A Unified Method for Extracting Simple and Multiword Verbs with Valence Information and Application for Hungarian

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

Abstract

We present a method for extracting verb-centered constructions (VCCs) from corpora. In our framework, simple and multiword verbs, with or without valence are all VCCs. They are treated uniformly, from e.g. *to breathe* till e.g. *to take something into consideration*. In order to extract VCCs we represent the corpus as a sequence of clauses that contain a verb together with all its NP dependents. The method is a generalization of a former subcategorization frame extraction method. It is based on cumulative counting of frequent subframes: small frequency counts are inherited to one of the longest available subframes using random selection. The method finds out automatically the number of elements in VCCs; and it detects automatically whether a content word is integral part of the VCC (forming a multiword verb), or just the verb-dependent relation is important (forming a valence slot of the verb). Significance of our method lies in its capability to deal with multiword verbs and (their) valence simultaneously. The paper includes evaluation for Hungarian, we obtain precision values above 80% using n-best lists evaluation. The representation and the method is in essence language independent, it could be applied to other languages as well.

1 Introduction

Multiword expressions (MWEs) consist of several words, but semantically act as one unit, having non-compositional (idiomatic) meaning [12, 9]. Their meaning cannot be deduced, although the meaning of each part is known. Nevertheless, it is necessary to know their meanings if we want to deal with semantics in any field of natural language processing. Being a borderline case between grammar and lexicon, importance of MWES was underestimated until quite recently [12]. In fact, number of MWEs is large, one fifth of all verbs can be part of a MWE in runnig text [6].

In NLP applications MWEs are usually stored in a lexical resource together with their meaning, thus the main task (called lexical acquisition) is to build up such a lexicon. The traditional collocation-based approach of collecting/extracting MWEs is based on the fact that words in MWEs appear more frequently together than expected. The strength of association between these words can be measured using particular statistical *association measures* [3]. Most of them are worked out to handle exactly *two* words (bigrams), but this is too limiting, because there are longer MWEs, obviously, and there are cases when we do not even know the number of words in the MWEs beforehand.

Conventional classes ot MWEs [12, 9, 6], which can be located along a scale from most idiomatic to most literal meaning are shown below, with examples:

1. fully rigid expressions – *ad hoc*;
2. idioms – *kick the bucket*;
3. verb particle constructions (VPCs) – *hand in*;
4. support verb constructions (SVCs) – *take a walk*;
5. institutionalized phrases – *traffic light*.

It can be noticed that the [verb + NP/PP dependent(s)] – or *verb frame* – structure is very common among MWEs, we find MWEs of this general type in every classes mentioned above. These verb centered MWEs are called *multiword verbs (MWVs)*. Since they cover substantial part of all MWEs, we will deal with this broad class aiming to have a comprehensive picture of MWEs in general.

Like common verbs, some multiword verbs also has one or more arguments (e.g. the *of*-phrase in *get rid of*). To our knowledge, these two research paths – MWEs and valence – have not crossed each other in the literature until recently. Our present aim is to develop a framework that is suitable to handle both aspects, extracting also verb-centered MWEs which are multiword and have valence at the same time.

Accordingly, our target are the *verb-centered constructions (VCCs)*. They consist of a verb, zero or more additional NPs and zero or more valence slots; and the verb together with the NPs (if any) has a (to some degree) non-compositional/idiomatic meaning. If the core meaning of the construction is changing when we change the content word at the head of NP(s), the meaning is considered idiomatic.

Let us see example (1) and introduce the notion of *content* and *relational* units. Hungarian *-bA* ('into' in English) is a relational unit which relates a locative to the verb. Hungarian *-t* is also a relational unit which marks the direct object. The content unit in the object relation is *orr* ('nose' in English). If we change this content unit, the original meaning of this construction changes. So, according to the definition this example is a VCC, moreover a full-grown VCC: a multiword verb with one valence.

(1) beleüt orr-t -bA
knock·in nose-OBJ IN

International Conference RANLP 2009 - Borovets, Bulgaria, pages 399–403

≈ 'meddle with something'[1]

In this paper we introduce a VCC extraction method which fulfils the following two flexibility requirements: (1) the number of units is not restricted to a fixed number, the algorithm detects the number of units within a multiword expression processed; (2) the algorithm detects whether there are any integral content unit in the VCC – forming a multiword verb –, or just some relational units are relevant – forming some valence slots of the verb.

2 Related work

Within the MWE literature there is a significant amount of research in the field of multiword verb extraction. The target is almost always a specific type of VCCs, e.g. verb-particle constructions [2], or verb+noun idiomatic constructions [5]. There is also a MWV-collection and MWV-annotated corpus for Estonian (a language closely related to Hungarian) [6]. A paper studies valence of MWVs, but only one predefined type of valence, namely whether a MWV is transitive or not [1].

There are two important publications concerning Hungarian MWEs. In the first one ⟨verb+noun+casemark⟩ triplets were investigated [7]. These triplets also constitute a specific VCC type, namely multiword verbs without valence. The other paper presents an analysis of different aspects of extracting MWEs, and experiments with a particular extraction method based on rigidity if MWEs [9].

The basic idea of our method comes from a former verb subcategorization frame extraction method [15]. Subsequent further development or application of this method is not known from the literature. For evaluation, we use the n-best lists, as described in [4].

3 Unified representation

The representation is rather staightforward, we must represent the verb, the relational units and the content units. The solution can be imagined as a one-level-deep dependency structure: the verb becomes the head, the content units (the lemma of the heads of NPs/PPs beside the verb) become the dependents, and the relational units becomes the dependency relations in between. This is a kind of *mixed* clause model: the dependency structure is only one level deep, the dependents are phrases, they are not associated with internal dependency structure, just represented by their heads instead.

[1] We provide Hungarian examples with English glosses in this form. The first line contains the MWV, the verb is shown always first. The *-t* and *-bA* are casemarks. *orr-t* is not a real wordform but the lemma and the casemark (the content and the relational unit) separated by a dash for didactic purposes. Note: the upper case letter (e.g. in *-bA*) signs a vowel alternation point where the exact vowel is determined by Hungarian vowel harmony. The second line contains the word-by-word translation. The uppercase codes means relations, which can be SUBJ, OBJ or a preposition. The dot (·) separates two words, which has a one-word counterpart in the other language. The third line contains the overall English translation.

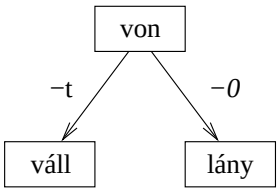

Fig. 1: *Dependency tree of sentence (2). Content unit* váll *'shoulder' is in object relation, and* lány *'girl' is in subject relation. Hungarian casemark for subject is zero suffix depicted as* -0.

Such a way, we can represent not only all kinds of VCCs but also *clause skeletons (CSs)* (i.e. the verb, the relational units and the content units in a particular clause). The dependency tree visualization of example (2) can be seen in Fig. 1.

(2) A lány váll-t von.
the girl shoulder-OBJ shrug
'The girl shrugs her shoulder.'

This model can also be seen as a flat database structure: we have labeled *positions*, which are filled or not. To be clear, these positions are not physical positions in the original clause, they have nothing to do with word order, they just record the existence of some dependent phrases and their relations to the verb. We call a position *fixed*, if there is a particular content word. Similarly, we call a position *free*, if it can be filled by several words from a broad word class. All position is fixed in clause skeletons. MWVs has fixed positions, and valences correspond to free positions (see Fig. 2). An example of a simple verb with one valence is shown in Fig. 3.

Hungarian is an agglutinative language with a relatively free word order. The surface dependencies between the verb and its NP dependents are expressed by relation markers at the end of NPs. Relation markers can be *casemarks* (e.g. *-bA* in example (1)) or *postpositions* (e.g. *mellett* 'beside'). It should be noted that using this model the VCCs need not be ordered nor continuous, so we can also represent free word order languages.

The above outlined representation seems to be language independent, in essence it only relies on the existence of predicate-argument structure. Using positions dictated by the processed language it abstracts away from actual language specific markers express-

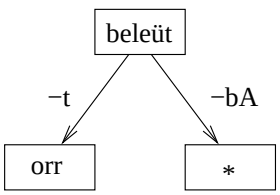

Fig. 2: *Representation of example (1), a VCC with one fixed and one free position (depicted as* ∗*).*

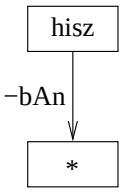

Fig. 3: *Dependency tree of the subcategorization frame* hisz -bAn *'believe IN'.*

ing the relations between the verb and its dependents: separate words (e.g. prepositions), bound morphemes (e.g. the Hungarian casemark *-bAn*), or even relational units which appear as order restrictions (e.g. the relation between the English verb and subject) included.

4 Method

According to our unified representation, verb subcategorization frames (SCFs) (i.e. verbs with some valences) constitute a subset of VCCs. We took an SCF extraction method [15], and worked out the details of extending it to our data structure, namely the unified VCC representation.

The main idea is: we should initially store not just the relational units but also the content units, and we should allow the algorithm to get rid of the content units, where they are just some words filling in a valence slot. Outline of our algorithm is the following (see text below for details):

1. We take all CSs of the corpus with frequency counts. We perform *alternating omission* of content units on all CSs (they are „fully fixed"), to have verb frames with some free positions.

2. We sort the resulting verb frame list according to *length*.

3. Starting with the longest one we discard CSs with frequency less than 5, and add their frequency to a *one-unit-shorter* frame on the list. If there are several such frames which could inherit the frequency, we choose randomly among them. Choosing at random was suggested by the original paper as the best performing possibility [15].

4. Intended VCCs are the final remaining verb frames, ranked by cumulative frequency.

Alternating omission means that (1) for every CS we add a "free" variant with all the relations kept and all the content words deleted; and (2) for CSs of length two we add two "partially free" variants (that means once we keep one lemma and delete the other, then keep the other lemma and delete the first). To make it clear, from CS of sentence (2) we could obtain these verb frames:

(3) von -0 -t

(4) von lány-0 -t

(5) von -0 váll-t

Input:

```
3 take consideration-INTO future-OBJ
3 take consideration-INTO information-OBJ
3 take consideration-INTO refraction-OBJ
3 take consideration-INTO rarity-OBJ
3 take consideration-INTO preference-OBJ
```

Result:

```
15 take consideration-INTO OBJ
```

Fig. 4: *An English example illustrating the method in operation. We obtain the single true VCC from a hypothetical simple input CS-list. Notation: every row consists of a frequency count, then a verb frame in unified representation (a verb followed by content unit + relation pairs).*

It is alternating omission (or frames in the initial list which contain fixed and free positions both) that makes possible to have not just fully-fixed MWVs but VCCs with free positions also in the resulting list of our algorithm. Regarding our example, the correct VCC is obviously (5).

The definition of length fits the intuitive length of an VCC, namely how many units belong to it (beside the verb): we count relational and content units both. In other words, we count fixed positions doubly, as they correspond to a relational unit plus a content unit together. Thus, the length of a VCC is: number of free positions + number of fixed positions · 2. (The VCCs shown in examples (1) and (5) both have length of 3; the length of the SCF on Fig. 3 is 1.) Taking this definition into account, a frame is "one-unit-shorter" if it has one less free positions *or* it has a free position instead of a fixed one.

Compared to the original method our contribution is the idea of storing all content units, the alternating omission procedure, and the suitable definition of length for VCCs.

To illustrate how the method provides true VCCs let us see the VCC in Fig. 2. It will be on the resulting list because in the corpus clauses whose main verb is *beleüt*, the *-t* position is usually filled by *orr* (so its frequency can cumulate), but the *-bA* position is much more variable (so words in this position are more easily dropped out). To make it completely clear, let us see an English example in Fig. 4. As we see, the infrequent content units are dropped out, and we obtain the desired true VCC.

5 Evaluation

To test our VCC extraction method we need a corpus equipped with a one-level-deep dependency annotation for verbs and NPs. We use the 187 million word Hungarian National Corpus, which is morphosyntactically tagged and disambiguated [14]. We lean on an automatic approximation of the dependency annotation described in [13].

In our case, because of storing all content units, size of VCC candidate list grows large, even to some mil-

Table 1: *Results. Average precision values by type and by n (of n-best list). The ± percentages point out two values corresponding to the two annotators. Most important numbers are shown in grey. Cohen's κ measuring inter-annotator agreement is shown in the last column; it corresponds to the rightmost percentage value in every row. In the 'total' line we evaluate the first 500 candidates of the whole list. Type distribution of these 500 candidates is: [1:01] – 307; [0:00] – 131; [2:02] – 33; [3:11] – 21; [2:10] – 8.*

type	$n =$ 50	100	150	200	500	Cohen's κ
[0:00]	83.0% ± 5.0%	82.0% ± 4.0%				0.53
[1:01]	94.0% ± 2.0%	92.0% ± 1.0%	92.0% ± 0.7%	91.8% ± 0.8%		0.77
object	99.0% ± 1.0%	97.0% ± 1.0%	98.0% ± 0.7%	98.0% ± 0.5%		0.75
other	79.0% ± 1.0%	79.5% ± 0.5%	78.7% ± 1.3%	79.8% ± 1.8%		0.68
[2:10]	58.0% ± 6.0%	44.0% ± 3.0%				0.64
subject	20.0% ± 6.0%	19.0% ± 6.0%				0.43
other	83.0% ± 1.0%	80.5% ± 1.5%				0.33
[2:02]	77.0% ± 7.0%	66.5% ± 8.5%				0.63
[3:11]	94.0% ± 0.0%	88.5% ± 3.5%	87.0% ± 3.0%	83.3% ± 3.3%		0.59
[4:20]	51.0% ± 7.0%	39.0% ± 5.0%				0.50
total	94.0% ± 0.0%	93.5% ± 1.5%	89.3% ± 1.3%	89.5% ± 1.5%	88.9% ± 1.3%	0.65

lion entries. Manual annotation of a list of such size is not feasible, so we cannot create P-R graphs (or calculate MAP values) [3], we can only recline upon the n-best lists method [4, 3] for evaluation. It consists of the following steps: the list of initial candidates is sorted by the extraction method; first n candidates is considered by human annotators; and *precision* = the number of true positive MWEs from the first n candidates.

Results obtained by using different evaluation methods cannot be compared directly, but we can state as a rule of thumb that values obtained from n-best lists is broadly comparable with the maximum values of P-R graphs, which are obviously larger then MAP values. We usually found n-best lists results of 50-70% in the literature. Maximum values of P-R graphs in [4] are between 55-65%. In a recent paper which compares several association measures, the best MAP value is 69% (with a baseline of 52%) [10] elsewhere a MAP value of 57% can be reached using the classic χ^2 measure [11]. Concerning the Hungarian language we mention the earlier result of 54% obtained by using n-best lists for $n = 250$ [9].

We evaluate our method using n-best lists with two annotators. We take the resulting list first as a whole (all VCC types together) to have a picture about overall performance, and then by type to map the strength and weaknesses ot the method. By *type* we mean the number of fixed and free positions a VCC has. We use the following notation for types: first comes the length (followed by a colon), then the number of fixed and free positions respectively. For example type `[2:10]` means one fixed positions (typical MWV), type of the VCC shown in Fig. 2 is `[3:11]` (typical full-grown VCC) and type of the VCC shown in Fig. 3 is `[1:01]` (typical SCF).

Applying the method to the 8000 most frequent verbs in the Hungarian National Corpus, it provides a list of 47000 possible VCCs using a cutoff-threshold of 50. We evaluated types having at most two positions. Beforehand, we filtered out candidates where the lemma was a pronoun or a named entity (trivial non-VCCs), and candidates which were erroneous because of some earlier processing step, as we wanted to evaluate only the VCC extraction step. We annotated real VCCs among the first $n = 500$; then per type among the first $n = 200$ or 100.

According to the definition of VCCs the annotation criterion was this: a candidate is a true positive VCC if and only if (1) there is no fixed positions or the verbal part (verb + occurrent fixed positions) has a (to some degree) non-compositional/idiomatic meaning; and (2) the (possibly multiword) verb truly has such a subcategorization frame which is present and this frame is complete.

Results obtained are summarized in Table 1. Compared with the results found in literature (see percentages in the text above) our results are fairly good. Inter-annotator agreement measured by Cohen's κ is also fair enough, it is mostly above 0.6, reaching 0.8 two times. We can say that our annotation criterion gives a solid foundation for annotators.

We comment the most important results (shown in grey in Table 1) in the following discussion. In type `[1:01]` we have the highest inter-annotator agreement. We get best results in the case of simple transitive verbs, with precision values coming close to 100 percent. Results of type `[1:01]` SCFs having one non-object position (see e.g. Fig. 3) are around 80 percent. Concerning to typical MWVs having one fixed position (type `[2:10]`), if the fixed position is the subject position, the expression usually have compositional meaning (typically with verb *van* 'be' acting as a copula). Conversely, if the fixed position is non-subject (see e.g. Fig. 5) we obtain far better results, but κ values are low here.

Full-grown VCCs (type `[3:11]` structures) are in the focus of this paper, these are the valence bearing multiword verbs. Number and significance of these expressions is high, and (with a moderate inter-annotator agreement) our method performs considerably well on them (see Table 1). This type does not

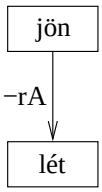

Fig. 5: *Example MWV of type [2:10]* – jön lét-rA *'come existence-INTO'.*

Table 2: *First five real VCCs of type [3:11]. 'X' means that the particular Hungarian casemark do not have an exact English counterpart.*

1. van szó -rÓl
   be word-SUBJ ABOUT
   ≈ 'something is said'
2. tesz lehető-vÁ -t
   make possible-X OBJ
   'make something possible'
3. van szükség -rA
   be need-SUBJ ONTO
   'something is wanted'
4. vesz ész-rA -t
   take mind-ONTO OBJ
   'became aware of something'
5. kerül sor -rA
   come line-SUBJ ONTO
   ≈ 'something takes place'

belong to SCFs nor to simple MWVs, as it contains free and fixed positions both. Being such a borderline case, they usually get out of field of vision, however they are as important as other MWVs having idiomatic meaning often. The main message is: handling SCFs and MWVs in a uniform general way our approach can also collect these kind of expressions. You can see first five real VCCs of type `[3:11]` in Table 2.

6 Application

The resulting list of VCCs has already been used in two projects. VCCs with fixed positions together with manual translations was integrated into the lexical resource of a Hungarian-to-English machine translation system (which is available at `http://www.webforditas.hu`). During building the Hungarian WordNet the verbal synsets was also enriched with VCCs [8].

Most frequent VCCs are also obviously important in language teaching. We are planning to create semi-automatically a "Verbal expression frequency dictionary" for Hungarian. We expect that the manual lexicographic work can be reduced using the result list of VCCs grouped by verb as a starting point.

7 Conclusion

We presented a new approach to extract all types of verb-centered constructions from corpora. Significance of our method lies in its capability of extracting structures which are in the grey area between verb subcategorization frames and multiword verbs having fixed and free positions (valences) both (see Table 2). The method matches the two requirements of flexibility stated at the beginning of this paper: it extracts VCCs with two or more units alike; it extracts VCCs with (even mixed) free and fixed positions alike. Performance of the method is good enough to automatically create reliable lexical resources of VCCs from corpora.

References

[1] T. Baldwin. The deep lexical acquisition of english verb-particle constructions. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414, 2005.

[2] T. Baldwin and A. Villavicencio. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan, 2002.

[3] S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, 2005.

[4] S. Evert and B. Krenn. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics*, Toulouse, France, 2001.

[5] A. Fazly and S. Stevenson. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the EACL*, pages 337–344, Trento, Italy, 2006.

[6] H.-J. Kaalep and K. Muischnek. Multi-word verbs of Estonian: a database and a corpus. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, pages 23–26, Marrakech, Morocco, 2008.

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

[8] J. Kuti, K. Varasdi, Á. Gyarmati, and P. Vajda. Hungarian WordNet and representation of verbal event structure. *Acta Cybernetica*, 18(2):315–328, 2007.

[9] C. Oravecz, V. Nagy, and K. Varasdi. Lexical idiosyncrasy in MWE extraction. In *Proceedings of the 3rd Corpus Linguistics conference*, Birmingham, 2005.

[10] P. Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, pages 54–57, Marrakech, Morocco, 2008.

[11] C. Ramisch, P. Schreiner, M. Idiart, and A. Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC2008 workshop: Towards a Shared Task for Multiword Expressions*, pages 50–53, Marrakech, Morocco, 2008.

[12] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for NLP. In *Proceedings of 3rd CICLING*, pages 1–15, Mexico City, Mexico, 2002.

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

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

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