Longest-commonest match for word sketches

Vít Baisa, †† Adam Kilgarriff, † Pavel Rychlý, †† Miloš Jakubíček ††

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

July 23, 2015
CL 2015, Lancaster
Outline

1. Sketch Engine
2. Word sketches & Longest-commonest match
3. Conclusions
Sketch Engine

- corpus management system
- web service (including API)
- platform for providing language resources
- widely used for
  - lexicography purposes
  - 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

Vít Baisa, Adam Kilgarriff, Pavel Rychlý, Miloš Jakubíček

CL 2015
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, ...
Users

- Lexicographers
- Researchers
- Teachers
- Language Learners
- Translators
- Terminologists
- Copywriters
Learning how words are used

- reading concordances
Reading concordances

Query **resource 13,141** (117.14 per million)

1. **bnccd#1**: room is now a well established and valuable resource for undergraduates in particular. Susan resources are as much if not more in its members.
2. **bnccd#1**: at Brown University. resource The college's resources. Somerville Lawyers <p> The College Librarian.
3. **bnccd#1**: to make maximum use of its extra resources. Somerville Lawyers <p> The College Librarian.
4. **bnccd#1**: mini pupillages. The new group provides a valuable new resource for those turning to law later in their resources which should be provided if new purpose built.
5. **bnccd#1**: financial trouble. Please help. With additional resources the Fund could do more. The Secretary of resources to new short term projects rather than.
6. **bnccd#1**: projects, and begin a survey of the space and resources which should be provided if new purpose built.
7. **bnccd#1**: because trust funds usually direct their resources to new short term projects rather than.
8. **bnccd#2**: expertise which exists already in the Corine resource, a project which ends this year, is in resources of tropical forests. <p> Thomas Eisner.
9. **bnccd#2**: Forests " Trained amateurs" to tap rainforest resources <p> American biologists have called for the resources to run major conferences and seminars. <p>
10. **bnccd#2**: populations to help tap potential medical resources, people, information) The enhancement of resources of tropical forests. <p> Thomas Eisner.
11. **bnccd#8**: The College has the ideal environment and resources to run major conferences and seminars. <p>
12. **bnccd#8**: international context (managing markets, resources, people, information) The enhancement of resources of tropical forests. <p> Thomas Eisner.
13. **bnccd#8**: exchanges and joint research. <p> COLLEGE resources THE LIBRARY <p> The Library forms part of resources of tropical forests. <p> Thomas Eisner.
14. **bnccd#8**: corporate benefit of information systems and resources of tropical forests. <p> Thomas Eisner.
15. **bnccd#8**: and implementation of organisational human resources of tropical forests. <p> Thomas Eisner.
16. **bnccd#10**: geographers are able to use the library resources of the Bodleian, other departments of the resources of tropical forests. <p> Thomas Eisner.
17. **bnccd#10**: distribution of human settlements; natural resources of tropical forests. <p> Thomas Eisner.
18. **bnccd#10**: inform students of the specialist educational resources of tropical forests. <p> Thomas Eisner.
19. **bnccd#10**: Coin Room, offers a particularly important resources for both research and teaching in Classical resources and accommodation; (c) junior academic
20. **bnccd#10**: (a) annual estimates; (b) allocation of resources for both research and teaching in Classical resources and accommodation; (c) junior academic
Learning how words are used

- reading concordances
  - slow and tedious
Learning how words are used

- reading concordances
  - slow and tedious
- reviewing collocations / word sketches
  - quick overview – “one-page summary of…”
### Word sketch

#### resource *(noun)*  
**British National Corpus** freq = **12658** (112.8 per million)

<table>
<thead>
<tr>
<th>modifier</th>
<th>6477</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>scarce</td>
<td>163</td>
<td>9.53</td>
</tr>
<tr>
<td>natural</td>
<td>321</td>
<td>8.94</td>
</tr>
<tr>
<td>limited</td>
<td>187</td>
<td>8.86</td>
</tr>
<tr>
<td>financial</td>
<td>249</td>
<td>8.3</td>
</tr>
<tr>
<td>mineral</td>
<td>89</td>
<td>8.19</td>
</tr>
<tr>
<td>additional</td>
<td>107</td>
<td>7.92</td>
</tr>
<tr>
<td>valuable</td>
<td>74</td>
<td>7.86</td>
</tr>
<tr>
<td>extra</td>
<td>88</td>
<td>7.53</td>
</tr>
<tr>
<td>human</td>
<td>134</td>
<td>7.38</td>
</tr>
<tr>
<td>renewable</td>
<td>33</td>
<td>7.31</td>
</tr>
<tr>
<td>adequate</td>
<td>49</td>
<td>7.28</td>
</tr>
<tr>
<td>non-renewable</td>
<td>25</td>
<td>6.97</td>
</tr>
<tr>
<td>existing</td>
<td>53</td>
<td>6.68</td>
</tr>
<tr>
<td>finite</td>
<td>22</td>
<td>6.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>object_of</th>
<th>3285</th>
<th>2.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>allocate</td>
<td>194</td>
<td>9.58</td>
</tr>
<tr>
<td>pool</td>
<td>39</td>
<td>8.43</td>
</tr>
<tr>
<td>exploit</td>
<td>64</td>
<td>8.23</td>
</tr>
<tr>
<td>divert</td>
<td>38</td>
<td>7.86</td>
</tr>
<tr>
<td>deploy</td>
<td>31</td>
<td>7.67</td>
</tr>
<tr>
<td>devote</td>
<td>44</td>
<td>7.64</td>
</tr>
<tr>
<td>concentrate</td>
<td>62</td>
<td>7.35</td>
</tr>
<tr>
<td>utilise</td>
<td>22</td>
<td>7.28</td>
</tr>
<tr>
<td>conserve</td>
<td>17</td>
<td>7.09</td>
</tr>
<tr>
<td>lack</td>
<td>37</td>
<td>7.0</td>
</tr>
<tr>
<td>reallocate</td>
<td>13</td>
<td>6.98</td>
</tr>
<tr>
<td>mobilise</td>
<td>13</td>
<td>6.83</td>
</tr>
<tr>
<td>mobilize</td>
<td>13</td>
<td>6.79</td>
</tr>
<tr>
<td>distribute</td>
<td>29</td>
<td>6.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>1906</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>allocation</td>
<td>135</td>
<td>9.42</td>
</tr>
<tr>
<td>implication</td>
<td>46</td>
<td>7.09</td>
</tr>
<tr>
<td>management</td>
<td>153</td>
<td>6.98</td>
</tr>
<tr>
<td>defense</td>
<td>7</td>
<td>6.68</td>
</tr>
<tr>
<td>Stonier</td>
<td>6</td>
<td>6.65</td>
</tr>
<tr>
<td>utilisation</td>
<td>7</td>
<td>6.63</td>
</tr>
<tr>
<td>committee</td>
<td>132</td>
<td>6.49</td>
</tr>
<tr>
<td>centre</td>
<td>158</td>
<td>6.4</td>
</tr>
<tr>
<td>allocator</td>
<td>5</td>
<td>6.4</td>
</tr>
<tr>
<td>depletion</td>
<td>6</td>
<td>6.21</td>
</tr>
<tr>
<td>pack</td>
<td>17</td>
<td>6.2</td>
</tr>
<tr>
<td>investigator</td>
<td>8</td>
<td>6.17</td>
</tr>
<tr>
<td>column</td>
<td>20</td>
<td>6.16</td>
</tr>
<tr>
<td>constraint</td>
<td>14</td>
<td>6.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>subject_of</th>
<th>512</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>devote</td>
<td>28</td>
<td>7.69</td>
</tr>
<tr>
<td>consume</td>
<td>4</td>
<td>5.36</td>
</tr>
<tr>
<td>tie</td>
<td>6</td>
<td>4.87</td>
</tr>
<tr>
<td>last</td>
<td>4</td>
<td>4.6</td>
</tr>
<tr>
<td>back</td>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>stretch</td>
<td>4</td>
<td>4.29</td>
</tr>
<tr>
<td>result</td>
<td>6</td>
<td>3.93</td>
</tr>
<tr>
<td>depend</td>
<td>6</td>
<td>3.84</td>
</tr>
<tr>
<td>limit</td>
<td>5</td>
<td>3.59</td>
</tr>
<tr>
<td>match</td>
<td>3</td>
<td>3.58</td>
</tr>
<tr>
<td>share</td>
<td>6</td>
<td>3.55</td>
</tr>
<tr>
<td>earn</td>
<td>3</td>
<td>3.55</td>
</tr>
<tr>
<td>enable</td>
<td>7</td>
<td>3.54</td>
</tr>
<tr>
<td>remain</td>
<td>12</td>
<td>3.5</td>
</tr>
</tbody>
</table>
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ...easy bits
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ... easy bits
- ... but not always as informative
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ... easy bits
- ... but not always as informative
  - put – objects of p. – phone (BNC)
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ...easy bits
- ...but not always as informative
  - put – objects of p. – phone (BNC)
  \[\Rightarrow \text{put the phone down}\]
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ...easy bits
- ...but not always as informative
  - put – objects of p. – phone (BNC)
  - put the phone down
  - resource – nouns modified by r. – department (SkELL)
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ... easy bits
- ... but not always as informative
  - put – objects of p. – phone (BNC)
    ⇒ put the phone down
  - resource – nouns modified by r. – department (SkELL)
    ⇒ human resources department
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ... easy bits

- ... but not always as informative
  - put – objects of p. – phone (BNC)
    ⇒ *put the phone down*
  - resource – nouns modified by r. – department (SkELL)
    ⇒ *human resources department*
  - resource – words and/or r. – conservation (SkELL)
Reading collocations – drawbacks?

- easier than reading concordances
  - resource – modifiers of r. – scarce
  - resource – verbs with r. as object – allocate
  - ... easy bits
- ... but not always as informative
  - put – objects of p. – phone (BNC)
    \[\Rightarrow \text{put the phone down}\]
  - resource – nouns modified by r. – department (SkELL)
    \[\Rightarrow \text{human resources department}\]
  - resource – words and/or r. – conservation (SkELL)
    \[\Rightarrow \text{state department of conservation and natural resources}\]
Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- $\Rightarrow$ slows down the workflow
Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- \( \Rightarrow \) slows down the workflow

- make it easier by providing most common concordance examples for the headword–relation–collocation triple
Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
  - ⇒ slows down the workflow

- make it easier by providing most common concordance examples for the headword–relation–collocation triple
  - “commonest match”
Understanding word sketches

- sometimes it is necessary to see concordance examples to understand the word sketch
- ⇒ slows down the workflow

- make it easier by providing most common concordance examples for the headword–relation–collocation triple
- “commonest match”
- ...the longest available ⇒ **longest-commonest match (LCM)**
Longest-commonest match (LCM)

- headword – relation – collocation
- resource – noun modifier – department
  $\Rightarrow$ human resources department
LCM algorithm

- get all concordance matches starting with headword and ending with collocate (or vice versa if applies)
- does any match occur in at least 1/4 of the concordance lines?
  - yes – it is an LCM candidate, try extending the phrase to the left or right and repeat with this subset of concordances
  - no – quit
- simple, fast, language independent
**LCM algorithm**

- get all concordance matches starting with headword and ending with collocate (or vice versa if applies)
- does any match occur in at least 1/4 of the concordance lines?
  - yes – it is an LCM candidate, try extending the phrase to the left or right and repeat with this subset of concordances
  - no – quit
- simple, fast, language independent
- ... but no fast enough for online computation ⇒ LCM preindexed as part of word sketches
  - triples: (headword, relation, collocation)
  - 5-tuples: + (headword position, collocation position)
  - n-tuples: + (longest commonest match words)
## LCM integration in Sketch Engine

<table>
<thead>
<tr>
<th>Word</th>
<th>Type</th>
<th>Count</th>
<th>Frequency</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>put</td>
<td>verb</td>
<td>67367</td>
<td></td>
<td></td>
</tr>
<tr>
<td>down</td>
<td>part_trans</td>
<td>2534</td>
<td>9.94</td>
<td>put down</td>
</tr>
<tr>
<td>forward</td>
<td>modifier</td>
<td>1720</td>
<td>11.56</td>
<td>put forward</td>
</tr>
<tr>
<td>up</td>
<td>part_intrans</td>
<td>1176</td>
<td>7.83</td>
<td>to put up with</td>
</tr>
<tr>
<td>just</td>
<td>modifier</td>
<td>832</td>
<td>8.77</td>
<td>just put</td>
</tr>
<tr>
<td>in</td>
<td>part_intrans</td>
<td>796</td>
<td>8.88</td>
<td>put in</td>
</tr>
<tr>
<td>pressure</td>
<td>object</td>
<td>519</td>
<td>8.20</td>
<td>put pressure on</td>
</tr>
<tr>
<td>then</td>
<td>modifier</td>
<td>438</td>
<td>8.05</td>
<td>and then put</td>
</tr>
<tr>
<td>head</td>
<td>object</td>
<td>387</td>
<td>6.92</td>
<td>put his head</td>
</tr>
<tr>
<td>thing</td>
<td>object</td>
<td>382</td>
<td>6.57</td>
<td>put things</td>
</tr>
<tr>
<td>end</td>
<td>object</td>
<td>327</td>
<td>6.80</td>
<td>to put an end to</td>
</tr>
<tr>
<td>off</td>
<td>part_intrans</td>
<td>320</td>
<td>7.76</td>
<td>be put off</td>
</tr>
<tr>
<td>place</td>
<td>pp_in-p</td>
<td>241</td>
<td>6.61</td>
<td>put in place</td>
</tr>
<tr>
<td>right</td>
<td>np_adj_comp</td>
<td>217</td>
<td>7.96</td>
<td>put it right</td>
</tr>
<tr>
<td>phone</td>
<td>object</td>
<td>213</td>
<td>7.47</td>
<td>put the phone down</td>
</tr>
<tr>
<td>hand</td>
<td>part_out-a_obj</td>
<td>167</td>
<td>5.92</td>
<td>put out a hand</td>
</tr>
<tr>
<td>lot</td>
<td>object</td>
<td>147</td>
<td>6.30</td>
<td>put a lot of</td>
</tr>
<tr>
<td>use</td>
<td>pp_to-p</td>
<td>145</td>
<td>6.33</td>
<td>put to good use</td>
</tr>
<tr>
<td>kettle</td>
<td>object</td>
<td>142</td>
<td>7.15</td>
<td>put the kettle on</td>
</tr>
</tbody>
</table>
LCM examples

- *baby* in BNC
- *baby* in SkELL
- *test* in SkELL
- *resource* in SkELL
not a classification task

some LCM is clearly better than no LCM

what should the threshold be?
  - two annotators
  - options 1/2, 1/4, 1/6
  - 1/6 the best, people prefer longer matches

primitive filtering of LCM (e.g. punctuation) easy to do but might not be desirable
understanding collocations can be made easier by providing the most common concordance examples
examples = Longest-commonest match
simple algorithm
integrated into Sketch Engine, anywhere where a word sketch is (soon)