New features in Corpus Architect (corpus management)

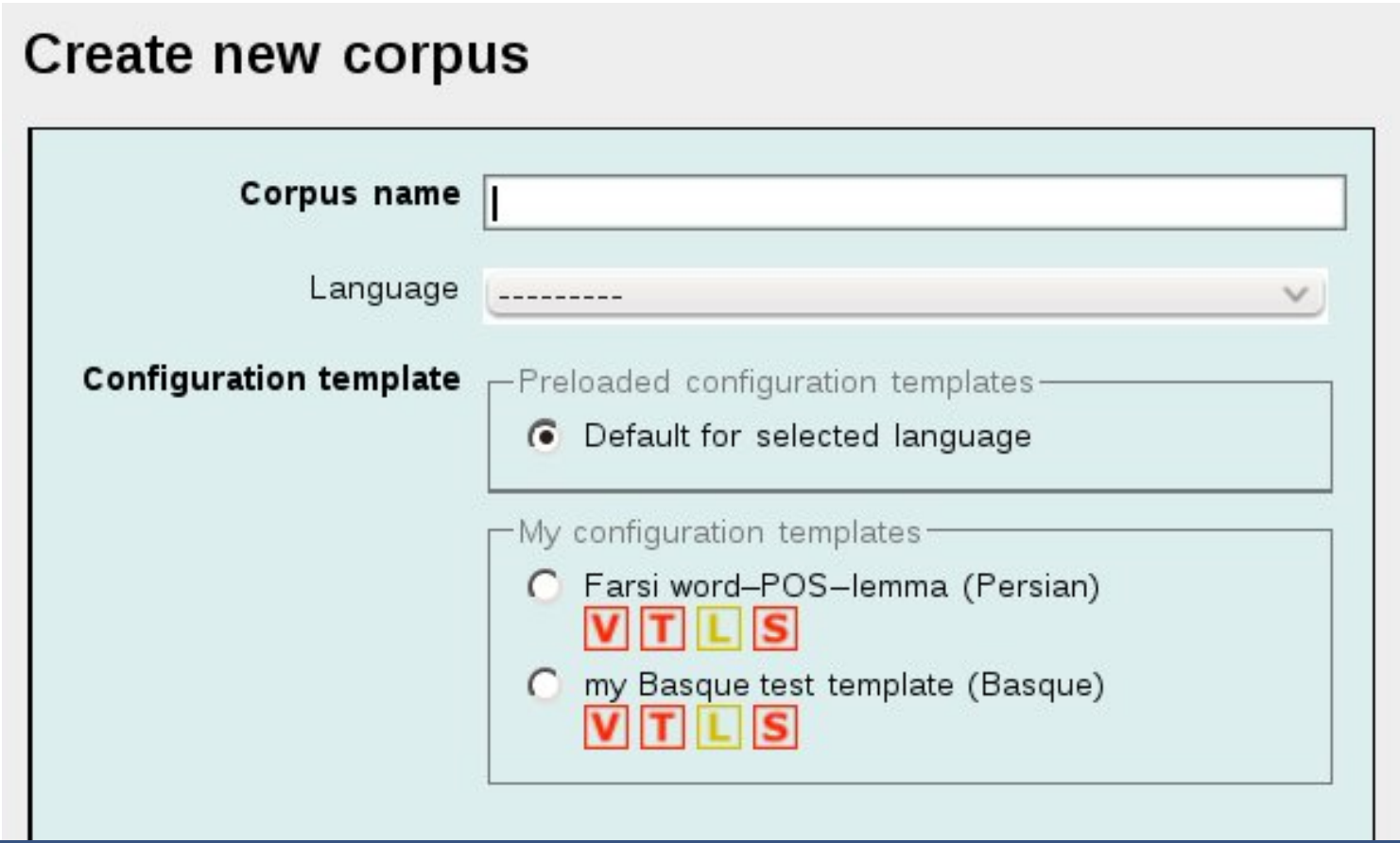
Vít Suchomel

vit.suchomel@sketchengine.co.uk

4th Sketch Engine Workshop
Tallinn, October 16, 2013

Nicer user interface

- Creation of user corpora simplified – automated steps:
  - The most complete processing chain available
  - The most up-to-date sketch grammar



Nicer user interface

- Better display of corpora and file names
  - The original file and corpus names displayed
  - No more underscores to replace non latin characters

Keywords, terms

- Term extraction for English, French, Chinese, Japanese, Korean, Spanish, Russian
  - Corpus processing chain reviewed and separated from CA
  - New/reviewed term grammars applied to reference corpora
- Keyword and term extraction results merged
  - One page, keywords + multi word terms
- API in development
  - input text, output keywords

Keywords & terms extracted from a domain specific corpus

Keywords

- dioxide (415.2, 427)
- trophic (264.9, 33)
- greenhouse (238.4, 282)
- ecology (237.7, 196)
- methane (233.5, 108)
- arrhenius (232.2, 25)
- photosynthesis (230.6, 46)
- callendar (215.4, 22)
- ecosystems (211.4, 114)
- warming (193.8, 504)
- keeling (192.5, 23)
- carbon (186.8, 558)
- n't (177.1, 17)
- gases (173.9, 159)
- -oct- (169.3, 28)
- vapor (151.3, 72)
- mutualism (75.6, 8)
- radiative (75.0, 12)
- gasses (75.0, 12)
- lca (74.4, 10)
- biotic (74.2, 10)
- acidification (74.1, 9)
- above-ground (73.6, 9)
- holism (73.5, 9)
- felzer (73.5, 7)
- carbonic (72.4, 9)
- loa (71.5, 10)
- biogeography (71.2, 9)
- organisms (70.4, 86)
- mauna (69.7, 10)
- flowering (68.4, 23)
- emitted (68.2, 27)

Terms

- carbon dioxide (567.1)
- greenhouse effect (515.0)
- water vapor (486.8)
- global warming (298.8)
- industrial ecology (261.6)
- infrared radiation (170.9)
- carbon cycle (169.0)
- surface temperature (161.0)
- elevated carbon (156.4)
- elevated carbon dioxide (156.4)
- greenhouse gas (135.8)
- climate system (134.1)
- food web (124.3)
- amount of carbon dioxide (116.8)
- other greenhouse (114.2)
- global temperature (109.1)

Configuration templates in Corpus Architect administration

Name	Default grammar	Reference corpus	Terms grammar	Terms corpus
TreeTagger for English	english-penn_tt-2.5.wsdef.txt	enTenTen08 (trial)	english-penn_tt-terms-2.3.wsdef.m4	enTenTen12 [sample 40M] with term grammar (trial)
TreeTagger for French	french-tt-1.0.wsdef.txt	frTenTen12 (trial)	french-tt-terms-1.0.wsdef.m4	frTenTen12 [sample] with term grammar (trial)
Stanford Chinese Segmenter and Tagger	chinese-universal-with-tags.1.0.wsdef.m4	zhTenTen11 (trial)	chinese-stanford-terms-1.0.wsdef.m4	zhTenTen11 [sample 10M] with term grammar (trial)
Mecab (+Unidic2 +LUW) for Japanese	japanese-mecab-unidic2-1.7.wsdef.txt	jpTenTen11 [LUW, sample] (trial)	japanese-mecab_unidic2_luw-terms-1.0.wsdef.m4	jpTenTen11 [LUW, sample] with term grammar (trial)
HanNanum for Korean	korean-universal-with-tags.1.0.wsdef.m4	koTenTen12 (trial)	korean-hannanum_simplified-terms-1.0.wsdef.m4	koTenTen12 [sample] with term grammar (trial)
TreeTagger for Russian	russian-tt-1.0.wsdef.txt	ruTenTen11 [sample 50M] (main)	russian-tt-terms-1.0.wsdef.m4	ruTenTen11 [sample 50M] with term grammar (restricted)
Freeling for Spanish	spanish-freeling-1.0.wsdef.txt	esTenTen11 [Eu + Am, Freeling, sample 50M] (trial)	spanish-freeling-terms-1.0.wsdef.m4	esTenTen11 [Eu + Am, Freeling, sample 50M] with term grammar

Single sign on

- Service provider software set up
  - Use case: University students are granted access based on authentication at their university, no additional authentization with Sketch Engine required
- Ready in UK academic identity fedaration

WebBootCaT search engine

- New word search service in MS Azure Marketplace used
- Service availability improvements

Future work

- improve usability
- finish term extraction
- finish corpus comparison
- implement corpus homogeneity

Conclusion

Nicer interface, parallel corpora support, keyword & term extraction,... $\Rightarrow$ much work done!