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

1. Sketch Engine
2. SkELL
3. Visualizations
4. 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
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
Users

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

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

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

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

TOO LITTLE TIME
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

**deputy** 28.0 hits per million

1. He has been *deputy* governor since 1995.
2. He had been *deputy* head of the bank since 2002.
3. A *deputy* chair has not been appointed.
4. This was a *deputy* head of state.
5. He became *deputy* leader again from 1990 to 1994.
6. Thus the case was listed before a *deputy*.
7. Barry agreed and sent his *deputy* home.
8. He is the *deputy* Mayor's deputy.
9. Always include *deputies* to cover key roles.
10. The second speaker was a *deputy* head teacher.
11. We tried talking to the *deputy* head.
12. He gave up his *deputy* 's seat.
13. He has been *deputy* chairman of its board since 1999.
14. He was formerly the *deputy* chief accountant.
15. She has been a *deputy* headteacher for five years.
Concordance II

- simple query (word or lemma, case insensitive)
- 40 top sentences shown
- no references
- no structure tags
- no details, contexts
# Word Sketch 1

**mother** *noun*

<table>
<thead>
<tr>
<th>verbs with mother as subject</th>
<th>words and/or mother</th>
<th>modifiers of mother</th>
<th>verbs with mother as object</th>
<th>nouns modified by mother</th>
<th>adjectives with mother</th>
</tr>
</thead>
<tbody>
<tr>
<td>die</td>
<td>father</td>
<td>expectant</td>
<td>breastfeed</td>
<td>tongue</td>
<td>ill</td>
</tr>
<tr>
<td>breastfeed</td>
<td>sister</td>
<td>lone</td>
<td>love</td>
<td>goddess</td>
<td>unwell</td>
</tr>
<tr>
<td>remarry</td>
<td>grandmother</td>
<td>widowed</td>
<td>marry</td>
<td>grandmother</td>
<td>pregnant</td>
</tr>
<tr>
<td>tell</td>
<td>baby</td>
<td>teenage</td>
<td>bereave</td>
<td>father</td>
<td>alive</td>
</tr>
<tr>
<td>cry</td>
<td>wife</td>
<td>unmarried</td>
<td>grieve</td>
<td>hen</td>
<td>dead</td>
</tr>
<tr>
<td>abandon</td>
<td>daughter</td>
<td>birth</td>
<td>nurse</td>
<td>earth</td>
<td>anxious</td>
</tr>
<tr>
<td>smoke</td>
<td>brother</td>
<td>pregnant</td>
<td>divorce</td>
<td>sister</td>
<td>hypertensive</td>
</tr>
<tr>
<td>suffer</td>
<td>aunt</td>
<td>kill</td>
<td>lode</td>
<td>lode</td>
<td>illiterate</td>
</tr>
<tr>
<td>refuse</td>
<td>infant</td>
<td>murder</td>
<td>womb</td>
<td>liquor</td>
<td>distraught</td>
</tr>
<tr>
<td>live</td>
<td>stepfather</td>
<td>tell</td>
<td>liquor</td>
<td>Eleanor</td>
<td>heartbroken</td>
</tr>
<tr>
<td>insist</td>
<td>housewife</td>
<td>loving</td>
<td>lode</td>
<td>Mary</td>
<td>HIV-positive</td>
</tr>
<tr>
<td>weep</td>
<td>sibling</td>
<td>single</td>
<td>age</td>
<td></td>
<td>resident</td>
</tr>
<tr>
<td></td>
<td></td>
<td>foreign-born</td>
<td>domineer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Word Sketch II

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

mother noun

Alternatives: verb

parent
wife
father
woman
child
friend
girl
man
sister
family
daughter
person
people
husband
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