The Verb Argument Browser

Bálint Sass

Pázmány Péter Catholic University, Budapest, Hungary
sass.balint@itk.ppke.hu

Abstract. We present a special corpus query tool – the Verb Argument Browser – which is suitable for investigating argument structure of verbs. It can answer the following typical research question: What are the salient words which can appear in a free position of a given verb frame? In other words: What are the most important collocates of a given verb (or verb frame) in a particular morphosyntactic position? At present, the Hungarian National Corpus is integrated, but the methodology can be extended to other languages and corpora. The application has been of significant help in building lexical resources (e.g. the Hungarian WordNet) and it can be useful in any lexicographic work or even language teaching. The tool is available online at `http://corpus.nytud.hu/vab` (username: `tsd`, password: `vab`).

1 Introduction

We "...shall know a word by the company it keeps." [1] Traditionally, corpora are investigated through concordances based on the above firthian statement. Nowadays, when corpus sizes can reach 10^9 tokens, manual processing (reading-through) of concordances is not feasible. We need corpus query systems which summarize the information from corpora and present it to the user (e.g. the linguist) in an effective and comfortable way. Since the corpus-based COBUILD dictionary was published, it has been known that such tools can give substantive support to lexicographic work.

One of the first such tools is the *Sketch Engine* [2], which is able to collect salient collocates of a word in particular predefined relations: such as salient objects of a verb, or salient adjectives preceding a noun. The Verb Argument Browser is also a corpus-summarization tool, it collects salient collocations from corpora. Compared to the Sketch Engine it implements a smaller-scale approach focusing only on verbs and arguments[1], but it has one major advantage. Namely, it can treat not just a single word but a whole verb frame (a verb together with some arguments) as one unit in collocation extraction. In other words, instead of collecting salient objects of a verb, it can collect for example salient objects of a given subject–verb pair, or even salient locatives of a given subject–verb–object triplet and so on. In such a way we can outline the salient patterns of a verb "recursively".

[1] As there is a confusion in terminology, it should be noted that throughout this paper the term *argument* will mean complements and adjuncts together.

For the time being the Hungarian National Corpus is integrated, but the methodology can be extended to other languages, if a shallow parsed, adequately processed corpus is available. The simple sentence model to which such a corpus must adhere, is described in section 2. Section 3 considers the processing steps which have been done in the case of the Hungarian National Corpus. The Verb Argument Browser tool is presented in section 4 and the final section gives a variety of different application possibilities.

2 Sentence Model

The company of the verb are its arguments. In an abstract level, a simple sentence can be seen as one verb (in a given tense or mood or possibly modified some way) and the *set* of arguments belonging to the verb. Arguments are usually (nominal) phrases. An NP argument can be represented as the lemma of the head-word of the phrase plus the morphosyntactic function of the phrase.

Some languages take care of the order of (some) arguments. For example, in SVO languages the subject and object functions are determined by their places in the sentence. Some languages use cases and mark arguments with case markers, which usually allows fairly free argument order. For example, Hungarian has about twenty different cases. In Hungarian – just like different complements and adjuncts – the subject and the object is also marked with specific case markers, and can occur almost everywhere in the sentence. The place and order of different complements and adjuncts are far more variable also in SVO languages and they can be put as prepositional phrases (as in English) or prepositional phrases combined with cases (as in German). Order, case markers, prepositions: these are ways to determine a morphosyntactic *position* in a sentence. It is possible to automatically collect all arguments occurring in a sentence by using a (language-specific) syntactic parser.

Based on the above considerations, the model of a simple Hungarian sentence will look the following:

verb + phrase(lemma+case)–set

Let us take an example sentence: ‘A lány vállat vont’. (The girl shoulder-OBJ pull-PAST. = The girl shrugged her shoulder.)[2] The representation according to the above model is: ‘von lány∅ váll-t’ (‘shrug girl-SUBJ shoulder-OBJ’). In this kind of description first comes the verb, then the arguments in the form: lemma, dash, case marker. There are two positions in this example sentence: a subject or nominative position (with zero case marker), and a direct object or accusative position (with case marker: ‘-t’).

An *argument frame* consists of a verb and a list of arguments which occurs (or can occur) with the verb together. Every sentence is an instance of an argument frame. The *argument structure* of a given verb is the union of all its (important)

[2] We will give literal word-to-word glosses and overall translation in English in parentheses, in the form you can see in this example.

argument frames. An example can be seen in Fig. 1. The figure also illustrates that different frames often represent different meanings, and subsequently in different frames the verb should be translated differently (usually to another verb in the target language).

'von kétség-bA -t'	('pull doubt-INTO OBJ'	= to question sg)
'von váll-t'	('pull shoulder-OBJ'	= to shrug one's shoulder)
'von -t'	('pull OBJ'	= to pull sg)
'von felelősség-rA -t'	('pull responsibility-ONTO OBJ'	= to call sy to account)

Fig. 1. Four main frames of the argument structure of verb 'von' (pull). '-t', '-bA' and '-rA' are case markers. (Note: the upper case letter signs a vowel alternation point where the exact vowel is determined by Hungarian vowel harmony.)

The frame 'von felelősség-rA -t' ('pull responsibility-ONTO OBJ' = to call sy to account) exemplifies the notion of *free* and *fixed* positions. The object position is free in both languages, that means we can choose a word from a broad class of words to fill it, the meaning of the verbal construction remains the same. Conversely, the '-rA' position (the 'to' position in English) is fixed, the lemma is unambiguously defined, it is 'felelősség' (account), and we cannot change it without changing the meaning. If we treat lexically fixed NPs (as 'to account') as a part of the verb itself, we can call such constructions *multi word verbs*. They are real verbs: they usually have separate meaning, and even a full-fledged argument structure. Multi word verbs are typical units, which can be investigated using the Verb Argument Browser.

3 Preparing the Corpus

To format a corpus for the Verb Argument Browser one should build the representation of the corpus sentences according to the model desribed in section 2.

Working with the Hungarian National Corpus we have taken the following steps. The corpus is already morphosyntactically tagged and disambiguated [3]. Natural language sentences are usually compound/complex, so the first task was to split up the sentences into units consisting of one verb frame. For completing this task we used a rule based method based on [4]. The rules were regular expressions, which mark clause boundaries on the basis of particular punctuation and conjunction patterns. Main principle was that every clause must contain one and only one verb.

Verbal forms were normalized: separated verbal prefixes were attached to the verb to form the proper lemma of the verb, verbal affixes which do not change the verb frame were removed. Noun phrases were detected by chunking with a cascaded regular grammar. The lemma and the case of the head word of the phrase were recorded. In Hungarian cases and postpositions work the same way apart from the fact that the postpositions are written as a separate word. Accordingly, postpositions were treated as case markers.

4 The Verb Argument Browser Tool

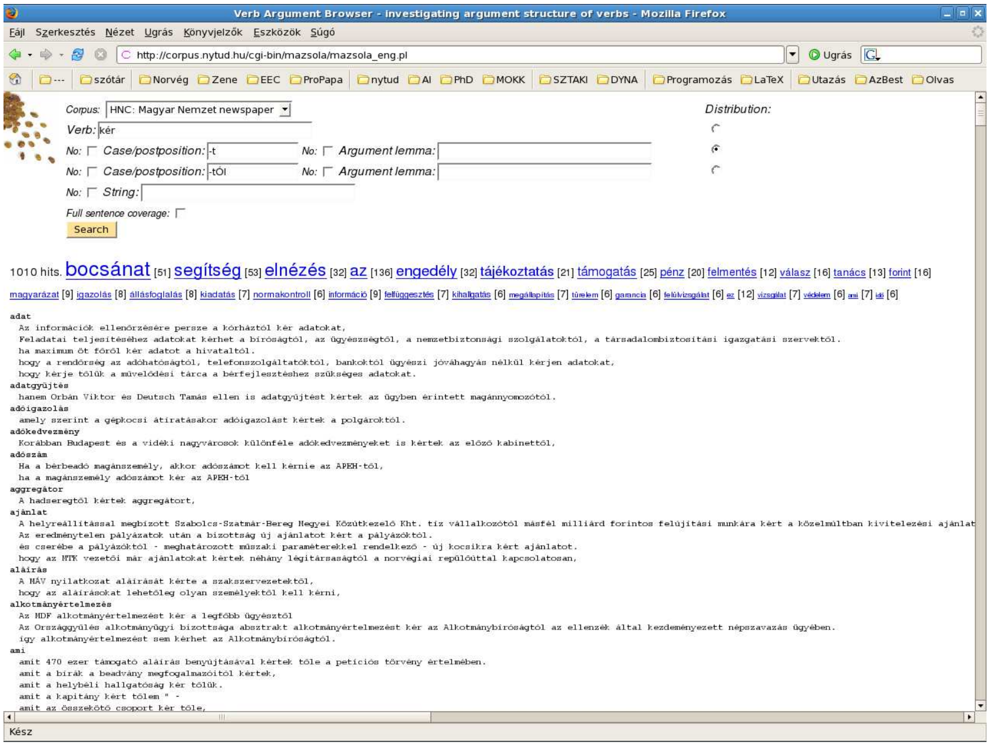
The Verb Argument Browser is useful in answering the following typical research question: What are the salient lemmas which can appear in a free position of a given verb frame? In other words: What are the most important collocates of a given verb (or verb frame) in a particular morphosyntactic position? It should be emphasized that one can give not just a verb but a whole verb frame (or a part of it), and query the lemmas occurring in an additional free position.

Mutual information is a classical measure for collocation extraction, it has a disadvantage of ranking rare items too high. To eliminate this there is an appropriate adjustment to multiply by the logarithm of frequency yielding the *salience* measure [5]. Such a measure can be used to extract collocations mentioned above: one part of the collocation is a verb (or verb frame), and the other part is a lemma in a free argument position. According to this measure a lemma in this free position is salient if it occurs more frequently with the frame than it is expected, and the lemma itself is frequent.

The user interface is shown at the top of Fig. 2. The corpus can be selected in the first field. In the present version, the whole Hungarian National Corpus (187 million words) is integrated. Response times are only a few seconds even at such corpus sizes. Three smaller subcorpora are also available representing three different genres: the 11 million word corpus of the daily paper ‘Magyar Nemzet’, the 12 million word corpus of the forums of the biggest Hungarian portal ‘Index’ and the 11 million word corpus of selected modern Hungarian fiction. There is an option to investigate short sentences exclusively. These sentences originally consists of one verb frame, so the above splitting method (see section 3) was not applied to them.

In the second field the verb stem can be entered. The next two lines are for the two arguments, either free (specified by a case/postposition), or fixed (specified by both case/postposition and lemma). Using the negation checkboxes it is possible to exclude some cases or some lemmas within a case. On the right side you can check the free position, which you are interested in and want to study its salient lemmas. The query can be refined by regular expression full text search below in the *String* field.

In the example in Fig. 2 we are searching for the salient lemmas of the direct object position of the frame ‘kér -t -tÓl’ (‘ask OBJ INDIR-OBJ’ = ask sy sg). We can see these lemmas in blue at the top of the answer screen (Fig. 2). Salience is demonstrated by order and also by font size. Salient direct objects are: ‘bocsánat’ (forgiveness), ‘segítség’ (help), ‘elnézés’ (also forgiveness), ‘engedély’ (permission) and so on. The number in brackets shows the freqency of the lemma. If we click the lemma we can access the corpus examples listed below this “table of contents”.


Verb Argument Browser - investigating argument structure of verbs - Mozilla Firefox
Fájl Szerkesztés Nézet Ugrás Könyvjelzők Eszközök Súgó
Ugrás
--- szótár Norvég Zene EEC ProPapa nytud AI PhD MOKK SZTAKI DYNA Programozás LaTeX Utazás AzBest Olvas
Search
Kész

http://corpus.nytud.hu/cgi-bin/mazsola/mazsola_eng.pl

Corpus: HNC: Magyar Nemzet newspaper

Verb: kér

No: □ *Case/postposition:* -t *No:* □ *Argument lemma:*

No: □ *Case/postposition:* -tÖl *No:* □ *Argument lemma:*

No: □ *String:*

Full sentence coverage: □

Distribution: ○ ● ○

1010 hits. bocsánat [51] segítség [53] elnézés [32] az [136] engedély [32] tájékoztatás [21] támogatás [25] pénz [20] felmentés [12] válasz [16] tanács [13] forint [16] magyarázat [9] igazolás [8] állásfoglalás [8] kiadatás [7] normakontroll [6] információ [9] felfüggesztés [7] kihallgatás [6] megállapítás [7] türelem [6] garancia [6] felülvizsgálat [6] ez [12] vizsgálat [7] védelem [6] ami [7] idő [6]

```
adat
 Az információk ellenőrzésére persze a kórháztól kér adatokat,
 Feladatai teljesítéséhez adatokat kérhet a bíróságtól, az ügyészségtől, a nemzetbiztonsági szolgálatoktól, a társadalombiztosítási igazgatási szervektől.
 ha maximum öt főről kér adatot a hivataltól.
 hogy a rendőrség az adóhatóságtól, telefonszolgáltatóktól, bankoktól ügyészi jóváhagyás nélkül kérjen adatokat,
 hogy kérje tőlük a művelődési tárca a bérfejlesztéshez szükséges adatokat.
adatgyűjtés
 hanem Orbán Viktor és Deutsch Tamás ellen is adatgyűjtést kértek az ügyben érintett magánnyomozótól.
adóigazolás
 amely szerint a gépkocsi átíratásakor adóigazolást kértek a polgároktól.
adókedvezmény
 Korábban Budapest és a vidéki nagyvárosok különféle adókedvezményeket is kértek az előző kabinettől,
adószám
 Ha a bérbeadó magánszemély, akkor adószámot kell kérnie az APEH-től,
 ha a magánszemély adószámot kér az APEH-től
aggregátor
 A hadseregtől kértek aggregátort,
ajánlat
 A helyreállítással megbízott Szabolcs-Szatmár-Bereg Megyei Közútkezelő Kht. tíz vállalkozótól másfél milliárd forintos felújítási munkára kért a közelmúltban kivitelezési ajánlat
 Az eredménytelen pályázatok után a bizottság új ajánlatot kért a pályázóktól.
 és cserébe a pályázóktól - meghatározott műszaki paraméterekkel rendelkező - új kocsikra kért ajánlatot.
 hogy az MTK vezetői már ajánlatokat kértek néhány légitársaságtól a norvégiai repülőúttal kapcsolatosan,
aláírás
 A MÁV nyilatkozat aláírását kérte a szakszervezetektől,
 hogy az aláírásokat lehetőleg olyan személyektől kell kérni,
alkotmányértelmezés
 Az MDF alkotmányértelmezést kér a legfőbb ügyésztől
 Az Országgyűlés alkotmányügyi bizottsága absztrakt alkotmányértelmezést kér az Alkotmánybíróságtól az ellenzék által kezdeményezett népszavazás ügyében.
 így alkotmányértelmezést sem kérhet az Alkotmánybíróságtól.
ami
 amit 470 ezer támogató aláírás benyújtásával kértek tőle a petíciós törvény értelmében.
 amit a bírák a beadvány megfogalmazóitól kértek,
 amit a helybéli hallgatóság kér tőlük.
 amit a kapitány kért tőlem " -
 amit az összekötő csoport kér tőle,
```

Fig. 2. Answer screen of the Verb Argument Browser. Input form is located at the top.

An illustrative example can be to compare verbs' argument structure in different text genres. If we look at the frame 'ad -t' ('give OBJ' = to give sg) in daily paper and internet forum texts, it becomes clear that these two genres have expressly distinct language. Common lemmas in the object position include 'hang' ('voice' = to give voice to sg) and 'lehetőség' ('chance' = to give sy a chance). At the same time 'hír' ('news' = to report) is diagnostic to daily paper text while 'igaz' ('justice' = to take sy's side) is diagnostic to the internet forum text.

The Verb Argument Browser is available at the following URL:
`http://corpus.nytud.hu/vab` (username: `tsd`; password: `vab`). Some clickable examples can be found on the page to test the tool.

5 Applications

The Verb Argument Browser has been applied in two projects concerning the Hungarian language.

During creating the verbal part of the Hungarian WordNet [6] the tool was of significant help in two fields. On the one hand, it is important, how many different meanings a given verb has. Different frames often imply different meanings thus by analysing the argument structure of a given verb one can determine the different synsets where this verb should get into. On the other hand, the Hungarian WordNet was enriched with verbal multi-word units, which are typically a verb together with one fixed argument. Such units are directly provided by the tool, as we saw.

An online demo of a new Hungarian to English machine translation system is available at the following URL: `http://www.webforditas.hu`. During the manual development of the lexical database of this system, an essential task was to collect important lemmas in a position of a given verb frame. From the machine translation point of view important lemmas are the ones for which the verb frame has a special translation. In most cases they are exactly the same as the salient lemmas provided by our tool.

At the end we mention some other possible applications. The Verb Argument Browser can be useful in any lexicographic work, where one of the languages involved is Hungarian. Using this tool, inductive statements can be derived from language data, and an outline of a dictionary entry can emerge from these statements. It gives opportunity to take the empirical frequency of given frames into consideration: it will be possible to determine truly freqent meanings of a verb, or to order the meanings according to frequency.

Similar argument structure can entail similar meaning. To test this hypothesis it is worth generating verb classes based on argument stucture similarity and analysing semantic coherence of these classes [7, 8].

According to a proposition in theoretical linguistics, which is gaining attention nowadays [9, 10], the aim of grammar is not simply to separate grammatical from ungrammatical but it must explain why certain – sometimes ungrammatical – utterances occur and why certain grammatical utterances does not occur at all. The tool can provide evidence for research in this direction.

It can be of extensive use in language teaching (e.g. by showing the differences in argument structure of verbal synonyms) and also in creation of a verb frame frequency dictionary as automatic generation of verb argument profiles can be built on top of the Verb Argument Browser.

References

1. Firth, J.R.: A synopsis of linguistic theory 1930-1955. Studies in linguistic analysis (1957) 1–32
2. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Proceedings of EURALEX, Lorient, France (2004) 105–116
3. 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
4. Sass, B.: Igei vonzatkeretek az MNSZ tagmondataiban [Verb frames in the clauses of the Hungarian National Corpus]. In: Proceedings of the 4th Magyar Számítógépes Nyelvészeti Konferencia [Hungarian Conference on Computational Linguistics] (MSZNY2006), Szeged, Hungary (2006) 15–21
5. Kilgarriff, A., Tugwell, D.: Word Sketch: Extraction and display of significant collocations for lexicography. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation, Toulouse (2001) 32–38
6. Kuti, J., Varasdi, K., Gyarmati, Á., Vajda, P.: Hungarian WordNet and representation of verbal event structure. Acta Cybernetica **18**(2) (2007) 315–328
7. Gábor, K., Héja, E.: Clustering Hungarian verbs on the basis of complementation patterns. In: Proceedings of the ACL-SRW'07 conference, Prague (2007)
8. Sass, B.: First attempt to automatically generate Hungarian semantic verb classes. In: Proceedings of the 4th Corpus Linguistics conference, Birmingham (2007)
9. Stefanowitsch, A.: Negative evidence and the raw frequency fallacy. Corpus Linguistics and Linguistic Theory **2**(1) (2006) 61–77
10. Sampson, G.R.: Grammar without grammaticality. Corpus Linguistics and Linguistic Theory **3**(1) (2007) 1–32