

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 version 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 (IRefIV)
- 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
ouvrés
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	...	ID;LVC;LVC	cont;cont;cont	cont;cont;cont	1;3;2	faire le objet;faire le objet festival;faire le objet exposition
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.



Querying the data in KonText

enter a new query; when logged in, also access your recent queries

kon text Query Corpora Save Concordance Filter Frequency Collocations View Help Login

log in to unlock all available functions

the Parseme corpora are in the monolingual group

all queries mentioned on this page are CQL queries

see a list of all available attributes; most are taken directly from the .conllu files, the last four contain the VMWE annotation

Specify content

Specify query according to the meta-information

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

kon text Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: Parseme VMWE 1.0 - Czech | Query: LVC, first (2,887 hits)

Hits: 2,887 | 1 p.m. 3,464.98 (related to the whole 'parseme_cs_a') | ARF 1,711.19 | Res 1 / 73

Line selection: simple | Display options

Attributes: Lemmas, Node forms, Doc IDs, Text Types, Custom...

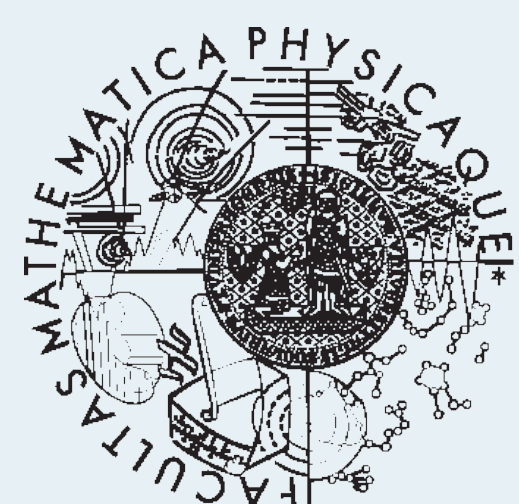
Train musí v případě nouze jet vlastním autem . Tím pádem naše zaměstnanci nedosahují takových kvalit , jaké potřebují . Stejný majitel našich soukromých firem a nakonec i představitel firem zahraničních tvrdí , že každý stupeň nad 20 ° C znamená roce už bude stát gigajoul nejméně 220 Kč . Pak Draze a levně Výroba tepla je nákladná . Ředitel Nováček více ? Ředitel Paprskář navrhuje na pohled nepopulární opatření ke - 10 tisíc Kč ročně . Opatření by nutilo ke oproti ostatním státům odhadují odborníci za ČR . Opatření ke . Optimalizovat topení Zlepšení izolačních vlastností obytných domů může přinést měřidel je velmi nákladná . Lepší je podle předseda názoru předložit nepřijatelné podmínky , ale jeho monopolní postavení družstvo přinutí , tel. : (069) 611 2526 . jeho dotazy v nejbližším čase zodpoví . Zároveň s hovorem hrazené denními zálohami . Ve stejný den je termín pro měsíční stáž , která byla mojí první návštěvou Kanady ,

name problém se silniční daní . Víme , že jste na má i řada našich soukromých podnikatelů . Mají či nemají dojem , že v této republice nejsou schopní lidé . zvýšení spotřeby energie o 6 % . Nedivte se , že dá cena tepla v tryskový let . Podle prognózy Moravskoslezských zásadní nesouhlas se způsobem regulační politiky Ministerstva financí ČR . současné situace . Zdražit ceny paliv o 100 % . změně chování spotřebitele paliv , kteří ještě stále spíše investují do spotřeby jsou prozatím značně nákladná . Zateplení 1 m 2 spotřeby tepla pro otop zhruba o dvacet procent , říká snížení spotřeby tepla pro otop zhruba o dvacet procent , říká věnovat pozornost optimalizaci provozu topných soustav . Prozatím jsou některé byty uzavřít . Jinak nebude mít družstevník teplo ani dotaci . Informace o výjimkách ze zákazu podá odbor městského hospodářství , tel referent na svůj monitor potřebné informace o volajícím zákazníkovi . placení Daně z přidané hodnoty : Podání příznání a daň za měl jsem o této zemi určité představy , zejména o ohromné

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>



CHARLES UNIVERSITY
Faculty of mathematics
and physics



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

