New word sketch data format

Miloš Jakubíček

Lexical Computing

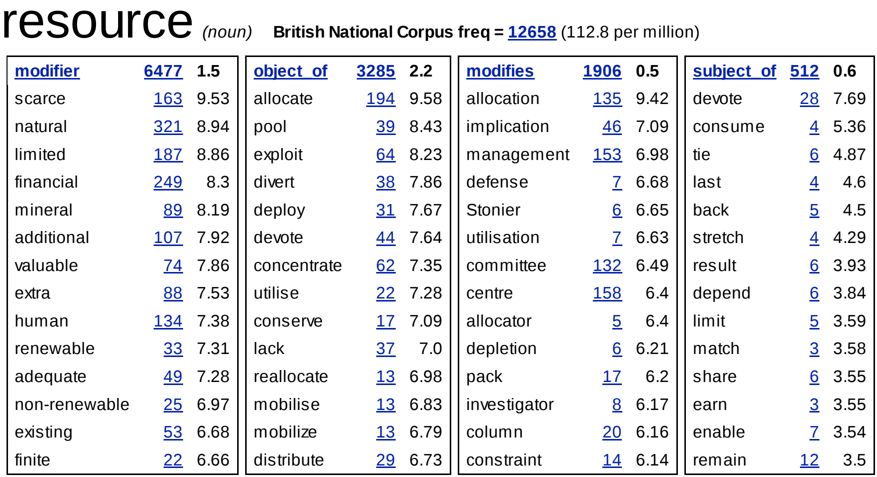
milos.jakubicek@sketchengine.co.uk

6th Sketch Engine Workshop
Herstmonceux, August 10, 2015

Word sketches

- you know them everybody

Word sketch



resource *(noun)* **British National Corpus freq = 12658** (112.8 per million)

modifier	6477	1.5
scarce	163	9.53
natural	321	8.94
limited	187	8.86
financial	249	8.3
mineral	89	8.19
additional	107	7.92
valuable	74	7.86
extra	88	7.53
human	134	7.38
renewable	33	7.31
adequate	49	7.28
non-renewable	25	6.97
existing	53	6.68
finite	22	6.66

object_of	3285	2.2
allocate	194	9.58
pool	39	8.43
exploit	64	8.23
divert	38	7.86
deploy	31	7.67
devote	44	7.64
concentrate	62	7.35
utilise	22	7.28
conserve	17	7.09
lack	37	7.0
reallocate	13	6.98
mobilise	13	6.83
mobilize	13	6.79
distribute	29	6.73

modifies	1906	0.5
allocation	135	9.42
implication	46	7.09
management	153	6.98
defense	7	6.68
Stonier	6	6.65
utilisation	7	6.63
committee	132	6.49
centre	158	6.4
allocator	5	6.4
depletion	6	6.21
pack	17	6.2
investigator	8	6.17
column	20	6.16
constraint	14	6.14

subject_of	512	0.6
devote	28	7.69
consume	4	5.36
tie	6	4.87
last	4	4.6
back	5	4.5
stretch	4	4.29
result	6	3.93
depend	6	3.84
limit	5	3.59
match	3	3.58
share	6	3.55
earn	3	3.55
enable	7	3.54
remain	12	3.5

Word sketch format 1 (2003)

- triples: headword – relation – collocation
- in a corpus: include headword position

⇒ 4-tuple (h, r, c, h_{pos})

Word sketch format 2 (2010)

- include collocation position

⇒ 5-tuple $(h, r, c, h_{pos}, c_{pos})$

Word sketch format 3 (2014)

- include longest-commonest match (**Vít Baisa, Wed 10.20**)
- more scalable – for corpora over 100 billion words

⇒ n-tuple $(h, r, c, h_{pos}, c_{pos}, LCM_1, \ldots, LCM_n)$

Word sketch format 4 (2015)

- score computation changes
- see `https://www.sketchengine.co.uk/statistics-used-in-sketch-engine/`

Conclusions

New word sketch format being introduced:

- biggest changes since 2003
- affects local installations
- changes score computations
- is backward compatible, part of Manatee 2.125
- smaller indices, faster access
- corpora in SkE now being rebuilt