Verb Argument Browser

Argument frames in the Hungarian National Corpus

Bálint Sass
sass.balint@nytud.mta.hu

Research Institute for Linguistics, Hungarian Academy of Sciences

Corpus resources for quantitative and psycholinguistic analysis
Eger, 2-3 June 2014

PREVIEW

- Verb Argument Browser – a specific corpus query tool for investigating argument structure of verbs.
- original version:
  - for Hungarian

    based on the old version of the Hungarian National Corpus
  - not just arguments

    all NP and PP dependents of verbs

    subjects, objects, complements, adjuncts included
  - investigating

    verb subcategorization frames, institutionalized phrases, light verb constructions, idiomatic verbal expressions, figures of speech . . .

    common property: verb + some NP/PP dependents
  - *examples*
- language independence

1. Sentence model

2. Verb phrase constructions as collocations

3. Usage & examples

4. Applications

5. Language independence

1. Sentence model
2. Verb phrase constructions as collocations
3. Usage & examples
4. Applications
5. Language independence

Sentence model

- *Basic unit:* simple sentence or clause.

A	lány	váll-at	von.
the	girl	shoulder-ACC	pull.

’The girl shrugs her shoulder.’

- clause = verb + set of NP/PP dependents → verb frame

verb=von	*NOM*=lány	*ACC*=váll
verb=shrug	*SUBJ*=girl	*OBJ*=shoulder

- *Dependent types*: defined . . .
  - syntactically: word order (in English)
  - morphologically: case markers (in Hungarian)

Sentence model

- *Basic unit:* simple sentence or clause.

A	lány	váll-at	von.
the	girl	shoulder-ACC	pull.

'The girl shrugs her shoulder.'

- clause = verb + set of NP/PP dependents → verb frame

verb=von	*NOM*=lány	*ACC*=váll
verb=shrug	*SUBJ*=girl	*OBJ*=shoulder

- *Dependent types*: defined . . .

– syntactically:	word order	(in English)
– morphologically:	case markers	(in Hungarian)

Dependent types

- in Hungarian: 20 different case markers
  in English: usually prepositions

case marker	case	abbr.	English
-∅	nominative	NOM	word order
-t	accusative	ACC	word order
-bAn	inessive	INE	*in*-phrase
-rÓl	delative	DEL	*from*-phrase^1
-bÓl	elative	ELA	*from*-phrase^2
…			

EXAMPLES

Az	emberek	az	időjárás-ról	beszélnek.
the	people	the	weather-DEL	talk.

’People talk about the weather.’

verb=beszél	*NOM*=ember	*DEL*=időjárás
verb=talk	*SUBJ*=people	*ABOUT*=weather

Péter	fél	az	ismeretlen-től.
Peter	fear	the	unknown-ABL.

’Peter fears of the unknown.’

verb=fél	*NOM*=Péter	*ABL*=ismeretlen
verb=fear	*SUBJ*=Peter	*OF*=unknown

Examples

Az	emberek	az	időjárás-ról	beszélnek.
the	people	the	weather-DEL	talk.

’People talk about the weather.’

verb=beszél	*NOM*=ember	*DEL*=időjárás
verb=talk	*SUBJ*=people	*ABOUT*=weather

Péter	fél	az	ismeretlen-től.
Peter	fear	the	unknown-ABL.

’Peter fears of the unknown.’

verb=fél	*NOM*=Péter	*ABL*=ismeretlen
verb=fear	*SUBJ*=Peter	*OF*=unknown

Examples

Az	emberek	az	időjárás-ról	beszélnek.
the	people	the	weather-DEL	talk.

’People talk about the weather.’

verb=beszél	*NOM*=ember	*DEL*=időjárás
verb=talk	*SUBJ*=people	*ABOUT*=weather

Péter	fél	az	ismeretlen-től.
Peter	fear	the	unknown-ABL.

’Peter fears of the unknown.’

verb=fél	*NOM*=Péter	*ABL*=ismeretlen
verb=fear	*SUBJ*=Peter	*OF*=unknown

FIXED AND FREE DEPENDENTS

Hogy	jöttek	lét-re	az	első	csillagok?
how	came	existence-SUB	the	first	stars?

’How the first stars came into existence?’

verb=jön	*SUB*=lét	*NOM*=csillag
verb=come	*INTO*=existence	*SUBJ*=star

- *fixed dependent:*
  cannot change the content word
  without changing the meaning of the VPC
- *free dependent:*
  can change the content word
  without changing the meaning of the VPC

Fixed and free dependents

Hogy	jöttek	lét-re	az	első	csillagok?
how	came	existence-SUB	the	first	stars?

’How the first stars came into existence?’

verb=jön	*SUB*=lét	*NOM*=csillag
verb=come	*INTO*=existence	*SUBJ*=star

- *fixed dependent:*
  cannot change the content word
  without changing the meaning of the VPC
- *free dependent:*
  can change the content word
  without changing the meaning of the VPC

MULTI-WORD VERBS

- *multi-word verb:* verb stem + some fixed dependent(s)

lét-re	jön
existence-SUB	come

’come into existence’

Multi-word verbs have. . .

- separate meaning
- own argument structure

rész-t	vesz	bAn
part-ACC	take	INE

’take part in SOMETHING’

Typical units to be investigated using the VAB.

Sentence model

sentence = verb + set of dependents
dependent = type + content word

i.e.

verb=jön	*SUB*=lét	*NOM*=csillag
verb=come	*INTO*=existence	*SUBJ*=star

In this way we can investigate VPCs independently from the particular word order in which they appear in the corpus.

CORPUS PREPARATION

Input: Hungarian National Corpus
(POS-tagged and disambiguated)

- clause boundary detection
  – regexps based on conjunction and punctuation patterns
- verb normalization
  – e.g. separated verbal prefixes attached
- noun phrase chunking
  → case and lemma of the head of dependent phrases

→ representation according to the model

1 Sentence model

2 Verb phrase constructions as collocations

3 Usage & examples

4 Applications

5 Language independence

VERB PHRASE CONSTRUCTIONS AS COLLOCATIONS

A specific kind of VPCs:

’take part in SOMETHING’

- fixed dependent (object) + free dependent (*in*-phrase)
- multi-word verb with argument structure

These kind of expressions are

1. subcategorization frames

   and
2. collocations

   at the same time.

The idea behind the VAB is: **treat VPCs as collocations.**

VERB PHRASE CONSTRUCTIONS AS COLLOCATIONS

We search for collocations in the space of these structures:

verb=jön	*SUB*=lét	*NOM*=csillag
verb=come	*INTO*=existence	*SUBJ*=star

IDEA

Apply an association measure (designed for bigrams) taking . . .

- the content word of a particular dependent – as one unit,
- *all other* parts of the verb frame – as the other unit

of the collocation.

Verb phrase constructions as collocations

We search for collocations in the space of these structures:

verb=jön	*SUB*=lét	*NOM*=?
verb=come	*INTO*=existence	*SUBJ*=?

Idea

Apply an association measure (designed for bigrams) taking . . .

- the content word of a particular dependent – as one unit,
- *all other* parts of the verb frame – as the other unit

of the collocation.

If we choose the subject's content word as the first unit,

VERB PHRASE CONSTRUCTIONS AS COLLOCATIONS

We search for collocations in the space of these structures:

verb=jön	*SUB*=lét	*NOM*=?
verb=come	*INTO*=existence	*SUBJ*=?

IDEA

Apply an association measure (designed for bigrams) taking . . .

- the content word of a particular dependent – as one unit,
- *all other* parts of the verb frame – as the other unit

of the collocation.

If we choose the subject's content word as the first unit, we will query the most important or usual subjects of this construction, namely *what* is used to come into existence.

Verb phrase constructions as collocations

The Verb Argument Browser can answer the following typical research question:

- What are the salient words which can appear as a particular dependent of a given verb frame?
- What are the most important collocates of a given verb (or verb frame) as a particular dependent?

Association measure: *salience* (adjusted mutual information)

$$S(x,y) = log_2 f(x) \cdot log_2 N \frac{f(x,y)}{f(x) \cdot f(y)}$$

Verb phrase constructions as collocations

Consequence:

The Verb Argument Browser can treat not just a single word but a whole verb frame (a verb together with some arguments) as one unit in collocation extraction.

It can collect . . .

- salient subjects of a verb,
- salient objects of a given verb–subject pair,
- salient locatives of a given verb–subject–object triplet . . .

1 Sentence model

2 Verb phrase constructions as collocations

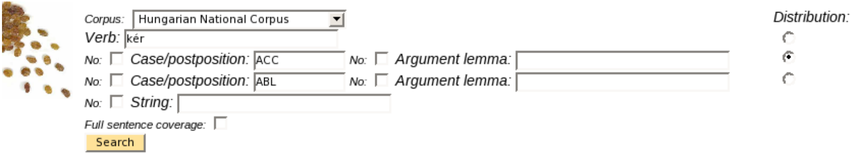
3 Usage & examples

4 Applications

5 Language independence

USAGE

- *corpus:* Hungarian National Corpus (187 million words)
- *response times:* a few seconds



Corpus: Hungarian National Corpus ▾ *Distribution:*

Verb: kér ○

No: □ *Case/postposition:* ACC *No:* □ *Argument lemma:* ●

No: □ *Case/postposition:* ABL *No:* □ *Argument lemma:* ○

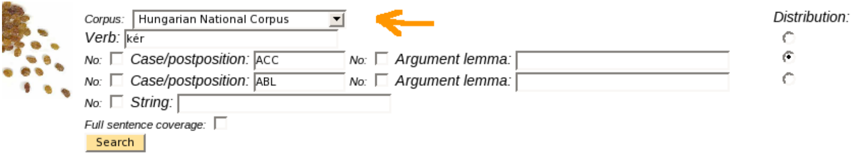
No: □ *String:*

Full sentence coverage: □

Search

USAGE

- *corpus:* Hungarian National Corpus (187 million words)
- *response times:* a few seconds



Corpus: Hungarian National Corpus

Verb: kér

No: □ *Case/postposition:* ACC *No:* □ *Argument lemma:*

No: □ *Case/postposition:* ABL *No:* □ *Argument lemma:*

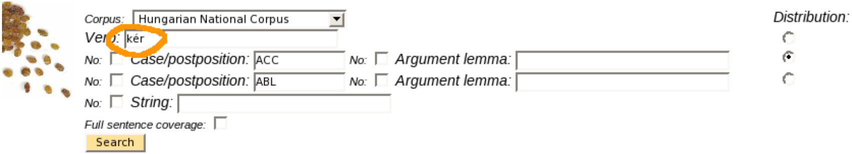
No: □ *String:*

Full sentence coverage: □

Distribution: ○ ● ○

USAGE

- *corpus:* Hungarian National Corpus (187 million words)
- *response times:* a few seconds



Corpus: Hungarian National Corpus

Verb: kér

No: □ Case/postposition: ACC No: □ Argument lemma:

No: □ Case/postposition: ABL No: □ Argument lemma:

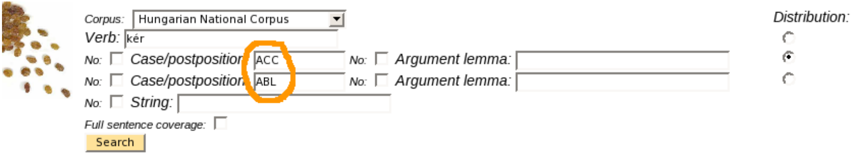
No: □ String:

Full sentence coverage: □

Distribution: ○ ● ○

USAGE

- *corpus:* Hungarian National Corpus (187 million words)
- *response times:* a few seconds



Corpus: Hungarian National Corpus

Verb: kér

No: □ *Case/postposition:* ACC *No:* □ *Argument lemma:*

No: □ *Case/postposition:* ABL *No:* □ *Argument lemma:*

No: □ *String:*

Full sentence coverage: □

Search

Distribution:

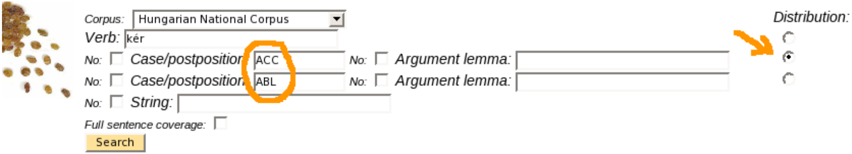
○

●

○

USAGE

- *corpus:* Hungarian National Corpus (187 million words)
- *response times:* a few seconds



Corpus: Hungarian National Corpus

Verb: kér

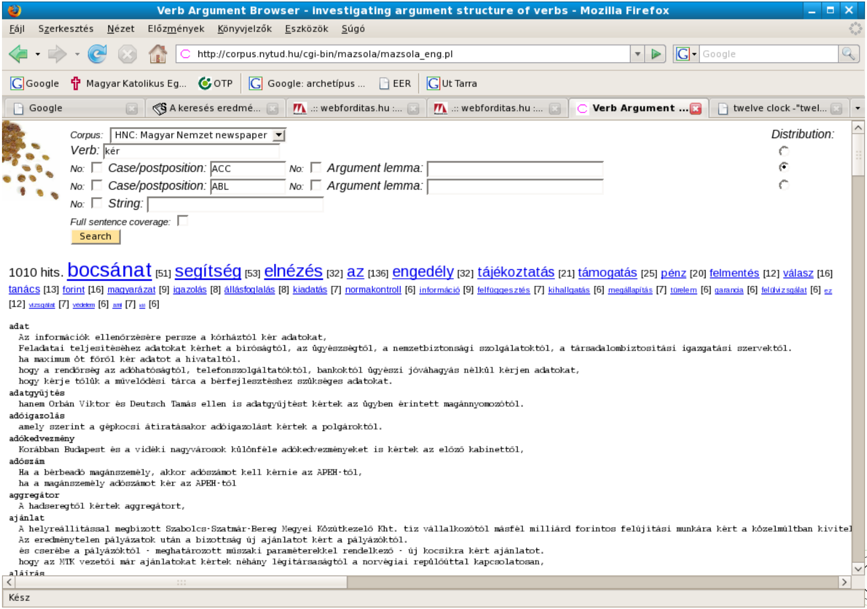
No: □ *Case/postposition:* ACC *No:* □ *Argument lemma:*

No: □ *Case/postposition:* ABL *No:* □ *Argument lemma:*

No: □ *String:*

Full sentence coverage: □

Distribution: ○ ● ○


Verb Argument Browser - investigating argument structure of verbs - Mozilla Firefox
Fájl Szerkesztés Nézet Előzmények Könyvjelzők Eszközök Súgó
http://corpus.nytud.hu/cgi-bin/mazsola/mazsola_eng.pl
Google
Google Magyar Katolikus Eg... OTP Google: archetípus ... EER Ut Tarra
Google A keresés eredmé... .:: webforditas.hu :... .:: webforditas.hu :... Verb Argument ... twelve clock -"twel...
Kész

Corpus: HNC: Magyar Nemzet newspaper

Verb: kér

No: □ Case/postposition: ACC No: □ Argument lemma:

No: □ Case/postposition: ABL No: □ Argument lemma:

No: □ String:

Full sentence coverage: □

Search

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 teljesitéséhez adatokat kérhet a biróságtól, az ügyészségtől, a nemzetbiztonsági szolgálatoktól, a társadalombiztositási 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.
adatgyüjtés
  hanem Orbán Viktor és Deutsch Tamás ellen is adatgyűjtést kértek az ügyben érintett magánnyomozótól.
adóigazolás
  amely szerint a gépkocsi átiratásakor 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 helyreállitással megbizott Szabolcs-Szatmár-Bereg Megyei Közútkezelő Kht. tiz vállalkozótól másfél milliárd forintos felújitási munkára kért a közelmúltban kivitel
  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
```

QUERY: 2 FREE DEPENDENTS

kér	-t	-tól
ask	ACC	ABL

’ask for SOMETHING from SOMEBODY’

verb=kér	*ABL=?*	*ACC=?*
verb=ask	*from=?*	*for=?*

QUERY: 2 FREE DEPENDENTS

kér	-t	-tól
ask	ACC	ABL

’ask for SOMETHING from SOMEBODY’

verb=kér	*ABL=*?	*ACC=*?
verb=ask	*from=*?	*for=*?

Result: (Most salient objects:)

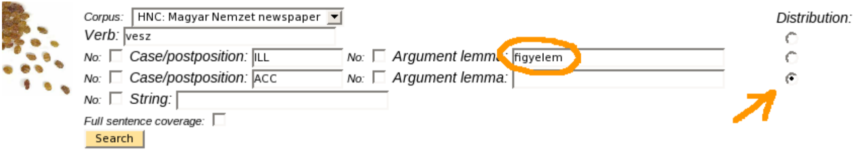
- bocsánat – ’forgiveness’
- segítség – ’help’
- elnézés – also ’forgiveness’
- engedély – ’permission’
- . . .

QUERY: 1 FIXED + 1 FREE DEPENDENT

vesz	figyelem-bA	-t
take	consideration-ILL	ACC

'take SOMETHING into consideration'

verb=vesz	*ILL*=figyelem	*ACC*=?
verb=take	*INTO*=consideration	*OBJ*=?



Corpus: HNC: Magyar Nemzet newspaper

Verb: vesz

No: □ Case/postposition: ILL No: □ Argument lemma: figyelem

No: □ Case/postposition: ACC No: □ Argument lemma:

No: □ String:

Full sentence coverage: □

Distribution: ○ ○ ●

QUERY: 1 FIXED + 1 FREE DEPENDENT

vesz	figyelem-bA	-t
take	consideration-ILL	ACC

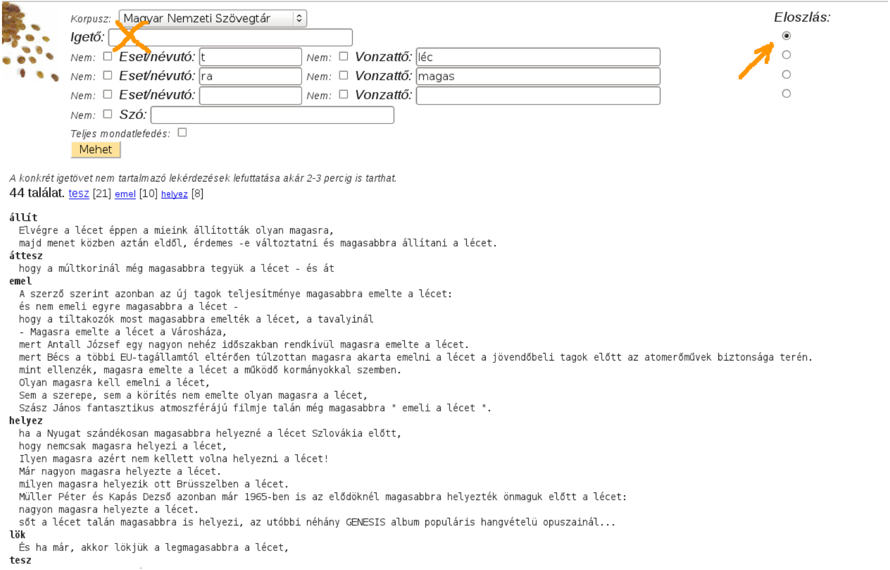
’take SOMETHING into consideration’

verb=vesz	*ILL*=figyelem	*ACC*=?
verb=take	*INTO*=consideration	*OBJ*=?

Result: (Most salient direct objects:)

- szempont – ’point of view’
- érdek – ’interest’
- vélemény – ’opinion’
- . . .

A TRICK: QUERYING THE VERB



Korpusz: Magyar Nemzeti Szövegtár

Igető:

Nem: □ Eset/névutó: t Nem: □ Vonzattő: léc

Nem: □ Eset/névutó: ra Nem: □ Vonzattő: magas

Nem: □ Eset/névutó: Nem: □ Vonzattő:

Nem: □ Szó:

Teljes mondatlefedés: □

Mehet

Eloszlás: ● ○ ○ ○

A konkrét igetövet nem tartalmazó lekérdezések lefuttatása akár 2-3 percig is tarthat.

44 találat. tesz [21] emel [10] helyez [8]

```
állít
  Elvégre a lécet éppen a mieink állították olyan magasra,
  majd menet közben aztán eldől, érdemes -e változtatni és magasabbra állítani a lécet.
áttesz
  hogy a múltkorinál még magasabbra tegyük a lécet - és át
emel
  A szerző szerint azonban az új tagok teljesítménye magasabbra emelte a lécet:
  és nem emeli egyre magasabbra a lécet -
  hogy a tiltakozók most magasabbra emelték a lécet, a tavalyinál
  - Magasra emelte a lécet a Városháza,
  mert Antall József egy nagyon nehéz időszakban rendkívül magasra emelte a lécet.
  mert Bécs a többi EU-tagállamtól eltérően túlzottan magasra akarta emelni a lécet a jövendőbeli tagok előtt az atomerőművek biztonsága terén.
  mint ellenzék, magasra emelte a lécet a működő kormányokkal szemben.
  Olyan magasra kell emelni a lécet,
  Sem a szerepe, sem a körítés nem emelte olyan magasra a lécet,
  Szász János fantasztikus atmoszférájú filmje talán még magasabbra " emeli a lécet ".
helyez
  ha a Nyugat szándékosan magasabbra helyezné a lécet Szlovákia előtt,
  hogy nemcsak magasra helyezi a lécet,
  Ilyen magasra azért nem kellett volna helyezni a lécet!
  Már nagyon magasra helyezte a lécet.
  milyen magasra helyezik ott Brüsszelben a lécet.
  Müller Péter és Kapás Dezső azonban már 1965-ben is az elődöknél magasabbra helyezték önmaguk előtt a lécet:
  nagyon magasra helyezte a lécet.
  sőt a lécet talán magasabbra is helyezi, az utóbbi néhány GENESIS album populáris hangvételü opuszainál...
lök
  És ha már, akkor lökjük a legmagasabbra a lécet,
tesz
```

A SIMPLE QUERY

ad	-t
give	ACC

'give SOMETHING'

verb=ad	*ACC*=?
verb=give	*OBJ*=?

A SIMPLE QUERY

ad	-t
give	ACC

'give SOMETHING'

verb=ad	*ACC*=?
verb=give	*OBJ*=?

Result: (Most salient direct objects:)

- hang – 'voice' → 'to give voice to SOMETHING'
- hír – 'news' → to give news ∼ 'to report'
- igaz – 'true' → to give true ∼ 'to take sy's side'
- . . .

→ **multi-word verbs**

Another simple query

üt	∅
beat	NOM

’SOMETHING beats’

verb=üt	*NOM*=?
verb=beat	*SUBJ*=?

ANOTHER SIMPLE QUERY

üt	∅
beat	NOM

'SOMETHING beats'

verb=üt	*NOM*=?
verb=beat	*SUBJ*=?

Result: (Some salient subjects:)

- óra – 'clock' → 'The clock strikes twelve.'
- forint → 10 Ft beat his palm. ∼ 'He receives 10 Ft.'
- kő – 'stone' → Üsse kő!
  – Let a stone beat it! ∼ 'It does not matter.'
- ...

→ **multi-word verbs, figures of speech**

COLLECTING MWVS

Important property of the Verb Argument Browser:

For any specific dependent, the tool provides constructions where *this dependent is fixed*, if there is any such construction.
(e.g. light verb constructions, idiomatic verbal expressions, figures of speech)

+ 'take' + into → 'consideration', 'account' . . .

– 'eat' + OBJ → just some kinds of food
in this case we obtain frequent words with literal meaning, often forming a semantically coherent class

COLLECTING MWVS

Important property of the Verb Argument Browser:

For any specific dependent, the tool provides constructions where *this dependent is fixed*, if there is any such construction.
(e.g. light verb constructions, idiomatic verbal expressions, figures of speech)

+ 'take' + into → 'consideration', 'account' . . .
– 'eat' + OBJ → just some kinds of food
  in this case we obtain frequent words with literal meaning, often forming a semantically coherent class

Warning: VPCs with fixed position(s) are frequent.

They are not to be ignored, it is necessary to deal with them.

E.g. when doing something with verbs, do not forget about multi-word verbs.

Corpus size ↔ annotation richness

- compared to…
  – large raw or POS-tagged corpora ($\sim$ big data)
  – small syntactically annotated corpora ($\sim$ rich information)
- a VAB uses (and works well with) corpora which are
  big enough + have "some" syntactic information annotated
- using this approach
  corpus-driven information can be gathered
  about some *"higher level"* phenomena
  (i.e. the predicate-argument structure in our case)
  based on querying a quite *large* piece of text

→ a big corpus (size $\gg$ treebanks)
with *some shallow* syntactic annotation
can be a valuable resource. :)

1 Sentence model

2 Verb phrase constructions as collocations

3 Usage & examples

4 Applications

5 Language independence

APPLICATIONS

Past…

- lexical database development of a Hungarian to English machine translation system – handling MWVs
- searching for MWVs to include them into the Hungarian WordNet
- lexicography

Future…

- language teaching
- determining the frequency of particular VPCs to be used in experiments in psycholinguistic research
- linguistic research
  - studying verb synonyms
  - classifying verbs based on argument structure similarities
  - studying selectional preferences of verbs

1 Sentence model

2 Verb phrase constructions as collocations

3 Usage & examples

4 Applications

5 Language independence

Language independence

Claim: the approach is language independent.

- To extend the methodology to other languages all we need is a corpus represented according to our sentence model.
- Can we create such a representation?
- Essentially, the representation relies on the very fact that there are *some kind of predicate-argument structure* in the languages.
- All we should do is segment the text into sentences/clauses (containing a verb and its dependents) and specify the relationship between the verb and the particular dependents.

A PILOT VAB FOR SERBIAN

Just to show that the approach works. :)

Serbian: dependents are defined by
case markers *or* case marker + preposition combinations.

- much smaller corpus (Intera)
- much simpler (pre)processing
    1. clause boundary detection = just split at punctuations
    2. verb identification = take the last verb + attach *se* if occurs
    3. noun phrase chunking = extract PPs according to this simple pattern: *a preposition + possible not-nouns + a noun*
    4. no case information:
       all NPs without preposition
       go to a big class (called ANYCASE)

Representation of a Serbian clause

Example clause from the corpus:

Svako	ima	pravo	na	rad.
Everyone	has	right	to/for	work.

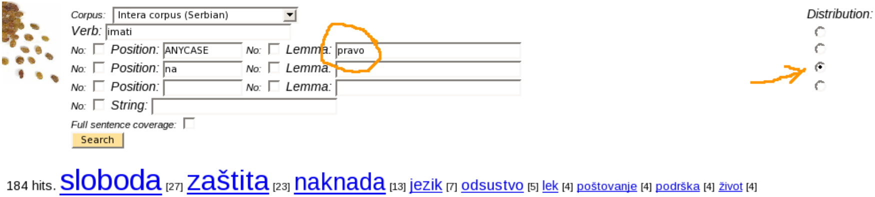
'Everyone has the right to work.'

Representation of the example clause:

verb=imati	*ANYCASE*=pravo	*na*=rad
verb=have	*OBJ*=right	*to/for*=work

EXAMPLE: A MULTIWORD VERB

Query:
’imati pravo na’ (’have right to/for’)



Corpus: Intera corpus (Serbian)
Verb: imati
No: □ Position: ANYCASE No: □ Lemma: pravo
No: □ Position: na No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ String:
Full sentence coverage: □
Search

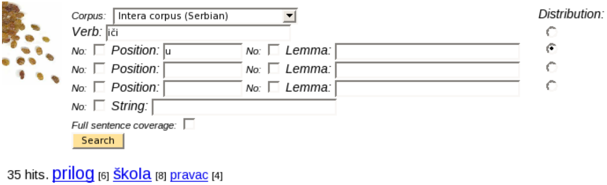
Distribution:
○
○
●
○

184 hits. sloboda [27] zaštita [23] naknada [13] jezik [7] odsustvo [5] lek [4] poštovanje [4] podrška [4] život [4]

freedom, protection, compensation, language . . .

Example: discovering multiword verbs

Query:
’ići u’ (’go in’)



Corpus: Intera corpus (Serbian)
Verb: iči
No: □ Position: u No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ String:
Full sentence coverage: □
Search

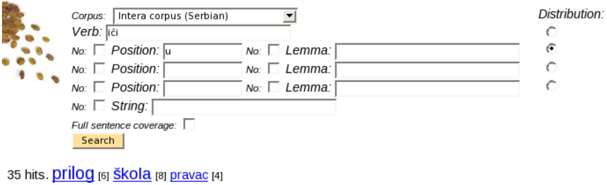
Distribution:
○
●
○
○

35 hits. prilog [6] škola [8] pravac [4]

benefit, school, direction

EXAMPLE: DISCOVERING MULTIWORD VERBS

Query:
’ići u’ (’go in’)



Corpus: Intera corpus (Serbian)
Verb: iči
No: □ Position: u No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ Position: No: □ Lemma:
No: □ String:
Full sentence coverage: □

Distribution:
○
●
○
○

35 hits. prilog [6] škola [8] pravac [4]

benefit, school, direction

prilog does not fit into this little semantic class.
→ This phenomenon is a good indicator of being a MWV!

’ići u prilog’ is a MWV.
Meaning: ∼ *support* (?) – literally: ’go in benefit’ (?)

LANGUAGE INDEPENDENCE

The methodology can be extended to other languages,
and a fully functioning VAB can be created
if a shallow parsed, adequately processed corpus is available.

Availability

Available for you:

- Hungarian version:

`http://corpus.nytud.hu/vab`

- Serbian version:

`http://corpus.nytud.hu/vabs`

username: `eger`; password: `vab`

Availability

Available for you:

- Hungarian version:

```
http://corpus.nytud.hu/vab
```

- Serbian version:

```
http://corpus.nytud.hu/vabs
```

username: `eger`; password: `vab`

Thank you for your attention!