Informatics 1: Data & Analysis
Lecture 12: Corpora

Ian Stark

School of Informatics
The University of Edinburgh

Friday 3 March 2017
Semester 2 Week 6
What is a Corpus?

In empirical linguistics, a corpus is a large body of written or spoken text, constructed to be representative of some language or language variety.

What is the plural of corpus?

- Corpori
- Corpora
- Corpuscle

Time: 10 seconds
What does the XPath expression `//*`/text() identify in this tree?

- The single word “I”.
- The set of strings {“spoken-text”, “sentence”, “word”, “punctuation”}
- The set of strings {“I”, “feel”, “really”, “awake”, “.”}
Thelma & Louise, 1991

Which of these lists the complete text of all sentences containing an adverb?

- `//sentence[word[@pos="adverb"]]/text()`
- `//word[@pos="adverb"]/text()`
- `//word[@pos="adverb"]/..//text()`
- `//*[word[@pos="adverb"]/word/text()]`
We start with technologies for modelling and querying *semistructured data*.

- Semistructured Data: Trees and XML
- Schemas for structuring XML
- Navigating and querying XML with XPath

**Corpora**

One particular kind of semistructured data is large bodies of written or spoken text: each one a *corpus*, plural *corpora*.

- Corpora: What they are and how to build them
- Applications: corpus analysis and data extraction
Remote Working (again)

Much coursework can be done on your own machines, but sometimes it’s important to be able to connect to and use DICE systems.

You can always do this by going into the Forrest Hill labs, open 24/7.

There are also many things you can access remotely:

- Files over the web. https://ifile.inf.ed.ac.uk
- Command line. ssh student.ssh.inf.ed.ac.uk then ssh student.login
  (On Microsoft Windows, use PuTTY to reach student.ssh.inf.ed.ac.uk)
- Desktop. Graphical login with NX to nx.inf.ed.ac.uk

You can also use Virtual DICE, tunnel X Windows, access files over AFS, and connect by VPN to internal networks at Informatics and the University.

http://computing.help.inf.ed.ac.uk
Written or spoken natural language has plenty of internal structure: it consists of words, phrases and sentences, governed by spelling and grammatical rules, and so forth.

Nevertheless, on a computer, it is standardly represented as a text file: a simple sequence of characters.

This is an example of unstructured data: the data format itself has no structure imposed on it. (Above the level of character encoding.)

Often, however, it is useful to annotate text by marking it up with additional information about its linguistic or semantic content.

Text with this kind of markup is a widespread and substantial example of semistructured data.
What is a Corpus?

The word *corpus* (plural *corpora* or *corpuses*) is Latin for “body”.

In literature a *corpus* is a collection of written texts, in particular the complete works of a single author, or a body of writing on a single subject.

In *computational linguistics* and in *theoretical linguistics* a *corpus* is a body of written or spoken text used for study of a particular language or language variety.

This application domain depends on the following features in a corpus:

- Representative sampling
- Finite size
- Machine-readable form
- Use as a standard reference

The following slides expand on these: all are important for a corpus to be a useful linguistic resource.
Representative Sampling

Sampling

Corpora provide data for *empirical linguistics*: the scientific investigation of real-world use of language, through proposing and testing hypotheses.

However, any corpus can only contain a *sample* of language use — although it might be very large, it will usually be dwarfed by the actual language in the wild.  

(XKCD: What if? #34)

Representative

For meaningful linguistic analysis, the sample in a corpus needs to be *representative*: it should contain a similar mix of text to the language variant for which it is being developed.

For example, the complete works of Shakespeare is an appropriate corpus for analysing how Shakespeare used language; but would not give a representative sample for studying Elizabethan English.
It’s natural that corpora should be finite. Most also have a fixed size. When building a corpus it is usually decided at the outset how the language is to be sampled and how much data to include. Once the samples have been taken, the corpus content is fixed.

There are exceptions to this: monitor corpora capture the continuing growth and change of a language. They remain finite, but may extend in size over time.

This finite size rule for corpora contrasts with the study of grammars in theoretical linguistics. These are sets of rules, such as context-free grammars, which generate potentially infinite collections of sentences.
Historically, the word “corpus” referred to a body of printed (or even written) text.

Now, corpora are almost universally machine-readable — that is, stored on and transferred between computers.

Machine-readable corpora have several distinctive features in comparison with books of printed text.

- They can be huge in size, up to billions of words.
- They can be searched and analysed efficiently.
- They can be made available to many users simultaneously, at large distances.
- They can easily (and sometimes automatically) be annotated with additional useful information.
A corpus is often a *standard reference* for the language variety it represents.

Having a corpus as a standard reference allows competing theories about the language variety to be compared against each other on the same sample data.

For this, the corpus has to be widely available to researchers, fitting their shared requirements and used by them in practice.

The likely usefulness of a corpus as a standard reference depends on all the preceding three features: representativeness, fixed finite size and machine readability.
A corpus is — in general — a widely available fixed-sized body of machine-readable text, appropriately sampled to properly represent a certain language variety.

Any particular corpus, however, may not have all of these characteristics.
Some Notable English Language Corpora

- The Brown Corpus of American English was compiled at Brown University and published in 1967. It contains around 1,000,000 words.

- The British National Corpus (BNC), published in the mid-1990’s, is a 100,000,000-word text corpus intended to representative of written and spoken British English from the late 20th century.

- The Corpus of Contemporary American English (COCA) is a 530,000,000-word monitor corpus of texts since 1990 covering five genres: spoken conversation; fiction; popular magazines; newspapers; and academic journals.

- The Oxford English Corpus (OEC) is an English corpus used by the makers of the Oxford English Dictionary. It is the largest text corpus of its kind, containing around 2,500,000,000 words. 100,000,000 of them are “the”
Some Notable English Language Corpora

- The Brown Corpus of American English was compiled at Brown University and published in 1967. It contains around 1 megaword.

- The British National Corpus (BNC), published in the mid-1990’s, is a 100-megaword text corpus intended to representative of written and spoken British English from the late 20th century.

- The Corpus of Contemporary American English (COCA) is a semigigaword monitor corpus of texts since 1990 covering five genres: spoken conversation; fiction; popular magazines; newspapers; and academic journals.

- The Oxford English Corpus (OEC) is an English corpus used by the makers of the Oxford English Dictionary. It is the largest text corpus of its kind, containing around 2.5 gigaword.

  Including 100 mega“the”
**corpus.byu.edu**

**corpora, size, queries = better resources, more insight**

---

Created by Mark Davies, BYU. **Overview**, search types, looking at variation, corpus-based resources, **updates**.

The most widely used online corpora -- more than **130,000** distinct researchers, teachers, and students each month.

<table>
<thead>
<tr>
<th>English</th>
<th># words</th>
<th>language/dialect</th>
<th>time period</th>
<th>compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOW Corpus</td>
<td>3.9 billion+</td>
<td>20 countries / Web</td>
<td>2010-yesterday</td>
<td></td>
</tr>
<tr>
<td>Global Web-Based English (GloWbE)</td>
<td>1.9 billion</td>
<td>20 countries / Web</td>
<td>2012-13</td>
<td></td>
</tr>
<tr>
<td>Wikipedia Corpus</td>
<td>1.9 billion</td>
<td>English</td>
<td>-2014</td>
<td>Info</td>
</tr>
<tr>
<td>Hansard Corpus (British Parliament)</td>
<td>1.6 billion</td>
<td>British</td>
<td>1803-2005</td>
<td>Info</td>
</tr>
<tr>
<td>Corpus of Contemporary American English (COCA)</td>
<td>520 million</td>
<td>American</td>
<td>1990-2015</td>
<td>***</td>
</tr>
<tr>
<td>Corpus of Historical American English (COHA)</td>
<td>400 million</td>
<td>American</td>
<td>1810-2009</td>
<td>**</td>
</tr>
<tr>
<td>Corpus of US Supreme Court Opinions <strong>NEW</strong></td>
<td>130 million</td>
<td>American</td>
<td>1790s-present</td>
<td></td>
</tr>
<tr>
<td>TIME Magazine Corpus</td>
<td>100 million</td>
<td>American</td>
<td>1923-2006</td>
<td></td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td>100 million</td>
<td>American</td>
<td>2001-2012</td>
<td>*</td>
</tr>
<tr>
<td>British National Corpus (BYU-BNC)*</td>
<td>100 million</td>
<td>British</td>
<td>1980s-1993</td>
<td>**</td>
</tr>
<tr>
<td>Strathy Corpus (Canada)</td>
<td>50 million</td>
<td>Canadian</td>
<td>1970s-2000s</td>
<td></td>
</tr>
<tr>
<td>CORE Corpus</td>
<td>50 million</td>
<td>Web registers</td>
<td>-2014</td>
<td></td>
</tr>
</tbody>
</table>
Created by Mark Davies, BYU. **Overview**, search types, looking at variation, corpus-based resources, **updates**.

The most widely used online corpora -- more than 130,000 distinct **researchers**, teachers, and students each month.

<table>
<thead>
<tr>
<th>English</th>
<th># words</th>
<th>language/dialect</th>
<th>time period</th>
<th>compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOW Corpus</td>
<td>3.9 billion+</td>
<td>20 countries / Web</td>
<td>2010-<strong>yesterday</strong></td>
<td></td>
</tr>
<tr>
<td>Global Web-Based English (GloWbE)</td>
<td>1.9 billion</td>
<td>20 countries / Web</td>
<td>2012-13</td>
<td></td>
</tr>
<tr>
<td>Wikipedia Corpus</td>
<td>1.9 billion</td>
<td>English</td>
<td>-2014</td>
<td>Info</td>
</tr>
<tr>
<td>Hansard Corpus (British Parliament)</td>
<td>1.6 billion</td>
<td>British</td>
<td>1803-2005</td>
<td>Info</td>
</tr>
<tr>
<td>Corpus of Contemporary American English (COCA)</td>
<td>520 million</td>
<td>American</td>
<td>1990-2015</td>
<td>** ****</td>
</tr>
<tr>
<td>Corpus of Historical American English (COHA)</td>
<td>400 million</td>
<td>American</td>
<td>1810-2009</td>
<td>**</td>
</tr>
<tr>
<td>Corpus of US Supreme Court Opinions <strong>NEW</strong></td>
<td>130 million</td>
<td>American</td>
<td>1790s-present</td>
<td></td>
</tr>
<tr>
<td>TIME Magazine Corpus</td>
<td>100 million</td>
<td>American</td>
<td>1923-2006</td>
<td></td>
</tr>
<tr>
<td>Corpus of American Soap Operas</td>
<td>100 million</td>
<td>American</td>
<td>2001-2012</td>
<td>*</td>
</tr>
<tr>
<td>British National Corpus (BYU-BNC)*</td>
<td>100 million</td>
<td>British</td>
<td>1980s-1993</td>
<td>**</td>
</tr>
<tr>
<td>Strathy Corpus (Canada)</td>
<td>50 million</td>
<td>Canadian</td>
<td>1970s-2000s</td>
<td></td>
</tr>
<tr>
<td>CORE Corpus</td>
<td>50 million</td>
<td>Web registers</td>
<td>-2014</td>
<td></td>
</tr>
</tbody>
</table>
CORPUS OF AMERICAN SOAP OPERAS

100 MILLION WORDS, 1990-2012

BRIGHAM YOUNG UNIVERSITY
Two Kinds of Corpus

Individual corpora may be *unannotated* — just consisting of bare text — or *annotated*, usually with some linguistic or semantic information.

- Unannotated corpora are examples of *unstructured data*.
- Annotated corpora are examples of *semistructured data*.

The English language corpora listed earlier are all annotated.

From here on we will be looking almost exclusively at annotated corpora.
Building a Corpus

Two tasks are central to building an annotated corpus:

- Collect data — this involves balancing and sampling;
- Add information — perform the annotation.

**Balancing** ensures that the linguistic content of a corpus represents the full variety of the language sources for which the corpus is intended to provide a reference. For example, a balanced text corpus (as opposed to spoken) includes materials from sources such as books, newspapers, magazines, letters, etc.

**Sampling** ensures that the material is representative of the types of source. For example, sampling from newspaper text involves selecting texts randomly from different newspapers, different issues, and different sections of each newspaper. Note the randomness — this is essential.
Balancing

Balancing may operate across several dimensions of source material.

- **Language type**: Taking samples from some or all of:
  - edited text (e.g., articles, books, news wire);
  - spontaneous text (e.g., email, blog comments, letters);
  - spontaneous speech (e.g., conversations, dialogues);
  - scripted speech (e.g., formal speeches).

- **Genre**: Finer-grained resolution of material type (e.g., 18th century novels, scientific articles, movie reviews, parliamentary debates).

- **Domain**: What the material is about (e.g., crime, travel, biology, law).

- **Media**: The physical realization of a corpus (e.g., text, audio, transcribed speech, video).

Planning for a corpus involves fixing on which kinds of balancing are required, and how they will be realised.
Example Balanced Corpora

Brown Corpus

A balanced corpus of written American English:
- One of the earliest machine-readable corpora;
- Developed by Francis and Kučera at Brown in early 1960's;
- 1M words of American English texts printed in 1961;
- Sampled from 15 different genres.

British National Corpus:

A large, balanced corpus of British English:
- One of the main reference corpora for English today;
- 90M words text; 10M words speech;
- Text sampled from newspapers, magazines, books, letters, school and university essays;
- Speech recorded from volunteers balanced by age, region, and social class; also meetings, radio shows, phone-ins, etc.
Comparison of Some Classic Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Genre</th>
<th>Mode</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown Corpus</td>
<td>1M</td>
<td>general</td>
<td>text</td>
<td>American English</td>
</tr>
<tr>
<td>British National Corpus</td>
<td>100M</td>
<td>general</td>
<td>mixed</td>
<td>British English</td>
</tr>
<tr>
<td>Penn Treebank</td>
<td>1M</td>
<td>news</td>
<td>text</td>
<td>American English</td>
</tr>
<tr>
<td>Broadcast News Corpus</td>
<td>300k</td>
<td>news</td>
<td>speech</td>
<td>7 languages</td>
</tr>
<tr>
<td>MapTask Corpus</td>
<td>147k</td>
<td>dialogue</td>
<td>speech</td>
<td>British English</td>
</tr>
<tr>
<td>CallHome Corpus</td>
<td>50k</td>
<td>dialogue</td>
<td>speech</td>
<td>6 languages</td>
</tr>
</tbody>
</table>

These are largely pre-internet — the corpus landscape is changing
Pre-processing and Annotation

Going from raw linguistic data to an annotated corpus can be broken down into stages.

- **Pre-processing** identifies the basic units in the corpus, such as:
  - Tokenization;
  - Sentence boundary detection.

- **Annotation** adds information appropriate to the corpus, such as:
  - Parts of speech;
  - Syntactic structure;
  - Dialogue structure;
  - Prosody (rhythm, intonation and stress in speech).
Tokenization

**Tokenization:** Divide raw textual data into *tokens* such as words, numbers, punctuation marks.

**Word:** A continuous string of *alphanumeric* characters delineated by *whitespace* (space, tab, newline).

Unicode provides a lot of detailed information about individual characters to help classify them and support tokenization.

However, there remain many potentially difficult cases. For example:

- amazon.com, Micro$oft
- John’s, isn’t, rock’n’roll
- A final take-it-or-leave-it offer
- Led down a cul de sac
Sentence boundary detection: Identify the start and end of individual sentences.

Sentence: A string of words ending in a full stop, question mark or exclamation mark.

This is enough much of the time; and again, Unicode can help classify sentence-breaking punctuation elements.

There are still potential problem cases:

- Prof. Higgs and his eponymous boson.
- “Are you going?” John asked.
- Moonpig.com sell greeting cards.

Detecting word and sentence boundaries is particularly difficult for speech.
Corpus Annotation

**Annotation** adds information to the corpus that is not explicit in the data itself. This is often specific to a particular application; and a single corpus may be annotated in multiple ways.

**Annotation scheme** is a basis for annotation, made up of a *tag set* and *annotation guidelines*.

**Tag set** is an inventory of labels for markup.

**Annotation guidelines** tell annotators — domain experts — how a tag set should be applied. In particular, this is to ensure consistency across different annotators.

What if the annotator is a machine?

The next lecture will look at *part-of-speech* or *POS* annotation as an example of this.
What is a Corpus?

A corpus is a body of written or spoken text used for study of a particular language or language variety.

This depends on the following features in a corpus.

- Representative sampling
- Finite size
- Machine-readable form
- Use as a standard reference

All are important for a corpus to be a useful linguistic resource.

*The Google effect.* Web crawling and automated book scanning means that some new “corpora” are fantastically large (approaching a teraword) but don’t sample or balance, and may not be fixed. Comprehensiveness and scale enables some new kinds of research; lack of classic corpus features may limit others.
Homework (1/2)

Do This

Explore the Corpus of Contemporary American English.

- Go to http://corpus.byu.edu/coca/
- Read the text box on the right-hand side, beginning “The Corpus of Contemporary American English (COCA) is the largest freely-available corpus of English”. Then follow the link “large and balanced” and read about how COCA is built up.
- Click on “SEARCH” at the top left, select “Chart” below that, type “data” in the text box, click “See frequency by section”, and wait for a chart that shows how often the word “data” appears in different kinds of writing.
- Click on one of the blue bars to dig into specific occurrences: wait for a table showing uses of the word “data” in context. Pick one row and click on the left-hand index number to see the original source text.

Now try out the other SEARCH options, and read the box on the right about each one.
Read These

- **Drake Baer**
  A Data Scientist Discovered the Most Metal Word in the English Language
  *Science of Us, nymag, July 2016. https://is.gd/mostmetal*

- **Burr Settles**
  On “Geek” versus “Nerd”.
  *Slackpropagation, June 2013. https://is.gd/geeknerd*
Trends in Words Carrying Emotional Content

DOI: 10.1371/journal.pone.0059030

“We report here trends in the usage of “mood” words, that is, words carrying emotional content, in 20th century English language books, using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008...”
FIGURE 1: Historical periods of positive and negative moods. Difference between z-scores of Joy and Sadness for years from 1900 to 2000 (raw data and smoothed trend). Values above zero indicate generally ‘happy’ periods, and values below the zero indicate generally ‘sad’ periods.

DOI: 0.1371/journal.pone.0059030.g001
Text mining uncovers British reserve and US emotion

Writers' expressions of sentiment have grown apart in recent decades.

Philip Ball

21 March 2013

If you associate modern British fiction with the cool, detached tones of Martin Amis and Julian Barnes, and US fiction with Jonathan Franzen's emotional inner worlds or John Irving's sentimentality, it seems you have good reason. An analysis of the digitized texts of English-language books over the past century concludes that, since the 1980s, words that carry emotional content have become significantly more common in US books than in British ones.

The frequency of occurrence of certain words has evolved differently in US and British books...
In my mind, “geek” and “nerd” are related, but capture different dimensions of an intense dedication to a subject:

- **geek** - An enthusiast of a particular topic or field. Geeks are “collection” oriented, gathering facts and mementos related to their subject of interest. They are obsessed with the newest, coolest, trendiest things that their subject has to offer.

- **nerd** - A studious intellectual, although again of a particular topic or field. Nerds are “achievement” oriented, and focus their efforts on acquiring knowledge and skill over trivia and memorabilia.
On “Geek” Versus “Nerd”
Burr Settles

Settles sampled 2.6 megatweet during December 2012, evaluating pointwise mutual information (pmi) all words appearing in tweets with “geek” or “nerd”.

https://is.gd/geeknerd
http://slackprop.wordpress.com/2013/06/03/on-geek-versus-nerd/
Olivia Culpo with the Boston Accompagnietta, March 2012
(Picture: Taichi Fukumura, Wikimedia Commons, CC BY-SA 3.0)