’ (the mush is not eaten that hot = wait a minute!).

The Verb Argument Browser is described in section 3.2 in the dissertation, traits of it are worded in the following thesis:

Thesis 3

I created the Verb Argument Browser special corpus query system. It can be used to map the dependent structure of verbs, or to identify the essential dependents of verb or verb frames, complex verbs included. The Browser is a useful tool in corpus linguistics research, manual building of lexical resources, and when authentic examples are needed for some verb phrase constructions.

Publications related to the thesis:

(Sass és Pajzs, 2010b) (Sass, 2009b) (Sass, 2008) (Sass, 2006b)

The system can be applied to any corpus if it is equipped with the representation according to the model. The query interface of the original Hungarian version which includes the whole material of the

Hungarian National Corpus is freely available at <http://corpus.nytud.hu/vab>. It can be tried by the `vendeg` temporary user name and `mazsola` temporary password. Searching in a hundred-million word corpus response times are just a few seconds.



Present-day corpora have reached the magnitude where beside manual query tools there is a need for automatic tools which sum up the linguistic information available in corpora. From this viewpoint the Verb Argument Browser is a manual tool, it can present typical words filling in concrete dependent slots.

Main result of my dissertation is the automatic method that goes one very important step further: using a corpus it is able to determine which are the typical verb phrase constructions of a verb *at all*. It is able to determine „what are the relevant queries”, and „runs” these queries moreover. Thereby, we can collect all typical verb phrase constructions containing a given verb.

Detailed presentation and evaluation of the algorithm can be found in section 3.3 in the dissertation, its essence is summed up in the next thesis:

Thesis 4

I worked out a lexical acquisition method which is based on adding up frequencies of sentence skeletons in a special way. This method is capable of extracting characteristic verb phrase constructions of different kinds from a corpus which is represented according to the model (Thesis 1).

Publications related to the thesis:

(Sass, 2010d), (Sass és Pajzs, 2010b), (Sass, 2009c)

The novelty of the method lies in two facts. On the one hand, it adapts to the length (the number of units) of the verb phrase construction

resulting in expressions consisting of two and even more units. On the other hand, it is able to discover that in case of a given verb and a given dependent, only the relation marker is relevant (LFD), or the relation marker together with the concrete content unit (LXD). Accordingly it provides constructions containing LFDs and LXDs, or even both of them mixed. The examples mentioned at Thesis 1 belong to the latter group, they are complex verbs with valences: *'hasznot húz vmiből'* (pull benefit from sg = benefit from sg), *'igényt tart vmire'* (lay claim to sg) and *'lehetővé tesz vmit'* (make sg possible).

Applications

The list of verb phrase constructions provided by the algorithm is directly applicable in the creation of a dictionary of verb phrase constructions. Arranging the constructions around verbs we obtain automatically created raw dictionary entries. To reach the quality of a real dictionary some manual lexicographic work should be done. It is not a labour-intensive step, the manual lexicographic work is limited to construction checking and example selection, the dictionary can be created fast and with a small budget. The dictionary is a valence dictionary, a collocation dictionary and a frequency dictionary at the same time. The sophisticated indexes allow comparison of verb phrase constructions according to several aspects.

Steps of dictionary creation, the dictionary itself and its possible applications are covered in section 4.2 in the dissertation, its significance is stated in the following thesis:

Thesis 5

I created a dictionary of a new kind, whose basic units are not words but expressions: verb phrase constructions. The way from bare text to the raw dictionary entries lead using purely automatic natural language processing tools. The most important step is the algorithm for extracting typical verb phrase constructions (Thesis 4) which automates the dictionary material collection step. I showed that this lexical acquisition method is well suited

for dictionary creation: the final dictionary truly contains the valences and verbaél expressions that are typical in the Hungarian language. In conclusion, a new kind of learners' dictionary was created this way which highlights the most important verbal meanings, and allows the language learner to speak idiomatically not just grammatically correct.

Publications related to the thesis:

(Sass et al., 2010a) (Sass és Pajzs, 2010b) (Pajzs és Sass, 2010)
(Sass és Pajzs, 2010c)

How can we use such a dictionary to support language learning or if we want to say something in Hungarian? By the help of the dictionary the verb–noun collocations can be discovered: nouns which usually collocates with a given verb and also verbs which usually collocates with a given noun can be determined (using the index of fixed words). Consider that we want to speak Hungarian as an English native speaker. If we search for the translation of ‘*meet the requirements*’ knowing that ‘*requirement*’ is ‘*követelmény*’ in Hungarian, we will find the appropriate verb at ‘*követelmény*’ which is ‘*megfelel*’ (that is not the literal translation of ‘*meet*’ which would be ‘*találkozik*’).

The dictionary (Sass et al., 2010a) is available, it is published by the Tinta Publishing House.



The fact that an automatic language processing method is *language independent* gives strong significance to it. Language independence of my approach depends on the language independence of being able to create the representation. Tools and methods based on the representation (the corpus query system, the lexical acquisition algorithm, the automatic part of the dictionary creation as described in the previous theses) work automatically if the representation is at hand. As the representation relies only on the fact that there is predicate–argument

structure in human languages, it is expected that the representation can be created for several languages. This guess was supported by experiments with languages having a different structure compared to Hungarian, namely Danish and Serbian.

Language independence of my approach is covered in section 5.1 in the dissertation, the next thesis contains the results of this section:

Thesis 6

I showed that the unified representation (Thesis 1) is language independent, it can be created for several languages. This result essentially depends on that utterances generally can be decomposed into units (clauses) which contain a verb and its dependents, and the dependency relationship between the verb and a dependent can be specified. The Verb Argument Browser (Thesis 3) for a language can be prepared with little effort having the representation. The algorithm for extracting typical verb phrase constructions (Thesis 4) can run on any corpus represented according to the model, thereby the collection of verb phrase constructions is feasible independently of language. Ultimately, the dictionary (Thesis 5) can also be created based on this algorithm by investing a limited amount of manual lexicographic work.

Publications related to the thesis:
(Sass, 2009d)

Using my method new learners' dictionaries – similar to the Hungarian version described in the previous thesis – can be prepared for new foreign languages which is popular among language learners in Hungary.

The model (Thesis 1) can be extended in several ways, specifically, some complex structures can be traced back to the 1-level deep dependency tree shown in Fig. 1 (page 10). The most interesting question is: can we produce a representation which is made of a *parallel* corpus, and consequently contains parallel clauses and parallel verb phrase constructions (constructions and their translations); but at the same time it has formally the same structure as the original model so the lexical acquisition method can take it as input. In this way, we could gain a method which can extract parallel constructions applying the original extraction method: we would obtain translations of verb phrase constructions.

Extensions of the model are discussed in section 5.2 and 5.3 in the dissertation, I report about application of the method to parallel constructions in section 5.4, the last thesis sums up this promising direction.

Thesis 7

I showed that the common representation of a parallel clause (two clauses in two different languages corresponding to each other) can be produced as a 1-level deep dependency tree as the original model requires: the central unit becomes a pair consisting of the two verbs (in the two languages), and the dependents are assigned to this central unit as a combined set. Such a way a representation (for parallel corpora) formally similar to the original representation (for monolingual corpora) can be obtained. The lexical acquisition method (Thesis 4) can run on this representation directly extracting bilingual, parallel verb phrase constructions. The method is able to correlate bilingual constructions with each other that are asymmetric, that means have a completely different structure in the two languages.

Publications related to the thesis:
(Sass, 2010d)

I conducted the investigations about parallel verb phrase constructions on a Dutch–French parallel corpus. For example, in the result I obtained the asymmetric pair of Dutch *'nemen deel aan'* and French *'participer à'* (both means *'take part in'*). We see where a complex verb is used in Dutch, expressed by one word, a simple verb in French. In the future, this method can be used in the creation of new bilingual dictionaries which facilitate language learning through matching verb phrase constructions extracted from language use.

It is a task for the future to work out the details of that kind of bilingual dictionary creation, my work is an important step in this direction.

Acknowledgements

I am grateful to my wife, *Dóri*, to my children, *Mici*, *Csöpi*, *Lencsi* and *Jáni* and to my extended family for their constant support and encouragement.

I would like to say thank you to my supervisor, *Gábor Prószéky*; to my boss, *Tamás Váradi*; to my nearest colleague, *Csaba Oravecz*; to my lexicographer colleague, *Júlia Pajzs*; and to the the leaders of the Doctoral School, *Tamás Roska* és *Péter Szolgay* for professional support and help.

I would like to express my gratitude to my friends, colleagues and everybody who helped and supported me by their work, ideas, advices and striking insights during the past years and during the time of thesis writing.

Author's publications

Book

Sass Bálint – Váradi Tamás – Pajzs Júlia – Kiss Margit 2010a. *Magyar igei szerkezetek – A leggyakoribb vonzatok és szókapcsolatok szótára [Hungarian Verb Phrase Constructions – A Dictionary of Most Frequent Verb Frames and Collocations]*. Tinta, Budapest.

Journal article

Sass Bálint – Pajzs, Júlia 2010b. Igei szerkezetek gyakorisági szótára – félautomatikus szótárkészítés nyelvtechnológiai eszközök segítségével [Frequency Dictionary of Verb Phrase Constructions – Semi-automatic Lexicography Using NLP Tools]. *Alkalmazott Nyelvtudomány [Applied Linguistics]*, 2010(1–2):5–32.

Book chapter

Sass Bálint 2006a. Extracting Idiomatic Hungarian Verb Frames. In Salakoski, Tapio – Ginter, Filip – Pyysalo, Sampo – Pahikkala, Tapio (eds.): *Advances in Natural Language Processing*, 303–309. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 4139.

-
- Sass Bálint 2008. The Verb Argument Browser. In Sojka, Petr – Horák, Aleš – Kopeček, Ivan – Pala, Karel (eds.): *Text, Speech and Dialogue*, 187–192. Springer, Berlin Heidelberg New York. Lecture Notes in Computer Science, Vol. 5246.
- Sass Bálint 2009a. Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára [Corpus Linguistic Tool for Investigating Argument Structure of Hungarian Verbs]. In Sinkovics Balázs (ed.): *LingDok 8. – Nyelvész-doktoranduszok dolgozatai [Papers of PhD students in Linguistics]*, 143–155. JATEPress, Szeged.
- Sass Bálint 2009b. „Mazsola” – eszköz a magyar igék bővítményszerkezetének vizsgálatára [Mazsola – a Tool for Investigating Argument Structure of Hungarian Verbs]. In Váradi Tamás (ed.): *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiából [A Selection of Papers of Hungarian Student Conference on Applied Linguistics]*, 117–129, RIL HAS, Budapest.
- Sass Bálint – Pajzs Júlia 2010c. FDVC – Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions. In Granger, Sylviane – Paquot, Magali (eds.): *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009, Cahiers du CENTAL 7. Presses universitaires de Louvain*, 263–272, Louvain-la-Neuve, Belgium.

Proceedings of International Conferences

- Pajzs Júlia – Sass Bálint 2010. Towards Semi-automatic Dictionary Making. In *Proceedings of the XIV. EURALEX International Congress*, 453–462.
- Sass Bálint 2007. First Attempt to Automatically Generate Hungarian Semantic Verb Classes. In *Proceedings of the 4th Corpus Linguistics conference*, Birmingham.

Sass Bálint 2009c. A Unified Method for Extracting Simple and Multi-word Verbs with Valence Information and Application for Hungarian. In *Proceedings of RANLP 2009*, 399–403, Borovets, Bulgaria.

Sass Bálint 2009d. Verb Argument Browser for Danish. In *Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009*, 263–266, Odense, Denmark.

Proceedings of Hungarian Conferences

Sass Bálint 2005. Vonzatkeretek a Magyar Nemzeti Szövegtárban [Verb Frames in the Hungarian National Corpus]. In Alexin Zoltán – Csenedes Dóra (ed.): *III. Magyar Számítógépes Nyelvészeti Konferencia [3rd Hungarian Conference on Computational Linguistics] (MSZNY2005)*, 257–264, Szeged.

Sass Bálint 2006b. Igei vonzatkeretek az MNSZ tagmondataiban [Verb Frames in the Clauses of the Hungarian National Corpus]. In Alexin Zoltán – Csenedes Dóra (ed.): *IV. Magyar Számítógépes Nyelvészeti Konferencia [4th Hungarian Conference on Computational Linguistics] (MSZNY2006)*, 15–21, Szeged.

Sass Bálint 2010d. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból [Extracting Parallel Verb Phrase Constructions from Parallel Corpus]. In Tanács Attila – Vincze Veronika (ed.): *VII. Magyar Számítógépes Nyelvészeti Konferencia [7th Hungarian Conference on Computational Linguistics] (MSZNY2010)*, 102–110, SZTE, Szeged.