Abbreviated syntax

The abbreviated syntax is more economical and often (but not always!) more intuitive.

The XPath abbreviations are:

- The syntax `child::` may be omitted from a location step altogether. (The child axis is chosen as default.)
- The syntax `@` is an abbreviation for: `attribute::`
- The syntax `//` is an abbreviation for:

  `/descendant-or-self::node()/`
- The syntax `..` is an abbreviation for: `parent::node()`
- The syntax `.` is an abbreviation for: `self::node()`

Queries and alternatives

Consider again the last query above:

Find the name of every region in which there is a mountain.

An alternative location path for this is:

```
//Region[Feature/@type='Mountain']/Name/text()
```

Similarly, consider:

Find the name of countries containing a feature called Everest.

Two queries for this are:

```
//Feature[text()='Everest']/../../Name/text()
//Country[.//Feature/text()='Everest']/Name/text()
```

One subtle point

A subtle point with XPath is illustrated by the second solution above to:

Find the name of countries containing a feature called Everest.

While the given query (repeated below) is correct,

```
//Country[.//Feature/text()='Everest']/Name/text()
```

the following (natural) attempt would be incorrect:

```
//Country[//Feature/text()='Everest']/Name/text()
```

The problem is that the location path `//Feature/text()` starts with a `/` character, and this means that XPath interprets this path as starting at the root node, whereas the path needs to start at the current node.

The omission of a necessary '`.`' character at the start of a predicate expression is a common source of errors in XPath.

More on XPath

In practice, when using XPath, one often needs to prefix the location path with a pointer to the given XML document; e.g.,

```
doc("gazetter.xml")//Feature[@type='Mountain']/text()
```

Other features in XPath include: navigation based on document order, position and size of context, treatment of namespaces, a rich language of expressions.

For full details on XPath and XQuery see the W3C specification:

`http://www.w3.org/TR/xpath`

A tutorial can be found at:

`http://www.w3schools.com/xpath/`

Part II — Semistructured Data

XML:

II.1 Semistructured data, XPath and XML

II.2 Structuring XML

II.3 Navigating XML using XPath

Corpora:

II.4 Introduction to corpora

II.5 Querying a corpus

Recommended reading

The recommended reading for the material on corpora is:

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

This book is written for a linguistics audience.

Nevertheless, Chapter 2, from the start of chapter to end of §2.2.2, will provide excellent background for the material covered in the lectures.

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 linguistic 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

Furthermore, corpora usually have a *fixed* and *finite* 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* (beyond the scope of this course) are sometimes an exception to the *fixed size* rule: they capture the continuing growth and change of a language.

While the *finite size* rule for a corpus is natural, 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. (In this 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 (and sometimes automatically) 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 finite 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.

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 a 100,000,000-word corpus of written and spoken American English since 1990.

  The ANC currently contains 22,000,000 million words and is published, with annotations, as XML.
- 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.

Two forms of corpus

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

Unannotated corpora are examples of *unstructured data*.

Annotated corpora are examples of *semistructured data*.

The four English language corpora on slide II: 77 are all annotated.

Annotations are extremely useful for many purposes. They will play an important role in future lectures.

Building a corpus

To build a corpus we need to perform two tasks:

- Collect corpus data — this involves *balancing* and *sampling*
- In the case of an annotated corpus, add meta-information —- this is called *annotation*

Balancing ensures that the linguistic content of a corpus represents the full variety of the language sources that the corpus is intended to provide a reference for. For example, a balanced text corpus includes texts from many diffeerent types of source; e.g., books, newspapers, magazines, letters, etc.

Sampling ensures that the material is representative of the types of source. For example, sampling from newspaper text: select texts randomly from different newspapers, different issues, different sections of each newspaper.

Balancing

Things to take into account when balancing:

- *language type*: may wish to include samples from some or all of:
  - edited text (e.g., articles, books, newswire);
  - spontaneous text (e.g., email, blog comments, letters);
  - spontaneous speech (e.g., conversations, dialogs);
  - scripted speech (e.g., formal speeches).
- *genre:* fine-grained type of material (e.g., 18th century novels, scientific articles, movie reviews, parliamentary debates)
- *domain*: what the material is about (e.g., crime, travel, biology, law);

Examples of balanced corpora

Brown Corpus: a balanced corpus of written American English:

- one of the earliest machine-readable corpora;
- developed by Francis and Kucera at Brown in early 1960's;
- 1M words of American English texts printed in 1961;
- sampled from 15 different genres.

British National Corpus: large, balanced corpus of British English.

- one of the main reference corpora for English today;
- 90M words text; 10M words speech;
- text part 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 standard corpora

Corpus	Size	Genre	Modality	Language
Brown Corpus	1M	balanced	text	American English
British National Corpus	100M	balanced	text/speech	British English
Penn Treebank	1M	news	text	American English
Broadcast News Corpus	300k	news	speech	7 languages
MapTask Corpus	147k	dialogue	speech	British English
CallHome Corpus	50k	dialogue	speech	6 languages