

# DIACRAN: a framework for diachronic analysis

Adam Kilgarriff, Ondřej Herman, Jan Bušta,  
Vojtěch Kovář, Vít Baisa, Miloš Jakubíček



August 13, 2015  
eLex 2015, Herstmonceux, UK

# Outline

- 1 Sketch Engine
- 2 DIACRAN
- 3 Conclusions

# 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

**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**

**Overlaps**

**1st hit in doc**

**Frequency**

**Node tags**

**Node forms**

**Doc IDs**

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

Page  of 825  [Next](#) | [Last](#)

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

Page  of 825  [Next](#) | [Last](#)

Lexical  Computing

# Word sketch

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

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



# Diachronic analysis

Main goal: neologism finding → lexicography

Neologisms:

- new lexemes
- new senses

# Diachronic analysis

Main goal: neologism finding → lexicography

Neologisms:

- new lexemes – easy bits
- new senses – hard bits

# Diachronic analysis

## Needed:

- data
- algorithms

## Output:

- trend
- significance

# Neologisms: data

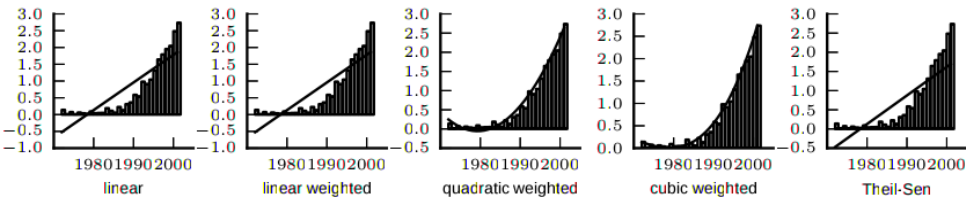
Corpora with accurate time annotation are a scarce resource

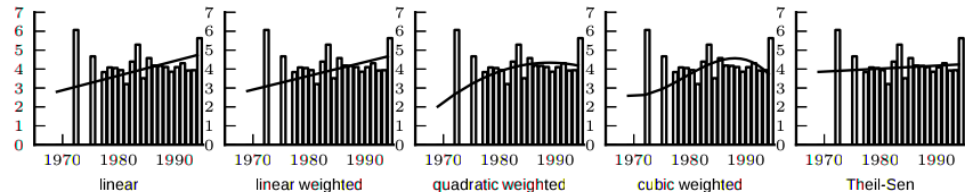
- COCA (© Mark Davies)
- BNC (but . . . )
- in-house data
- FeedsCorpus (2008–2014)
  - RSS feeds
  - so far English only, others to follow

# Neologisms: algorithms

- linear regression (and its variations)
- Mann-Kendall / Theil-Sen

Data much more important than algorithms.

Google ngrams: *globalization*













British National Corpus: *tree*

new configuration directive: DIACHRONIC "doc.year,doc.month"

### Top 1,000 trends sorted by trend

Corpus: **Corpus of Contemporary American English (COCA)** Method: **Mann-Kendall, Theil-Sen (all)**

Show  items

Word	Trend		p-value	Freq	Graph
<a href="#">blog</a>	3.4874	+	0.000000	<a href="#">1,976</a>	
<a href="#">blogged</a>	3.4874	+	0.000002	<a href="#">53</a>	
<a href="#">SODIUM</a>	3.2708	+	0.000002	<a href="#">993</a>	
<a href="#">website</a>	3.2708	+	0.000000	<a href="#">4,460</a>	
<a href="#">dcor</a>	3.2708	+	0.000002	<a href="#">180</a>	
<a href="#">bloggers</a>	3.2708	+	0.000002	<a href="#">402</a>	
<a href="#">jalapeo</a>	3.2708	+	0.000001	<a href="#">275</a>	
<a href="#">earbuds</a>	3.2708	+	0.000001	<a href="#">94</a>	
<a href="#">Googling</a>	3.2708	+	0.000001	<a href="#">54</a>	
<a href="#">blogger</a>	3.2708	+	0.000000	<a href="#">366</a>	
<a href="#">Googled</a>	3.2708	+	0.000004	<a href="#">79</a>	
<a href="#">Crme</a>	3.2708	+	0.000001	<a href="#">132</a>	
<a href="#">modele</a>	-3.0776	-	0.000007	<a href="#">18</a>	



# Evaluation

- work in progress
- neologism data obtained from major UK publishing houses

# Current work

- neologisms: new lexical items vs. new senses
- so far: new lexical items
- to be continued with: new senses
- new senses = new contexts  $\Rightarrow$  word sketches as input to regression

# Conclusions

- diachronic analysis to become part of Sketch Engine
- data more important than algorithms
- ongoing work on new sense detection