

# Sketch Engine development: Done and To-do

---

# Done (in last 18 months):

---

## ☐ Corpus Architect

- Replaces the home page, CorpusBuilder, WebBootCat, Account mgt system
  - shortcut for WebBootCaT: still to do
  - Cleaner
  - Easier to maintain, modify
  - Jan (Honza) Pomikalek
-

# New servers

---

- New ISP
  - Security certificates
  - Improved redundancy
  - Faster
    - we hope – an ongoing challenge as corpora get bigger etc
  - Milos Jakubicek
-

# Breaking the 2b ceiling

---

- Re-engineering to use 64-bit integers
  - Pavel Rychly
-

# Lists

---

- ☐ FindX
  - ☐ "Simple Maths for Keywords"
  - ☐ Full ARF lists
  
  - ☐ Vojtěch Kovář
-

Corpus: **preloaded/bawe2**  
Subcorpus: **ArtsHum**

Reference corpus: **preloaded/bawe2**  
Reference subcorpus: **SocSci**

lemma	<i>preloaded/bawe2:ArtsHum</i>			<i>preloaded/bawe2:SocSci</i>			Score
	Freq	ARF	ARF/mill	Freq	ARF	ARF/mill	
poem	<a href="#">894</a>	70.5	114.0	0	0.0	0.0	12.4
god	<a href="#">1623</a>	200.1	323.6	<a href="#">109</a>	26.3	32.9	7.8
historian	<a href="#">522</a>	73.5	118.8	<a href="#">17</a>	6.3	7.9	7.2
poet	<a href="#">272</a>	43.8	70.8	<a href="#">2</a>	1.6	2.0	6.7
verb	<a href="#">407</a>	44.9	72.5	<a href="#">4</a>	2.0	2.5	6.6
speaker	<a href="#">641</a>	70.2	113.5	<a href="#">26</a>	7.3	9.2	6.4
greek	<a href="#">896</a>	93.7	151.4	<a href="#">35</a>	12.3	15.3	6.4
narrator	<a href="#">277</a>	32.5	52.5	0	0.0	0.0	6.2
noun	<a href="#">245</a>	35.6	57.6	<a href="#">4</a>	1.0	1.3	6.0
tense	<a href="#">355</a>	42.4	68.5	<a href="#">4</a>	2.5	3.1	6.0
poetry	<a href="#">364</a>	48.5	78.4	<a href="#">7</a>	3.9	4.9	5.9
adjective	<a href="#">220</a>	30.1	48.7	0	0.0	0.0	5.9
roman	<a href="#">708</a>	68.1	110.1	<a href="#">28</a>	8.7	10.8	5.8
native	<a href="#">536</a>	75.8	122.5	<a href="#">33</a>	10.9	13.6	5.6

# Word sketches

---

- ☐ Gramrel ordering
  - ☐ Two words in a wordsketch table
  - ☐ Trinary relations on a separate page
  - ☐ Clustering
- 
- ☐ Pavel Rychly
-

experience [117](#) 2.15  
Scripture [114](#) 6.29

along [7](#) 0.04

always [51](#) 1.36  
generally [50](#) 3.11

<a href="#">359</a> 1.9	<a href="#">PP_X</a> <a href="#">8572</a>	<a href="#">PP_PP</a> <a href="#">3562</a> 2.3	<a href="#">and_or</a> <a href="#">6481</a> 1.1
<a href="#">176</a> 4.16	<a href="#">PP_as-i</a> <a href="#">3695</a> 22.2	<a href="#">as_of</a> <a href="#">953</a> 6.09	<a href="#">analyse</a> <a href="#">515</a> 7.91
<a href="#">129</a> 0.45	<a href="#">PP_in-i</a> <a href="#">2020</a> 2.2	<a href="#">in_of</a> <a href="#">585</a> 3.37	<a href="#">understand</a> <a href="#">378</a> 5.07
<a href="#">85</a> 2.48	<a href="#">PP_by-i</a> <a href="#">1543</a> 5.8	<a href="#">by_as</a> <a href="#">257</a> 7.45	<a href="#">apply</a> <a href="#">255</a> 4.74
<a href="#">83</a> 4.02	<a href="#">PP_with-i</a> <a href="#">330</a> 1.0	<a href="#">as_for</a> <a href="#">164</a> 5.29	<a href="#">read</a> <a href="#">241</a> 4.11
<a href="#">59</a> 4.17	<a href="#">PP_for-i</a> <a href="#">180</a> 0.3	<a href="#">in_with</a> <a href="#">152</a> 3.69	<a href="#">use</a> <a href="#">215</a> 1.13
<a href="#">57</a> 2.27	<a href="#">PP_from-i</a> <a href="#">123</a> 0.5	<a href="#">as_to</a> <a href="#">123</a> 5.88	<a href="#">translate</a> <a href="#">183</a> 6.94
<a href="#">53</a> 4.21	<a href="#">PP_at-i</a> <a href="#">99</a> 0.4	<a href="#">by_of</a> <a href="#">112</a> 2.79	<a href="#">present</a> <a href="#">158</a> 3.96
<a href="#">52</a> 7.31	<a href="#">PP_through-i</a> <a href="#">90</a> 2.3	<a href="#">as_in</a> <a href="#">105</a> 4.93	<a href="#">evaluate</a> <a href="#">141</a> 5.79
<a href="#">51</a> 3.09	<a href="#">PP_on-i</a> <a href="#">82</a> 0.2	<a href="#">by_in</a> <a href="#">76</a> 3.7	<a href="#">explain</a> <a href="#">103</a> 3.47
<a href="#">44</a> 1.7	<a href="#">PP_within-i</a> <a href="#">75</a> 2.5	<a href="#">as_on</a> <a href="#">60</a> 5.98	<a href="#">record</a> <a href="#">90</a> 3.87
<a href="#">42</a> 1.49	<a href="#">PP_into-i</a> <a href="#">69</a> 0.9	<a href="#">in_as</a> <a href="#">56</a> 3.9	<a href="#">describe</a> <a href="#">87</a> 3.0
<a href="#">40</a> 1.53	<a href="#">PP_to-i</a> <a href="#">48</a> 0.1	<a href="#">as_by</a> <a href="#">46</a> 6.27	<a href="#">implement</a> <a href="#">82</a> 4.12
<a href="#">39</a> 1.64	<a href="#">PP_of-i</a> <a href="#">36</a> 0.0	<a href="#">in_by</a> <a href="#">46</a> 3.81	<a href="#">collect</a> <a href="#">76</a> 3.95
<a href="#">39</a> 1.08	<a href="#">PP_without-i</a> <a href="#">36</a> 2.7	<a href="#">in_to</a> <a href="#">43</a> 2.49	<a href="#">identify</a> <a href="#">67</a> 2.6
<a href="#">37</a> 2.32	<a href="#">PP_than-i</a> <a href="#">30</a> 0.7	<a href="#">within_of</a> <a href="#">33</a> 3.88	<a href="#">discuss</a> <a href="#">56</a> 2.72



# Header fields

---

- Available for sorting, filtering
  - Can be hierarchical
    - Science::physics
  - Vojtěch Kovář
-

# Corpus encoding

---

- Add attributes without re-encoding corpus
  - Vojtěch Kovář
-

# Read-only accounts

---

- ☐ No usernames
  - ☐ Authentication by IP-range
  - ☐ For university-wide accounts
    - For any staff/students
    - Standard for Univ Library subscriptions
    - New pricing model
  - ☐ Jan Pomikalek
-

- 
- Tickbox lexicography
  - Evaluation infrastructure
  - Vojtěch Kovář
-

# To-do

---

- Online payment
    - Monthly subscription
    - Two cups of coffee
    - Paypal
  - Words and tokens
    - Give (sub)corpus sizes in words
  - Sketch diff by subcorpus
    - Contrast same word in different subcorp
-

- 
- Multiword sketches
    - Click on the collocate
    - See sketch for the collocation
  - GDEX by language, corpus
  - COBUILD-style *picture*
  - Unified approach for long processes
  - Preprocess for lgs without spaces between words
    - Chinese, Japanese
    - Show user a list of possible tokenisations first
-

# New interface

---

- ❑ Top bar
  - ❑ Customisable
  - ❑ Integrates Sketch Engine and Corpus Architect
-

