Parts-of-speech and the Lexicon in Natural Language

Informatics 2A: Lecture 16

Shay Cohen

School of Informatics
University of Edinburgh

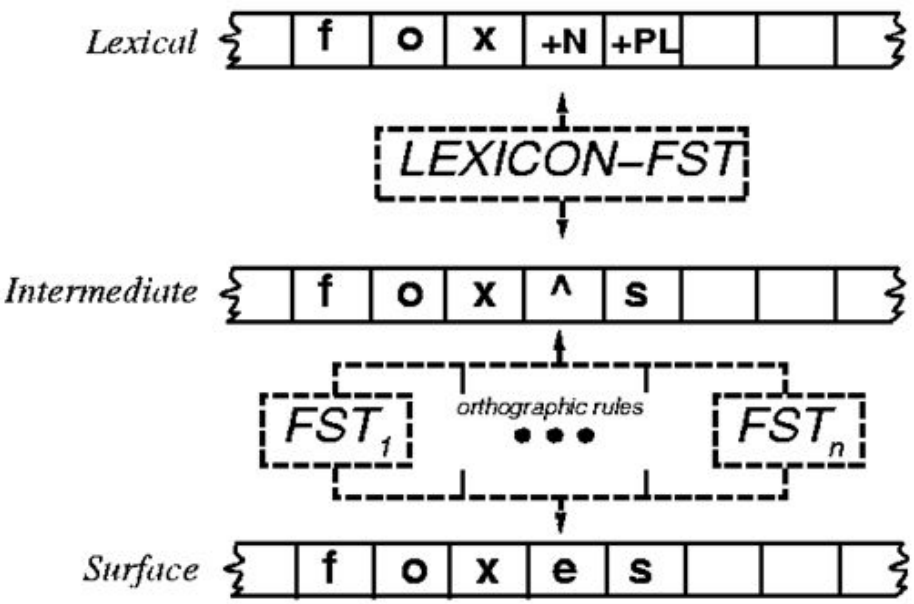
23 October 2017

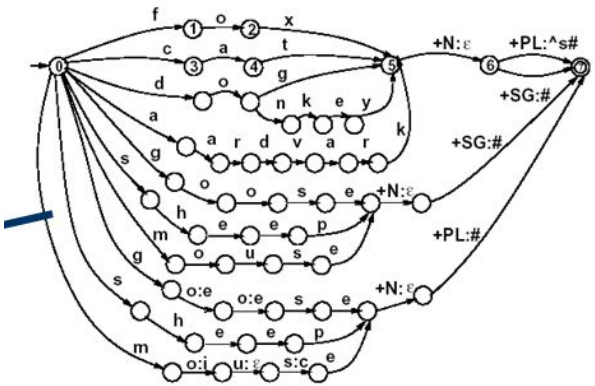
Last class

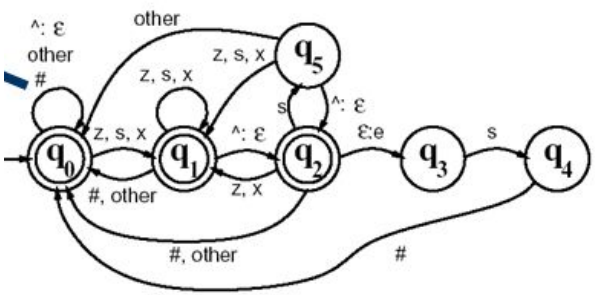
We discussed morphological analysis (parsing, generation and recognition).

We described a finite-state transducer for analysing the morphological properties of nouns

This FST is a cascade of two FSTs: one for translating a string from an *lexical form* to an *intermediate form*, and one for translating a string from an intermediate form to a *surface form*



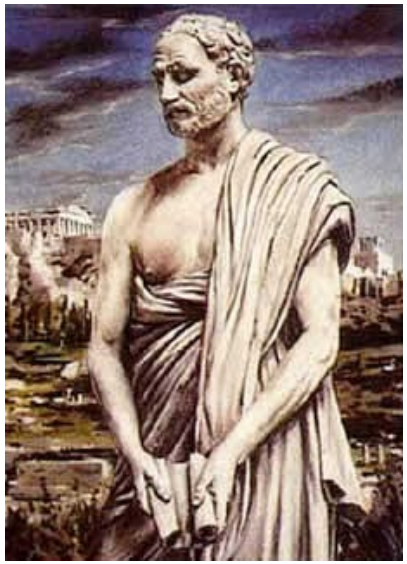




1. Word classes and POS tags
2. Some specific word classes
3. Lexical ambiguity and word frequency

Reading: Jurafsky & Martin, Chapter 5.

Parts of speech in NL grammar



Linguists have been classifying words for a long time . . .

- Dionysius Thrax of Alexandria (c. 100 BC) wrote a grammatical sketch of Greek involving 8 parts-of-speech:

nouns	verbs	pronouns	prepositions
adverbs	conjunctions	participles	articles

- Thrax's list and minor variations on it dominated European language grammars and dictionaries for 2000 years.
- (Anyone sees an important POS missing?)

Modern tagsets

In modern (English) NLP, larger (and more fine-grained) tagsets are preferred. E.g.

Penn Treebank	45 tags	`http://bit.ly/1gwbird`
Brown corpus	87 tags	`http://bit.ly/1jG9i2P`
C7 tagset	146 tags	`http://bit.ly/1Mh36KX`

Trade-off between complexity and precision . . . and whatever tagset we use, there'll be some words that are hard to classify.

Why do we need so many tags? We will see soon.

Distributional equivalence

Recall that for prog langs, a parser typically works entirely with tags produced by the lexer (e.g. IDENT, NUM). It won't care whether an identifier is `x` or `y`, or whether a numeral is `0` or `5`.

Consequence: `x` and `y` have the same *distribution*: `x` can occur wherever `y` can, and vice versa.

The idea of POS tags is much the same: group the words of a language into classes of words with the same (or similar) distributions. E.g. the words

crocodile pencil mistake

are very different as regards meaning, but grammatically can occur in the same contexts. So let's classify them all as nouns.
(More specifically, as *singular, countable, common nouns*.)

Distributional equivalence

We can operationalize the idea of distributional equivalence by using tests: can one word substitute for another?

- Kim saw the elephant before we did.
- Kim saw the movie before we did.
- Kim saw the mountain before we did.
- Kim saw the error before we did.

Tests can be too strict:

- (*) Kim saw the Sam before we did
- (*) Kim arrived the movie before we did

Criteria for classifying words

When should words be put into the same class?

Three different criteria might be considered . . .

- Distributional criteria: Where can the words occur?
- Morphological criteria: What form does the word have? (E.g. -tion, -ize). What affixes can it take? (E.g. -s, -ing, -est).
- Notional (or semantic) criteria: What sort of concept does the word refer to? (E.g. nouns often refer to 'people, places or things'). More problematic: less useful for us.

We'll look at various parts-of-speech in terms of these criteria.

Open and closed classes in natural language

There's a broad distinction between open and closed word classes:

- Open classes are typically large, have fluid membership, and are often stable under translation.
- Four major open classes are widely found in languages worldwide: *nouns*, *verbs*, *adjectives*, *adverbs*.
  - Virtually all languages have at least the first two.
  - All Indo-European languages (e.g. English) have all four.
- Closed classes are typically small, have relatively fixed membership, and the repertoire of classes varies widely between languages. E.g. *prepositions* (English, German), *post-positions* (Hungarian, Urdu, Korean), *particles* (Japanese), *classifiers* (Chinese), etc.
- Closed-class words (e.g. of, which, could) often play a structural role in the grammar as function words.

Nouns

Notionally, nouns generally refer to living things (*mouse*), places (*Scotland*), non-living 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. For example, nouns can appear with possession: “his car”, ”her idea”.

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.

Different types of verbs have different distributional properties. For example, base form verbs can appear as infinitives: “to jump”, ”to learn”.

Example of noun and verb classes

Nouns:

- Proper nouns: names such as Regina, IBM, Edinburgh
- Pronouns: he, she, it, they, we
- Common nouns
  - Count nouns: e.g. goat
  - Mass nouns: e.g. snow (? snows)

Verbs can be in base form, past tense, gerund... Also, consider auxiliary verbs.

Why do we need so many tags?

What is the part of speech tag for "walking"? Use **linguistic tests**.

Why do we need so many tags?

What is the part of speech tag for "walking"? Use **linguistic tests**.
Verb tests:

Example

Walking quickly is awkward
Quickly walking is awkward

Why do we need so many tags?

What is the part of speech tag for "walking"? Use **linguistic tests**.
Verb tests:

Example

Walking quickly is awkward
Quickly walking is awkward

Noun tests:

Example

Walking is awkward
Her walking is awkward
Fast walking is awkward

Why do we need so many tags?

What is the part of speech tag for "walking"? Use **linguistic tests**.
Verb tests:

Example

Walking quickly is awkward
Quickly walking is awkward

Noun tests:

Example

Walking is awkward
Her walking is awkward
Fast walking is awkward

"Walking" has both properties of both noun and verb. A separate tag for gerunds?

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*.

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.

Other classes (closed)

- **prepositions:** on, under, over, near, by, at, from, to, with
- **determiners:** a, an, the
- **conjunctions:** and, but, or, as, if, when
- **particles:** up, down, on, off, in, out, at, by
- **numerals:** one, two, three, first, second, third

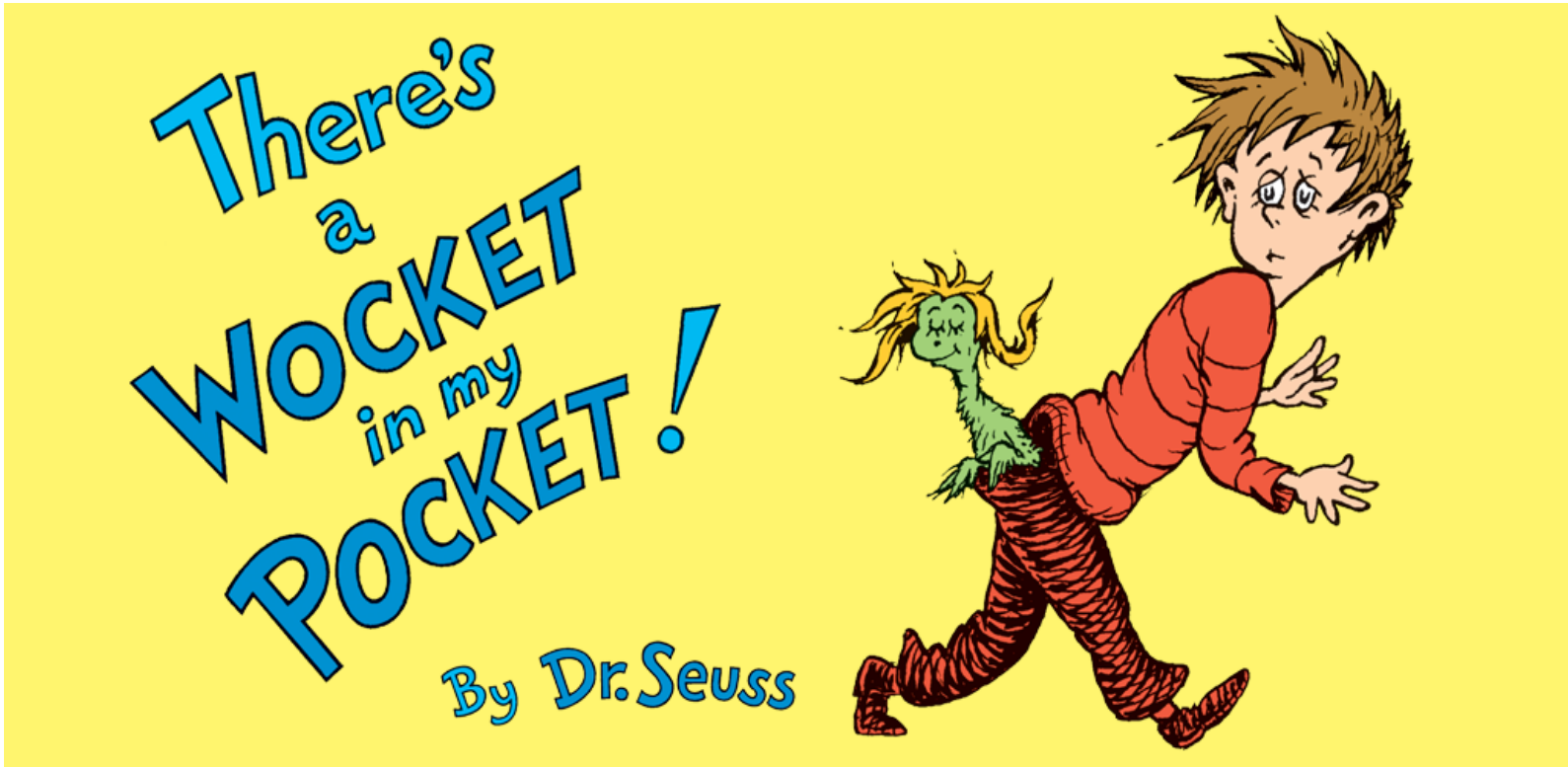
Importance of formal and distributional criteria

Often in reading, we come across unknown words. (Especially in computing literature!)

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 which (open) class an unknown word belongs to.

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


There's
a
WOCKET
in my
POCKET!
By Dr. Seuss

Example of POS inference

Those **zorls** you **splarded** were **malgy**.

What is the part of speech of the word **malgy**?

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

Example of POS inference

The highly-valued share plummeted over the course of the busy week .

Can you decide on the tags of each word?

Example of POS inference

The highly-valued share plummeted over the course of the busy week .

Can you decide on the tags of each word?

The/ highly-valued/ share/ plummeted/ over/ the/ course/ of/ the/ busy/ week/ .

Example of POS inference

The highly-valued share plummeted over the course of the busy week .

Can you decide on the tags of each word?

The/DT highly-valued/JJ share/NN plummeted/VBD over/IN the/DT course/NN of/IN the/DT busy/JJ week/NN ./.

The tagging problem

Given an input text, we want to tag each word correctly:

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

There/EX was/VBD still/JJ lemonade/NN in/IN the/DT bottle/NN ./.

(Many Brown/Penn tags are quite counterintuitive!)

- In the above, number and bottle are nouns not verbs — but how does our tagger tell?
- In the second example, still could be an adjective or an adverb — which seems more likely?

These issues lead us to consider word frequencies, which serve as the basis of the idea of *model estimation*. (among other things).

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

Word Frequencies in Different Languages

Ambiguity by part-of-speech tags:

Language	**Type-ambiguous**	**Token-ambiguous**
English	13.2%	56.2%
Greek	<1%	19.14%
Japanese	7.6%	50.2%
Czech	<1%	14.5%
Turkish	2.5%	35.2%

Taken from real data for treebanks annotated with their POS tags

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

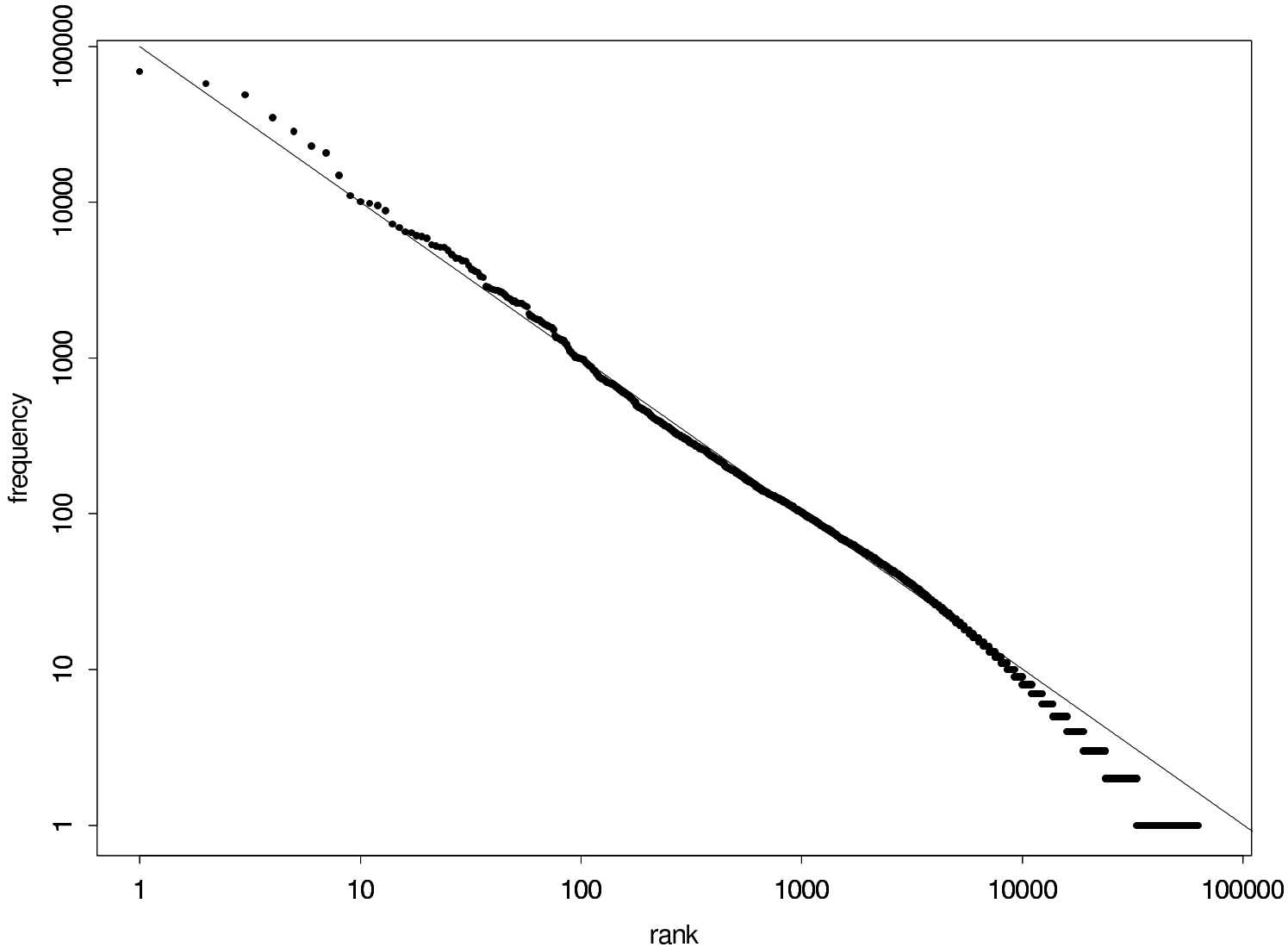
Zipf's law

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: $f \cdot r = k$.

Zipf's law for the Brown corpus



Zipf's law

According to 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.

(It's not fully understood why Zipf's law works so well for word frequencies.)
In fact, many other kinds of data conform closely to a Zipfian distribution:

- Populations of cities.
- Sizes of earthquakes.
- Amazon sales rankings.

Rule-based tagging

Old POS taggers used to work in two stages, based on hand-written rules: the first stage identifies a set of possible POS for each word in the sentence (based on a lexicon), and the second uses a set of hand-crafted rules in order to select a POS from each of the lists for each word.

Example:

1. If a word belongs to the set of determiners, tag is at `DT`.
2. Tag a word next to a determiner as an adjective (`JJ`), if it ends with `-ed`.
3. If a word appears after `is` or `are` and it ends with `ing`, tag it as a verb `VBG`.

Why do we need POS tags?

- They are often an essential ingredient in natural language applications
- Usually appear at the "bottom" of the pipeline
- For example: most of the syntactic variability (we will learn about that later) is determined by the sequence of POS tags in a sentence. POS tags are easier to predict than the full syntax, and therefore, by predicting the POS tags, we pave the way for identification of full phrases: noun phrases, verb phrases, etc.

Extreme POS ambiguity...

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo

Extreme POS ambiguity...

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo

Bison from Buffalo, which bison from Buffalo bully, themselves bully bison from Buffalo.

Extreme POS ambiguity...

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo

Bison from Buffalo, which bison from Buffalo bully, themselves bully bison from Buffalo.

If police police police police, who police police police? Police police police police police police

Look it up!