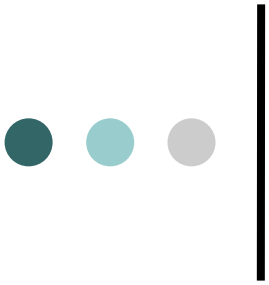
Corpus Evaluation

Adam Kilgarriff

Lexical Computing Ltd

- Then
  - Very few corpora
  - Use what's there
- Now
  - Corpora to spec
  - Choice
  - Need to evaluate



- Intrinsic
  - See what it looks like
- Extrinsic
  - Embed in a task
  - How well do you do at the task
  - ***Better***
    - It all depends what you want it for

it all depends what you want it for ***but***

- ‘general English (/French/Chinese/ …)’
  - Many purposes
  - Not specialist sublanguage
- A decent construct?
  - Not sure but it has form
    - General language dictionaries
    - *“how good is a corpus, for making them?”*

General truths

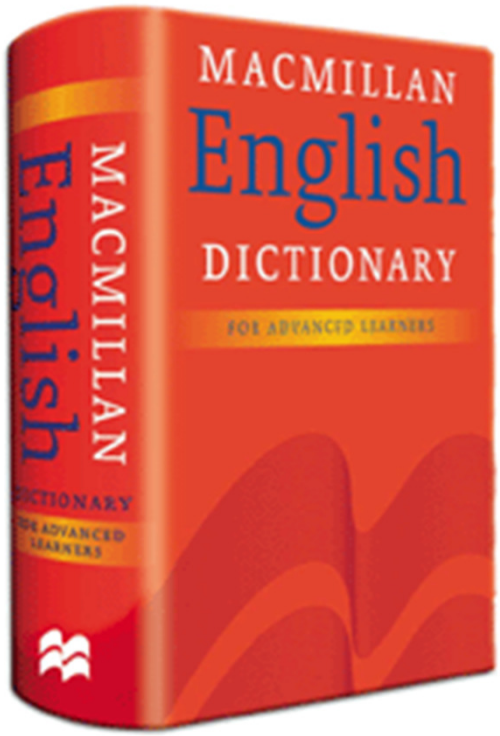
- Duplicates bad
- Noise bad
- Big good
- Diverse (good coverage of varieties within research scope, not dominated by any one variety) good

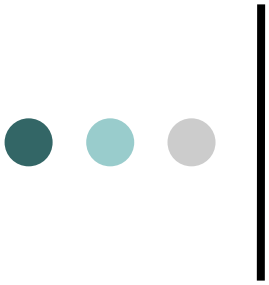
word sketch

A corpus-derived one-page summary of a word's grammatical and collocational behaviour

Macmillan English Dictionary For Advanced Learners

Ed: Rundell, 2002





- 11 years
  - 1999-2010
- Feedback
  - Good but anecdotal
- Formal evaluation

Goal

- Collocations dictionary
  - Model: Oxford Collocations Dictionary
  - Publication-quality
- Ask a lexicographer
  - For 42 headwords
    - For 20 best collocates per headwords
  - ***"should we include this collocation in a published dictionary?"***

Sample of headwords

- Nouns verbs adjectives, random
- **High (Top 3000)**
- *N* space solution opinion mass corporation leader
- *V* serve incorporate mix desire
- *Adj* high detailed open academic
- **Mid (3000- 9999)**
- *N* cattle repayment fundraising elder biologist sanitation
- *V* grieve classify ascertain implant
- *Adj* adjacent eldest prolific ill
- **Low (10,000- 30,000)**
- *N* predicament adulterer bake bombshell candy shellfish
- *V* slap outgrow plow traipse
- *Adj* neoclassical votive adulterous expandable

Precision and recall

- We tested **precision**
- **Recall** is harder
  - How do we find all the collocations that the system should have found?

Four languages, three families

- Dutch
  - ANW, 102m-word lexicographic corpus
- English
  - UKWaC, 1.5b web corpus
- Japanese
  - JpWaC, 400m web corpus
- Slovene
  - FidaPlus, 620m lexicographic corpus

User evaluation

- Evaluate whole system
  - Will it help with my task
    - Eg preparing a collocations dictionary
- Contrast: developer evaluation
  - Can I make the system better?
    - Evaluate each module separately
    - Current work

Components

- Corpus
- NLP tools
  - Segmenter, lemmatiser, POS-tagger
- Sketch grammar
- Statistics

Practicalities

- Interface
  - Good, Good-but
    - Merge to **good**
  - Maybe, Maybe-specialised, Bad
    - Merge to **bad**
- For each language
  - Two/three linguists/lexicographers
  - If they disagree
    - Don't use for computing performance

Results

- Dutch 66%
- English 71%
- Japanese 87%
- Slovene 71%

  - Two thirds of a collocations dictionary can be gathered automatically

<*world, final*> problem

- Is it good?
  - Superficially no
  - Look at concordances:
    - *World cup finals*
- Solution
  - ‘Commonest string’

Next step

- Recall
  - 200 collocates per headword
  - Selected from
    - All the corpora we have
    - Various parameter settings
  - Plus just-in-time evaluation for 'new' collocates
- Then
  - For a sample of headwords
    - *These are the collocations we should get*

From sketches to corpora

- Hold other inputs constant
  - Just one varies
  - Evaluate that one
- Hold tools, stats, grammar constant
  - ***evaluate corpora***

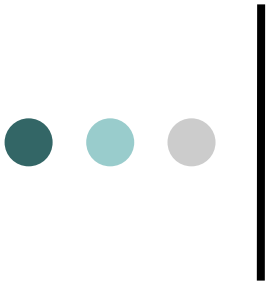
Criteria

- Duplicates bad
- Noise bad
- Big good
- Diverse (good coverage of varieties within research scope, not dominated by any one variety) good

- We think so

Over next year

- Build test sets
- Textbook cases
  - English
    - BNC vs UKWaC vs OEC vs Gigaword
  - Dutch
    - ANW corpus vs web corpus
- web crawling, deduplication
  - Which parameters give best results?



Thank you

http://www.sketchengine.co.uk