Sketch Engine for Terminology

Miloš Jakubíček

Lexical Computing

Brighton, UK & Brno, CZ
milos.jakubicek@sketchengine.co.uk

November 6, 2015
11th International Conference on Terminology and Artificial Intelligence
University of Granada

Where we are
&
Where we go

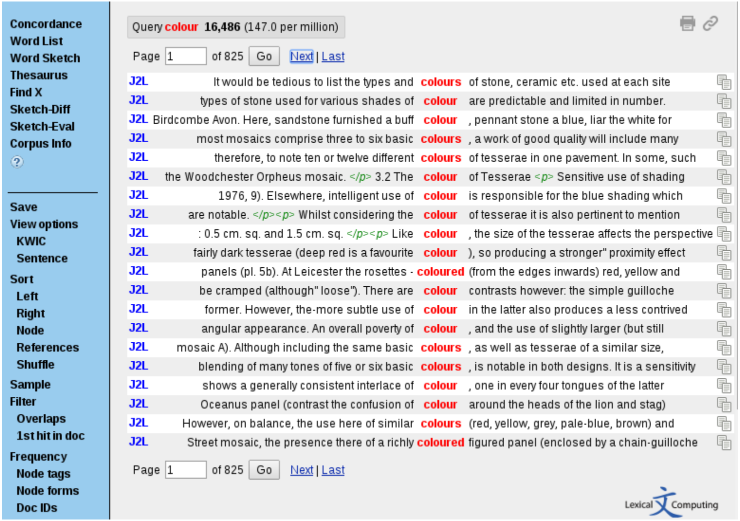
Sketch Engine

- corpus management system
- web service (including API)
- platform for providing language resources
- widely used for
  - lexicography purposes
    - Harper Collins, Oxford University Press, Cambridge University Press, Macmillan, . . .
  - linguistic and language technology teaching and research at universities
    - more than 100 academic institutions worldwide
    - dozens of thousands of individuals
  - language modelling (IT/LT companies)

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

Concordance search



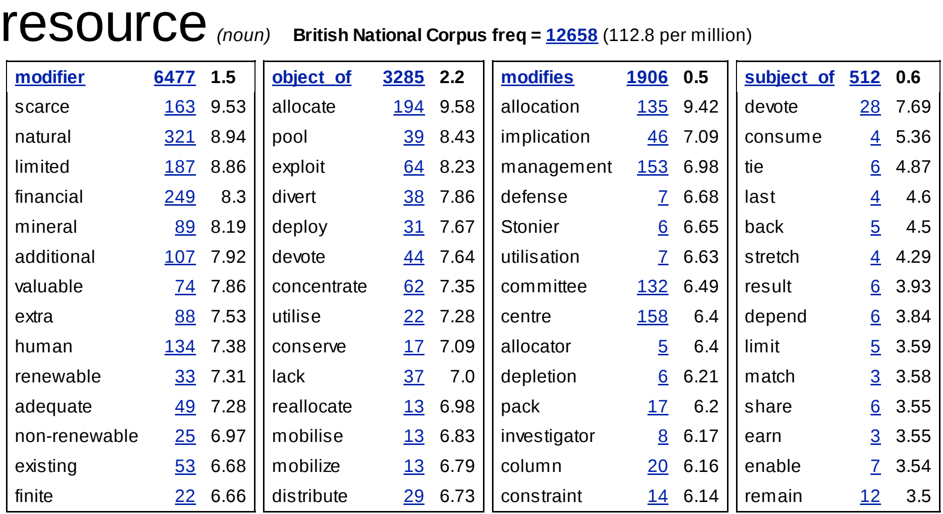
Query colour 16,486 (147.0 per million)

Page 1 of 825 Go Next | Last

J2L	It would be tedious to list the types and	colours	of stone, ceramic etc. used at each site
J2L	types of stone used for various shades of	colour	are predictable and limited in number.
J2L	Birdcombe Avon. Here, sandstone furnished a buff	colour	, pennant stone a blue, liar the white for
J2L	most mosaics comprise three to six basic	colours	, a work of good quality will include many
J2L	therefore, to note ten or twelve different	colours	of tesserae in one pavement. In some, such
J2L	the Woodchester Orpheus mosaic. </p> 3.2 The	colour	of Tesserae <p> Sensitive use of shading
J2L	1976, 9). Elsewhere, intelligent use of	colour	is responsible for the blue shading which
J2L	are notable. </p><p> Whilst considering the	colour	of tesserae it is also pertinent to mention
J2L	: 0.5 cm. sq. and 1.5 cm. sq. </p><p> Like	colour	, the size of the tesserae affects the perspective
J2L	fairly dark tesserae (deep red is a favourite	colour	), so producing a stronger" proximity effect
J2L	panels (pl. 5b). At Leicester the rosettes -	coloured	(from the edges inwards) red, yellow and
J2L	be cramped (although" loose"). There are	colour	contrasts however: the simple guilloche
J2L	former. However, the-more subtle use of	colour	in the latter also produces a less contrived
J2L	angular appearance. An overall poverty of	colour	, and the use of slightly larger (but still
J2L	mosaic A). Although including the same basic	colours	, as well as tesserae of a similar size,
J2L	blending of many tones of five or six basic	colours	, is notable in both designs. It is a sensitivity
J2L	shows a generally consistent interlace of	colour	, one in every four tongues of the latter
J2L	Oceanus panel (contrast the confusion of	colour	around the heads of the lion and stag)
J2L	However, on balance, the use here of similar	colours	(red, yellow, grey, pale-blue, brown) and
J2L	Street mosaic, the presence there of a richly	coloured	figured panel (enclosed by a chain-guilloche

Page 1 of 825 Go Next | Last

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 June 2015 more than **400 corpora** for **82 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
- 60+ languages with a PoS-tagged corpus
- 42 languages with word sketches
- 26 languages with integrated tagger for tagging user corpora
- parallel corpora: EUROPARL, DGT, OPUS, . . .

Users

- Lexicographers
- Researchers
- Teachers
- Language Learners
- Translators
- Terminologists
- Copywriters

Sketch Engine – where we go

- Sketch Engine after Adam Kilgarriff
- more questions than answers, of course

Research Agenda in a Nutshell

- Building Very Large Text Corpora from the Web
- Parallel and Distributed Processing of Very Large Corpora
- Corpus Heterogeneity and Homogeneity
- Corpus Evaluation
- Corpora and Language Teaching
- Language Change over Time
- Corpus Data Visualization
- Terminology Extraction

Building Very Large Text Corpora from the Web

- well-studied domain
- but many ongoing challenges including:
  - text type identification (genres on the web)
  - spam fighting
  - text normalization and cleaning
  - dealing with low-resourced languages
  - diachronic analysis (timestamping)

Parallel and Distributed Processing of Very Large Corpora

- targeting corpus size of ca 100 billion words
- trivial parallelization often not possible
- compile-time:
    - corpus virtualization
- run-time
    - asynchronous processing all over web pages
    - reimplementation of the database backend (Manatee) in Go language
        - native support for concurrency

Corpus Heterogeneity and Homogeneity

- what is in the corpus?
- how is corpus X similar to corpus Y? (link)
- assumes we know how much X and Y are homogenous
- text type induction, clustering, . . .

Corpus Evaluation

- is corpus X better than corpus Y?
  - assumes: better for a purpose
- 2012: collocation dictionary task (En, Cz)
  - word sketch evaluation
  - sketch grammar vs. parser comparison
  - ...
- next run of the task to come soon

Corpora and Language Teaching

- biggest problem with Sketch Engine: *too many buttons*
- SkELL – `http://skell.sketchengine.co.uk`
  - English only
  - Russian coming very soon
  - more to come on demand

Language Change over Time

- neologisms finding
- so far: new lexemes (link)
- now: new/changed senses based on word sketches
- data is the problem, not the algorithms

Corpus Data Visualization

- work by Lucia Kocincova
- to be integrated into Sketch Engine and continued
- preview

Terminology extraction

Automatic terminology extraction
→ given a domain corpus, find all terms in it

Terms and Terminology
→ term as a concept is plausible only within a fixed domain

Terminology extraction

Why use corpora for terminology extraction?

- to work faster
    - → allow people to focus on intellectually demanding tasks, leave the easy bits to computer
- to work better
    - → data-driven evidence instead of linguistic introspection

Terminology is a fast moving target.

Terminology extraction

What is a “term”?

- unithood
  - which words form a grammatically well-defined unit?
  - → simplifying assumption: terms are noun phrases
- termhood
  - does it belong to the domain?
  - → keyword formula: ratio of relative frequencies in contrast to a general language corpus

Unithood

Recognizing noun phrases in corpora

- exploiting the Sketch Grammar formalism: CQL queries matching noun phrases

Term grammar example: English

```
=terms
*COLLOC "%(2.lc)_%(1.lc)"

  2:[tag=="NN" | tag=="JJ" | tag=="VVG"]  1:[tag=="NN"]

*COLLOC "%(3.lc)_%(2.lc)_%(1.lc)"

  3:[tag=="NN" | tag=="JJ" | tag=="VVG"]
        2:[tag=="NN" | tag=="JJ" | tag=="VVG"]
        1:[tag=="NN"]
```

Term grammar example: German

```
=terms

define(‘adj’,‘[kind="ADJA"]’)
define(‘subs’,‘[kind="N"]’)

...

# kleines Haus
*COLLOC "%(2.adj_stem)%(1.gender_ending)_%(1.lemma_cap)-x"
2:adj 1:subs & 1.case = 2.case
```

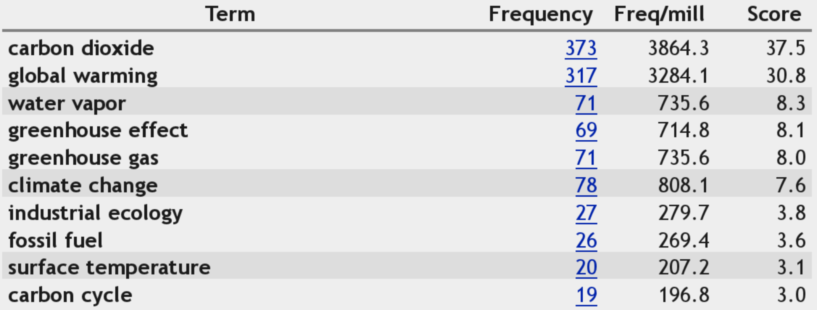
Termhood

- so called “simplemath” formula

$$\frac{f_f + N}{f_r + N}$$

- used for general keyword extraction
- varying N influences whether rare of frequent words are preferred

Output example



Term	Frequency	Freq/mill	Score
carbon dioxide	373	3864.3	37.5
global warming	317	3284.1	30.8
water vapor	71	735.6	8.3
greenhouse effect	69	714.8	8.1
greenhouse gas	71	735.6	8.0
climate change	78	808.1	7.6
industrial ecology	27	279.7	3.8
fossil fuel	26	269.4	3.6
surface temperature	20	207.2	3.1
carbon cycle	19	196.8	3.0

Languages covered (13)

- Chinese
- Czech
- Dutch
- English
- French
- German
- Italian
- Japanese
- Korean
- Polish
- Portuguese
- Russian
- Spanish

- Background: WIPO

Demo: Iterative Building of Domain Corpora

1. bootstrap a corpus via WebBootCat – use seed words
2. extract terms, reuse them as seed words in step 1

Challenges

- compatible focus and reference corpora
- lemmatization – word lemmas vs. term lemmas (French, Czech, German, ...)
- example: “klein Haus” vs. “kleines Haus”
- → new technical attributes (e.g. gender_lemma)
- coverage vs. term grammar accuracy
- evaluation

Bilingual Terminology Extraction

- given to parallel corpora
- find terms in both and align according to their translation
- now experimental in Sketch Engine, staged as technological preview
- example

Terminology Checking

- given a translation and termbase
- can we check whether the translation uses the terms in a consistent manner?
- lots of linguistic processing needed (morphology)
- work in progress

Integration with existing tools

- API available
- plugin development
- others, e. g. Intelliwebsearch
- *What?* – not *How?* is the question here

Conclusions

- very many ongoing developments
- × but for now mainly: keep things going
- bringing corpora to masses
  - translators and terminologists
  - teachers and learners
  - more languages, more corpora, more tools