Querying Multi-word Expressions Annotation with CQL

Natalia Klyueva, Anna Vernerová, Behrang QasemiZadeh
natalia.klyueva@polyu.edu.hk, vernerova@ufal.mff.cuni.cz, zadeh@phil.hhu.de

Motivation

Querying corpora with multi-word expression (MWE) annotation using a concordance system.

PARSEME data

PARSEME corpus vesrion 1.0:
Bulgarian, Czech, Farsi, French, German, Greek, Hebrew, Hungarian, Italian, Lithuanian, Maltese, Polish, Brazilian Portuguese, Romanian, Slovene, Spanish, Swedish and Turkish

- idioms (ID)
- light verb constructions (LVC)
- inherently reflexive verbs (IReflV)
- verb-particle constructions (VPC)
- OTH (other)

Two files: .conllu and .parsemetsv

1	Tím	ten	DET	PD	Case=Ins|Gender=Masc,Neut|Number=Sing|PronType=Dem	…	1	Tím	_	_
2	pádem	pád	NOUN	NN	Animacy=Inan|Case=Ins|Gender=Masc|Negative=Pos|Number=Sing	…	2	pádem	_	_
3	máme	mít	VERB	VB	Mood=Ind|Negative=Pos|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	…	3	máme	_	1:LVC
4	problém	problém	NOUN	NN	Animacy=Inan|Case=Acc|Gender=Masc|Negative=Pos|Number=Sing	…	4	problém	_	1
5	se	s	ADP	RV	AdpType=Voc|Case=Ins	…	5	se	_	_
6	silniční	silniční	ADJ	AA	Case=Ins|Degree=Pos|Gender=Fem|Negative=Pos|Number=Sing	…	6	silniční	_	_
7	daní	daň	NOUN	NN	Case=Ins|Gender=Fem|Negative=Pos|Number=Sing	…	7	daní	nsp	_
8	.	.	PUNCT	Z:	_	…	8	.	_	_

Format conversion

Attributes representing the MWE annotation:

- **mwe** MWE type, e.g. LVC, ID
- **mwe_order** with values **first**, **cont** (continuation)
- **mwe_order_new** with values **first**, **cont**, **last**
- **mwe_id** id of the MWE, unique within a sentence
- **mwe_lemma** concatenation of lemmas

1	Tím	ten	DET	PD	…	_	_	_	_	_
2	pádem	pád	NOUN	NN	…	_	_	_	_	_
3	máme	mít	VERB	VB	…	LVC	first	first	1	mít problém
4	problém	problém	NOUN	NN	…	LVC	cont	last	1	mít problém
5	se	s	ADP	RV	…	_	_	_	_	_
6	silniční	silniční	ADJ	AA	…	_	_	_	_	_
7	daní	daň	NOUN	NN	…	_	_	_	_	_
8	.	.	PUNCT	Z:	…	_	_	_	_	_

Figure 1: Simple occurrence of VMWE in Czech

Les	…	_	_	_	_	_
œuvres	…	_	_	_	_	_
d’	…	_	_	_	_	_
Abbas	…	_	_	_	_	_
Kiarostami	…	_	_	_	_	_
font	…	ID;LVC;LVC	first;first;first	first;first;first	1;3;2	faire le objet;faire le objet festival;faire le objet exposition
régulièrement	…	_	_	_	_	_
l’	…	ID;LVC;LVC	cont;cont;cont	cont;cont;cont	1;3;2	faire le objet;faire le objet festival;faire le objet exposition
objet	…	ID;LVC;LVC	cont;cont;cont	last;cont;cont	1;3;2	faire le objet;faire le objet festival;faire le objet exposition
d’	…	_	_	_	_	_
expositions	…	LVC	cont	last	2	faire le objet exposition
ou	…	_	_	_	_	_
de	…	_	_	_	_	_
festival	…	LVC	cont	last	3	faire le objet festival
,	…	_	_	_	_	_
…						

Figure 2: Representation of overlapping VMWEs

Query systems

The converted vertical format is suitable for any instance of the Manatee-open corpus search engine. The data is currently available through two front-ends. Both use the same query language, namely CQL.

kontext NoSketch Engine

Querying the data in KonText

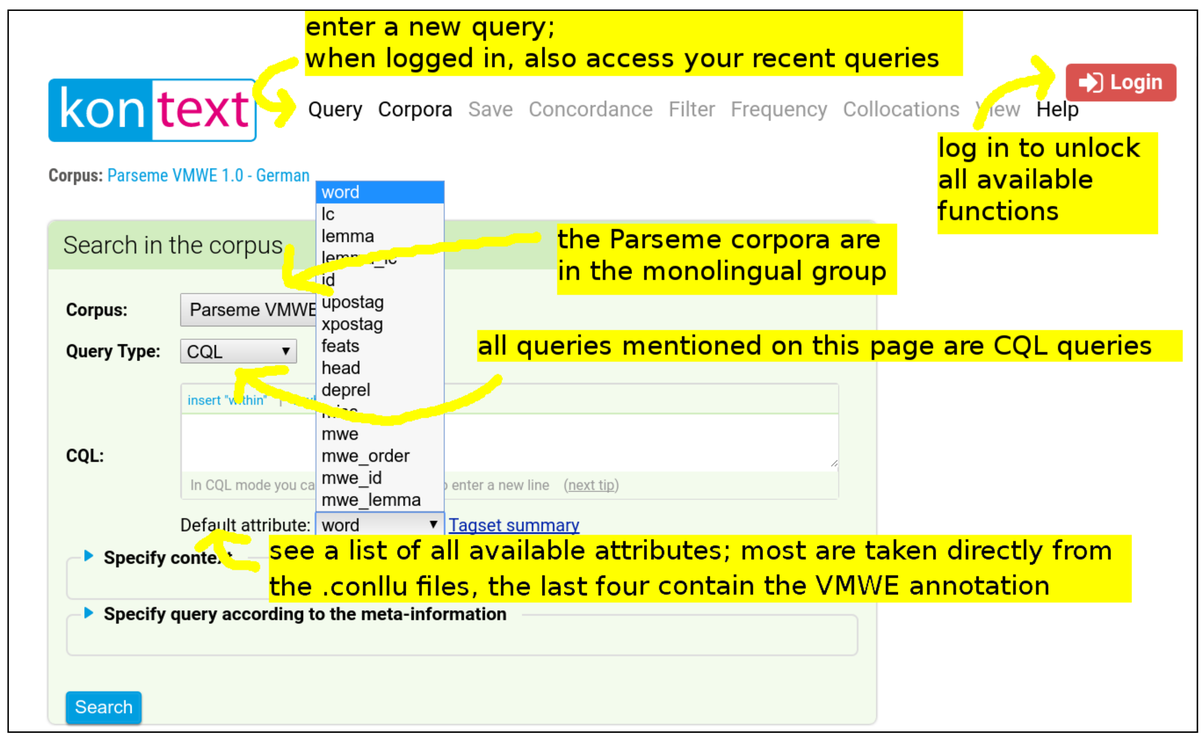

Figure 3: List of attributes for CQL query

- `[mwe_order=”first”]`
  find the first token in each MWE annotated in the corpus
- `[lemma=”faire” & mwe!=”_”]`
  find a particular word annotated with any category of MWE
- `[mwe=”LVC” & upostag=”VERB”]`
  find the verbs in light verb constructions
- `[mwe_lemma=”faire partie”]`
  search by a concatenation of the lemmas of words belonging to the MWE (in the order in which they appear in the text)
- `[mwe_order_new=”first”][mwe_order_new=”cont”]* [mwe_order_new=”last”]`
  display and highlight continuous MWEs consisting of at least two words
- `1:[mwe_order_new=”first”] []* 2:[mwe_order_new=”last”] & 1.mwe_id=2.mwe_id within <s/>`
  matches each MWE consisting of at least two words together with any words lying between its first and last word

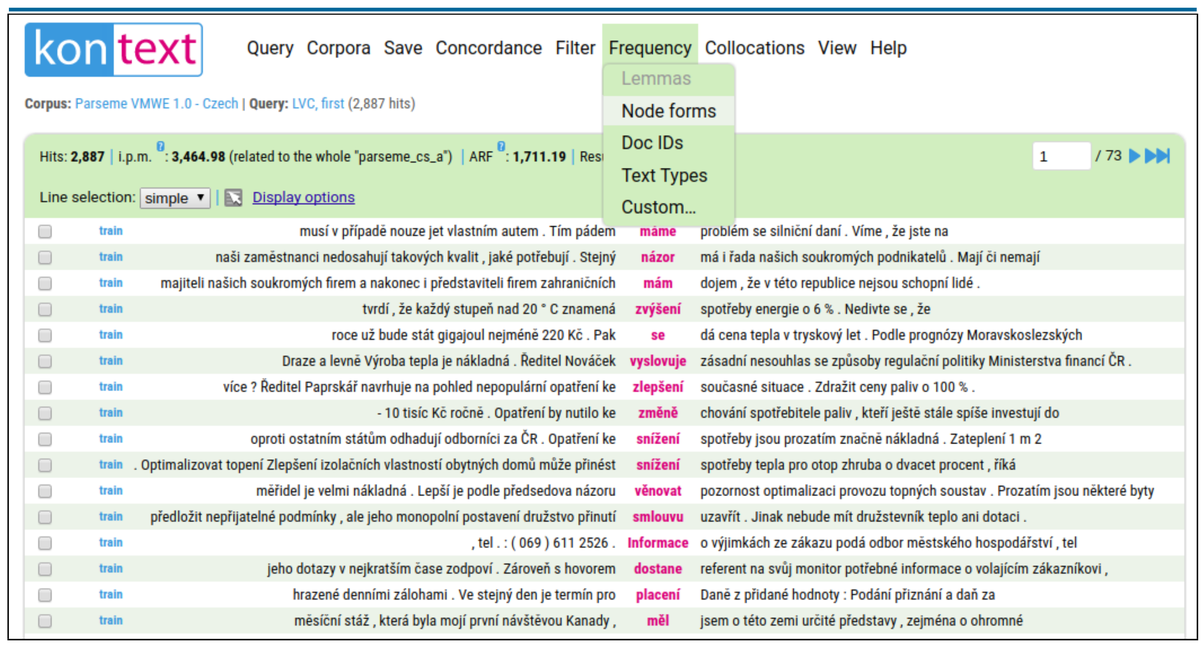

Figure 4: Click ”Frequency->Node forms” to sort according to a token frequency

Find Out More

http://ufal.mff.cuni.cz/lindat-kontext/parseme-mwe

LINDAT CLARIN · ÚFAL · Mathematica Physicaque Facultas · CHARLES UNIVERSITY Faculty of mathematics and physics · The Hong Kong Polytechnic University 香港理工大學 · Heinrich Heine Universität Düsseldorf