1. What word is the least metal, according to research on the *Degenerate State* blog?

The *British National Corpus* (BNC) contains hundreds of millions of words from many different sources. Building such a corpus requires balancing and sampling to ensure it is representative.

What is the meaning of the following, as used here?

2. Balancing

3. Sampling

4. Representative
XML

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
Corpus Annotation

The last lecture introduced *preprocessing* steps of identifying tokens and sentence boundaries. Now we look to add further information to the data.

**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** describe how a tag set should be applied.

Originally, annotation was largely done by domain experts, or people trained by domain experts. Now, annotation is largely done by computer.
Part-of-Speech (POS) Annotation

Tagging by *part-of-speech* (POS) is the most basic linguistic annotation. Each token is assigned a code indicating its part of speech. This might be a very simple classification:

- Noun (such as “claw” or “hyphen”);
- Adjective (“red”, “small”);
- Verb (“encourage”, “betray”).

Or it could be more refined:

- Singular common noun (“elephant”, “table”);
- Comparative adjective (“larger”, “neater”);
- Past participle (“listened”, “written”).

Even simple POS tagging can, for example, disambiguate some *homographs* like “wave” (verb) and “wave” (noun).
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

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
<th>CLAWS5</th>
<th>Brown</th>
<th>Penn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>happy, bad</td>
<td>AJ0</td>
<td>JJ</td>
<td>JJ</td>
</tr>
<tr>
<td>Adverb</td>
<td>often, badly</td>
<td>PNI</td>
<td>CD</td>
<td>CD</td>
</tr>
<tr>
<td>Determiner</td>
<td>this, each</td>
<td>DT0</td>
<td>DT</td>
<td>DT</td>
</tr>
<tr>
<td>Noun</td>
<td>aircraft, data</td>
<td>NN0</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>Noun singular</td>
<td>goose, book</td>
<td>NN1</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>Noun plural</td>
<td>geese, books</td>
<td>NN2</td>
<td>NN</td>
<td>NN</td>
</tr>
<tr>
<td>Noun proper singular</td>
<td>London, Michael</td>
<td>NP0</td>
<td>NP</td>
<td>NNP</td>
</tr>
<tr>
<td>Noun proper plural</td>
<td>Greeks, Methodists</td>
<td>NP0</td>
<td>NPS</td>
<td>NNPS</td>
</tr>
</tbody>
</table>
POS Tagging

**Idea:** Tag parts of speech by looking up words in a dictionary.

**Problem:** Ambiguity: words can carry several possible POS.

Time flies like an arrow (1) / Fruit flies like a banana (2)

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

**Combinatorial explosion:** 2 × 2 × 3 = 12 POS sequences for (1).

To resolve this kind of ambiguity, we need more information. One way is to investigate the meaning of words and sentences — their *semantics*.

Perhaps unexpectedly, it turns out that impressive improvements are possible using only the *probabilities* of different parts of speech.
Probabilistic POS Tagging

**Observation:** Words can have more than one POS, but one may be more frequent than the others.

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

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

**Sample POS tagger output**

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness

Charles Dickens, *A Tale of Two Cities*
Probabilistic POS Tagging

**Observation:** Words can have more than one POS, but one may be more frequent than the others.

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

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

**Sample POS tagger output**

```plaintext
It/PP was/VBD the_DT best/JJS of_IN times/NNS ,/it/PP was/VBD the_DT worst/JJS of_IN times/NNS ,/it/PP was/VBD the_DT age/NN of_IN wisdom/NN ,/it/PP was/VBD the_DT age/NN of_IN foolishness/NN ,/it/PP was/VBD the_DT epoch/NN of_IN belief/NN ,/it/PP was/VBD the_DT epoch/NN of_IN incredulity/NN ,/it/PP was/VBD the_DT season/NN of_IN Light/NP ,/it/PP was/VBD the_DT season/NN of_IN Darkness/NN
```
Data and Metadata

One important application of markup languages like XML is to separate data from metadata:

**Data** is the thing itself.
- In a corpus this is the samples of text.

**Metadata** is data about the data.
- In a corpus this includes information about source of text as well as various kinds of annotation.

At present XML is the most widely used markup language for corpora, replacing various others including the *Standard Generalized Markup Language* (SGML).

The example on the next slide is taken from the BNC, which was first released as XML in 2007 (having been previously formatted in SGML).
The Mamur Zapt and the girl in the Nile

Text J10 from the 100,000,000-word British National Corpus is a detective novel.

It starts like this:

CHAPTER 1

‘But,’ said Owen, ‘where is the body?’
<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="PUN">’</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 BNC Example

- The \texttt{wtext} element stands for \textit{written text}. Its attribute \texttt{type} indicates the kind of text (here \texttt{FICTION}).

- Element \texttt{head} tags a portion of header text (here, a chapter heading).

- The \texttt{s} element tags sentences. Sentences are numbered via the attribute \texttt{n}.

- The \texttt{w} element tags words. The attribute \texttt{pos} is a basic part-of-speech tag, with more detailed information given by the \texttt{c5} attribute containing the CLAWS code.

- The attribute \texttt{hw} represents the \textit{head word}, also known as the \textit{lemma} or \textit{root form} of the word. For example, the root of “said” is “say”.

- The \texttt{c} element tags punctuation.
Syntactic Annotation

Moving above the level of individual words, *parsing* and *syntactic annotation* give information about the structure of sentences.

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

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

Phrase markers group hierarchically into a *syntax tree*.

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

Try out the Stanford Parser online at [http://nlp.stanford.edu:8080/parser](http://nlp.stanford.edu:8080/parser)
Example Syntax Tree

The following is from the Penn Treebank corpus.
Syntax Tree in XML

Here is the same syntax tree expressed in XML:

```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>
```

Some choices made in this XML coding: phrase markers are represented by XML elements; while POS tags are given by attribute values.

Note that, as a result of this, the tree on the previous slide is *not* quite the same as the XML element tree for this document.
Syntax Tree in XML

Here is the same syntax tree expressed in XML:

Some choices made in this XML coding: phrase markers are represented by XML elements; while POS tags are given by attribute values.

Note that, as a result of this, the tree on the previous slide is not quite the same as the XML element tree for this document.
## Exam Date and Time

**Informatics 1: Data & Analysis Main Diet Examination**

- **Date**: Tuesday 16 May 2017
- **Time**: 0930–1130
- **Place**: The Pleasance Sports Hall

---

http://www.ed.ac.uk/student-administration/exams
Ricardo Almeida, Inf1-DA Teaching Assistant, has analysed the complete set of last year’s tutorial attendance records and the final exam results.

- Students who attended every tutorial all passed at the first attempt.

- Across all students, each extra tutorial attended correlates with an additional 8.7 marks in the final exam result. (Significance > 99.99%)

Of course, students who attended tutorials also did the exercises, followed the lectures, did the reading, and everything else involved in course participation. So it isn’t just being at tutorials. But it all helps.
Unicode Characters and UTF-8
1992-09-02 New Jersey drive

1

\[ \text{ASCII} \]

2

\[ 110 - 5 - 11 \text{ bits} \]

3

\[ 1110 - 4 - 10 - 6 - 101 - 6 - 16 \text{ bits} \]

4

\[ 11110 - 3 - 10 - 6 - 101 - 6 - 101 - 6 - 21 \text{ bits} \]

5

\[ 111110 - 2 - 1 - 1 - 1 - 26 \text{ bits} \]

6

\[ 1 - \ldots - \text{FG FF} \]
+ ASCII
+ Stateful
+ Self-Synchronizing
+ Robust: duplication/deletion
+ Searchable
+ Backwards traversal

- variable length
- finding Nth char
- ambiguity
Character Encodings on the Web

Share of web pages with different encodings

Google measurements

Graphic: Chris55, Wikimedia Commons, CC BY-SA 3.0
Applications of Corpora

Answering empirical questions in linguistics and cognitive science:

- Corpora can be analyzed using statistical tools;
- Hypotheses about language processing and 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 these systems have to handle;
- Algorithms can find and extract regularities from corpus data;
- Text-based or speech-based computer applications can learn automatically from corpus data.
Once we have an annotated corpus, we can begin to use it to find out information and answer questions. For now, we start with the following:

- The basic notion of a *concordance* in a text.
- Statistics of word *frequency* and *relative frequency*, useful for linguistic questions and natural language processing.
- Word groups: *Unigrams, bigrams* and *n-grams*.

The next lecture will look at more substantial examples of detecting *collocations* and the *machine translation* of natural language.
Concordances

**Concordance:** all occurrences of a given word, shown in context.

More generally, a concordance may extend to all matches for some query expression.

- Specialist concordance programs will generate these from a given *keyword*.
- This might can specify word, annotation (POS, etc.) or more complex information (e.g., using regular expressions).
- Results are typically displayed as *keyword in context* (kwic): a matched keyword in the middle of a line with a fixed amount of context to left and right.
This is a concordance for all forms of the word “remember” in the works of Dickens, generated by the *Corpus Query Processor cqp*.

’s cellar. Scrooge then *remembered* to have heard that ghost, for your own sake, you *remember* what has passed between e-quarters more, when he *remembered*, on a sudden, that the corroborated everything, *remembered* everything, enjoyed eve urned from them, that he *remembered* the Ghost, and became c ht be pleasant to them to *remember* upon Christmas Day, who *remembered* the prediction of old Ja its festivities; and had *remembered* those he cared for at a wn that they delighted to *remember* him. It was a great sur ke ceased to vibrate, he *remembered* the prediction of old Ja as present myself, and I *remember* to have felt quite uncom ...
Frequencies

Frequency information obtained from corpora can be used to investigate characteristics of the language represented.

- **Token count $N$**: the number of tokens (words, punctuation marks, etc.) in a corpus; i.e., the size of the corpus.
- **Type count**: the number of types of token in a corpus.
- **Absolute frequency $f(t)$ of type $t$**: the number of tokens of type $t$ in a corpus.
- **Relative frequency of type $t$**: the absolute frequency of $t$ scaled by the token count, i.e., $f(t)/N$.

Here “tokens of type $t$” might mean a single word, or all its variants, or a particular part of speech.
Frequency Example

Here is a comparison of frequency information between two sources: the BNC and the Sherlock Holmes story *A Case of Identity* by Sir Arthur Conan Doyle.

<table>
<thead>
<tr>
<th></th>
<th>BNC</th>
<th>A Case of Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token count N</td>
<td>100,000,000</td>
<td>7,006</td>
</tr>
<tr>
<td>Type count</td>
<td>636,397</td>
<td>1,621</td>
</tr>
<tr>
<td>f(“Holmes”)</td>
<td>890</td>
<td>46</td>
</tr>
<tr>
<td>f(“Sherlock”)</td>
<td>209</td>
<td>7</td>
</tr>
<tr>
<td>f(“Holmes”)/N</td>
<td>0.00000089</td>
<td>0.0066</td>
</tr>
<tr>
<td>f(“Sherlock”)/N</td>
<td>0.00000209</td>
<td>0.000999</td>
</tr>
</tbody>
</table>
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.
- Result: list of *unigram* frequencies — frequencies of individual words.
The unigram rankings are different, but we can see similarities. For example, the definite article “the” is the most frequent word in both corpora; and prepositions like “of” and “to” appear in both lists.

<table>
<thead>
<tr>
<th>BNC</th>
<th>A Case of Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,184,914</td>
<td>the</td>
</tr>
<tr>
<td>3,997,762</td>
<td>be</td>
</tr>
<tr>
<td>2,941,372</td>
<td>of</td>
</tr>
<tr>
<td>2,125,397</td>
<td>a</td>
</tr>
<tr>
<td>1,812,161</td>
<td>in</td>
</tr>
<tr>
<td>1,372,253</td>
<td>have</td>
</tr>
<tr>
<td>1,088,577</td>
<td>it</td>
</tr>
<tr>
<td>917,292</td>
<td>to</td>
</tr>
<tr>
<td>350</td>
<td>the</td>
</tr>
<tr>
<td>212</td>
<td>and</td>
</tr>
<tr>
<td>189</td>
<td>to</td>
</tr>
<tr>
<td>167</td>
<td>of</td>
</tr>
<tr>
<td>163</td>
<td>a</td>
</tr>
<tr>
<td>158</td>
<td>I</td>
</tr>
<tr>
<td>132</td>
<td>that</td>
</tr>
<tr>
<td>117</td>
<td>it</td>
</tr>
</tbody>
</table>
The notion of unigram generalises:

- **Bigrams** — pairs of adjacent words;
- **Trigrams** — triples of adjacent words;
- **n-grams** — n-tuples of adjacent words.

These larger clusters of words carry more linguistic significance than individual words; and, again, we can make use of these even before finding out anything about their semantic content.
### n-grams example

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

<table>
<thead>
<tr>
<th>bigrams</th>
<th>trigrams</th>
<th>4-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 of the</td>
<td>5 there was no</td>
<td>2 very morning of the</td>
</tr>
<tr>
<td>23 in the</td>
<td>5 Mr. Hosmer Angel</td>
<td>2 use of the money</td>
</tr>
<tr>
<td>21 to the</td>
<td>4 to say that</td>
<td>2 the very morning of</td>
</tr>
<tr>
<td>21 that I</td>
<td>4 that it was</td>
<td>2 the use of the</td>
</tr>
<tr>
<td>20 at the</td>
<td>4 that it is</td>
<td>2 the King of Bohemia</td>
</tr>
</tbody>
</table>

Note that frequencies of even the most common $n$-grams naturally get smaller with increasing $n$. As more word combinations become possible, there is an increase in *data sparseness*.
Here is a concordance for all occurrences of bigrams in the Dickens corpus in which the second word is “tea” and the first is an adjective.

This query use the POS tagging of the corpus to search for adjectives.

[pos="J.*"] [word="tea"]

87773: now, notwithstanding the <hot tea> they had given me before
281162: ’ ’ Shall I put a little <more tea> in the pot afore I go ,
565002: o moisten a box-full with <cold tea> , stir it up on a piece
607297: tween eating , drinking , <hot tea> , devilled grill , muffi
663703: e , handed round a little <stronger tea> . The harp was there ;
692255: e so repentant over their <early tea> , at home , that by eigh
1141472: rs. Sparsit took a little <more tea> ; and , as she bent her
1322382: s illness ! Dry toast and <warm tea> offered him every night
1456507: of robing , after which , <strong tea> and brandy were administ
1732571: rsty . You may give him a <little tea> , ma’am , and some dry t
**Annotation:** tag sets, annotation guidelines. Metadata.

**POS (part-of-speech) tagging:** automatic by dictionary lookup, most frequently seen, and POS sequences.

**Syntactic annotation:** phrase markers, syntax trees, also automatic.

**Concordance:** keyword-in-context (kwic)

**Frequencies:** token count $N$, type count, absolute frequency $f(t)$, relative frequency $f(t)/N$.

**Unigrams; bigrams; n-grams:** one-word, two-word, and $n$-word sequences within a corpus.
Homework

None today. But, arriving this week:

Tutorial 6: Corpus Querying

Analyse the complete works of Charles Dickens with the Corpus Query Processor (CQP) tool.

Coursework Assignment

Past exam paper: out this week, due in two on Thursday 23 March.

Your tutor marks and returns this in Week 11 with feedback.

These marks are not directly part of your final grade for Inf1-DA — this formative assessment is entirely for your feedback and learning. Because of this you can freely share help on the questions, discuss on Piazza, and talk about your work with other students. Please do.