Tasks
- Design
- Collection
- Encoding

Design
- Size
  - Very large resource needed for rare items
  - good statistical generalisations
- Types of text

Why size matters (1)

Zipf’s Law: $r \times f = C$

<table>
<thead>
<tr>
<th>Word (pos)</th>
<th>$r$</th>
<th>$f$</th>
<th>$r \times f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the (det)</td>
<td>1</td>
<td>6187267</td>
<td>6187267</td>
</tr>
<tr>
<td>to (prep)</td>
<td>10</td>
<td>917579</td>
<td>9175790</td>
</tr>
<tr>
<td>as (adv)</td>
<td>100</td>
<td>91583</td>
<td>9158300</td>
</tr>
<tr>
<td>playing (vb)</td>
<td>1000</td>
<td>9738</td>
<td>9738000</td>
</tr>
<tr>
<td>amateur (adj)</td>
<td>10,000</td>
<td>741</td>
<td>7410000</td>
</tr>
<tr>
<td>reptilian</td>
<td>100,000</td>
<td>20</td>
<td>2000000</td>
</tr>
</tbody>
</table>

Why size matters (2)

Zipfian factors also affect:
- specific meanings
- specific combinations

Case study: break (verb)

<table>
<thead>
<tr>
<th>BNC frequency</th>
<th>19.512</th>
</tr>
</thead>
<tbody>
<tr>
<td># meanings/multiwords (MED 2002)</td>
<td>77</td>
</tr>
<tr>
<td>Meaning 1 ('core' meaning)</td>
<td>4.000</td>
</tr>
<tr>
<td>Meaning 5 (break the news/news breaks):</td>
<td>750</td>
</tr>
<tr>
<td>Meaning 11</td>
<td>50</td>
</tr>
<tr>
<td>Combinations: break even, break into a sweat, it won’t break the bank...</td>
<td>fewer and fewer</td>
</tr>
</tbody>
</table>
How large is large enough?

for what?
- grammar studies
- sociolinguistics
- lexicography
- NLP
- grammar induction
- lexical acquisition

Corpus size: another view

I don’t think there can be any corpora, however large, that contain information about all of the areas of English ... that I want to explore.

[but]
Every corpus that I’ve had a chance to examine, however small, has taught me facts that I couldn’t imagine finding out about in any other way.

Fillmore 1992: 35

Types of text

Why size matters (1)

Zipf’s Law: \[ r \times f = C \]

<table>
<thead>
<tr>
<th>Word (pos)</th>
<th>( r )</th>
<th>( f )</th>
<th>( r \times f )</th>
</tr>
</thead>
<tbody>
<tr>
<td>the (det)</td>
<td>1</td>
<td>6187267</td>
<td>6187267</td>
</tr>
<tr>
<td>to (prep)</td>
<td>10</td>
<td>917579</td>
<td>9175790</td>
</tr>
<tr>
<td>as (adv)</td>
<td>100</td>
<td>91583</td>
<td>9158300</td>
</tr>
<tr>
<td>playing (vb)</td>
<td>1000</td>
<td>9738</td>
<td>9738000</td>
</tr>
<tr>
<td>amateur (adj)</td>
<td>10,000</td>
<td>741</td>
<td>7410000</td>
</tr>
<tr>
<td>reptilian</td>
<td>100,000</td>
<td>20</td>
<td>2000000</td>
</tr>
</tbody>
</table>

Zipf’s Law: implications

the:
- 6%
- 100 most frequent: 45%
- 10,000 most frequent: 88%
- the rest... 12%

Why size matters (2)

Zipfian factors also affect:
- specific meanings
- specific combinations

Case study: break (verb)

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC frequency</td>
<td>19.512</td>
</tr>
<tr>
<td># meanings/multiwords (MED 2002)</td>
<td>77</td>
</tr>
<tr>
<td>Meaning 1 ('core' meaning)</td>
<td>4.000</td>
</tr>
<tr>
<td>Meaning 5 (break the news/news breaks)</td>
<td>750</td>
</tr>
<tr>
<td>Meaning</td>
<td>50</td>
</tr>
<tr>
<td>Combinations: break even, break into a sweat, it won’t break the bank</td>
<td>fewer and fewer</td>
</tr>
</tbody>
</table>
How large is large enough?

for what?
- grammar studies
- sociolinguistics
- lexicography
- NLP
- grammar induction
- lexical acquisition

Corpus size: another view

I don't think there can be any corpora, however large, that contain information about all of the areas of English ... that I want to explore.

[but]

Every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.

Fillmore 1992: 35

Corpus creation Kilgarriff: Barcelona Oct 2005

One project

<table>
<thead>
<tr>
<th>Text category</th>
<th>Irish</th>
<th>Hiberno-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words: actual</td>
<td></td>
<td>Words: actual</td>
</tr>
<tr>
<td>Books - imaginative</td>
<td>7,600,000</td>
<td>6,000,000</td>
</tr>
<tr>
<td>Books - informative</td>
<td>8,400,000</td>
<td>7,000,000</td>
</tr>
<tr>
<td>Newspapers</td>
<td>4,500,000</td>
<td>5,300,000</td>
</tr>
<tr>
<td>Periodicals</td>
<td>2,600,000</td>
<td>700,000</td>
</tr>
<tr>
<td>Official/Govt</td>
<td>1,200,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Broadcast</td>
<td>400,000</td>
<td>0</td>
</tr>
<tr>
<td>Websites</td>
<td>5,500,000</td>
<td>5,000,000</td>
</tr>
<tr>
<td>TOTALS</td>
<td>30,200,000</td>
<td>25,000,000</td>
</tr>
</tbody>
</table>

Corpus creation Kilgarriff: Barcelona Oct 2005

Types of text

<table>
<thead>
<tr>
<th>mode</th>
<th>medium</th>
<th>genre</th>
<th>variety</th>
<th>domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>written</td>
<td>book</td>
<td>novel</td>
<td>(for English)</td>
<td>cooking</td>
</tr>
<tr>
<td>spoken</td>
<td>newspaper</td>
<td>play</td>
<td>US English</td>
<td>philosophy</td>
</tr>
<tr>
<td>in between</td>
<td>email</td>
<td>academic</td>
<td>UK English</td>
<td>romance</td>
</tr>
<tr>
<td></td>
<td>movie</td>
<td>instruction</td>
<td>Indian English</td>
<td>history</td>
</tr>
<tr>
<td></td>
<td></td>
<td>weblog</td>
<td></td>
<td>sport</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>business</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>linguistics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dermatology</td>
</tr>
</tbody>
</table>

Text type and language use

<table>
<thead>
<tr>
<th>WStJ financial assets such as</th>
<th>bonds</th>
<th>shares or other securities</th>
</tr>
</thead>
<tbody>
<tr>
<td>WStJ to yield on 30-year Treasury bonds</td>
<td>at 8.25% and that the...</td>
<td></td>
</tr>
<tr>
<td>WStJ made a fortune in junk bond trading...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WStJ NewS: how short-covalent bonds and long ionic bonds...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NewS Two of the hydrogen bonds are due to its own atoms...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-f forming material bonds with her new baby...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-f would strengthen the bonds that united the small co...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fact he undid the bonds and tore off the gag...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

lovely conversation 437
academic writing 9
whole corpus 64
police news text 960
whole corpus 275
‘Burstiness’

- BNC:
  - mucosal - 675, 9 texts
  - nome - 330, 3 texts
  - urge (noun) - 670, 414 texts

Design and “High quality Irish”

- Irish is ‘smaller’ than 150 years ago
- Many documents are translations
- Learners’ errors, inelegant prose
- Samuel Johnson: “writers of the first reputation”

Con
- Who judges?
- Risk of literary or backward-looking bias
- Lexicographers need corpus to translate
  - Boot the computer as well as the babbling brook
- Trench and the OED: “an historian, not a critic”
- Will a quality filter limit corpus breadth (and size)?

Quality: outcome

- Wide range of text types wanted
- Particular effort to gather native speaker non-translations

Collection

- Use existing
- Ask publishers
- Web

Use existing

- PAROLE corpora for a range of languages
- Distributors
  - LDC, ELRA
- Ask on corpora

Ask publishers

- The junkmail problem
- Appeals to national pride
- Charm and persistence
Web
- Fast becoming the usual place to look
  - Kilgarriff and Grefenstette, CL 2003
- Lots of data for many languages

Web issues
- Formats
  - conversion from pdf etc needed
- Character representation
  - Not many pages “do the right thing”
- Navigational material: “click here”
- Lists
- Mixed languages
- Duplication

Encoding
- Clean-up
- Linguistic processing
- Delivery formalism

Clean-up
- Deletion of:
  - Title pages, table of contents, tables, figures, footnotes, endnotes, page headers and footers, crosswords, TV listings, sports results, team listings …

Linguistic processing
- Lemmatize
  - give, giving, gives, given, gave => give (verb)
- Part-of-speech tagging
  - bank (verb) or bank (noun)?
- English, Spanish: tools are available
- Smaller languages
  - Perhaps a person-year to develop
  - Finite state methods, constraint grammar

Delivery formalism
- Both
  - XML Corpus Encoding Standards (XCES)
  - For longevity, interchange format
- And
  - Loaded into a corpus query tool
Summary: corpus creation requires

- **Design**
  - Size, text types, proportions

- **Collection**
  - Use-existing, ask-copyright-holders, web

- **Encoding**
  - Clean-up, XML, linguistic
Web as Corpus

Adam Kilgarriff
Lexicography Masterclass Ltd
Lexical Computing Ltd
University of Sussex

Outline

- Is the web a corpus?
- History
- Survey
- Representativeness
- Studying the web
- Search engine: proposal

Is the web a corpus?

- McEnery and Wilson textbook:
  "In principle, any collection ... but in context of modern linguistics, connotation include:
  - Sampling and representativeness
  - Finite size
  - Machine-readable form
  - Standard reference"
- most NLP corpora do not fit
- Different questions:
  - Is it a corpus?
  - Is it a good corpus for the research question?
- Def
  - A corpus is a collection of texts, when viewed as an object
    of language or literary study

Is the web a corpus?

Yes

History

You can't help noticing

- spellchecker: speculater or speculator?
- word research: chocolatier
Higher tech

- Resnik’s Linguistic Search Engine
- Fletcher’s Kwicfinder
- Baroni and Bernadini’s BootCat

Approaches

- Use Google hit counts
  - Markert, Keller and Lapata
- Use snippets
- Use google, then download pages
- Spider from relevant starting sites

(Marco Baroni’s analysis)

Project with Oxford University Press

- BNC: 1980s text is too old
  - New corpus needed
  - Web: cheap and quick (12 month project)
- copyright
- specification:
  - X million words of text type Y
- identify good hubs, then spider
- automatic classifiers
- tidying up
  - delete duplicates, non-text, navigational material
- 100M word corpus delivered (in 2003) and in use

Project with Irish Government

- New English-Irish Dictionary
  - Main contractor: Lexicography MasterClass
  - Corpora needed:
    - Hiberno-English
    - Irish (a “small” language)
  - Strategies
    - Re-use existing resources
    - Ask publishers, newspapers, authors for text
    - Web
  - Web corpus
    - Methods: as before
    - Delivery July 2004
    - 15M Irish corpus
    - 20M Hiberno-English corpus

but it’s not representative

- Representativeness
  - Theory
  - Sublanguages

Theory

A random sample of a population is representative of it. Observations on the sample support inferences about the population (within confidence bounds).
Theory

A random sample of a population is representative of it...

- **What is the population?**
  - production and reception
  - speech and text
  - background language
  - 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

Summary

- The web is not representative
- **but nor is anything else**
- Text type variation
  - under-researched, lacking in theory
  - 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
- **Sampling is non-trivial**
- Snapshots

Classifiers

Manual starter set of web pages for each text type
Starter set of text types → Taxonomy of text types → Build text classifier

Linguist revises/extends taxonomy
Classify New samples: Check misfits
Take new random sample
“What is out there” revisited

- 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
  - use the web as a source of data for studying language
  - apply language technology for web use (IR, QA, MT)
  - use the web as infrastructure

Outline

- Is the web a corpus?
- History
- Survey
- Representativeness
- Studying the web
- Search engine: proposal

The Trouble with Google

- not enough instances (max 2000)
- not enough context
  - ca 10-word snippet around search term
- ridiculous sort order
  - search term in titles and headings
- linguistically dumb
  - not lemmatised
  - think/thinks/thinking/thought: four searches
  - not POS-tagged
  - mixes up beat (n) and beat (v)
  - and why not parsed

DIY

- Language researchers aren’t a mighty commercial interest
- Google won’t prioritise doing things our way
- let’s do it ourselves

Components

1. web crawler
2. filter/classifier
3. linguistic processing
4. database
5. statistical summaries
6. user interface

Scale and speed

- Commercial search engines
  - banks of computers
  - highly optimised code
  - but this is for performance
  - no downtime
  - instant responses to millions of simultaneous queries
- This proposal
  - crawling: once a year
  - downtime: acceptable
  - not so many users
Filters

- Reject
  - image, sound, video files etc (by file type)
  - timetables, price lists, navigation
    - hard: linguistic clues
    - we only want “stuff in sentences”
    - rejects non-text within pages

For text chunks

- Identify
  - language (solved problem)
  - Text type (genre and/or domain)

Pointers

- WACKY project
  - Marco Baroni, U Bologna
  - http://wacky.sslmit.unibo.it/
  - Mailing list
  - 1.7B words of web German
    - Others to follow

- Reference:
  - Special Issue, Computational Linguistics 29 (3) 2003
    Editors: Kilgarriff and Grefenstette