Verb Argument Browser for Danish

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

Aim

Previous work: original VAB for Hungarian [2].
A corpus query tool
for investigating argument structure of verbs.

Is the methodology language independent?
Present aim: to show yes it is language independent!

Method: build such a tool for a different language.

Two aspects to test:

1. Can we work out the required corpus representation?
2. Does the tool have the same properties,
does it work the same way as the original version?

Corpus Representation

Basic unit: clause = verb + its dependents
Dependent representation:
the lemma of the head + surface relationship to the verb

Surface relationships (or *positions*) can be defined by word order, prepositions, case markers etc.
Testbed: Danish – structure different from Hungarian

Positions for Danish:
‘subj’, ‘dobj’ and prepositional (i, til, på etc.) positions.

Example sentence & representation:
26 personer kom på hospitalet.
verb=komme subj=person på=hospital

How does VAB work?

Verb Argument Browser . . .

- answers this research question: *what are the most important collocates of a given verb (or verb frame) in a particular position.*
- performs collocation extraction using the *salience* association measure.
- is able to treat not just a single word but a whole verb frame (a verb together with some arguments) as one unit in collocation extraction.
→ to outline the salient patterns of a verb “recursively”

Accessibility

Available at http://corpus.nytud.hu/vabd
(temporary username: nodalida, password: vabd).
For free individual access please contact the author.

Application

With a larger corpus . . .
– useful corpus query tool for the language learner; it can help “students to write and speak idiomatically” [1].
– support for corpus-driven lexicographic work;
– support for building a multiword verb language resource.

A VAB can be created for (hopefully) any language. If you want to build a VAB for another language, please contact the author.

Converting a Treebank for VAB

To integrate a corpus into VAB we need to work out the representation:
– to extract the clauses;
– to identify the main verb in each clause;
– to identify the dependents, their heads and their relations to the verb.

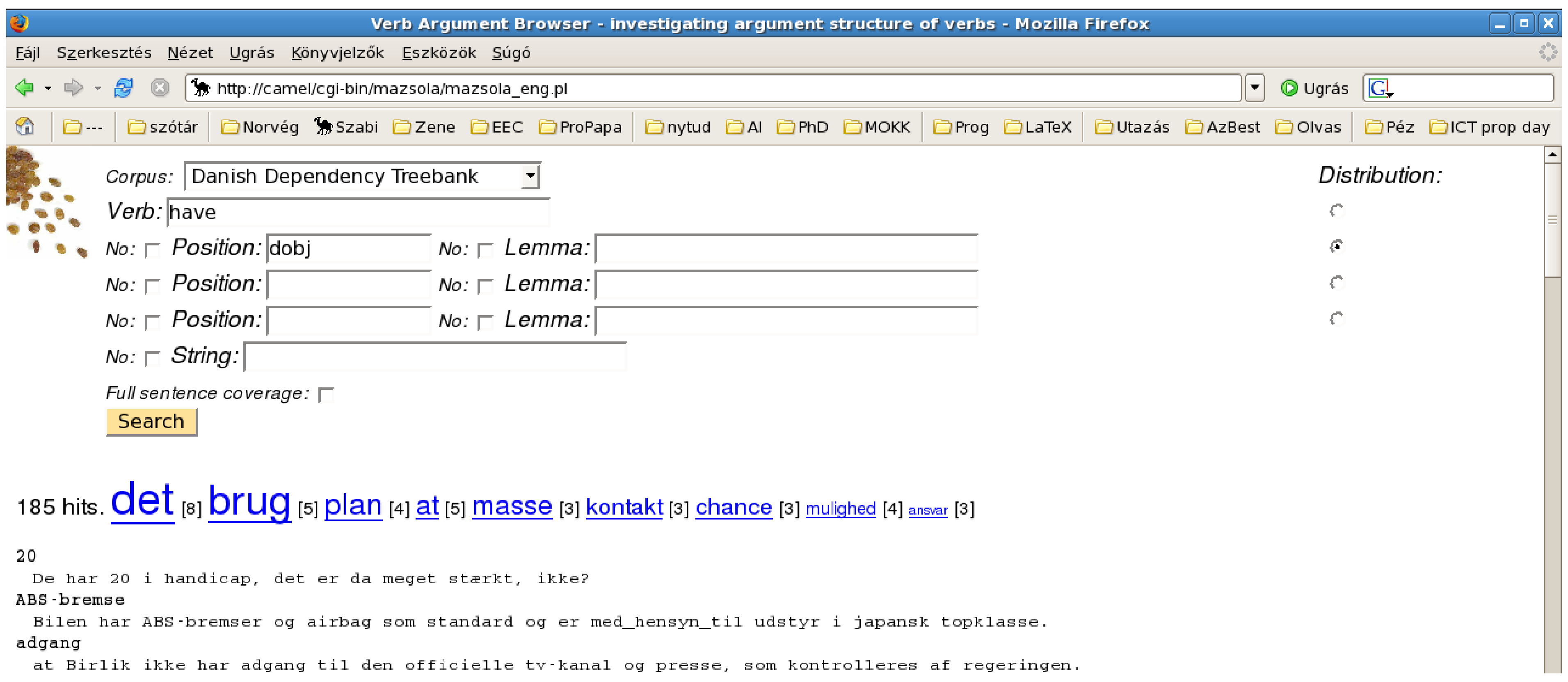
From a POS-tagged corpus – *task:* to develop a chunker with clause boundary detection.
From a treebank – *task:* to extract the information needed.
Treebanks are much smaller, but for our testing purposes they suffice: *Danish Dependency Treebank* [3].

DDT → VAB representation converting steps:
– approximate clause boundary detection: at comma + conjunction
– main verb detection – *task:* discard auxiliary verbs
– collecting dependents (positions + lemmas):
  – subject and object are straightforward;
  – PP-s: record the prepositions (as positions) + the head of the phrase
  – head identification – simple search in the tree

Example sentence & representation:
Med én røv kan man ikke sidde på to heste.
verb=sidde subj=man med=røv på=hest

→ Conclusion 1.
The language independency holds.
The representation can be created for a language different from Hungarian.
It relies only on the notions of clause, verb, dependent and relation between the verb and its dependents.

VAB for Danish


Verb Argument Browser - investigating argument structure of verbs - Mozilla Firefox
Fájl Szerkesztés Nézet Ugrás Könyvjelzők Eszközök Súgó
http://camel/cgi-bin/mazsola/mazsola_eng.pl
Ugrás
--- szótár Norvég Szabi Zene EEC ProPapa nytud AI PhD MOKK Prog LaTeX Utazás AzBest Olvas Péz ICT prop day
Distribution:
Search

Corpus: Danish Dependency Treebank
Verb: have
No: □ Position: dobj No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ String:
Full sentence coverage: □

185 hits. det [8] brug [5] plan [4] at [5] masse [3] kontakt [3] chance [3] mulighed [4] ansvar [3]

```
20
  De har 20 i handicap, det er da meget stærkt, ikke?
ABS-bremse
  Bilen har ABS-bremser og airbag som standard og er med_hensyn_til udstyr i japansk topklasse.
adgang
  at Birlik ikke har adgang til den officielle tv-kanal og presse, som kontrolleres af regeringen.
```

Usage: the verb stem + 3 dependents can be given by position, by lemma or both.
Question: What are the salient collocates of have in the direct object position?
Answer: brug, plan, masse, kontakt etc.

VAB provides two kinds of salient collocates [2]:

1. frequent words with literal meaning (often forming a semantically coherent class);
2. words that form *multiword verbs* with the verb.

We see the same behaviour in the Danish version:

verb	*position*	*frequent word*	*multiword verb*	
have	dobj	have plan	have brug (for)	‘to need sg’
få	dobj	få (nogle) krone	få lov (til)	‘to allow sg’
være	i	være i Danmark	være i tvivl (om) være i forbindelse (med)	‘to be in doubt about’ ‘to be in connection with”
være	på	være på hospital	være på vej (til) være på besøg	‘to be on the road’ ‘to visit’

→ Conclusion 2.
The language independency holds. The browser works the same way as the original.
It can be used to collect multiword verbs
and other important verb frames of the Danish language.

Note: Apart from verbs, prepositions and nouns can also be investigated. Typical frequent prepositional phrases: til gengæld ‘in exchange’; på (en) måde ‘in a way’; for (første) gang ‘at first’; ved bord ‘at the table’.

References

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

[2] Bálint 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.

[3] Matthias Trautner Kromann. The Danish Dependency Treebank and the DTAG treebank tool. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden, 2003.

NoDaLiDa 2009 Conference, 15-16 May 2009, Odense, Denmark