Ambiguity and the Lexicon in Natural Language

Informatics 2A: Lecture 14

Mirella Lapata

School of Informatics
University of Edinburgh

20 October 2010

Structural ambiguity: example

$$
\begin{aligned}
NP &\rightarrow NP\ VBG \\
NP &\rightarrow N\ PP \\
NP &\rightarrow N \\
PP &\rightarrow about\ NP \\
N &\rightarrow complaints \mid referees \\
VBG &\rightarrow multiplying
\end{aligned}
$$

Structural ambiguity: example

$$
\begin{aligned}
NP &\rightarrow NP\ VBG \\
NP &\rightarrow N\ PP \\
NP &\rightarrow N \\
PP &\rightarrow about\ NP \\
N &\rightarrow complaints \mid referees \\
VBG &\rightarrow multiplying
\end{aligned}
$$

Complaints about referees multiplying

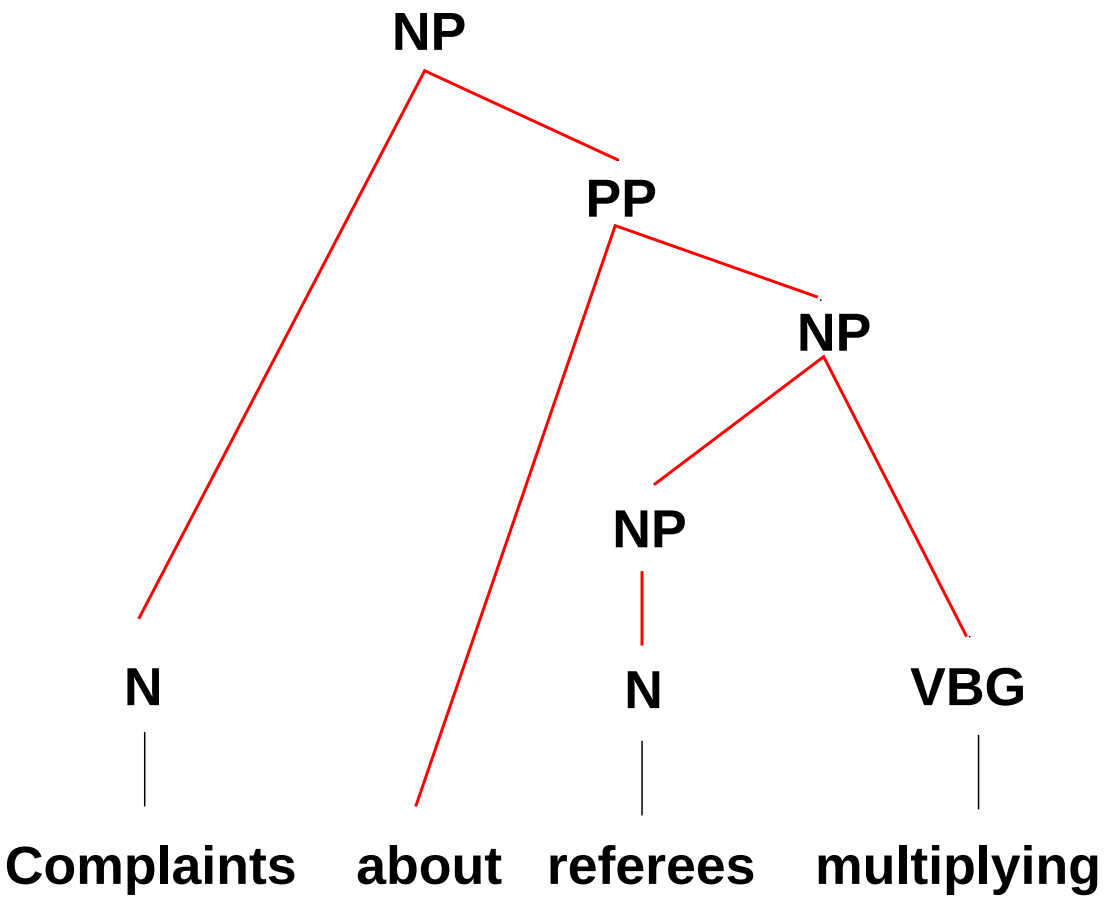
Structural ambiguity: example

$$
\begin{aligned}
NP &\rightarrow NP\ VBG \\
NP &\rightarrow N\ PP \\
NP &\rightarrow N \\
PP &\rightarrow about\ NP \\
N &\rightarrow complaints \mid referees \\
VBG &\rightarrow multiplying
\end{aligned}
$$

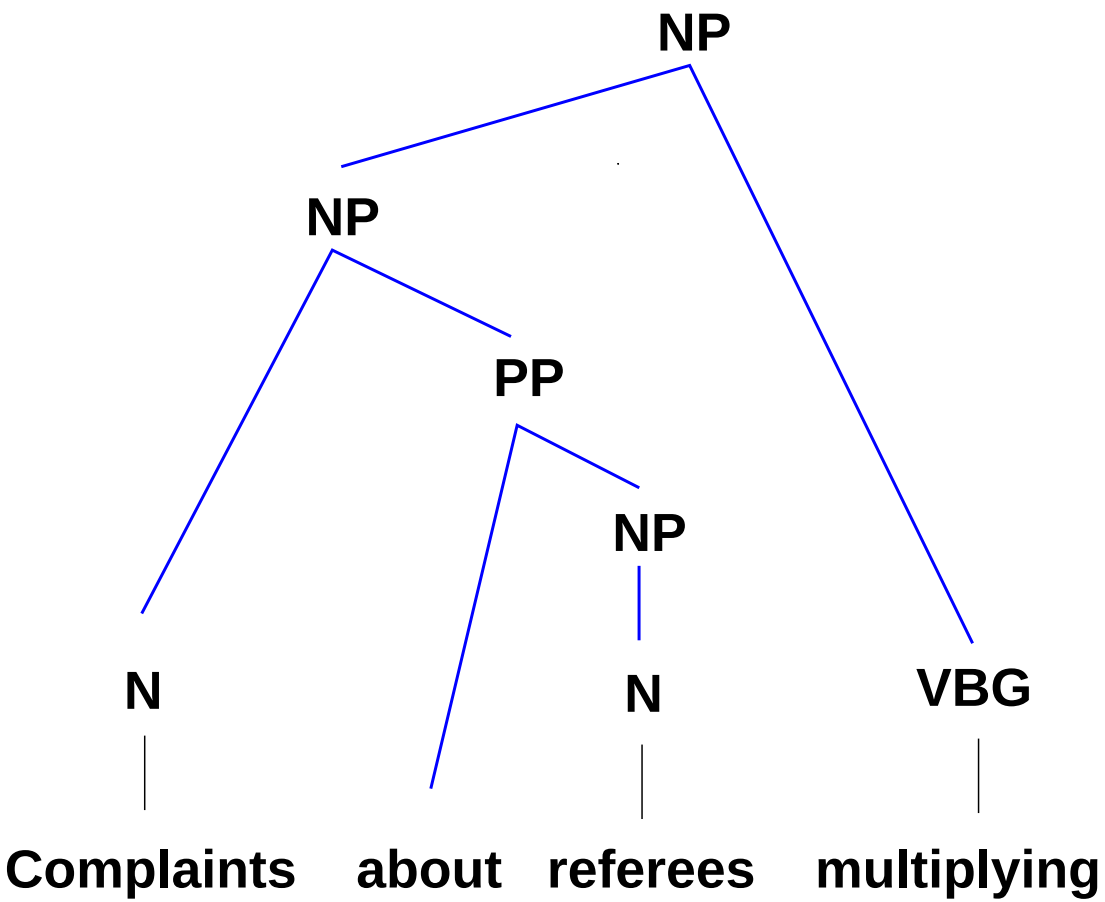
Complaints about referees multiplying

How many non-equivalent sets of derivations (i.e., different trees) are there for this string?

Headline announcing new complaints



Headline announcing new trend in complaints



Derivations and structural ambiguity

- Given a grammar, those strings that can be associated with more than one tree (i.e., non-equivalent derivations) are called structurally ambiguous.
- Of course, an agent who produces a structurally ambiguous string usually only has one meaning in mind, so only one of the structures corresponds to what s/he intended.

Example: Newspaper Headlines

stolen painting found by tree
lung cancer in women mushrooms
dealers will hear car talk at noon
juvenile court to try shooting defendant

Avoiding Ambiguity

The designers of formal languages (e.g., XML) or programming languages try to eliminate or reduce structural ambiguity.

For example, Python uses indentation to indicate **embedding** and no indentation to indicate **sequence**.

```
if a<b:
    c = 0
    a = a+1
```

vs.

```
if a<b:
    c = 0
a = a+1
```

Avoiding Ambiguity

- When we talk, we can use speech rate, pauses and emphasis to indicate what we intend.
- Also, one reading usually makes more sense in the circumstances than other readings do.
- These are both reasons why we don't normally notice that what we read, hear and/or say can have multiple analyses (and multiple meanings!).

Example

lung cancer in WOMEN | mushrooms
dealers will hear CAR_TALK at noon
the students are enjoying the lecture

Handling Ambiguity

- Given a string from a language, the role of a parser is to deliver either all its possible structures or its most likely structure.
- Later on, we'll look at various techniques that parsers use to do this efficiently.
- But structural ambiguity is not the only form of ambiguity in language.
- Natural Languages can also have part-of-speech ambiguity – ambiguity as to what class(es) (aka "parts of speech") a word belongs to.

Open and Closed Classes in Natural Languages

- NL grammars are largely specified in terms of the classes that words belong to.
- Several broad word classes are found in all Indo-European languages and many others: nouns, verbs, adjectives, adverbs.
- These are examples of open classes. They are typically large, and are often stable under translation.
- Other word classes are more specific to particular languages: prepositions (English, German), post-positions (Hungarian, Urdu, Korean), particles (Japanese), classifiers (Chinese), etc.
- These are examples of closed classes. They are typically small and often have structuring uses in grammar. Little correlation between languages.

Parts of Speech

How do we tell the part of speech of a word?

At least three different criteria can be used:

- Notional (semantic) criteria: What does the word refer to?
- Formal (morphological) criteria: What does the word look like?
- Distributional (syntactic) criteria: Where is the word found?

We will look at different parts of speech (POS) using these criteria.

Nouns

Notionally, nouns generally refer to living things (*mouse*), places (*Scotland*), things (*harpoon*), or concepts (*marriage*).

Formally, *-ness*, *-tion*, *-ity*, and *-ance* tend to indicate nouns. (*happiness, exertion, levity, significance*).

Distributionally, we can examine the contexts where a noun appears and other words that appear in the same contexts.

```
>>> from nltk.book import *
>>> text2.concordance('happiness')
```

hat sanguine expectation of **happiness** *which is happiness itself*
to inform her confidante , of her **happiness** *whenever she received a letter*
early in life to despair of such a **happiness** *. Why should you be less fortunate*
and it would give me such **happiness** *, yes , almost the greatest*

Nouns

Notionally, nouns generally refer to living things (*mouse*), places (*Scotland*), things (*harpoon*), or concepts (*marriage*).

Formally, *-ness*, *-tion*, *-ity*, and *-ance* tend to indicate nouns. (*happiness, exertion, levity, significance*).

Distributionally, we can examine the contexts where a noun appears and at other words that appear in the same contexts.

```
>>> from nltk.book import *
>>> text2.similar(happiness') #What else appears in such contexts?
```

heart, mind, time, behaviour, **kindness**, *feelings, attachment, fancy, spirits,* **joy**, *attention, it, mother,* **pleasure**, *name, eyes, and, disappointment, sake, interest*

Verbs

Notionally, verbs refer to actions (*observe*, *think*, *give*).

Formally, words that end in *-ate* or *-ize* tend to be verbs, and ones that end in *-ing* are often the present participle of a verb (*automate, calibrate, equalize, modernize; rising, washing, grooming*).

Distributionally, we can examine the contexts where a verb appears and at other words that appear in the same contexts, which may include their arguments.

```
>>> from nltk.book import *
>>> text2.concordance(marry') # Where 'marry' appears in S&S
>>> text2.similar(marry') # What else appears in such contexts?
```

Adjectives

Notionally, adjectives convey properties of or opinions about things that are nouns (*small*, *wee*, *sensible*, *excellent*).

Formally, words that end in *-al*, *-ble*, and *-ous* tend to be adjectives (*formal, gradual, sensible, salubrious, parlous*)

Distributionally, adjectives usually appear before a noun or after a form of *be*.

```
>>> from nltk.book import *
>>> text2.concordance('sensible') # Where sensible' appears in S&S
>>> text2.similar('sensible') # What else appears in such contexts?
```

Adverbs

Notionally, adverbs convey properties of or opinions about actions or events (*quickly*, *often*, *possibly*, *unfortunately*) or adjectives (*really*).

Formally, words that end in *-ly* tend to be adverbs.

Distributionally, adverbs can appear next to a verb, or an adjective, or at the start of a sentence.

```
>>> from nltk.book import *
>>> text2.concordance(highly') # Where 'highly' appears in S&S
>>> text2.similar(highly') # What else appears in such contexts?
```

The importance of formal and distributional criteria

Often in reading, we come across words unknown words.

bootloader, distros, whitelist, diskdrak, borked (`http://www.linux.com/feature/150441`) revved, femtosecond, dogfooding (`http://hardware.slashdot.org/`)

Even if we don't know its meaning, formal and distributional criteria help people (and machines) recognize what class an unknown word belongs to and what the sentence would mean, if we knew what the word meant.

I really wish mandriva would redesign the *diskdrak* UI. The "orphan" bit is *borked*.

Other Word Classes

Other word classes vary from language to language. English has

- determiners: *the*, *any*, *a*, ...
- prepositions: *in*, *of*, *with*, *without*, ...
- conjunctions: *and*, *because*, *after*, ...
- auxiliaries: *have*, *do*, *be*
- modals: *will*, *may*, *can*, *need*, *ought*
- pronouns: *I*, *she*, *they*, *which*, *where*, *myself*, *themselves*

English doesn't have clitics (like French *l'*) or particles (like Japanese *ga*). Russian lacks stand-alone reflexive pronouns.

N.B. Functions performed by words in one language may be performed by morphology in another one (e.g., reflexivity in Russian).

Types of Lexical Ambiguity

Part of Speech (PoS) Ambiguity: e.g., *still:*

1. *adverb*: at present, as yet
2. *noun*: (1) silence; (2) individual frame from a film; (3) vessel for distilling alcohol
3. *adjective*: motionless, quiet
4. *transitive verb*: to calm

Sense Ambiguity: e.g., *intelligence:*

1. Power of understanding
2. Obtaining or dispersing secret information; also the persons engaged in obtaining or dispersing secret information

Clicker Question

Do not **tweet** me Katy Perry lyrics. Do not **tweet** me anything Katy Perry related, unless it is something negative about her.

What is the part-of-speech of the word **tweet**?

1. adverb
2. noun
3. verb
4. adjective

Word Frequency – Properties of Words in Use

Take any corpus of English like the Brown Corpus or Tom Sawyer and sort its words by how often they occur.

word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440

Word Frequency – Properties of Words in Use

Take any corpus of English like the Brown Corpus or Tom Sawyer and sort its words by how often they occur.

word	Freq. (f)	Rank (r)	$f \cdot r$
two	104	100	10400
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000

Zipf's law

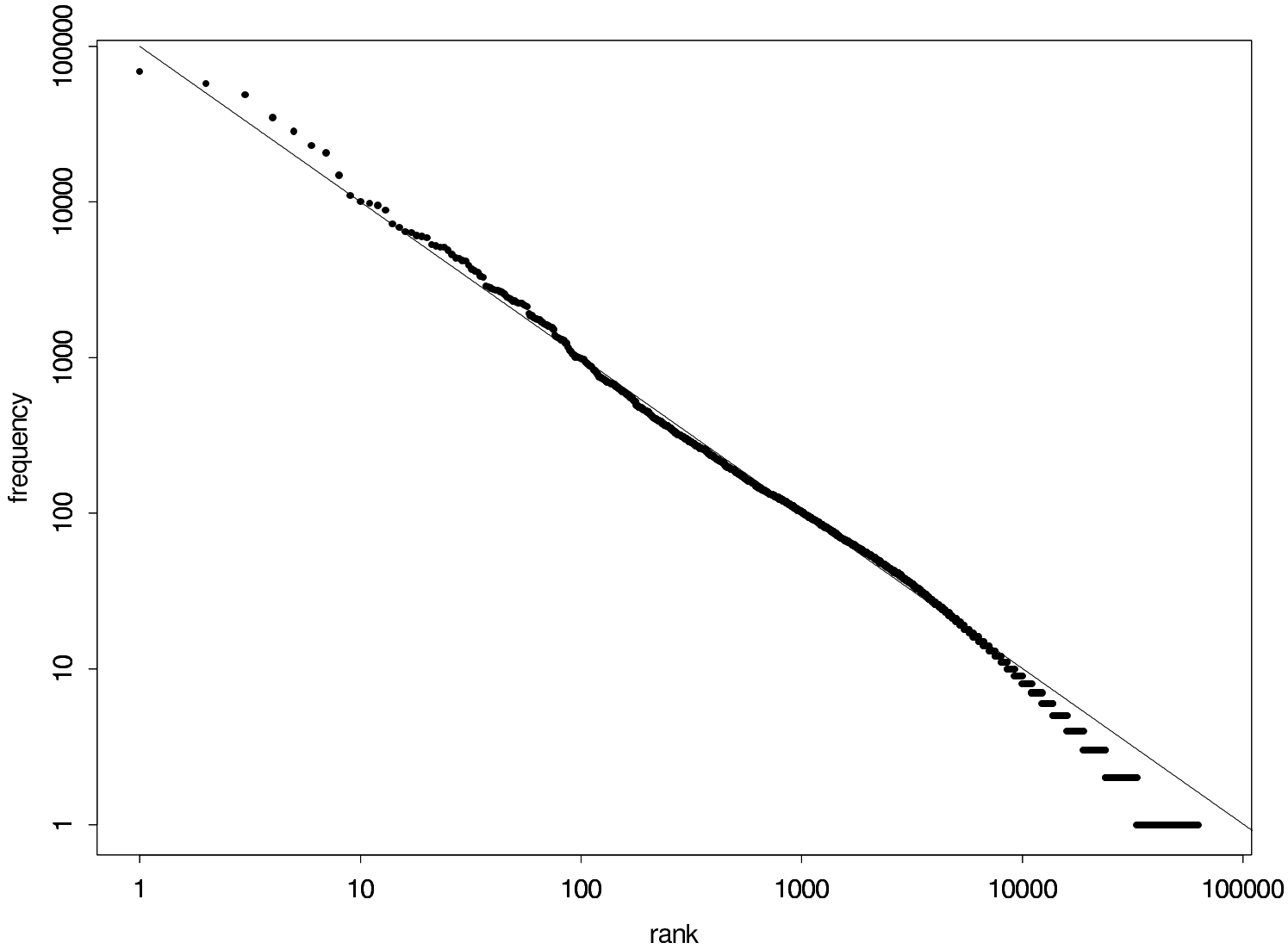
Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table (observation made by Harvard linguist George Kingsley Zipf).

Zipf's law states that: $f \propto \frac{1}{r}$

There is a constant k such that: $f \cdot r = k$

- Now frequently invoked for the web too!
  (See `http://www.nslij-genetics.org/wli/zipf/`)
- Income distribution amongst individuals
- Size of earthquakes

Zipf's law for the Brown corpus



Zipf's law

- There is a very small number of very common words
- There is a small-medium number of middle frequency words
- There is a very large number of words that are infrequent
- Mandelbrot refined Zipf's law.

$$f = P(r + p)^{-B} \quad \text{or} \quad \log f = \log P - B \log(r + p)$$

- Better fit at low and high ranks
- P, B, p are parametrised for particular corpora
- What happens when $B = 1$ and $p = 0$?

Summary

- Structural ambiguity occurs when a string can be associated with more than one structure (represented as trees).
- Words in a language fall into different classes (e.g., nouns, verbs, ajectives).
- To identify the class or part-of-speech (PoS) of a word, we can use notional, distributional, and/or formal criteria.
- Lexical ambiguity occurs when a word belongs to more than one part-of-speech class or has more than one sense.
- Words are found in a Zipfian distribution.

Reading: J&M (2nd edition) Chapter 5
NLTK Book: Chapter 3, Processing Raw Text

Next lecture: Part-of-speech tagging