Corpora for the coming decade

Adam Kilgarriff
Lexical Computing Ltd
Universities of Leeds, Sussex
Overview

- History
- Comparing Corpora
- The Web and Corpora
  - Demo
- Corpus Factory
- Simple Maths for Keywords
- Corpora for the Coming Decade
Corpus size since the 1960s

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown/LOB</td>
<td>10^6</td>
<td>10^7</td>
<td>10^8</td>
<td>10^9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COBUILD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no limits</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
‘Zipfian’ distribution of words

% of all texts

0 10 20 30 40 50 60 70 80 90 100
'the' 100 most frequent 3500 most frequent 7500 most frequent

Heidelberg 2010
Kilgarriff: Corpora
Comparing Corpora

• Basic science
  – Measure
  – Compare

• State of the art
  – “WSJ”, “medical abstracts”, “general”
  – atrocious
Wall St Journal vs. BNC?
Wall St Journal vs. BNC?

- Homogeneity
  - Self-similarity

- *Use same measure*
  - For homogeneity and similarity
  - (distance measure so:
    - Heterogeneity and distance
    - High number=different/ heterogeneity)
Thought experiment

<table>
<thead>
<tr>
<th>Corp1</th>
<th>Corp2</th>
<th>Distance</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>equal</td>
<td>equal</td>
<td>equal</td>
</tr>
<tr>
<td>2</td>
<td>equal</td>
<td>equal</td>
<td>high</td>
</tr>
<tr>
<td>3</td>
<td>high</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>4</td>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>5</td>
<td>high</td>
<td>low</td>
<td>higher</td>
</tr>
<tr>
<td>6</td>
<td>low</td>
<td>low</td>
<td>a little higher</td>
</tr>
<tr>
<td>7</td>
<td>high</td>
<td>high</td>
<td>a little higher</td>
</tr>
</tbody>
</table>
Measures

• Homogeneity
  – Divide randomly into halves
  – Measure distance between halves
  – Iterate, average

• Proposed measures
  • word frequency lists
    – Chi-square (normalise by DF): CBDF
    – Spearman Rank Correlation
  • From language modelling
    – Perplexity
How to evaluate measures

- Known-similarity corpora
  - Two text types
  - Eleven corpora
    - 100:0, 90:10, 80:20, 70:30, 60:40 ... 10:90, 0:100
  - Gold-standard judgements
    - 80:20 is-more-similar-to 70:30 than 90:10 is to 60:40

- What percentage of gold-standard judgements does each measure get right?
  - CBDF wins
    - best with 500 DF, 500-most-freq-words
### BNC 200,000-wd samples

|   | ACC | ART | BMJ | DMI | DNB | ENV | FAC | GRA | GUA | HAN | IND | NME | ACC | ART | BMJ | DMI | DNB | ENV | FAC | GRA | GUA | HAN | IND | NME |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|   | 4.6 | 21.4| 20.2| 21.6| 40.6| 22.7| 20.5| 27.8| 14.1| 24.1| 12.8| 21.2| 4.6 | 21.4| 20.2| 21.6| 40.6| 22.7| 20.5| 27.8| 14.1| 24.1| 12.8| 21.2| 4.6 |
Then and now

• Work done: 1995
  – Journal article 2001

• Then:
  – Theoretical interest
  – Beggars can't be choosers

• Now
  – Any number of corpora
    • to spec, from web
  – Practical importance
The Web and Corpora

- Is the web a corpus?
- Representativeness
- What is out there?
  - Web1T
- Googleology
- Web corpus types
  - Targeted sites: Oxford English Corpus
  - General: WaC family
  - WebBootCaT
You can’t help noticing

• *Replaceable or replacable?*
  – [http://googlefight.com](http://googlefight.com)
  – [http://looglefight.com](http://looglefight.com)
• Very very large
• Most languages
• Most language types
• Up-to-date
• Free
• Instant access
Is the web a corpus?

• Sinclair
  – in “Developing linguistic corpora, a guide to good practice. Corpus and Text – Basic Principles”
  “…not a corpus because
  • dimensions unknown, constantly changing
  • not designed from a linguistic perpective

• But
  – We can find out dimensions
  – Many corpora are not designed
    • “as much chatroom dialogue as I can get”

• Def: a corpus is a collection of texts
  – when viewed as an object of language research
Is the web a corpus?

Yes
but it’s not representative
Theory

A random sample of a population is representative of it.

Observations on sample support inferences about population

(within confidence bounds)
Theory

A random sample of a population is …

• *What is the population?*
  – production and reception
  – speech and text
  – copying
Theory

• Population not defined
• Representative sample not possible
sublanguage

- Language = core + sublanguages
- Options for corpus construction
  - none
  - some
  - all
- None
  - impoverished view of language
- Some: BNC
  - cake recipes and gastro-uterine disease
  - not car repair manuals or astronomy or ... 
- All: until recently, not viable
Representativeness

• The web is not representative

• *but nor is anything else*

• Text type variation
  – under-researched, lacking in theory
    • Atkins Clear Ostler 1993 on design brief for BNC; Biber 1988, Kilgarriff 2001

• Text type is an issue across NLP
  – Web: issue is acute because, as against BNC or WSJ, we simply don’t know what is there
What is out there?

• What text types are there on the web?
  – some are new: chatroom
  – proportions
    • is it overwhelmed by porn? How much?
  • Hard question
Classifiers

Starter set of text types, with examples

Taxonomy of text types

Linguist revises/extends taxonomy

Classify new samples: Check misfits

Build text classifier

Take new random sample

Marina Santini, Serge Sharoff
Comparing frequency lists

• Web1T vs BNC
  – Keywords of each vs other
Web-high (155 terms)

- 61 web and computing
  - config browser spyware url www forum
- 38 porn
- 22 US English (incl Spanish influence – los)
- 18 business/products common on web
  - poker viagra lingerie ringtone dvd casino rental collectible tiffany
  - NB: BNC is old
- 4 legal
  - trademarks pursuant accordance herein
Web-low

• Exclude British English, transcription/tokenisation anomalies

– herself stood seemed she looked yesterday sat considerable had council felt perhaps walked round her towards claimed knew obviously remained himself he him
Observations

• Pronouns and past tense verbs
  – Fiction
• Masc vs fem
• Yesterday
  – Probably daily newspapers
• Constancy of ratios:
  – He/him/himself
  – She/her/herself
• The web
  – a social, cultural, political phenomenon
  – new, little understood
  – a legitimate object of science
  – mostly language
    • we are well placed
  – a lot of people will be interested
• Let’s
  – study the web
  – source of language data
  – apply our tools for web use (dictionaries, MT)
  – use the web as infrastructure
Web corpus types

- Large, general corpora
- Small, specialised corpora
  - Specially for translators
  - BootCaT, WebBootCaT
Basic steps

• Gather pages
  – Google hits
  – Select and gather whole sites
  – General crawl
• Filter
• De-duplicate
• Linguistic processing
• Load into corpus tool
Filtering

• Non-text (sound, image etc) files
• Boilerplate (within file)
  – Copyright notices, navigation bars
  – “high markup” heuristic
• Not “text in sentences”
  – Look for function words
  – Lists?? Sports results?? Crossword puzzles??
• Spam, pornography
  – Tough
• De-duplication (also tough)
Corpus Factory

- Many languages
- General corpus, 100m+ words
  - Fast
  - High quality
  - Comparable across languages
Gather Seed Words

- Sharoff: used word lists from preexisting corpora
  - BNC for English
  - RNC for Russian
- Bottleneck: No pre-existing large general corpora for many languages.
  - That is why we are building them!
  - Seed words from many domains required.
Gather Seed words

- Wikipedia (Wiki) Corpora
  - many domains
  - free
  - 265 languages covered, more to come
- Extract text from Wiki.
  - Wikipedia2Text
- Tokenise the text.
  - Morphology of the language is important
  - Can use the existing word tokeniser tools.
Gather Seed words

- Thai Word Segmentation
  - Before tokenization
    ปัญหาของประเทศพม่าในภูมิภาคคืออะไร
    (Gloss: Burma's problems in the region)
  - After tokenization
    ปัญหา/ของ/ประเทศ/พม่า/ใน/ภูมิภาค/คือ/อะไร
    problem/of/Country/Burma/in/Region/is/?
- Used Swath word Segmentor.
Gather Seed words

- Most frequent are function words
  - Top 500 (roughly)
  - Use to identify connected text.
- Mid frequency as seeds
  - 1000th to 6000th words (roughly)
### Query Generation: cont..

**Table 2: Query length, hit counts at 90th percentile and Best Query Length**

<table>
<thead>
<tr>
<th></th>
<th>length= 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>1,300,000</td>
<td>3,580</td>
<td>74</td>
<td>5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Hindi</td>
<td>30,600</td>
<td>86</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Telugu</td>
<td>668</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Thai</td>
<td>724,000</td>
<td>1,800</td>
<td>193</td>
<td>5</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1,100,000</td>
<td>15,400</td>
<td>422</td>
<td>39</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Collection

- 30,000 queries
- Retrieve top 10 search hits of each query.
  - Yahoo Search API
- Download
Cleaning

- Body Text Extraction (Finn et al. 2001)
  - Boilerplate: rich in markup
  - Body text: middle of page, light in markup
  - 3 zones: High-low-high
  - Retain low
Filtering

- Wanted: “stuff in sentences”
  - Connected text
- Not wanted: anything else
  - Menus, directories, catalogues...
- Connected text
  - half of all tokens are very common words
- Discard pages failing test
Near Duplicate Detection

- Broder et al (1997) 'shingling'
- To be replaced by Pomikalek's methods (Pomikalek 2009)
## Web Corpus Statistics

<table>
<thead>
<tr>
<th>Language</th>
<th>Unique URLs collected</th>
<th>After filtering</th>
<th>After de-duplication</th>
<th>Web corpus size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Words</td>
</tr>
<tr>
<td>Dutch</td>
<td>97,584</td>
<td>22,424</td>
<td>19,708</td>
<td>739</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>108.6 m</td>
</tr>
<tr>
<td>Hindi</td>
<td>71,613</td>
<td>20,051</td>
<td>13,321</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30.6 m</td>
</tr>
<tr>
<td>Telugu</td>
<td>37,864</td>
<td>6,178</td>
<td>5,131</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.4 m</td>
</tr>
<tr>
<td>Thai</td>
<td>120,314</td>
<td>23,320</td>
<td>20,998</td>
<td>1.2 GB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>81.8 m</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>106,076</td>
<td>27,728</td>
<td>19,646</td>
<td>1.2 GB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>149 m</td>
</tr>
</tbody>
</table>
Evaluation

- For each of the languages, two corpora available:
  - Web and Wiki
  - Dutch: also a carefully designed lexicographic corpus.
- Hypothesis: Wiki corpora are ‘informational’
  - Informational --> typical written
  - Interactional --> typical spoken
Evaluation

- 1st, 2nd person pronouns
  - strong indicators of interactional language.
  - English: *I me my mine you your yours we us our*
- For each languages
  - Ratio: web:wiki
# Results

<table>
<thead>
<tr>
<th>Word</th>
<th>Web</th>
<th>Wiki</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>ผม</td>
<td>2935</td>
<td>366</td>
<td>8.00</td>
</tr>
<tr>
<td>ดิฉัน</td>
<td>133</td>
<td>19</td>
<td>7.00</td>
</tr>
<tr>
<td>ฉัน</td>
<td>770</td>
<td>97</td>
<td>7.87</td>
</tr>
<tr>
<td>คุณ</td>
<td>1722</td>
<td>320</td>
<td>5.36</td>
</tr>
<tr>
<td>ท่าน</td>
<td>2390</td>
<td>855</td>
<td>2.79</td>
</tr>
<tr>
<td>กระผม</td>
<td>21</td>
<td>6</td>
<td>3.20</td>
</tr>
<tr>
<td>ข้าพเจ้า</td>
<td>434</td>
<td>66</td>
<td>6.54</td>
</tr>
<tr>
<td>ตัว</td>
<td>2108</td>
<td>2070</td>
<td>1.01</td>
</tr>
<tr>
<td>กุ</td>
<td>179</td>
<td>148</td>
<td>1.20</td>
</tr>
<tr>
<td>ขึ้น</td>
<td>431</td>
<td>677</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>11123</strong></td>
<td><strong>4624</strong></td>
<td><strong>2.40</strong></td>
</tr>
</tbody>
</table>

*Table: 1st and 2nd person pronouns in Web and Wiki corpora per million words*
Corpora for the coming decade
How should they be different?

- Bigger

- Better
Bigger

- Motivation
  - Ample data for rare phenomena
  - Big subcorpora
  - For language modelling

- More like Google-scale
  - but without Google disadvantages
    - See *Googleology is Bad Science*, CL 2007
Better

• Less noise
• Fewer duplicates
• Richer markup
  – At word, sentence level
  – At document level (text type, subcorpora)
Divide and rule

• Bigger (+ cleaning + deduplication)
  – Big Web Corpus (BiWeC)
    • Currently 5.5b fully processed
    • Target 20b words
    • Jan Pomikalek, Pavel Rychly

• Better
  – New Model Corpus
New Model Corpus

• model
  1. small version: *model train*
  2. design: *data model*

• New Model Corpus
  – 1:100 scale model
  – To replace BNC as design model
BNC design model

• Most often used
  – Eg for other languages

• pre-web
  – $f(\text{blog})=0$

• Corpora now bigger, far quicker, far cheaper, different issues

• *BNC design model past its sell-by*
  – Kilgarriff Atkins Rundell, Corpus Lg 2007
New model

• Data
• Markup
Data

- From the web
- 100m words
- Small sample size
  - Copyright
  - ??Creative Commons Licence
Composition

• General crawl 50

• Targeted
  – Fiction 7
  – Blog 7
  – Newspaper (RSS feed) 7
  – Speech 10
    • Film transcripts, chatshow
  – Domain-specific 19
    • Business, medical, law
Markup

• Collaborative
  – We distribute data
  – Anyone applies their tools
    • Pos-tagger, parser, co-ref resolution, domain classifier, WSD, semantic classifier, time phrases, named entities...
  – We integrate, display in Sketch Engine
  – Research potential from multiple markup
Two strands

• Apply methods with good accuracy (and *fast*) to BiWeC
Two strands

• Apply methods with good accuracy (and fast) to BiWeC

• Bigger

• Better
Some plans

• Corpus similarity/homogeneity
  – Web service for measuring

• New General Service List
  • Replacing West (1953)
  • Words (English) you *always* need
    – Many corpora of different text types
    – 2000-wd samples
    – Which words occur in 95% of docs in every text type
Hierarchy of Domains

- Domains are in hierarchies
  - Science, physics, subatomic physics
- Domains: represented by corpora
- Can we find correlates in wordlists
- What we *could* find

<table>
<thead>
<tr>
<th></th>
<th>Core</th>
<th>science</th>
<th>physics</th>
<th>subatomic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>70</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Physics</td>
<td>70</td>
<td>5</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Subatomic</td>
<td>70</td>
<td>5</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>