## Longest-commonest match for word sketches

Vít Baisa, ${ }^{\dagger \ddagger}$ Adam Kilgarriff, ${ }^{\dagger}$ Pavel Rychlý, ${ }^{\dagger \ddagger}$ Miloš Jakubíček ${ }^{\dagger \ddagger}$

Lexical Computing, ${ }^{\dagger}$ Masaryk University ${ }^{\ddagger}$<br>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
■ 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 ( $\mathbf{8 0} \mathbf{G +}$ ) 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 $\mathbf{4 0 0}$ corpora for $\mathbf{8 2}$ languages:

- 100+ corpora having more than 100 million tokens

■ 30+ corpora having more than 1 billion tokens

- In 2010 a series of TenTen $\left(10^{10}\right)$ 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 bncdoc\#1 | room is now a well established and valuable | resource | for undergraduates in particular. Susan |
| :---: | :---: | :---: | :---: |
| 2 bncdoc\#1 | at Brown University. $\langle/ p\rangle\langle p\rangle$ The college's | resources | are as much if not more in its members |
| 3 bncdoc\#1 | Somerville to make maximum use of its extra | resources | . $</ p>$ Somerville Lawyers < $<$ > The College Librarian |
| 4 bncdoc\#1 | mini pupillages. The new group provides a valuable new | resource | for those turning to law later in their |
| 5 bncdoc\#1 | financial trouble. Please help. With additional | resources | the Fund could do more. The Secretary of |
| 6 bncdoc\#1 | projects, and begin a survey of the space and | resources | which should be provided if new purpose built |
| 7 bncdoc\#1 | because trust funds usually direct their | resources | to new short term projects rather than |
| 8 bncdoc\#2 | expertise which exists already in the Corine | resource | , a project which ends this year, is in |
| 9 bncdoc\#2 | Forests " Trained amateurs" to tap rainforest | resources | $<p>$ American biologists have called for the |
| 10 bncdoc\#2 | populations to help tap potential medical | resources | of tropical forests. $</ p><p>$ Thomas Eisner |
| 11 bncdoc\#8 | The College has the ideal environment and | resources | to run major conferences and seminars. </p> |
| 12 bncdoc\#8 | international context (managing markets, | resources | , people, information) The enhancement of |
| 13 bncdoc\#8 | exchanges and joint research. $</ p\rangle$ COLLEGE | RESOURCES | THE LIBRARY <p> The Library forms part of |
| 14 bncdoc\#8 | corporate benefit of information systems and | resources | '. College staff have regular contact with |
| 15 bncdoc\#8 | and implementation of organisational human | resource | policies in both the public and private |
| 16 bncdoc\#10 | geographers are able to use the library | resources | of the Bodleian, other departments of the |
| 17 bncdoc\#10 | distribution of human settlements; natural | resource | use and technological innovation; forms |
| 18 bncdoc\#10 | inform students of the specialist educational | resources | available in the different departments |
| 19 bncdoc\#10 | Coin Room, offers a particularly important | resource | for both research and teaching in Classical |
| 20 bncdoc\#10 | (a) annual estimates; (b) allocation of | 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 ${ }_{\text {(noun) }} \quad$ British National corpus freq $=\underline{12658}$ (112.8 per million)

| modifier | $\underline{6477}$ | 1.5 |
| :--- | ---: | :--- |
| scarce | $\underline{163}$ | 9.53 |
| natural | $\underline{321}$ | 8.94 |
| limited | $\underline{187}$ | 8.86 |
| financial | $\underline{\underline{249}}$ | 8.3 |
| mineral | $\underline{89}$ | 8.19 |
| additional | $\underline{107}$ | 7.92 |
| valuable | $\underline{74}$ | 7.86 |
| extra | $\underline{88}$ | 7.53 |
| human | $\underline{134}$ | 7.38 |
| renewable | $\underline{33}$ | 7.31 |
| adequate | $\underline{49}$ | 7.28 |
| non-renewable | $\underline{\underline{25}}$ | 6.97 |
| existing | $\underline{53}$ | 6.68 |
| finite | $\underline{22}$ | 6.66 |


| object of | 3285 | 2.2 | $\underline{\text { modifies }}$ | 1906 | 0.5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| allocate | 194 | 9.58 | allocation | 135 | 9.42 |
| pool | 39 | 8.43 | implication | 46 | 7.09 |
| exploit | 64 | 8.23 | management | 153 | 6.98 |
| divert | $\underline{38}$ | 7.86 | defense | 7 | 6.68 |
| deploy | 31 | 7.67 | Stonier | $\underline{6}$ | 6.65 |
| devote | 44 | 7.64 | utilisation | 7 | 6.63 |
| concentrate | 62 | 7.35 | committee | 132 | 6.49 |
| utilise | $\underline{22}$ | 7.28 | centre | 158 | 6.4 |
| conserve | 17 | 7.09 | allocator | $\underline{5}$ | 6.4 |
| lack | 37 | 7.0 | depletion | 6 | 6.21 |
| reallocate | 13 | 6.98 | pack | 17 | 6.2 |
| mobilise | 13 | 6.83 | investigator | 8 | 6.17 |
| mobilize | 13 | 6.79 | column | $\underline{20}$ | 6.16 |
| distribute | $\underline{29}$ | 6.73 | constraint | 14 | 6.14 |


| $\underline{\text { subject of }}$ | $\underline{512}$ | $\mathbf{0 . 6}$ |
| :--- | ---: | ---: | ---: |
| devote | $\underline{28}$ | 7.69 |
| consume | $\underline{4}$ | 5.36 |
| tie | $\underline{6}$ | 4.87 |
| last | $\underline{4}$ | 4.6 |
| back | $\underline{5}$ | 4.5 |
| stretch | $\underline{4}$ | 4.29 |
| result | $\underline{6}$ | 3.93 |
| depend | $\underline{6}$ | 3.84 |
| limit | $\underline{5}$ | 3.59 |
| match | $\underline{3}$ | 3.58 |
| share | $\underline{6}$ | 3.55 |
| earn | $\underline{3}$ | 3.55 |
| enable | $\underline{7}$ | 3.54 |
| remain | $\underline{12}$ | 3.5 |

## 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)
$\Rightarrow$ 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)
$\Rightarrow$ 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)
$\Rightarrow$ put the phone down
- resource - nouns modified by r. - department (SkELL)
$\Rightarrow$ 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)
$\Rightarrow$ put the phone down
■ resource - nouns modified by r. - department (SkELL)
$\Rightarrow$ 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)
$\Rightarrow$ put the phone down
■ resource - nouns modified by r. - department (SkELL)
$\Rightarrow$ human resources department

- resource - words and/or r. - conservation (SkELL)
$\Rightarrow$ 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

■ $\Rightarrow$ 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

■ $\Rightarrow$ slows down the workflow

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

■ "commonest match"
■ . . . the longest available $\Rightarrow$ 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 $\Rightarrow$ 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

| down | part_trans | $\underline{2.534}$ | 9.94 put down |
| :---: | :---: | :---: | :---: |
| forward | modifier | $\underline{1.720}$ | 11.56 put forward |
| up | part_intrans | $\underline{1.176}$ | 7.83 to put up with |
| just | modifier | 832 | 8.77 just put |
| in | part_intrans | 796 | 8.88 put in |
| pressure | object | 519 | 8.20 put pressure on |
| then | modifier | 438 | 8.05 and then put |
| head | object | 387 | 6.92 put his head |
| thing | object | 382 | 6.57 put things |
| end | object | 327 | 6.80 to put an end to |
| off | part_intrans | 320 | 7.76 be put off |
| place | $p p$ _in-p | $\underline{241}$ | 6.61 put in place |
| right | $n p \_a d j$ _comp | $\underline{217}$ | 7.96 put it right |
| phone | object | $\underline{213}$ | 7.47 put the phone down |
| hand | part_out-a_obj | 167 | 5.92 put out a hand |
| lot | object | 147 | 6.30 put a lot of |
| use | $p p \_t o-p$ | 145 | 6.33 put to good use |
| kettle | object | 142 | 7.15 put the kettle on |

## LCM examples

- baby in BNC
- baby in SkELL
- test in SkELL
- resource in SkELL


## LCM evaluation

- 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


## Conclusions

- 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)

