1. Another big problem with simple PCFGs

For a given structural ambiguity, say PP attachment

- A simple PCFG always makes the same choice

Consider the two alternative parses we would get from the Treebank grammar for "Mr Vinken is chairman of Elsevier":

2. PCFG example, cont'd

How did we get those two analyses?

- Here's a top-down parse trace using Penn Treebank rules:
So the only difference is the probabilities of the highlighted rules

• p4 and p5 versus p6 and p7

And those will always give the same answer

• Regardless of what the individual words and their probabilities are
• Because whatever the words
• p1, p2 and p3 are the same for both parses
• Compare, for instance, "chairman at the moment" with "chairman for life"

We really need to pay attention to the probability of those words being tightly or loosely connected

3. Paying attention to words

Improving our approach to probabilistic grammar requires paying more attention to individual words

• Let's explore how we might do that a bit more

Back to bigrams, but in a little more detail

• It turns out that not all bigrams are created equal

Here are the top 10 bigrams from *Moby Dick*:

- of the (1879)
- in the (1182)
- to the (731)
- from the (440)
- the whale (407)
- of his (373)
- and the (370)
Aside from the whale, these are all made up of very high-frequency closed-class items.

The highest bigram of open-class items doesn't come until position 27: sperm whale, with frequency 182.

None-the-less if feels as if there's something particularly interesting about that one.

4. Collocations

"You shall know a word by the company it keeps" (J. R. Firth)

One of the things we evidently know about our language is what words go with what:

- 'strong tea', not 'powerful tea'
- 'powerful enemies', not 'strong enemies'
- 'tall flagpole', not 'long flagpole'
- 'long railing', not 'tall railing'

Choosing the right word from among a set of synonyms is a common problem for second-language learners:

- 'un froid vive' is not (per Microsoft translate) 'a bright cold', but rather 'a piercing cold'
- Likewise 'универсальный магазин' is not 'universal store' but 'department store'

Or consider the old (linguists') joke:

- If a maternity dress is what a woman wears when she is pregnant, what is a paternity suit?

5. What measure to find collocations?

The name for an 'interesting' pair is collocation

How can we separate the interesting pairs from the dull ones?

We could try just throwing out the 'little words'

- Typically called 'stop words', see e.g. corpus.stopwords.word('english') in NLTK
- But that still isn't quite getting at what we want:
  sperm whale (182)
  white whale (106)
  moby dick (84)
  old man (81)
  captain ahab (62)
  right whale (57)
  mast head (49)
  mast heads (37)
  ye see (35)
  whale ship (34)

Some of these feel special ('right whale', 'moby dick'), but others ('old man' in particular) just seem ordinary.
• They are 'encoding transparent': we would expect their meaning to be formed from those words
• They're not 'terms' in the sense of 'terminology'

6. Normalising by expection: mutual information

What we want is some way of factoring in frequency more generally

• To take care of more than just some pre-determined list of stop words
• and to downgrade 'old man' and similar cases

Conditional and joint probability are the answer

• Remember that if two events are independent, the conditional probability is the same as the unconditional
  ◦ $P(Y|X)$ is just $P(Y)$
• so the joint probability is just the product of the two event probabilities
  ◦ $P(X,Y)$ is defined as $P(X)P(Y|X)$
  ◦ So if $X$ and $Y$ are independent, that's $P(X)P(Y)$

One way of getting at our intuition might be to say we're looking for cases where the two probabilities are not independent

Now the bigram frequency gives us an MLE of the joint probability directly

So the ratio of that probability, to what it would be if they were independent, would be illuminating:

<table>
<thead>
<tr>
<th>Mutual information</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_2 \left( \frac{P(X, Y)}{P(X)P(Y)} \right)$</td>
</tr>
</tbody>
</table>

7. Mutual information example

Let's compare the most frequent bigram ('of the') with the first interesting one we saw, 'sperm whale'
Simply put, the mutual information between 'sperm' and 'whale' is 5 binary orders of magnitude greater than that between 'of' and 'the'.

Why are we using log base 2?

- Because we traditionally measure information in bits

### 8. Collocations and machine translation

We can use the same approach to build a translation lexicon

Instead of bigrams within a single text

- Co-occurrence in aligned units of a bilingual text
- We treat sentences as sets of words
- And construct a distribution of pairs of words, one from each set in each sentence
- Here are the top 30 pairs from a corpus of 4600 'sentence' pairs from the European Parliament debates, French and English versions:
Pretty good

And would be better if we had done monolingual collocation detection first!

9. Words, heads and grammar

We mentioned the idea of the head of a constituent in earlier lectures

- Maybe organising our grammar in a completely different way would help with the attachment problem
- Because something like mutual information between heads might be just what we need

Approaches to grammar which focus on heads are called dependency grammars

The standard form of diagram shows where this name comes from:

Mr Vinken is chairman of Elsevier

(Green for the preferred attachment, red for the less likely one)
10. Dependency grammar

Dependency grammars don't have rules in the way that a CFG does

- But they do have categories
- And category-specific ways of identifying heads

A given approach to dependency grammar will also involve an inventory of relations

- Not just 'depends'
- But e.g. subj, obj, nmod, pmod:

![Dependency graph example](image)

Dependency graphs are not required to avoid crossings

- So they can directly represent long-distance dependencies
- And work well for case-marked, free-constituent-order languages

11. Lexicalised PCFG

It's possible to add some of the benefits of dependency grammar to PCFGs

- By decorating categories in the tree with their head

![Lexicalised PCFG example](image)

We can't do statistics directly on these augmented categories

- The space is too large to give anything other than very small numbers

But a range of techniques have been developed in the last few years to work around this

- Sharon will come back to this in a few weeks