Bigger and better and bigger and better

Computational corpus-driven linguistics: a research programme

Adam Kilgarriff
Lexical Computing Ltd/Lexicography MasterClass Ltd
University of Sussex

The week
- Mon: Research programme overview
  - Practical: using the Sketch Engine
- Tues (1): Corpus design
- Tues (2): Web as corpus
- Weds: Practical: creating and examining your own web corpus
- Thurs: Word senses: theoretical and lexicographic perspectives (incl practical)

Overview overview
- Research programme
- Examples
  - corpus lexicography/ word sketches
  - thesauruses
- Research programme revisited:
  - text and meaning

What is language?
- What is language?
- In our heads

What is language?
- In our heads
- In texts and sound signals
What is language?
- In our heads
- In texts and sound signals
- Both

Methodology
- Study language in our heads
  - Competence
  - Chomsky
  - "rationalist" (Descartes, Leibniz)

Methodology
- Study text
  - "empiricist" (Locke, Hume)
  - Physics: forces, matter
  - Chemistry: chemicals, bonds
  - Language: text, speech signals

It goes against the grain
- What is important about a sentence?
  - its meaning
- Corpus methodology:
  - Throw away individual sentence meaning
  - Find patterns

Computers and corpora
- Machine learning
  - finds patterns in data sets
- Corpora
  - bigger and bigger data sets
- Language technology tools
  - lemmatizers, POS-taggers, parsers
- A new way to find out about language
  - 15 years of rapid ascent
A virtuous circle

More data →
Corpus Design,
Web-as-corpus

• Part-of-speech tagging
• Parsing
• Lemmatizing

Pattern finding
Linguistic processing

gets richer each time round

How children learn language?

Example 1: corpus lexicography

- Where do you go when the dictionary does not tell you enough?
- where the lexicographers go
- corpus
- four ages of corpus lexicography

Age 1: Pre-computer

- Oxford English Dictionary:
  - 20 million index cards

Age 2: KWIC Concordances

- From 1980
- Computerised

Age 2: KWIC Concordance

- Oxford English Dictionary:
- 20 million index cards
Age 2: KWIC Concordances

- From 1980
- Computerised
- COBUILD project was innovator
- the coloured-pens method

Age 2: limitations

as corpora get bigger:

**too much data**
- 50 lines for a word: read all
- 500 lines: *could* read all, takes a long time
- 5000 lines: no

Age 3: Collocation statistics

- Problem: too much data - how to summarise?
- Solution:
  - list of words occurring in neighbourhood of headword, with frequencies
  - Sorted by salience

Collocation listing

For right collocates of save (>5 hits)

<table>
<thead>
<tr>
<th>word</th>
<th>f(x+y)</th>
<th>f(y)</th>
<th>word</th>
<th>f(x+y)</th>
<th>f(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>forests</td>
<td>6</td>
<td>170</td>
<td>life</td>
<td>36</td>
<td>4875</td>
</tr>
<tr>
<td>$1.2</td>
<td>6</td>
<td>180</td>
<td>dollars</td>
<td>8</td>
<td>1668</td>
</tr>
<tr>
<td>lives</td>
<td>37</td>
<td>1697</td>
<td>costs</td>
<td>7</td>
<td>1719</td>
</tr>
<tr>
<td>enormous</td>
<td>6</td>
<td>301</td>
<td>thousands</td>
<td>6</td>
<td>1481</td>
</tr>
<tr>
<td>annually</td>
<td>7</td>
<td>447</td>
<td>face</td>
<td>9</td>
<td>2590</td>
</tr>
<tr>
<td>jobs</td>
<td>20</td>
<td>2001</td>
<td>estimated</td>
<td>6</td>
<td>2387</td>
</tr>
<tr>
<td>money</td>
<td>64</td>
<td>6776</td>
<td>your</td>
<td>7</td>
<td>3141</td>
</tr>
</tbody>
</table>
**Mutual Information**

- Church and Hanks 1989
- How much more often does a word pair occur, than one might expect by chance: MI
- Adjust to emphasise higher-frequency collocates: MI x log joint frequency

**Collocation listing**

For right collocates of *save* (>5 hits)

<table>
<thead>
<tr>
<th>word</th>
<th>f(x+y)</th>
<th>f(y)</th>
<th>word</th>
<th>f(x+y)</th>
<th>f(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>forests</td>
<td>6</td>
<td>170</td>
<td>life</td>
<td>36</td>
<td>4875</td>
</tr>
<tr>
<td>$1.2</td>
<td>6</td>
<td>180</td>
<td>dollars</td>
<td>8</td>
<td>1668</td>
</tr>
<tr>
<td>lives</td>
<td>37</td>
<td>1697</td>
<td>costs</td>
<td>7</td>
<td>1719</td>
</tr>
<tr>
<td>enormous</td>
<td>6</td>
<td>301</td>
<td>thousands</td>
<td>6</td>
<td>1481</td>
</tr>
<tr>
<td>annually</td>
<td>7</td>
<td>447</td>
<td>face</td>
<td>9</td>
<td>2590</td>
</tr>
<tr>
<td>jobs</td>
<td>20</td>
<td>2001</td>
<td>estimated</td>
<td>6</td>
<td>2387</td>
</tr>
<tr>
<td>money</td>
<td>64</td>
<td>6778</td>
<td>your</td>
<td>7</td>
<td>3141</td>
</tr>
</tbody>
</table>

**Age 4: The word sketch**

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

**English word sketches**

- British National Corpus (BNC)
  - 100 M words, already POS-tagged
  - lemmatized
  - assisting => assist (v)
- parsed
- database of 70 million triples
  - <object, sip, coffee>  <subject, arrive, coffee>
  - <and-or, tea, coffee>  <modifier, coffee, instant>

**Macmillan English Dictionary**

For Advanced Learners

Ed: Rundell, 2002
Can I have them for my language please?

The Word Sketch Engine
- Input:
  - any corpus, any language
  - Lemmatised, part-of-speech tagged
  - specification of grammatical relations
- Word sketches integrated with
- Corpus query system
  - Supports complex searching, sorting etc
  - IMS-Stuttgart formalism (also for corpus input)
    - Corpus searches and grammar writing
    - Christ and Schulze 1994
- Demo

Functions
- KWIC concordance
  - Sorting, filtering etc
- Word sketch
- Automatic thesaurus
- Sketch difference
  - discriminate near-synonyms
- In development
  - key words in a subcorpus / text type
  - how word varies with text type

Grammar writing
- Uses CQL (Corpus query language)
  - Christ and Schulze, U. Stuttgart, 1994
- defining an object:
  \[ v \ (a|d|n|d|e|t|n|m|a|d|v)^* \ n \]
- rewritting in CQL with BNC/CLAWS-5 tags
  \[ [\text{tag}="V.*"] [\text{tag}="(A|JTV|D|I|O).**"] [\text{tag}="NN.**"] \]

Developer: Pavel Rychly, Brno
Users:
- OUP, CUP, Irish Govt for lexicography
- Universities for teaching and research
- ELT textbook authors
Demo:
- http://www.sketchengine.co.uk/
  - Self-registration for free account
A virtuous circle

Pattern finding

Linguistic processing

Corpus

Lexicon

gets richer each time round

Part-of-speech tagging

Parsing

Lemmatizing

More data

Word sketching

Example 2: thesauruses

What is a thesaurus?

a resource that groups words according to similarity

Thesauruses manual and automatic

- Manual
  - Roget, WordNets, many publishers
- Automatic
  - aka distributional
  - two words are similar if they occur in same contexts
  - Are they comparable?

Thesauruses in NLP

- sparse data

- does $x$ go with $y$?
  - don’t know, they have never been seen together
- New question: does $x$ + friends go with $y$ + friends
  - indirect evidence for $x$ and $y$
  - thesaurus tells us who friends are
  - “backing off”
Relevant in:
- Speech understanding
- Parsing
- PP-attachment
- conjunction scope
- Bridging anaphors
- Text cohesion
- Word sense disambiguation (WSD)
- Spelling correction

Conjunction scope
- Compare
  - old boots and shoes
  - old boots and apples
- Are the shoes old?

Conjunction scope
- Compare
  - old boots and shoes
  - old boots and apples
- Are the shoes old?
- Are the apples old?

Conjunction scope
- Compare
  - old boots and shoes
  - old boots and apples
- Are the shoes old?
- Are the apples old?
- Hypothesis:
  - wide scope only when words are similar

Automatic thesauruses
- Sparck Jones, Grefenstette, Lin
- Lots of recent work:
  - Weeds, Curran theses
  - Weeds and Weir, ACL 2003
  - McCarthy et al, ACL 04, COLING 04
  - Geffet and Dagan 2004
Word Sketch Engine thesaurus

- Similarity:
  - \langle \text{obj, drink, beer} \rangle
  - \langle \text{obj, drink, wine} \rangle
- one point similarity between \text{beer} and \text{wine}
- count all points of similarity between all pairs of words
- weight according to frequencies

Nearest neighbours

\text{zebra: giraffe buffalo hippopotamus rhinoceros gazelle antelope cheetah hippo leopard kangaroo crocodile deer rhino herbivore tortoise primate hyena camel scorpion macaque elephant mammoth alligator carnivore squirrel tiger newt chimpanzee monkey}

Nearest neighbours

<table>
<thead>
<tr>
<th>crane</th>
<th>winch</th>
<th>swan</th>
<th>heron</th>
</tr>
</thead>
<tbody>
<tr>
<td>winch</td>
<td>crane</td>
<td>heron</td>
<td>tern</td>
</tr>
<tr>
<td>heron</td>
<td>mast</td>
<td>crane</td>
<td>gull</td>
</tr>
<tr>
<td>tractor</td>
<td>rigging</td>
<td>gull</td>
<td>swan</td>
</tr>
<tr>
<td>truck</td>
<td>pump</td>
<td>tern</td>
<td>crane</td>
</tr>
<tr>
<td>swan</td>
<td>tractor</td>
<td>curlew</td>
<td>flamingo</td>
</tr>
</tbody>
</table>

Google sets

- http://labs.google.com/sets
- Input: zebra giraffe buffalo

Google sets

- http://labs.google.com/sets
- Input: zebra giraffe buffalo
- \text{kudu hyena impala leopard hippo waterbuck elephant cheetah eland}
A virtuous circle

More data → Corpus

Part-of-speech tagging

Linguistic processing

Lemma\[ification\]

Thesaurus

Pattern finding

Word sketching

Corpus

Linguistic processing

Lexicon

gets richer each time round

The long journey from text towards meaning

Raw text

Pure meaning

Rationalists

Empiricists

The long journey from text towards meaning

Raw text

Pure meaning

Rationalists

Empiricists

Lemmatizer

POS-tagger

Parser

Thesaurus

Frame elements/semantics

Semantic relations

Thesaurus

Rationalists

Empiricists