Verb Argument Browser for Danish

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

Abstract

The Verb Argument Browser is a linguistically relevant corpus query tool, which can be used for investigating argument structure of verbs. The original tool was developed for Hungarian corpora but the methodology is claimed to be language independent because of the dependecy grammar based representation. This paper examines this language independency applying the methodology to a language with different structure, namely: Danish. We will see that the methodology can be applied straightforwardly, and the resulting tool shows the same properties as the original version. The Verb Argument Browser for Danish is available at `http://corpus.nytud.hu/vabd` (username: `nodalida`, password: `vabd`).

1 Introduction

The Verb Argument Browser (VAB) is a corpus query tool which is suitable for investigating the argument structure of verbs (Sass, 2008). The paper cited defines the term *argument* as a phrase that appears in a syntactic relationship with the verb in a clause; and so we will use this term – as a synonym for *dependent* – for complements and adjuncts both.

In the VAB approach basic units are clauses: a verb together with its dependents. Dependents are represented by the lemma of their head and their surface relationship to the verb. These surface relationships are called *positions*, and can be defined by order (e.g. subject or direct object), by a preposition or a case marker etc. According to this terminology, in the sentence *"26 personer kom på hospitalet."*, we have the word *person* in subject position and *hospital* in *på* position.

The tool performs collocation extraction using the association measure *salience* (Kilgarriff and Tugwell, 2001). It can answer the following typical research question: what are the most important collocates of a given verb (or verb frame) in a particular position. The VAB has the important property that 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".

This dependency grammar based model outlined above should be suitable for the description of a broad class of languages. In (Sass, 2008) it is stated but not tested that "the methodology can be extended to other languages and corpora". The aim of the present paper is to test this statement.

We chose the Danish language as testbed because its structure is considerably different from Hungarian. They use different linguistic markers to express arguments. In brief, while Danish has fixed word order and a system of prepositions, Hungarian has a rich case system and its word order is relatively free.

2 Representation

As we mentioned, the basic unit of the VAB is the clause (a verb together with its dependents), and dependents are represented by their position (surface relationship to the verb) and the lemma of their head.

We can say that this is a kind of *mixed* clause model: a one-level-deep dependency structure, where the dependents are phrases. The verb has dependents in some particular relationship but dependents do not have internal dependency structure; they are treated as phrases instead, repre-

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 263–266

sented by their heads.

We can define positions as we like. Dealing with Danish we will have: subject position (*subj*), direct object position (*dobj*) and a position for every preposition (*i*, *til*, *på* etc.).

Thus, the above example in VAB input format looks like the following:

```
26 personer kom på hospitalet.
stem=komme subj=person på=hospital
```

We treat clausal dependents in two ways. As they are clauses per se, they are separate units in our representation: they have a verb and some dependents of it internally. From the main clause point of view they are dependents, so they are represented by position and the lemma of their head.

As we see the VAB is not just a classic concordancing tool – like e.g. (Dura, 2006) –, because it has a special corpus representation for the verb–argument structure which can also handle free word order.

3 Converting a Treebank for the VAB

To integrate a corpus into the VAB, the representation described above should be worked out. First, we need to extract the clauses then we need to identify the depedents, their heads and their relations to the verb.

There are two possibilities. On the one hand, we can set out from a POS-tagged corpus and develop a full-fledged chunking system with clause boundary detection. On the other hand, we can set out from a treebank and extract only the information needed. We chose the (obviously cheaper) second possibility. Although treebanks are usually about two orders of magnitude smaller than POS-tagged corpora, for our testing purposes they suffice. The chosen treebank is the 90000 word Danish Dependency Treebank (Trautner Kromann, 2003) because it is freely available with extensive documentation.

Converting the treebank we made the following steps:

- We detected clause boundaries with a simple rule: when we found a comma preceding a conjunction (with `CC` or `CS` msd code) then we split the sentence into two parts. Such a way we obtained approximate clauses.
- To detect the main verb we started from the root node of the tree of the given clause. If the root node was not a verb we descended the tree along `vobj` dependencies to search for the verb. If the verb found had a `vobj` dependent too, we selected that child node to discard auxiliary verbs and obtain the main verb which owns the formal dependents of the auxiliary verb semantically. This example shows the representation of a sentence with an auxiliary verb:

  ```
  Med én rœv kan man ikke sidde på to
  heste.
  stem=sidde subj=man med=rœv på=hest
  ```

- Collecting all first level dependencies, to identify subject and direct object was straightforward (by the `subj` and `dobj` relations). We detected prepositional phrases using other relations (e.g. `pobj`, `lobj`), and recorded the prepositions as they are the units which correspond to positions in the representation.
- We identified the heads of phrases descending the `nobj` and `possd` relations.

4 VAB for Danish

The answer screen of the resulting tool can be seen in Fig. 1. The user interface (at the top of Fig. 1) is built up determined by the representation: after entering the verb stem, three arguments (dependents) can be given by position, by lemma or both.

In the example in Fig. 1 we are searching for salient collocates of the Danish verb *have* in the direct object position. We can enter 'subj' or 'dobj' or any preposition into the position field. (The 'Distribution' radio button on the right determines the position in question. Setting it becomes relevant only if we specify two or three dependents.) The most salient collocates are shown in variable font size below the input form: *brug*, *plan*, *masse*, *kontakt* etc.

Salient collocates collected by the VAB tool can be divided into two parts (Sass, 2008):

1. frequent words with literal meaning, often forming a semantically coherent class – like kinds of food as the direct objects of *to eat*;
2. words that form a *multiword verb* together with the verb – like *part* in *take part in* or *rid* in *get rid of*.

We see that this holds for the Danish version, even at such a small corpus size. While *plan* is a frequent concept which people usually have, *have*

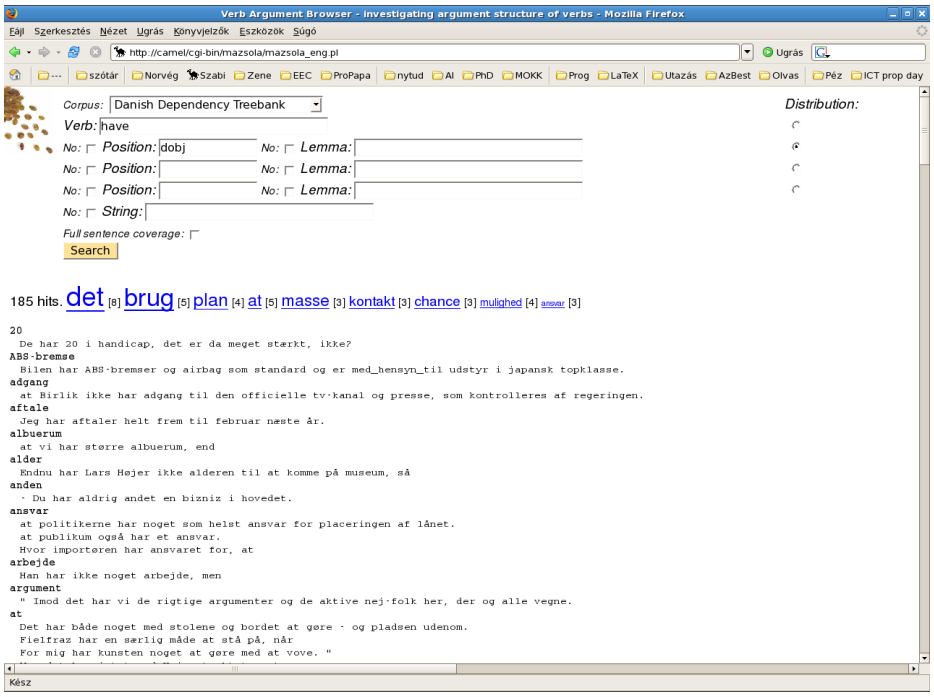

Figure 1: *have* + *dobj* (direct object) in the Danish VAB.

brug for (to need sg) is an authentic multiword verb.

Testing the tool with other (frequent) verbs and positions, we can get different multiword verbs, e.g. *være i tvivl om* (to be in doubt about), *være i forbindelse med* (to be in connection with), *være på vej* (to be on the road), *være på besøg* (to visit) or *få lov til* (to allow).

Apart from verbs, prepositions and nouns can also be investigated, if we leave the verb and/or the position field blank in the input form. This way we can discover important noun phrases, e.g. *ved bord* (at the table), *til gengæld* (in exchange) or *på en måde* (in a way).

5 Comparison with a Treebank Viewer

VAB for Danish can be seen as an alternative to the interactive treebank viewer for the Danish Dependency Treebank (available at: `http://treebank.dk/cdt-map/MapDep.html`) Important differences are:

- The treebank viewer is made to query exactly one graph edge; the VAB has a different approach, it can bring together several entities of a clause, e.g. the verb, the subject and the object.
- The VAB treats verbs and dependents as potential units in collocations, and applies a specific collocation extraction method, instead of just showing some parts of the corpus.
- For the VAB, a verb frame is "worth" exactly the same, not only as a top level structure of the sentence, but also as an embedded one. The VAB sums up all instances of a verb (frame), and does statistics on the whole list.
- The treebank viewer focuses on dependencies (relations); while in the VAB we can also specify and study the lemmas.
- In the treebank viewer we can see every edges of the tree; while in the flat representation of the VAB most of the edges are removed, only the verb and its one-level-deep dependents remain, together with their heads. Such a way VAB does some generalisation on verb frames.

In actual fact, the VAB can also be considered as a treebank viewer, however, not only (computational) linguists but e.g. language learners can benefit from its use.

6 Conclusion

The main message of this paper is that the language independency of the Verb Argument Browser approach (Sass, 2008) holds. The Danish version shows the same properties as the original Hungarian: it can be used to collect mulitword verbs and other important verb frames of the language.

Thus, the Danish version can also be used to support corpus-driven lexicographic work or can be used in corpus-driven language teaching, as it provides the most important verb phrase constructions. Using the VAB a special learners' dictionary can be compiled, which "helps students to write and speak idiomatically" (Hanks, 2008).

A VAB can be created for (hopefully) any language if we have the representation needed. We showed that a treebank can be converted to this representation with ease. The other approach of starting with a POS-tagged corpus and building a suitable chunker is more expensive but POS-tagged corpora are much larger so the resulting tool will have a more impressive coverage.

Most corpora are either large, and have no syntactic annotation (e.g. the so called "national corpora" with POS-tagging); or small with rich syntactic annotation (treebanks). A VAB would work well with a middle-sized chunked corpus, thus such tools set up a claim for a third type, which is in the middle in both respects and often missing: few ten million word shallow parsed corpora.

The Verb Argument Browser for Danish is available at `http://corpus.nytud.hu/vabd` (temporary username: `nodalida`, password: `vabd`). For free individual access or if you want to build a VAB for another language, please contact the author.

References

Elzbieta Dura. 2006. CULLER – a user-friendly conrpus query system. In *Proceedings of the Fourth International Workshop on Dictionary Writing Systems*, pages 47–52, Torino, Italy.

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

Adam Kilgarriff and David Tugwell. 2001. 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*, pages 32–38, Toulouse.

Bálint Sass. 2008. 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.

Matthias Trautner Kromann. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206