Part II — Semistructured Data

XML:

Corpora:

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 linguistic 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 analyzed.

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 linguistics, 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 preceding 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 different 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, dialogues);
  - 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 Kučera 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

Pre-processing and annotation

Raw data from a linguistic source can't be exploited directly. We first have to perform:

- *pre-processing:* identify the basic units in the corpus:
  - tokenization;
  - sentence boundary detection;
- *annotation:* add task-specific information:
  - parts of speech;
  - syntactic structure;
  - dialogue structure, prosody, etc.

Tokenization

Tokenization: divide the raw textual data into tokens (words, numbers, punctuation marks).

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

Example: potentially difficult cases:

- amazon.com, Micro$oft
- John's, isn't, rock'n'roll
- child-as-required-yuppie-possession

  (As in: "The idea of a child-as-required-yuppie-possession must be motivating them.")
- cul de sac

Sentence Boundary Detection

Sentence boundary detection: identify the start and end of sentences.

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

This is correct 90% of the time.

Example: potentially difficult cases:

- Dr. Foster went to Gloucester.
- He said “rubbish!”.
- He lost cash on lastminute.com.

The detection of word and sentence boundaries is particularly difficult for *spoken data*.

Corpus Annotation

Annotation: adds information that is not explicit in the data itself, increases its usefulness (often application-specific).

Annotation scheme: basis for annotation, consists of a tag set and annotation guidelines.

Tag set: is an inventory of labels for markup.

Annotation guidelines: tell annotators (domain experts) how tag set is to be applied; ensure consistency across different annotators.

Part-of-speech (POS) annotation

Part-of-speech (POS) tagging is the most basic kind of linguistic annotation.

Each linguistic token is assigned a code indicating its *part of speech*, i.e., basic grammatical status.

Examples of POS information:

- singular common noun;
- comparative adjective;
- past participle.

POS tagging forms a basic first step in the disambiguation of homographs.

E.g., it distinguishes between the verb “boot” and the noun “boot”.

But it does not distinguish between “boot” meaning “kick” and “boot” as in “boot a computer”, both of which are transitive verbs.

Example POS tag sets

- CLAWS tag set (used for BNC): 62 tags; (Constituent Likelihood Automatic Word-tagging System)
- Brown tag set (used for Brown corpus): 87 tags:
- Penn tag set (used for the Penn Treebank): 45 tags.

Category	Examples	CLAWS	Brown	Penn
Adjective	happy, bad	AJ0	JJ	JJ
Adverb	often, badly	PNI	CD	CD
Determiner	this, each	DT0	DT	DT
Noun	aircraft, data	NN0	NN	NN
Noun singular	woman, book	NN1	NN	NN
Noun plural	women, books	NN2	NN	NN
Noun proper singular	London, Michael	NP0	NP	NNP
Noun proper plural	Australians, Methodists	NP0	NPS	NNPS

POS Tagging

Idea: Automate POS tagging: look up the POS of a word in a dictionary.

Problem: POS ambiguity: words can have several possible POS's; e.g.:

Time flies like an arrow. (1)

time: singular noun or a verb;
flies: plural noun or a verb;
like: singular noun, verb, preposition.

Combinatorial explosion: (1) can be assigned $\mathbf{2 \times 2 \times 3 = 12}$ different POS sequences.

Need more information to resolve such ambiguities.

It might seem that higher-level meaning (semantics) would be needed, but in fact great improvement is possible using the *probabilities* of different POS.

Probabilistic POS tagging

Observation: words can have more than one POS, but one of them is more frequent than the others.

Idea: assign each word its most frequent POS (get frequencies from manually annotated training data). Accuracy: around 90%.

Improvement: use frequencies of POS sequences, and other context clues. Accuracy: 96–98%.

Example output from a POS tagger (not XML format!):

> Our/PRP$ enemies/NNS are/VBP innovative/JJ and/CC resourceful/JJ ,/, and/CC so/RB are/VB we/PRP ./. They/PRP never/RB stop/VB thinking/VBG about/IN new/JJ ways/NNS to/TO harm/VB our/PRP$ country/NN and/CC our/PRP$ people/NN, and/CC neither/DT do/VB we/PRP ./. (George W. Bush)

Use of markup languages

An important general application of markup languages, such as XML, is to separate *data* from *metadata*.

In a corpus, this serves to keep different types of information apart;

- *Data* is just the raw data.

  In a corpus this is the text itself.

- *Metadata* is data about the data.

  In a corpus this is the various annotations.

Nowadays, XML is the most widely used markup language for corpora.

The example on the next slide is taken from the BNC XML Edition, which was released only in 2007.
(The previous BNC World Edition was formatted in SGML.)

Example from the BNC XML Edition

```
<wtext type="FICTION">
 <div level="1">
 <head> <s n="1">
   <w c5="NN1" hw="chapter" pos="SUBST">CHAPTER </w>
   <w c5="CRD" hw="1" pos="ADJ">1</w>
 </s> </head>
 <p> <s n="2">
   <c c5="PUQ"> </c>
   <w c5="CJC" hw="but" pos="CONJ">But</w>
   <c c5="PUN">,</c> <c c5="PUQ">  </c>
   <w c5="VVD" hw="say" pos="VERB">said </w>
   <w c5="NP0" hw="owen" pos="SUBST">Owen</w>
   <c c5="PUN">,</c> <c c5="PUQ"> </c>
   <w c5="AVQ" hw="where" pos="ADV">where </w>
   <w c5="VBZ" hw="be" pos="VERB">is </w>
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="NN1" hw="body" pos="SUBST">body</w>
   <c c5="PUN">?</c> <c c5="PUQ"> </c>
 </s> </p>
 ....
 </div>
</wtext>
```

Aspects of this example

This example is the opening text of BNC text J10, which is the novel *The Mamur Zapt and the girl in the Nile* by Michael Pearce.

Some aspects of the tagging:

- The **`wtext`** element stands for *written text*. The attribute **`type`** indicates the genre.
- Element **`head`** tags a portion of header text (here, a chapter heading).
- The **`s`** element tags sentences (a chapter heading counts as a sentence). Sentences are numbered via the attribute **`n`**.
- The **`w`** element tags words. The attribute **`pos`** is a POS tag, with more detailed POS information given by the **`c5`** attribute, which contains the CLAWS code. The attribute **`hw`** represents the *root form* of the word (e.g., the root form of "said" is "say").
- The **`c`** element tags punctuation.

Syntactic annotation (parsing)

Syntactic annotation: information about the structure of sentences.
Prerequisite for computing meaning.

Linguists use phrase markers to indicates which parts of a sentence belong together:

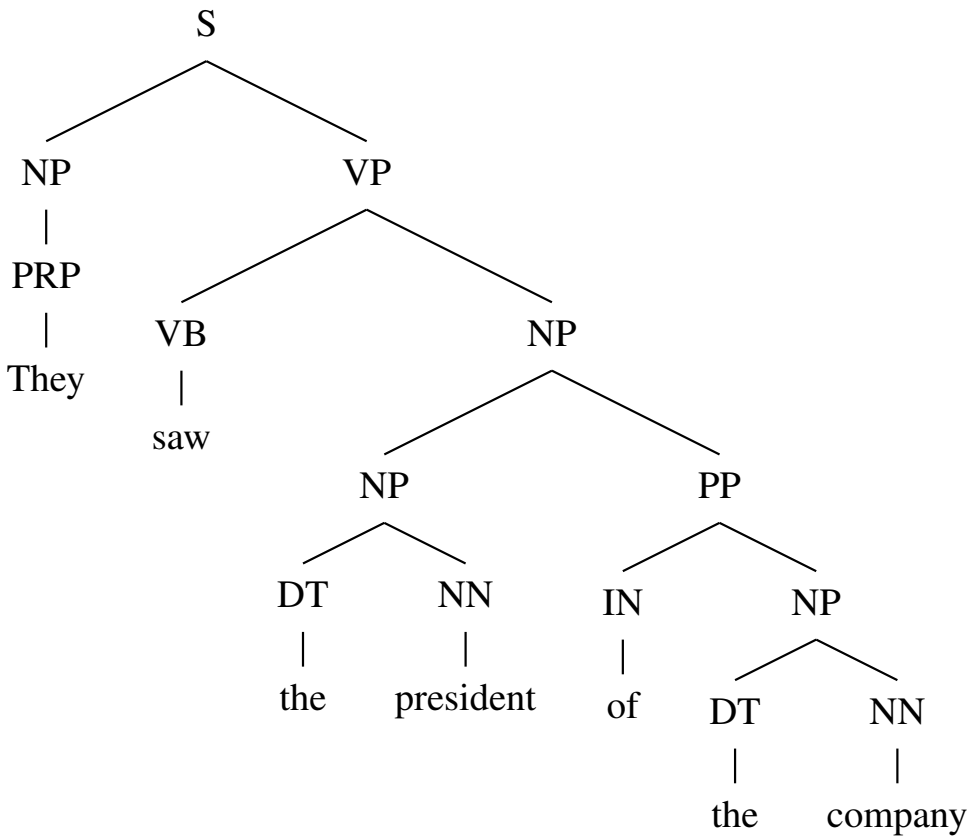
- noun phrase (NP): noun and its adjectives, determiners, etc.
- verb phrase (VP): verb and its objects;
- prepositional phrase (PP): preposition and its NP;
- sentence (S): VP and its subject.

Phrase markers group hierarchically in a *syntax tree*.

Syntactic annotation can be automated. Accuracy: around 90%.

Example syntax tree

Sentence from the Penn Treebank corpus:



The same syntax tree in XML:

```
<s>
 <np><w pos="PRP">They</w></np>
 <vp><w pos="VB">saw</w>
   <np>
     <np><w pos="DT">the</w> <w pos="NN">president</w></np>
     <pp><w pos="NN">of</w>
       <np><w pos="DT">the</w> <w pos="NN">company</w></np>
     </pp>
   </np>
 </vp>
</s>
```

Note the conventions used in the above document: phrase markers are represented as elements; whereas POS tags are given as attribute values.

N.B. The tree on the previous slide is *not* the XML element tree generated by this document.