Sketch Engine for language learning

Vít Baisa, †‡ Vít Suchomel, †‡ Adam Kilgarriff, † Miloš Jakubíček †‡

Lexical Computing, † Masaryk University ‡ Brighton, UK & Brno, CZ

July 24, 2015 CL 2015, Lancaster

Outline

Outline

- 1 Sketch Engine
- 2 SkELL
- 3 Visualizations
- 4 Conclusions

Sketch Engine SkELL Visualizations 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 SkELL Visualizations Conclusions

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

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, . . .

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

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

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

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

- Lexicographers
- Researchers
- Teachers
- Language Learners = general public
- Translators
- Terminologists
- Copywriters

Sketch Engine SkELL Visualizations Conclusions



Vít Baisa, Vít Suchomel, Adam Kilgarriff, Miloš Jakubíček



skell.sketchengine.co.uk

SkELL: lightweighting Sketch Engine

- Sketch Engine is a rather heavy weight tool
- SkELL = Sketch Engine for Language Learning
- http://skell.sketchengine.co.uk
- free of charge
- specific corpus material
- English so far, possibly others to follow
- showcased as a module for integration into 3rd party sites

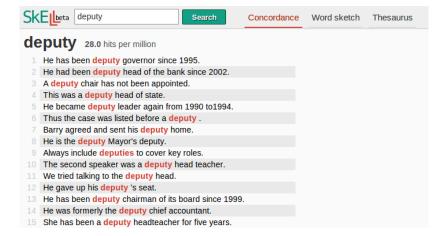
Features

- free access
- intuitive interface
- mobile version
- SkELL corpus
- only three functions
 - concordance
 - word sketch
 - thesaurus

Corpus for SkELL

- a mixture of both web based and edited content
- very extensive cleaning
- adding up to 1 billion words
- deduplicated on sentence level
- only sentences with GDEX score > 50%
- vertical sorted according to the GDEX score

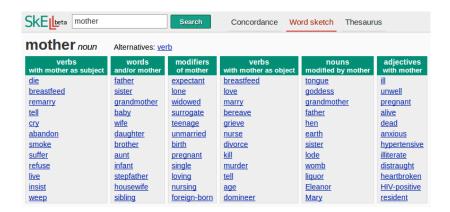
Concordance I



Concordance II

- simple query (word or lemma, case insensitive)
- 40 top sentences shown
- no references
- no structure tags
- no details, contexts

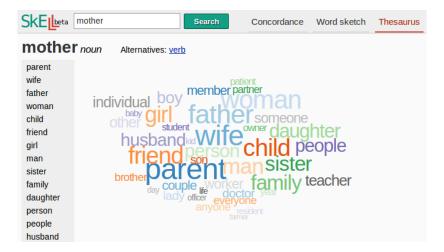
Word Sketch I



Word Sketch II

- up to 10 grammar relations
- names simplified
- up to 15 collocates per gramrel
- auto PoS (links to alternative PoSes)

Thesaurus



Visualizations of Corpus Data

- visualizations matter period
- very very many new visualizations heading to SkE (Kocincová, 2015)
- DEMO

Conclusions

"Bringing Corpora to the Masses" ???(Kilgarriff, 2003)

- http://skell.sketchengine.co.uk
- free and simple to use
- for integration with other sites