Natalia Klyueva,
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University, Czech Republic
kljueva@ufal.mff.cuni.cz

http://lindat.mff.cuni.cz/services/kontext

MWEs in a corpus query system

Corpus search engines like SketchEngine, IMS Corpus Workbench and others search in a linear representation of a sentence as "attributed strings". They display results of the searches as concordances. In comparison to that, treebank search engines like PML-TQ or INESS or Tundra allow sophisticated queries for tree structures of treebanks and they typically display results as trees (one at a time). The expressive power of CQL query language is limited in comparison with full treebank search engines, but the query language is simpler and intuitive.

Joint treebank: Number of words in different languages with the deprel attribute **mwe**.
[deprel="mwe"]

		doc.id	Freq	i.p.m.
1.	p / n	Swedish	1,105	16580.4
2.	p / n	French	1,058	8723.8
3.	p / n	Indonesian	1,026	10519.7
4.	p / n	Spanish	862	6474.7
5.	p / n	Italian	454	4300.3
6.	p / n	Danish	436	4900.0
7.	p / n	Bulgarian	329	5238.2
8.	p / n	Slovenian	305	3484.6
9.	p / n	Croatian	251	3184.6
10.	p / n	English	188	2462.6
11.	p / n	Basque	152	2293.6
12.	p / n	Czech	149	1913.8
13.	p / n	Romanian	53	5704.4
14.	p / n	German	52	597.2
15.	p / n	Dutch	23	494.3
16.	p / n	Polish	16	312.3
17.	p / n	Gothic	3	67.1
18.	p / n	Portuguese	1	7.8
19.	p / n	Old Church Slavonic	1	22.0
20.	p / n	Greek	1	21.1

Joint treebank: Number of words with the deprel attribute **name** featuring from 2 tokens **[deprel="name"]{2,9}**

		doc.id	Freq	i.p.m.
1.	p / n	Indonesian	3,668	37608.6
2.	p / n	Spanish	588	4416.6
3.	p / n	French	463	3817.7
4.	p / n	Norwegian	352	4307.5
5.	p / n	Danish	306	3439.0
6.	p / n	German	245	2813.8
7.	p / n	Basque	215	3244.2
8.	p / n	Hungarian	202	9728.4
9.	p / n	Croatian	136	1725.5
10.	p / n	English	116	1519.5
11.	p / n	Italian	107	1013.5
12.	p / n	Irish	50	2993.8
13.	p / n	Romanian	22	2367.9
14.	p / n	Bulgarian	13	207.0
15.	p / n	Slovenian	12	137.1
16.	p / n	Swedish	7	105.0
17.	p / n	Czech	1	12.8

Joint treebank: Number of words with the deprel attribute **compound**.
[deprel="compound"]

		doc.id	Freq	i.p.m.
1.	p / n	Hindi	10,288	101123.5
2.	p / n	Indonesian	5,947	60975.5
3.	p / n	English	3,738	48963.2
4.	p / n	Croatian	2,344	29739.8
5.	p / n	Irish	1,139	68199.5
6.	p / n	Basque	902	13610.6
7.	p / n	Czech	684	8785.7
8.	p / n	German	433	4973.0
9.	p / n	French	239	1970.7
10.	p / n	Italian	205	1941.7
11.	p / n	Spanish	138	1036.5
12.	p / n	Finnish	116	2744.3
13.	p / n	Portuguese	76	590.5
14.	p / n	Slovenian	68	776.9
15.	p / n	Dutch	26	558.7
16.	p / n	Hungarian	24	1155.8
17.	p / n	Swedish	15	225.1
18.	p / n	Estonian	13	1691.2
19.	p / n	Tamil	12	1896.0
20.	p / n	Romanian	12	1291.6
21.	p / n	Greek	12	105.4

UD English: Concordance with a tree. Query: **[deprel="mwe"]{2,3}**

UD French: Concordance with a tree. Query: **[deprel="name"]{3}**

English: phrasal verbs: distance between a verb and a particle:
[pos="VERB"] [pos!="VERB"]{2,5} [deprel="compound:prt"]

UD English: two-word MWEs

		word	word	Freq
1.	p / n	as	well	48
2.	p / n	due	to	41
3.	p / n	because	of	38
4.	p / n	such	as	27
5.	p / n	so	that	26
6.	p / n	more	than	20
7.	p / n	instead	of	18
8.	p / n	according	to	16
9.	p / n	rather	than	14
10.	p / n	prior	to	14
11.	p / n	of	course	14
12.	p / n	in	order	14
13.	p / n	less	than	14
14.	p / n	at	least	13
15.	p / n	up	to	11
16.	p / n	kind	of	10

three-word MWEs

		word	Freq
1.	p / n	as well as	22
2.	p / n	whether or not	8
3.	p / n	not to mention	3
4.	p / n	so as to	2
5.	p / n	as opposed to	2
6.	p / n	Not to mention	2

UD French: two-word MWEs

		word	word	Freq
1.	p / n	tant	que	54
2.	p / n	plus	de	42
3.	p / n	près	de	25
4.	p / n	lors	de	25
5.	p / n	ainsi	que	25
6.	p / n	moins	de	14
7.	p / n	plus	d'	13
8.	p / n	alors	que	13
9.	p / n	même	si	12
10.	p / n	grâce	à	12
11.	p / n	à	plus	11
12.	p / n	beaucoup	de	10
13.	p / n	tant	qu'	10
14.	p / n	pas	de	10

three-word MWEs

		word	Freq
1.	p / n	plus en plus	13
2.	p / n	jusqu' à ce	7
3.	p / n	tout à fait	5
4.	p / n	tout de même	5
5.	p / n	de plus en	5
6.	p / n	à partir de	3
7.	p / n	à ce que	3
8.	p / n	bel et bien	3
9.	p / n	un peu de	2

English: phrasal verbs: **[pos="VERB"][deprel="compound:prt"]** & statistics

		lemma	word	Freq
1.	p / n	show	up	22
2.	p / n	find	out	22
3.	p / n	end	up	17
4.	p / n	pick	up	17
5.	p / n	check	out	17
6.	p / n	go	on	11
7.	p / n	work	out	11
8.	p / n	figure	out	11
9.	p / n	turn	out	8
10.	p / n	give	up	8
11.	p / n	follow	up	7
12.	p / n	take	over	7
13.	p / n	make	up	7
14.	p / n	carry	out	6
15.	p / n	take	on	6
16.	p / n	take	off	6
17.	p / n	grow	up	6
18.	p / n	save	up	5
19.	p / n	call	in	5
20.	p / n	watch	out	4

particles that are used most frequently

		word	Freq
1.	p / n	up	220
2.	p / n	out	180
3.	p / n	off	44
4.	p / n	on	35
5.	p / n	down	32
6.	p / n	in	23
7.	p / n	over	12
8.	p / n	away	7
9.	p / n	back	5
10.	p / n	of	3

Manual:

https://github.com/czcorpus/kontext   **http://ufal.mff.cuni.cz/lindat-kontext**

LINDAT CLARIN

ÚFAL

ÚSTAV FORMÁLNÍ A APLIKOVANÉ LINGVISTIKY