Informatics 1B, 2008

School of Informatics, University of Edinburgh

Data and Analysis

Note 8

Introduction to Corpora

Alex Simpson

Part II — Semistructured Data

XML

Note 6 Semistructured data and XML

Note 7 Querying XML documents with XQuery

Corpora

Note 8 Introduction to corpora

Note 9 Building a corpus

Note 10 Querying a corpus

Natural language as data

Written or spoken natural language has plenty of *internal structure*: it consists of words, has phrase and sentence structure, etc.

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

This is an example of *unstructured data*: the data format itself has no structure imposed on it (other than the sequencing of characters).

Often, however, it is useful to annotate text by marking it up with additional information (e.g. linguistic information, semantic information).

Such marked-up text, is a widespread and very useful form of *semistructured data*.

What is a corpus?

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

It is used in (both computational and theoretical) linguistics as a word to describe *a body of text*, in particular a body of written or spoken text.

In practice, a *corpus* is a body of written or spoken text, from a particular language variety, that meets the following criteria.

1. sampling and representativeness;
2. finite size;
3. machine-readable form;
4. a standard reference.

Sampling and representativeness

In linguistics, corpora provide data for *empirical linguistics*

That is, corpora provide data that is used to investigate the nature of linguisitic practice (i.e., of real-world language usage), for the chosen language variety

For obvious practical reasons, a corpus can only contain a *sample* of instances of language usage (albeit a potentially large sample)

For such a sample to be useful for lingusitic analysis, it must be chosen to be *representative* of the kind of language practice being analysed.

For example, the complete works of Shakespeare would not provide a representative sample for Elizabethan English.

Finiteness

Again, for obvious practical reasons (e.g., so we can store it somewhere), a corpus should be *finite* in size.

Furthermore, corpora almost universally have a *fixed* size. It is decided at the outset how the language variety is to be sampled and how much data to include. An appropriate sample of data is then compiled, and the corpus content is fixed.

N.B. *Monitor corpora* (which are beyond the scope of this course) are an exception to the fixed size rule.

While the finite size rule for a corpus is obvious, it contrasts with theoretical lingustics, where languages are studied using *grammars* (e.g. context-free grammars) that potentially generate infinitely many sentences.

Machine readability

Historically, the word “corpus” was used to refer to a body of printed text.

Nowadays, corpora are almost universally machine (i.e. computer) readable. (Since this is an Informatics course, we are anyway only interested in such corpora.)

Machine-readable corpora have several obvious advantages over other forms:

- They can be huge in size (billions of words)
- They can be efficiently searched
- They can be easily annotated with additional useful information

Standard reference

A corpus is often a standard reference for the language variety it represents.

For this, the corpus has to be widely available to researchers.

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

The usefulness of a corpus as a standard reference depends upon all the preceeding three features of corpora: representativeness, fixed size and machine readability.

Summarizing

In practice, a *corpus* is generally a widely available fixed-sized body of machine-readable text, sampled in order to be maximally representable of the language variety it represents.

Note, however, not every corpus will have all of these characteristics.

Two forms of corpus

There are two forms of corpus: *unannotated*, i.e. consisting of just the raw language data, and *annotated*.

Annotations are extremely useful for many purposes, and are what connect corpora with the semistructured data theme of this part of the course. They will play an important role in future lectures.

However, the remainder of today's lecture applies equally to annotated and unannotated corpora.

Some prominent 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 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 *American National Corpus (ANC)* is an ongoing project to create an electronic text corpus of written and spoken American English since 1990. The aim is to create a 100,000,000-word corpus.

  The first release, made available (to subscribers only) in 2003, contains 11,000,000 words and was provided in XML format.
- 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 over 2,000,000,000 words. It is in XML format.

Applications of corpora

Answering *empirical questions* in linguistics and cognitive science:

- corpora can be analyzed using statistical tools;
- hypotheses about language processing and language acquisition can be tested;
- new facts about language structure can be discovered.

Engineering *natural-language systems* in AI and computer science:

- corpora represent the data that language processing system have to handle;
- algorithms exist to extract regularities from corpus data;
- text-based or speech-based computer applications can learn automatically from corpus data.

Simple questions corpora can answer

Assume a corpus that consists of the Arthur Conan Doyle story *A Case of Identity*.

Simple questions we could ask are:

- Find all lines containing the word “Holmes”.
- Find all lines beginning with the word “Holmes”.
- Find all lines starting with an upper case letter.

Question 1. Find all lines containing the word “Holmes”.

- My dear fellow.” said Sherlock Holmes as we sat on either
- a realistic efect,” remarked Holmes. “This is wanting in the
- said Holmes, taking the paper and glancing his eye down
- “I have seen those symptoms before,” said Holmes, throwing
- merchant-man behind a tiny pilot boat. Sherlock Holmes welcomed
- You’ve heard about me, Mr. Holmes,” she cried, “else how

. . .

Question 2. Find all lines beginning with the word “Holmes”.

- Holmes, when she married again so soon after father’s death,
- Holmes alone, however, half asleep, with his long, thin form
- Holmes. “He has written to me to say that he would be here at
- Holmes had been talking, and he rose from his chair now with a

. . .

Question 3. Find all lines starting with an upper case letter.

- A Case of Identity
- The husband was a teetotaler,
- there was no other woman
- Take a pinch of snuff, Doctor, and acknowledge that I
- The larger crimes are apt to be the simpler, for the
- And yet even here we may discriminate.
- When a woman has a secret
- Etherege, whose husband you found so easy when the

But is the kind of information provided by these three questions really useful?

Frequencies

Frequency information obtained from corpora is often useful for answering scientific or engineering questions.

Token count N: number of tokens (words, punctuation marks, etc.) in a corpus (i.e., size of the corpus).

Type count: number of different tokens in a corpus.

Absolute frequency $f(t)$ *of a type* t: number of tokens of type t in a corpus.

Relative frequency of a type t: absolute frequency of t normalized by the token count, i.e., $f(t)/N$.

Frequencies (example)

The British National Corpus (BNC) is an important reference.

Let's compare some counts from the BNC with counts from our sample corpus *A Case of Identity*

	BNC	A Case of Identity
Token count N	100,000,000	7,006
Type count	636,397	1,621
f(Holmes)	890	46
f(Sherlock)	209	7
f(Holmes)$/N$	.0000089	.0066
f(Sherlock)$/N$	.00000209	.000999

Unigrams

We can now ask questions such as: what are the most frequent words in a corpus?

- Count absolute frequencies of all word types in the corpus;
- tabulate them in an ordered list;
- results: list of *unigram* frequencies (frequencies of individual words).

The next slide compares unigram frequencies for BNC and *A Case of Identity*.

Unigrams (example)

BNC		A Case of Identity	
6184914	the	350	the
3997762	be	212	and
2941372	of	189	to
2125397	a	167	of
1812161	in	163	a
1372253	have	158	I
1088577	it	132	that
917292	to	117	it

N.B. The article “the” is the most frequent word in both corpora; prepositions like “of” and “to” appear in both lists; etc.

n-grams

The notion of unigram can be generalized:

- *bigrams* — adjacent pairs of words
- *trigrams* — triples of words
- n-*grams* — n-tuples of words.

As the value of n increases, the units become more linguistically meaningful.

n-grams (example)

Compute the most frequent n-grams in *A Case of Identity*, for $n = 2, 3, 4$.

bigrams		trigrams		4-grams	
40	of the	5	there was no	2	very morning of the
23	in the	5	Mr. Hosmer Angel	2	use of the money
21	to the	4	to say that	2	the very morning of
21	that I	4	that it was	2	the use of the
20	at the	4	that it is	2	the King of Bohemia

N.B. n-gram frequencies get smaller with increasing n. As more word combinations become possible, there is increased *data sparseness*.

Corpora in Informatics

Corpora are used extensively in two areas of informatics:

- *Natural Language Processing (NLP)* builds computer systems that understand or produce text. Example applications that rely on corpus data include:
  - *Summarization:* take a text and compress it, i.e., produce an abstract or summary. Example: Newsblaster.
  - *Machine Translation (MT):* take a text in a source language and turn it into a text in the target language. Example: Babel Fish.
- *speech processing* develops systems that understand or produce spoken language.

The techniques applied rely on probability theory, information theory and machine learning to extract statistical regularities from corpora.

There is still room for research!

Example translation by AltaVista Babel Fish.

O, my love is like a red, red rose,
That is newly sprung in June.

Robert Burns

English ⟶ Italian:

La O, il mio amore è come un rosso, colore rosso è aumentato,
che recentemente è balzato in giugno.

Italian ⟶ English:

Or, my love is like a red one, red color is increased,
than recently it is jumped in June.

Additional reading

There are no Data & Analysis lectures next week.

Instead there is *required reading*:

Corpus Linguistics
Tony McEnery & Andrew Wilson
Edinburgh University Press,
2nd Edition, 2001

Chapter 2: What is a Corpus and What is in It?
Required reading: from start of chapter to end of Section 2.2.1

Copies of the full chapter will be available from the shelves outside room 5.03 Appleton Tower from 3pm on Monday 18th February.