Sketch Engine as a Platform for Providing Corpora

Miloš Jakubíček

Lexical Computing

Brighton, United Kingdom

milos.jakubicek@sketchengine.co.uk

META-FORUM 2013
Berlin, September 20th, 2013

Premise

When building a language resource one aims at many users and many uses.

Premise

When building a language resource one aims at many users and many uses.

But:

- Deciding to use a language resource is a commitment.
- People want to explore before committing.

The Problem

If you have a language resource, how do you show it off?

- You give talks
- You send samples

The Problem

If you have a language resource, how do you show it off?

- You give talks
  - Fine, but only for first contact
- You send samples
  - Often not satisfactory by design
  - Cumbersome to deal with annotation schemes
  - Hard to compare with others

The Problem

If you have a language resource, how do you show it off?

- You give talks
  - Fine, but only for first contact
- You send samples
  - Often not satisfactory by design
  - Cumbersome to deal with annotation schemes
  - Hard to compare with others

Showcasing resources is central to a language resource programme.

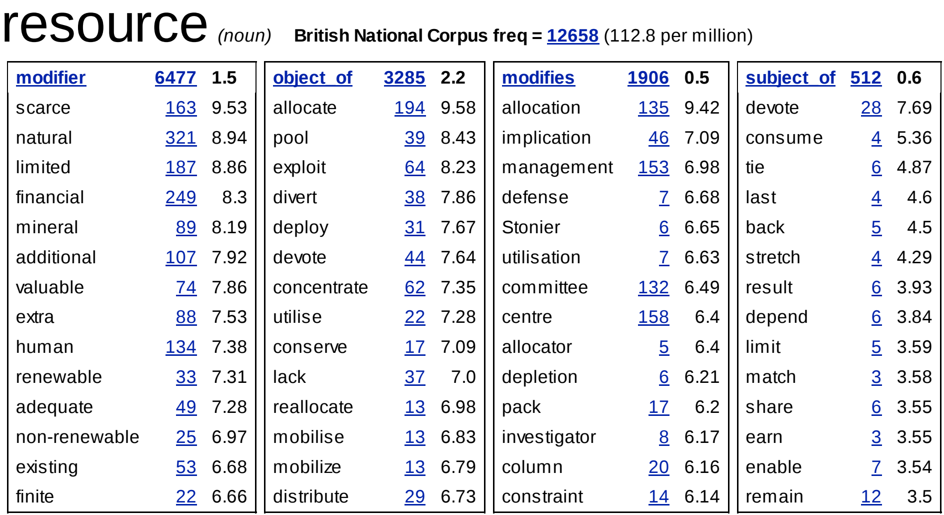
Sketch Engine

- corpus query system
- web service (including API)
- widely used for
  - lexicography purposes
    - Oxford University Press, Cambridge University Press, Harper Collins, Macmillan, ...
  - linguistic and language technology teaching and research at universities
    - about 100 academic institutions worldwide
    - thousands of individuals

Sketch Engine features

- **concordancing**, sorting, sampling, wordlists, collocation lists
- full **regular-expression** searching
- support for **parallel corpora**, virtual sub- and supercorpora
- handles **billion-word (80 G+)** corpora smoothly
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **Corpus Architect**: user corpora
  - uploaded by users
  - created by WebBootCaT

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

Sketch Engine languages

By September 2013 more than **400 corpora** for **70 languages**:

- 100+ corpora having more than 100 million tokens
- 30+ corpora having more than 1 billion tokens
  - In 2010 a series of TenTen (10^{10}) corpora started
- 56 languages with a PoS-tagged corpus
- 36 languages with word sketches
- 21 languages with integrated tagger for tagging user corpora

Why to use Sketch Engine as a commercial service?

Offering a showcasing programme for resources valuable for general audience, the company:

- takes the costs of providing the service and maintenance
- may gain additional customers

The resource developers:

- have their resources showcased at **no cost**,
- get access to their resources at **no cost**,
- get access to other resources in Sketch Engine **often at no cost too**,

A win-win scenario, a simple agreement easy to implement.

Case studies

- Slovene (FidaPLUS corpus, university consortium)
- Japanese (JpWaC corpus, Tokyo Institute of Technology)
- Chinese (Chinese GigaWord corpus, Academica Sinica)
- German, Italian, English (*WaC corpora, by M. Baroni)
- Czech (Czech National Corpus, ICNC)

- local installations possible as well (IP issues)
- large parts of SkE available as open-source NoSketch Engine

Conclusions

- Showcasing resources is central to a language resource programme.
- The premise: resources will be used by a number of groups
  - ⇒ If they cannot easily be assessed, they will not be.
- Sketch Engine is a leading corpus query tool, providing corpora online and offering a showcasing programme.
- We will be happy to extend the number of language resources, if you are interested, do talk to us!