

Working with metadata in Sketch Engine

Miloš Jakubíček



`milos.jakubicek@sketchengine.co.uk`

2nd BAAL Corpus Linguistics SIG event 2016
Coventry, December 10, 2016

Where we are & Where we go

- 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, **parallel corpora**, virtual sub- and supercorpora, GDEX
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **corpus architect**: building user corpora
 - from texts uploaded by users
 - from web by specifying keywords (WebBootCaT)
 - from TMX (→ parallel corpora)

Concordance search

Concordance
Word List
Word Sketch
Thesaurus
Find X
Sketch-Diff
Sketch-Eval
Corpus Info



Save
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample
Filter

Query **colour** 16,486 (147.0 per million)



Page of 825 [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)	

sketo

Parallel concordancing

Query **mother, Mutter** 1,063 (17.50 per million)

EUROPARL7, en

1830

I have to say, like motherhood, although we all support it, sometimes we do very little substantially and concretely to help the **mothers** or the people engaged in energy saving.

1859

Thus, when I look at this proposed instrument I do not see a measure to bring two peoples into the fold of **Mother** Europe.

579

On 14 December 1999, a five year old boy called Clément was found in a sleeping car on the Calais Ventimiglia train, crying for his **mother** who was taking him to visit his sick grandmother.

EUROPARL7, de

1817

Ich muß sagen, daß, ähnlich wie bei der Mutterschaft, wir dies zwar alle unterstützen, aber manchmal sehr wenig tun, um den **Müttern** oder denjenigen, die sich für die Energieeinsparung einsetzen, wirklich konkret zu helfen.

1879

Wenn ich nun das vorgeschlagene Instrument betrachte, sehe ich kein Mittel, mit dem man zwei Völker in die Arme von **Mutter** Europa holen könnte.

591

Am 14. Dezember letzten Jahres sitzt der kleine fünfjährige Clément weinend in einem der Schlafwagen des Zuges Calais Ventimiglia und verlangt nach seiner **Mutter**, die ihn zu einem Besuch bei der kranken Großmutter begleitet.

- **concordancing**, sorting, sampling, wordlists, collocation lists, **parallel corpora**, virtual sub- and supercorpora, GDEX
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **corpus architect**: building user corpora
 - from texts uploaded by users
 - from web by specifying keywords (WebBootCaT)
 - from TMX (→ parallel corpora)

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

modifier	6477	1.5	object of	3285	2.2	modifies	1906	0.5	subject of	512	0.6
scarce	163	9.53	allocate	194	9.58	allocation	135	9.42	devote	28	7.69
natural	321	8.94	pool	39	8.43	implication	46	7.09	consume	4	5.36
limited	187	8.86	exploit	64	8.23	management	153	6.98	tie	6	4.87
financial	249	8.3	divert	38	7.86	defense	7	6.68	last	4	4.6
mineral	89	8.19	deploy	31	7.67	Stonier	6	6.65	back	5	4.5
additional	107	7.92	devote	44	7.64	utilisation	7	6.63	stretch	4	4.29
valuable	74	7.86	concentrate	62	7.35	committee	132	6.49	result	6	3.93
extra	88	7.53	utilise	22	7.28	centre	158	6.4	depend	6	3.84
human	134	7.38	conserve	17	7.09	allocator	5	6.4	limit	5	3.59
renewable	33	7.31	lack	37	7.0	depletion	6	6.21	match	3	3.58
adequate	49	7.28	reallocate	13	6.98	pack	17	6.2	share	6	3.55
non-renewable	25	6.97	mobilise	13	6.83	investigator	8	6.17	earn	3	3.55
existing	53	6.68	mobilize	13	6.79	column	20	6.16	enable	7	3.54
finite	22	6.66	distribute	29	6.73	constraint	14	6.14	remain	12	3.5

Sketch-diff

perceptive	0	34	0.0	6.4	emotionally	0	111	0.0	8.6	being	0	208	0.0	6.1
thought-provoking	0	32	0.0	6.2	artificially	0	52	0.0	7.9	robot	0	77	0.0	6.1
adaptive	0	39	0.0	6.1	fiercely	0	26	0.0	7.0	agent	9	455	0.4	6.0
well-informed	0	24	0.0	6.0	moderately	0	11	0.0	5.7	guess	0	35	0.0	5.5
literate	0	26	0.0	5.9	reasonably	0	54	0.0	5.7	conversation	0	88	0.0	5.1
cultured	0	19	0.0	5.7	culturally	0	12	0.0	5.5	creature	11	137	2.4	5.9
rational	0	4	clever	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	intelligent	81	80	5.8 5.7
sensitive	8	134	2.0	5.9	wonderfully	20	9	5.4	4.5	fellow	52	14	5.1	3.1
thoughtful	14	121	5.0	7.7	very	1707	596	5.6	4.0	pass	67	9	5.2	2.2
affectionate	6	31	4.5	6.2	too	476	76	5.4	2.8	wordplay	21	0	5.8	0.0
clever	54	30	5.8	4.8	damn	12	0	5.6	0.0	chap	47	0	5.9	0.0
funny	233	103	7.0	5.7	awfully	15	0	6.1	0.0	twist	94	0	6.5	0.0
catchy	19	0	5.8	0.0	terribly	25	0	6.2	0.0	trick	166	0	6.7	0.0

Parallel word sketch

house

(noun)

ukWaC freq = [391,778](#) (251.18 per million)

Haus

(noun)

deTenTen [2013] freq = [7,264,685](#) (364.72 per million)

Use another candidate translation: [erinnern](#) [Hausarrest](#) [Ordnung](#) [Plenum](#) [hoch](#) [daran](#) [Parlament](#) [kehren](#) [mitteilen](#)

Click on collocates to access reciprocal bilingual search

object_of	96,897	2.10	VerbY+SubstXDat	1,476,115	3.60
			(obj_dat_of)		
terrace	1,667	9.04	wohnen	31,400	7.14
detach	1,737	8.94	fühlen	26,177	6.04
build	8,502	8.59	fahren	46,582	5.99
buy	3,998	8.10	kehren	7,991	5.64
board	853	7.94	finden	30,253	5.35
rent	935	7.90	schicken	9,019	5.34
sell	2,452	7.58	sitzen	15,913	5.25
demolish	650	7.54	leben	23,884	5.21
situate	1,061	7.39	kommen	144,447	5.04
own	1,284	7.25	holen	10,220	4.99
occupy	798	7.12	eilen	1,891	4.84
move	2,456	7.00	rennen	2,939	4.82

subject_of	58,690	1.90	SubstXNom+VerbY	322,711	0.80
			(subj_of)		
overlook	244	6.20	beherbergen	1,910	5.78
stand	734	6.12	verfügen	16,744	5.49
belong	306	6.09	brennen	3,018	5.48
rebuild	135	5.61	abbrennen	364	4.83
date	266	5.52	schmiegen	384	4.60
front	86	5.39	erstrahlen	562	4.56
burn	113	4.98	finden	15,624	4.47
line	84	4.93	bestechen	933	4.33
occupy	133	4.87	säumen	294	4.17
collapse	63	4.71	gruppieren	255	4.14
boast	75	4.68	liegen	32,608	3.99
survive	107	4.56	einstürzen	187	3.97

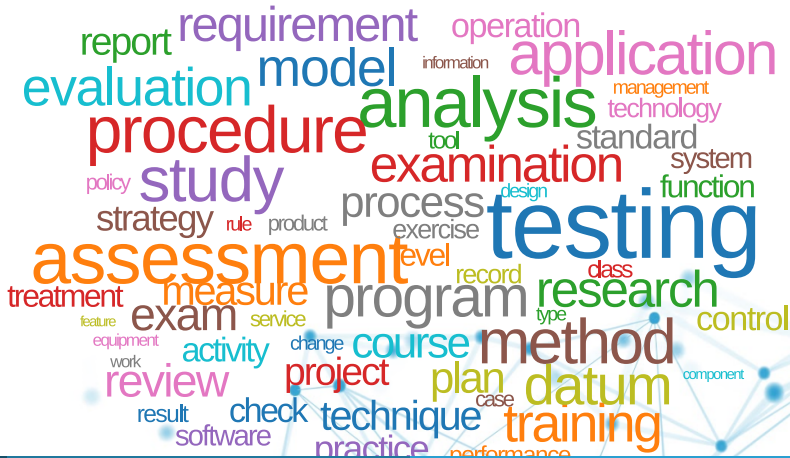
adj_subject_of	7,821	2.10	modifier	174,914	1.30	modifier	6,611,554	1.60	modifies	54,780	0.40
uninhabited	65	7.80	man	2,999	8.57	weg	159,999	7.18	price	5,868	8.28

Sketch Engine features

- **concordancing**, sorting, sampling, wordlists, collocation lists, **parallel corpora**, virtual sub- and supercorpora, GDEX
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **corpus architect**: building user corpora
 - from texts uploaded by users
 - from web by specifying keywords (WebBootCaT)
 - from TMX (→ parallel corpora)

test (*noun*) Alternative PoS: verb (freq: 941,372)
enTenTen [2012] freq = **1,915,482** (147.70 per million)

Lemma	Score	Freq
<u>testing</u>	0.520	558,727
<u>assessment</u>	0.410	640,347
<u>analysis</u>	0.399	1,196,660
<u>procedure</u>	0.382	1,311,372
<u>study</u>	0.380	3,090,402
<u>method</u>	0.373	2,760,051
<u>application</u>	0.366	3,171,582
<u>program</u>	0.365	6,442,955
<u>datum</u>	0.362	3,165,540
<u>evaluation</u>	0.360	468,130
<u>model</u>	0.357	2,557,538
<u>training</u>	0.354	2,486,409
<u>research</u>	0.354	3,171,715
<u>examination</u>	0.352	375,991
<u>requirement</u>	0.349	1,734,482





- **concordancing**, sorting, sampling, wordlists, collocation lists, **parallel corpora**, virtual sub- and supercorpora, GDEX
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **corpus architect**: building user corpora
 - from texts uploaded by users
 - from web by specifying keywords (WebBootCaT)
 - from TMX (→ parallel corpora)

























Terminology extraction

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

- **concordancing**, sorting, sampling, wordlists, collocation lists, **parallel corpora**, virtual sub- and supercorpora, GDEX
- **word sketches**: one-page summaries of a word's grammatical and collocational behaviour
- distributional **thesaurus**
- **keywords** extraction, **terms** extraction
- **corpus architect**: building user corpora
 - from texts uploaded by users
 - from web by specifying keywords (WebBootCaT)
 - from TMX (→ parallel corpora)

Corpus building

- + Add new file
- + Add web data (BootCaT)
- ⌚ Compile corpus
- 🔍 Open in SkE
- 📁 Extract keywords & terms
- ✏️ Configure corpus
- ✏️ Change sketch grammar
- ✏️ Set subcorpus definitions
- 📶 Download corpus
- 👛 Access privileges
- 🔍 View logs

#	Original file	Tokens	
1	Oceans-Acid-Carb...ide-4710223.html	471	  
2	Ocean_acidification.html	2,896	  
3	file_5.html	366	  
4	print_project_1082_28.html	1,137	  
5	ocean-acidification.html	1,329	  
6	clean-water-act-...ed-to-fight.html	794	  
7	more-one-way-lim...n-water-act.html	736	  
8	file_4.html	661	  

By 2016 more than **400 corpora** for **85 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, EUR-LEX...

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

Sketch Engine – where we go

- Sketch Engine after Adam Kilgarriff

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

corpus = data + metadata

- metadata = “headers” = “text types” = structures & structure attributes = ...
- structural markup: documents, paragraphs, sentences, ...
- linguistic annotation (PoS tagging, lemmatization)

“General language written corpus”

- synchronic, ballanced (?), manually built
- structural markup
- text types, genres, authors, ...

“Web corpus”

- genres??? (Sharoff, 2014, 2015, 2016)
- Sharoff & Suchomel, to appear

“Diachronic corpus”

- date and time annotation
- usually at the level of documents
- trends analysis
- neologisms

“Learner corpus”

- information about learners
- difficulty levels etc.
- errors & corrections

“Spoken corpus”

- utterance structuring
- linking to audio files

...and many others:

- historical corpora
- domain-specific corpora
- subtitle corpora
- legal corpora
- parliamentary corpora
- ...

- a more technical account
- input = vertical text (non well-formed XML)
- corpus = attributes + structures + structure attributes
- attributes
 - properties of corpus positions
 - word form, PoS tag, lemma, ...
- structures
 - arbitrary ranges in the corpus
 - documents, paragraphs, sentences, ...
- structure attributes
 - properties of structures
 - name, year, author, ...of a document or ...

```
<doc year="1600" author="William Shakespeare" name="Hamlet">
```

```
<s>
```

To	PP	to
----	----	----

be	VB	be
----	----	----

or	CJ	or
----	----	----

not	PA	not
-----	----	-----

to	PP	to
----	----	----

be	VB	be
----	----	----

```
<g/>
```

,	PU	,
---	----	---

that	PR	that
------	----	------

is	VB	be
----	----	----

the	DT	the
-----	----	-----

question	NN	question
----------	----	----------

```
</s>
```

```
</doc>
```

sketchengine.co.uk



What can you do with metadata in SkE

- search: CQL, text types
- analyse: frequency distribution
- display: concordance, word list
- define: subcorpora, word sketch highlights
- create: user corpus metadata

- `<doc>` vs. `</doc>` vs `<doc/>`
- `<doc year="2004">`
- `<doc id<="AA72">`
- `<doc/>` containing ...
- `<s/>` within `<doc/>` containing ...

- multivalued attributes

```
<doc author="John Smith,David Black,William White">
```

- hierarchical attributes

```
<doc type="Humanities,Humanities:Linguistics,Humanities:Linguistics:Syntax">
```

- nested structures

- learner error coding
- phrase syntax coding

Conclusions

- metadata as important as data
- big variability \Rightarrow many different concepts and options
- Sketch Engine covers lots of them
- hopefully in a more user-friendly way in the future ...

⚡ Word Sketch

⚡ Concordance

⚡ Thesaurus

⚡ Terms

⚡ List

⚡ n-grams

⚡ Trends

Word Sketch



>< Compare

[How to ?](#)

Options

Advanced

TickBox Lexicography

Part of speech:

auto ▾



Subcorpus:

none (whole corpus) ▾

[Create new subcorpus](#)

Items per column:

25 ▾ ▲



⚡ Word Sketch

⚡ Concordance

⚡ Thesaurus

⚡ Terms

⚡ List

⚡ n-grams

⚡ Trends

work

has **5,857,985** hits and frequency of **285.58**

verb noun objective

☒ categorized ☒ grouped ☒ examples ☒ hide numbers☒ Objects ☐ Modifiers ☒ Subjects

Modifiers	★	✕
▼ ▲		
perfect	almost perfect work	
bad	such a bad work	
unfinished	unfinished work	
common	very common work	

LOAD MORE

Modifiers	★	✕
▼ ▲ ▼ ▲		
perfect	31,458	6.15
bad	9,951	3.7
unfinished	3,458	2.48
common	1,458	1.1

LOAD MORE

Modifiers	★	✕
▼ ▲ ▼ ▲		
perfect	31,458	6.15
bad	9,951	3.7
unfinished	3,458	2.48
common	1,458	1.1

LOAD MORE

