

User interface enhancements, new corpus development and language support

Vít Baisa, Vojtěch Kovář, Jan Michelfeit, Vít Suchomel



`name.surname@sketchengine.co.uk`

6th Sketch Engine Workshop
Herstmonceux, August 10, 2015

Concordance – Token attributes as tooltips

Can be set in the view options.

...n noktadan kullanılır. `</p><p>` **Futbol** Oyunu İle İlgili İhlaleler / Cez
bir noktadan yapılır. `</p><p>` **Futbol** Topunun Oyunda ve Oyun D
üçüncülük ödülü ise ülkemizin **futbol** tarihindeki en büyük başarıla
büyük başarılarıdır. `</p><p>` **FUTBOL** `</p><p>` Uluslararası Futbol
UTBOL `</p><p>` Uluslararası **Futbol/futbol** Kurulu tarafından belir
el kural çerçevesinde oynana **futbol** , iki takım arasındaki on bir o
n popüler sporlardan biri olan **futbol** , günümüzde ise yaklaşık 25
okarak, gol atmaktır. `</p><p>` **Futbol** maçları; 45'er dakikalık iki de
k olarak 1962 yılında İngiltere **Futbol** Federasyonu tarafından sist

Concordance – Save as subcorpus by structures

Before: wordlist from KWIC

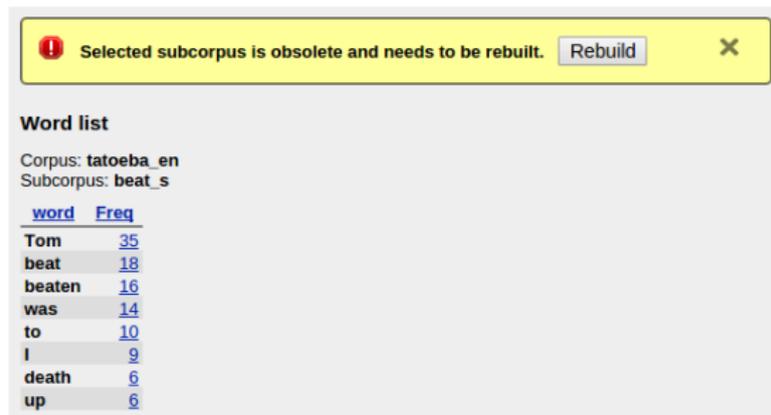
Now: the whole structure (containing KWIC) is stored

The screenshot shows a concordance tool interface. At the top, it displays "Query test: 624 (128.87 per million)". Below this, there are navigation buttons: "Page 1 of 32", "Go", "Next", and "Last". The main area shows a list of concordance lines, each starting with a file ID (e.g., "file2089999") and containing text with the word "test" highlighted in red. A dialog box is overlaid on the screen, titled "Save concordance as subcor...". The dialog has a close button (X) in the top right corner. It contains a text input field for "Subcorpus name:" with the value "test_sentences". Below this is a dropdown menu for "Save" with "s" selected, and a "structures" label. At the bottom of the dialog are two buttons: "Save" and "Cancel".

Concordance – Automated update of user subcorpora

In place: Subcorpora of user corpora defined by configuration file – updated during recompilation of the corpus.

New: Subcorpora of user corpora defined by selecting structure attributes or from concordance – updated when used for the first time after recompilation.



The screenshot shows a software interface with a yellow warning banner at the top. The banner contains a red exclamation mark icon, the text "Selected subcorpus is obsolete and needs to be rebuilt.", a "Rebuild" button, and a close button (X). Below the banner, the text "Word list" is displayed. Underneath, it shows "Corpus: tatoeba_en" and "Subcorpus: beat_s". A table follows with two columns: "word" and "Freq". The table lists the following words and their frequencies: Tom (35), beat (18), beaten (16), was (14), to (10), I (9), death (6), and up (6).

Selected subcorpus is obsolete and needs to be rebuilt.

Word list

Corpus: **tatoeba_en**
Subcorpus: **beat_s**

<u>word</u>	<u>Freq</u>
Tom	35
beat	18
beaten	16
was	14
to	10
I	9
death	6
up	6

Word Sketches – Merged sketches improved

In place: Merged sketches for multiple lemmata, e.g. 'colour, color' or 'red, green, blue'.

New: Merged sketches for all parts of speech, e.g. 'beat' as both noun and verb.

beat

tatoeba_en freq = 479 (98.92 per million)

object 177 2.60

drum	7	9.88
odd	4	9.03
shit	5	9.02
egg	5	7.48
team	5	6.74
dog	9	6.22
Mary	12	3.92
Tom	35	3.26

modifier 115 1.90

severely	4	9.63
fast	8	8.69
badly	4	8.67
finally	6	8.02
probably	4	7.06
never	5	4.33
not	46	3.65

pro_subject 85 1.00

she	13	4.19
he	18	3.59
we	10	3.57
they	4	3.20
I	20	2.23
you	15	2.03

part_around_obj 17 325.00

bush	17	12.41
------	----	-------

subject 131 2.50

heart	21	8.54
rain	5	7.22
team	5	6.79
guy	4	6.05
nothing	5	5.37
father	4	4.41
Tom	26	2.84

part_trans 44 16.50

around	17	9.16
up	20	6.13
out	6	4.32

part_intrans 19 4.60

up	16	5.81
----	----	------

pp_to 8 1.40

pulp	4	12.83
death	4	6.37

np_adj_comp 6 3.10

black	4	7.54
-------	---	------

pro_object 161 5.50

them	32	7.25
him	38	6.55
me	46	5.96
her	13	5.91
you	21	2.51
it	6	2.00

Word Sketches – Lemma coverage

‘Show lemma coverage’ in word sketch advanced options.
Useful for sketch grammar development.

tell <small>(verb)</small> British National Corpus freq = 73,213 (652.63 per million) Coverage: 85.76%											
subject	13,888	4.00	unary rels			object	19,474	3.00	modifier		
doctor	156	7.31	Sfin	16,025	6.30	story	1,309	9.80	yesterday		
instinct	74	7.14	np_np	12,482	38.40	truth	608	9.43	please		

Word Sketches – Adjustable number of columns

For your wide screens – adjust in advanced settings.



Translation & terminology – interoperability exports

TMX export for parallel user corpora

Strategic_TMX_English: Download corpus

Format plain text
 vertical
 TMX

Aligned corpus

Export to TBX and CSV in Keyword/term extraction

termeval_music_en: Extracted keywords and terms

[Change extraction options](#) [Download keywords: TBX CSV](#), [Download terms: TBX CSV](#)

Keywords				Terms					
		Score	F	RelF		Score	F	RelF	
<input type="checkbox"/>	fugue	3,039.27	376	3,818	<input type="checkbox"/>	ww ww	293.59	28	17
<input type="checkbox"/>	œce	1,005.18	96	0	<input type="checkbox"/>	scale degree	260.17	25	124
<input type="checkbox"/>	bach	961.45	337	34,593	<input type="checkbox"/>	scale length	223.96	22	427
<input type="checkbox"/>	fugues	921.64	93	734	<input type="checkbox"/>	œ œ	199.74	19	0
<input type="checkbox"/>	œ	666.49	88	4,962	<input type="checkbox"/>	ww ww ww	189.28	18	0
<input type="checkbox"/>	chord	507.70	254	54,923	<input type="checkbox"/>	musical experience	173.29	20	2,772
<input type="checkbox"/>	webern	428.93	43	661	<input type="checkbox"/>	major scale	167.29	19	2,524
<input type="checkbox"/>	notation	353.83	113	30,390	<input type="checkbox"/>	ww ww ww ww	157.90	15	0
<input type="checkbox"/>	final	323.65	32	484	<input type="checkbox"/>	basso continuo	135.49	13	154

- Not everything is precomputed
- Some computations may take minutes (hours)
- Computation = background job

- Server-client architecture
 - The job can be run on a dedicated server (sharing the data)
 - Running in the background
 - User notification by e-mail
 - Shared computation result (in case of shared corpora)
- User/Administrator view
 - Users can see all their jobs with links leading to the results
 - Administrators can handle processes of all users
 - I/O and CPU priority
 - Pause/resume or stop jobs

'My jobs' overview

Sketch Engine [Send feedback](#) corpus: Araneum Anglicum Africanum Maius [2015]

Concordance
Word List
Word Sketch
Thesaurus
Sketch-Diff
Corpus Info
My jobs
All jobs
?

Home
User guide

My background jobs

Auto-reload every 10 seconds: ([reload now](#))
Show completed jobs:
Show/hide columns: [ID](#) [Corpus](#) [Description](#) [User](#) [Started](#) [Estimation](#) [Status](#) [Progress](#)

Corpus	Description	Started	Estimation	Status
arTenTen12	(Sub)corpus statistics (Freq)	2015-03-23 16:14:26	0:13:40	Completed
enTenTen [2012]	(Sub)corpus statistics (Frekvence)	2015-03-18 17:49:42	0:00:45	Failed
London English Corpus	(Sub)corpus statistics (APF)	2015-03-23 12:55:10	0:00:00	Completed
London English Corpus	(Sub)corpus statistics (APF)	2015-03-23 12:55:36	0:00:01	Completed
czTenTen [2012]	(Sub)corpus statistics (Frekvence)	2015-03-18 17:54:23	0:08:01	Completed
Araneum Anglicum Africanum Maius [2015]	(Sub)corpus statistics (APF)	2015-08-10 13:27:30	0:00:27	Running
arTenTen12	(Sub)corpus statistics (APF)	2015-08-10 13:04:31	0:19:53	Completed
Brown Family [old tagging]	N-grams computation ('penntag' attribute)	2015-08-10 12:36:17	0:00:54	Completed
Susanne	N-grams computation ('status' attribute)	2015-08-10 12:36:14	0:00:01	Completed
Brown Family [old tagging]	N-grams computation ('tempo' attribute)	2015-08-10 12:36:13	0:01:00	Completed

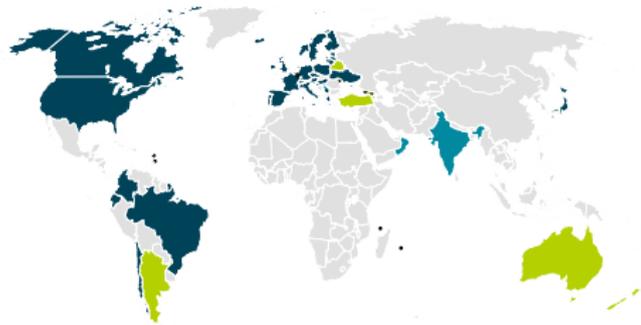
User accounts – Single Sign On coverage

In place: Anonymous sign on for institutions having a SkE site licence within UK Access Management Federation.

New: Coverage extended to all eduGAIN members.

Top five universities in SkE by number of recent SSO accesses:
Coventry, Birmingham, Antwerpen, Portsmouth, Queen Mary.

eduGAIN member countries (source: edugain.org)



User accounts – Single Sign On full accounts

Before: Anonymous 'read only' profiles (similar to IP based authentication).

New option: Full profiles (including custom settings, user built corpora and subcorpora).

Pricing: 2015 discount – organisation licence price + 50 %.

User accounts – Administration improvement

In place: Local administrators can create users and allocate user space.

New: Local administrators can activate and deactivate users.

User accounts

6 users

Login name	Full name	E-mail	Space allocated	Space used	Created	Last access	Actions
[REDACTED]	[REDACTED]	[REDACTED]	1,000,000	14 %	3 December 2014	5 August 2015	 
[REDACTED]	[REDACTED]	[REDACTED]	200,000,000	51 %	12 June 2009		 
[REDACTED]	[REDACTED]	[REDACTED]	1,000,000	10 %	22 February 2011		 
[REDACTED]	[REDACTED]	[REDACTED]	10,000,000	0 %	31 October 2009		 
[REDACTED]	[REDACTED]	[REDACTED]	100,000	0 %	11 June 2009	31 July 2015	 
[REDACTED]	[REDACTED]	[REDACTED]	100,000	0 %	15 June 2009	18 May 2015	 

[+ Add new user](#)

Deactivated users

Login name	Full name	E-mail	Space allocated	Space used	Created	Last access	Actions
[REDACTED]	[REDACTED]	[REDACTED]	100,000	0 %	9 October 2014		 
[REDACTED]	[REDACTED]	[REDACTED]	17,000,000	0 %	16 January 2013		 

All corpora – Corpus information page

Attributes, structures, sizes, charts (soon), links, errors, . . .

Corpus *czTenTen [2012]* – statistics and info

Czech web corpus crawled by SpiderLing in 2011 and Heritrix in 2010. Encoded in UTF-8, cleaned, deduplicated. Tagged by Desamb.

Counts	
Tokens	5,069,447,935
Words	4,175,089,440
Sentences	283,378,227
Paragraphs	98,625,666
Documents	9,213,821

General info	
Language	Czech
Encoding	UTF-8
Compiled	07/02/2015 15:13:52
Tagset	Description
Corpus description	Document
Sketch Grammar	Definition

Lexicon sizes	
word	18,725,879
lemma	13,976,481
tag	12,035
gender_lemma	10,716,606
lc	15,838,691
lemma_lc	12,250,726
k	15
g	7
n	4
c	9
p	4
m	10

Tags legend (tagset)	
noun	k1.*
adjective	k2.*
pronoun	k3.*
numeral	k4.*
verb	k5.*
adverb	k6.*
preposition	k7.*
conjunction	k8.*

Structures and attributes

doc	9,213,821
Top level domain	59
wordcount	18,597
Second level domain	147,536
Web domain	228,804
url	9,213,821

Corpus sanity report

Error Gramrel lexicon verification ended with errors
Warning Specified TAGSETDOC not accessible
Warning Specified INFOHREF not accessible

All corpora – Access to restricted corpora on demand

Restricted corpus: not accessible, shown in the list of corpora.
Access policy is displayed (e.g. a form to fill and email to the author).

English	CHILDES English Corpus	22,693,506	 
English	Corpus of English Dialogues 1560–1760	1,151,171	 
English	DGT_English	59,106,576	 
English	Dog	16	  

All corpora – Tabbed list of Corpora

Before: A long list of corpora from all sources.

Now: Separated tabs.

Corpora: **Recent** My own Grammar development Shared with me Featured Parallel All

Filter by language: all

Language	Name	Words	
Arabic	arTenTen [2012, Stanford tagger]	7,475,624,779	 
Arabic	arTenTen12	5,794,161,583	 
Azerbaijani	Turkic web - Azerbaijani	94,267,206	 
English	British Academic Written English Corpus (BAWE)	6,968,089	 
English	British National Corpus	96,048,950	 
English	Early English Books Online	826,296,048	 
English	English Corpus for SkELL 3.3	1,277,868,301	 
English	enTenTen [2012]	11,191,860,036	 
English	enTenTen [2013]	19,717,205,676	 
English	Environment	61,111,806	  
English	OEC 	2,073,319,589	 
English	Oxford Children's Corpus 2013 with BeebOx 	106,937,344	 
English	Oxford Children's Corpus 2014 	159,324,873	 
English	Oxford Corpus of Academic English (April 2012)	71,372,972	 
English	Strategic terms_en	197,600	  
English	UKWaC super_sensed	315,402,632	 

Manage corpus (Corpus Architect) \iff Search corpus (Bonito)

The image displays two screenshots from the Corpus Architect software interface. The left screenshot shows a sidebar menu with the following items: Concordance, Word List, Word Sketch, Thesaurus, Sketch-Diff, Sketch-Eval, Corpus Info, **Manage corpus** (highlighted with a red box), and My jobs. The main area shows a 'Simple query:' input field, a 'Query type:' dropdown menu set to 'simple', and several empty input fields for 'Lemma:', 'Phrase:', 'Word Form:', and 'Character:'. The right screenshot shows a 'Search corpus ?' header above a list of search options: Concordance, Word List, Keywords & terms, Word Sketch, Thesaurus, Sketch-Diff, and **Corpus Info** (selected with a black circle).

User corpora – Recently used corpora

User corpora were added to the recent corpora list.

Language	Corpus Name	Size	Actions
Norwegian	noTenTen [2015]	1,698,481,187	 
Portuguese	portuguese_test_1	14,565	  
Portuguese	ptTenTen [2011, Freeling v3]	3,900,501,097	 
Portuguese	ptTenTen [Freeling v3]	3,900,501,097	 

The link behind the corpus name:

- Preloaded corpus → Search corpus
- Not compiled user corpus → Manage corpus
- Compiled user corpus → Search corpus

New icons for both links on the right side:



User corpora – Editable file metadata

Assign attribute – value pairs to files.

Horse

horse

[+ Add new file](#) | [+ Add data from web using WebBootCaT](#) | [🔄 Compile corpus](#) | [🔍 Search corpus](#)

#	Original file	Plain text	Vertical	Tokens	Owner	
	 horse					
1	ABC-Official-Rac...ules-6-16-13.pdf	✓	✓	6,709	Vit Suchomel	  
2	ABC---Racing-Rules-10122011.pdf	✓	✓	3,969	Vit Suchomel	  
3	afineromance.html	✓	✓	813	Vit Suchomel	  
4	Al_breeding_onfarm.pdf	✓	✓	3,239	Vit Suchomel	  
5	American_Quarter_Horse	✓	✓	2,850	Vit Suchomel	  

Horse: andalusian.html: Edit metadata

Attribute	Value	
author	Anna Sewell	
audience	Children	
emotinions	much	
		

User corpora – Advanced configuration

Corpus templates, word sketches, subcorpora, GDEX

Before: File upload.

Now: File upload and direct editing in a text area.

My sketch grammars

Filename	Name	Templates	
grammar.bt	Grammar1	TreeTagger for English	  
test_grammar_4.bt	Grammar1	TreeTagger for English	  
mnc_grammar.bt	mnc_grammar.bt	TreeTagger for English	  
japanese-mecab-unidic2-1.1.wsdef.bt	japanese-mecab-unidic2-1.1.wsdef.bt	ChaSen	  

Edit sketch grammar

Name

Corpus template
 Tokeniser for Spanish
 TreeTagger for Bulgarian
 TreeTagger for Dutch
 TreeTagger for English

Hold down "Control", or "Command" on a Mac, to select more than one.

Content

```
=adj_re1  
1: [tag="N.*"] 2: [tag="J.*"] [tag="J.*"]?
```

Every corpus using this sketch grammar will need to be recompiled in order to reflect your changes!

Sharing corpora with all users within the same organisation.

- Manage corpus
- Show corpus files
- Compile corpus
- Configure corpus
- Set sketch grammar
- Set subcorpora
- Download corpus
- Share corpus
- View logs

Full access
allows accessing the corpus in the Sketch Engine, adding data to the corpus, changing the corpus configuration, changing sketch grammar and recompiling the corpus.

Users with full access
Entered pattern matches against user name, full name, e-mail and organisation.

Groups with full access LCL

Site licences with full access LCC
 Lexical Computing
 SiBoI/Port corpus group
 Test SSO

New language support

- Arabic Stanford tagging (arTenTen12, integrated in CA)
- Czech by Majka (czTenTen12)
- Finnish word sketch definition by Tarja Heinonen
- Hebrew tagging (heTenTen14, custom data on demand)
- Serbian by Hunpos
- Swahili tagging by TreeTagger
- Swedish tagging by Hunpos
- By customers' demand – basic support for Nepali devanagari, Scottish Gaelic, Tatar, Yoruba, . . .

Recently added:

Source	Corpus	Size [words]
Books	Early English Books Online <small>by TCP</small>	826 M
Web crawl	en, es, fr, it, ru Araneum <small>by Vlado Benko</small>	880 M each
Web search	noTenTen15	1.70 G
Web search	Nigerian WaCs: Hausa, Yoruba, Igbo	8 M all
Combined	New SkELL corpus	1.26 G

Near future:

- Common Crawl collection in processing by Common Crawl Foundation
- Web crawled corpora in Portuguese, Indonesian, Czech

- Showcasing resources is central to a language resource programme.
- 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!