UNIB20005

Language and Computation

Categorizing and Tagging Words

Steven Bird, Department of Computing and Information Systems

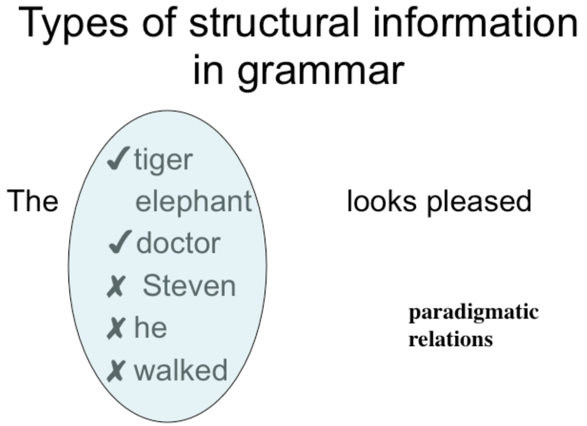
Chapter 4: Writing Structured Programs

- not covered in lectures or regular workshops
- this chapter goes back to the basics of programming;
  (emphasis on functions)
- please read sections 4.1-4.4
- if you need help with anything
  - ask tutor; attend extra workshop; post a question to discussion forum

What is a lexical category?

Review from Lesley's slides

- aka word class, part of speech
- **Noun**: a thing, an entity: *tree, friendship, floor*
- **Verb**: an action, a state: *go, sleep, give*
- **Adjective**: a property of a noun: *red, kind, easy*
- **Adverb**: a property of a verb: *soon, easily, angrily*
- **Preposition**: a relation, often spatial: *in near with*
- diagnostics:
  syntactic (distributional, "syntagmatic");
  morphological (internal, "paradigmatic")
  -- see the linguistic sections of chapter 5,
  especially section 5.7



Types of structural information in grammar

The ✓tiger elephant ✓doctor ✗ Steven ✗ he ✗ walked looks pleased

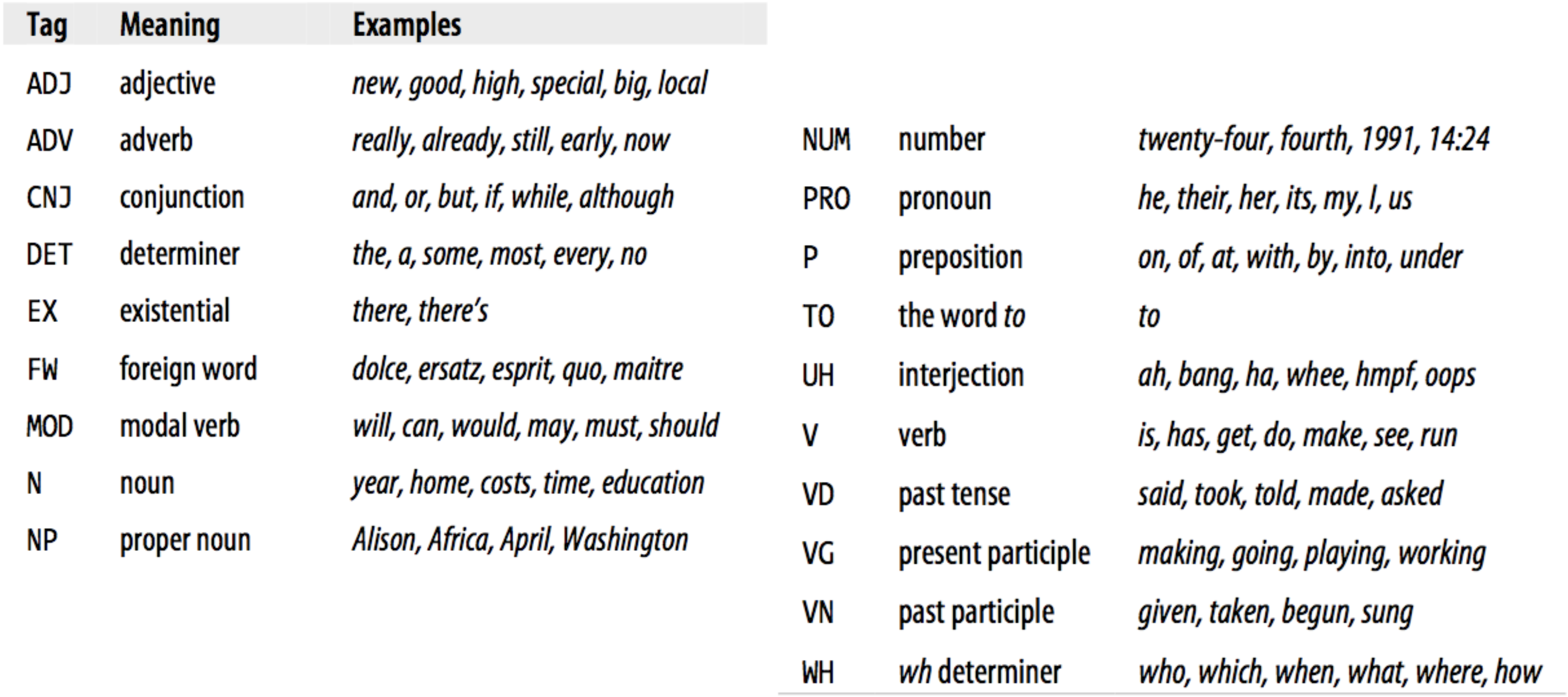
paradigmatic relations

Lexical category ambiguities

- British left waffles on Falkland Islands
- Lung cancer in women mushrooms
- Clinton wins on budget, but more lies ahead
- Juvenile court to try shooting defendant
- Deer kill 17,000

Simplified Part-of-Speech Tagset

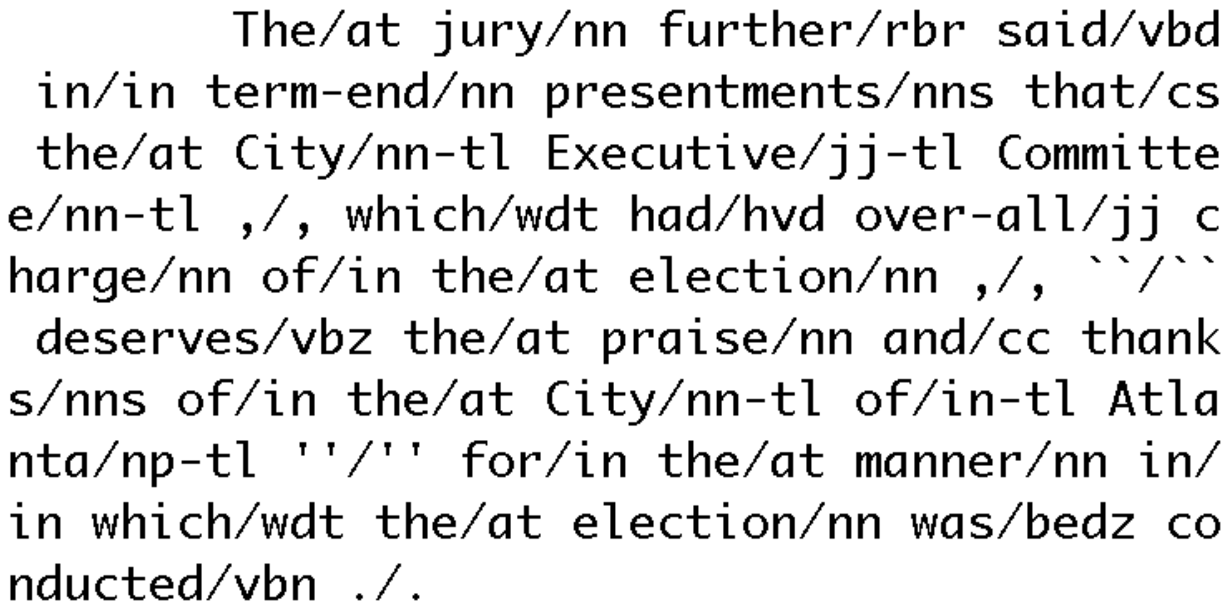
Table 5-1. Simplified part-of-speech tagset



Tag	Meaning	Examples
ADJ	adjective	*new, good, high, special, big, local*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
EX	existential	*there, there's*
FW	foreign word	*dolce, ersatz, esprit, quo, maitre*
MOD	modal verb	*will, can, would, may, must, should*
N	noun	*year, home, costs, time, education*
NP	proper noun	*Alison, Africa, April, Washington*
NUM	number	*twenty-four, fourth, 1991, 14:24*
PRO	pronoun	*he, their, her, its, my, I, us*
P	preposition	*on, of, at, with, by, into, under*
TO	the word *to*	*to*
UH	interjection	*ah, bang, ha, whee, hmpf, oops*
V	verb	*is, has, get, do, make, see, run*
VD	past tense	*said, took, told, made, asked*
VG	present participle	*making, going, playing, working*
VN	past participle	*given, taken, begun, sung*
WH	*wh* determiner	*who, which, when, what, where, how*

Tagged Corpora

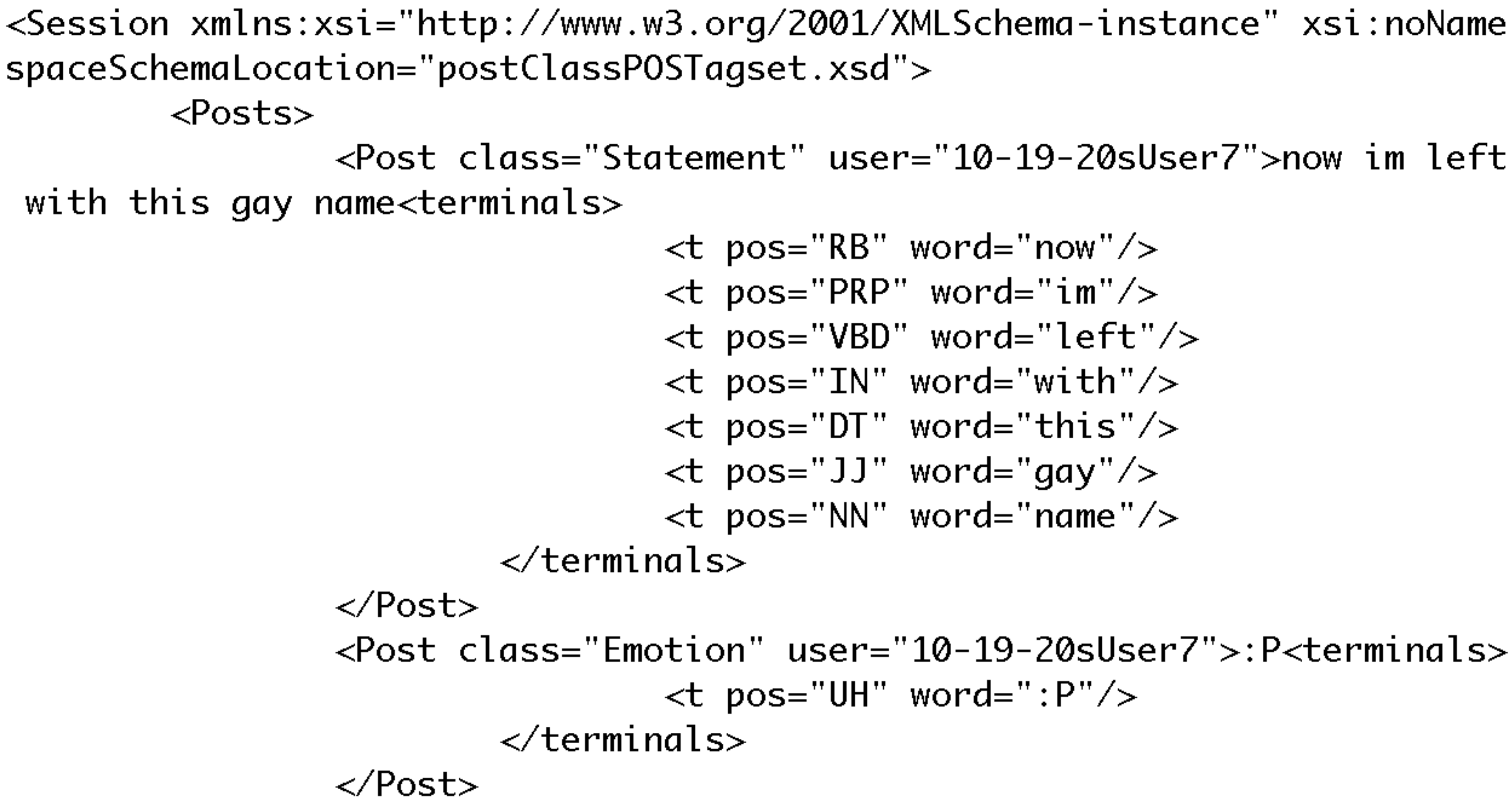
Tagged Corpora: Brown Corpus



```
        The/at jury/nn further/rbr said/vbd
 in/in term-end/nn presentments/nns that/cs
 the/at City/nn-tl Executive/jj-tl Committe
e/nn-tl ,/, which/wdt had/hvd over-all/jj c
harge/nn of/in the/at election/nn ,/, ``/``
 deserves/vbz the/at praise/nn and/cc thank
s/nns of/in the/at City/nn-tl of/in-tl Atla
nta/np-tl ''/'' for/in the/at manner/nn in/
in which/wdt the/at election/nn was/bedz co
nducted/vbn ./.
```

- nltk.corpus.brown.tagged_words(simplify_tags = True)

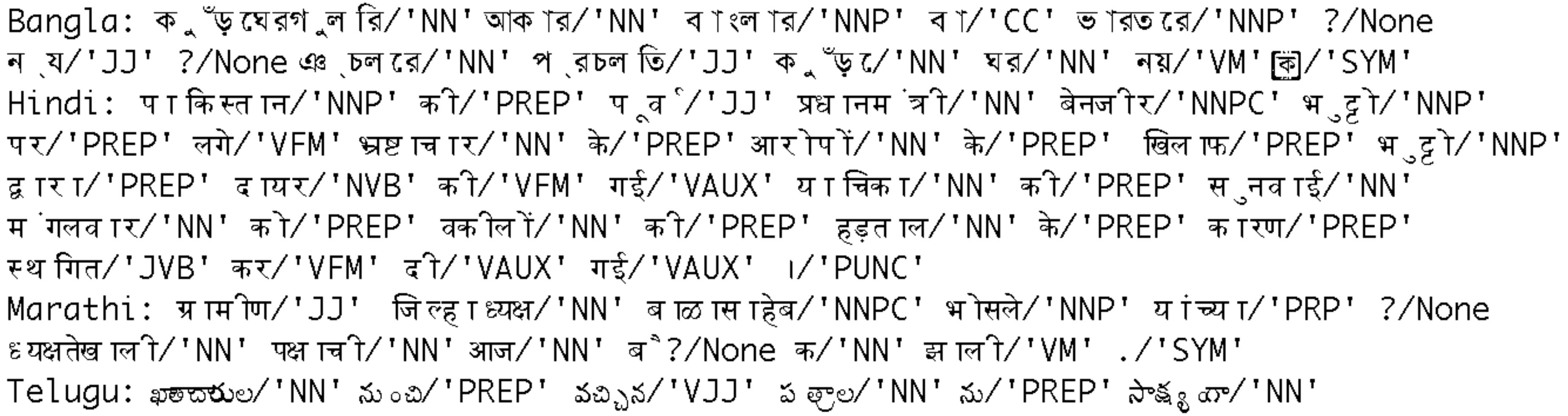
Tagged Corpora: NPS Chat Corpus



```
<Session xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noName
spaceSchemaLocation="postClassPOSTagset.xsd">
        <Posts>
                <Post class="Statement" user="10-19-20sUser7">now im left
 with this gay name<terminals>
                                <t pos="RB" word="now"/>
                                <t pos="PRP" word="im"/>
                                <t pos="VBD" word="left"/>
                                <t pos="IN" word="with"/>
                                <t pos="DT" word="this"/>
                                <t pos="JJ" word="gay"/>
                                <t pos="NN" word="name"/>
                        </terminals>
                </Post>
                <Post class="Emotion" user="10-19-20sUser7">:P<terminals>
                                <t pos="UH" word=":P"/>
                        </terminals>
                </Post>
```

- nltk.corpus.nps_chat.tagged_words(simplify_tags = True)

Tagged Corpora: Indian Language Corpus



```
Bangla: কুঁড়েঘরগুলির/'NN' আকার/'NN' বাংলার/'NNP' বা/'CC' ভারতের/'NNP' ?/None
ন্য/'JJ' ?/None ঞ্চলের/'NN' প্রচলিত/'JJ' কুঁড়ে/'NN' ঘর/'NN' নয়/'VM' ক/'SYM'
Hindi: पाकिस्तान/'NNP' की/'PREP' पूर्व/'JJ' प्रधानमंत्री/'NN' बेनजीर/'NNPC' भुट्टो/'NNP'
पर/'PREP' लगे/'VFM' भ्रष्टाचार/'NN' के/'PREP' आरोपों/'NN' के/'PREP' खिलाफ/'PREP' भुट्टो/'NNP'
द्वारा/'PREP' दायर/'NVB' की/'VFM' गई/'VAUX' याचिका/'NN' की/'PREP' सुनवाई/'NN'
मंगलवार/'NN' को/'PREP' वकीलों/'NN' की/'PREP' हड़ताल/'NN' के/'PREP' कारण/'PREP'
स्थगित/'JVB' कर/'VFM' दी/'VAUX' गई/'VAUX' ।/'PUNC'
Marathi: ग्रामीण/'JJ' जिल्हाध्यक्ष/'NN' बाळासाहेब/'NNPC' भोसले/'NNP' यांच्या/'PRP' ?/None
ध्यक्षतेखाली/'NN' पक्षाची/'NN' आज/'NN' बै?/None क/'NN' झाली/'VM' ./'SYM'
Telugu: ఖాతాదారుల/'NN' నుంచి/'PREP' వచ్చిన/'VJJ' పత్రాల/'NN' ను/'PREP' సాక్ష్యా/'NN'
```

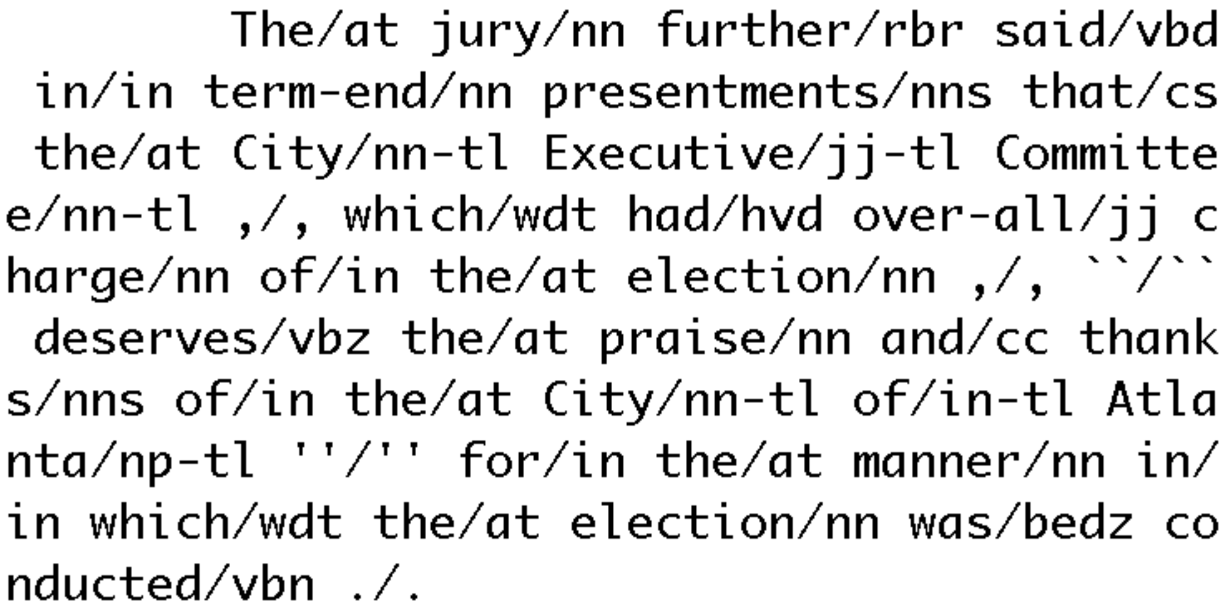
- nltk.corpus.indian.tagged_words(simplify_tags = True)

Part of Speech Tagging

- Lexical categories, word classes, parts of speech:
  - important part of human lexical knowledge
  - how do we get computers to work with lexical categories?

- **POS Tagging:**
  - early step in accessing meaning
  - early example of a computational model of language
  - statistical evidence for categorical judgements

What is the most frequently occurring lexical category in English text?



```
       The/at jury/nn further/rbr said/vbd
 in/in term-end/nn presentments/nns that/cs
 the/at City/nn-tl Executive/jj-tl Committe
e/nn-tl ,/, which/wdt had/hvd over-all/jj c
harge/nn of/in the/at election/nn ,/, ``/``
 deserves/vbz the/at praise/nn and/cc thank
s/nns of/in the/at City/nn-tl of/in-tl Atla
nta/np-tl ''/'' for/in the/at manner/nn in/
in which/wdt the/at election/nn was/bedz co
nducted/vbn ./.
```

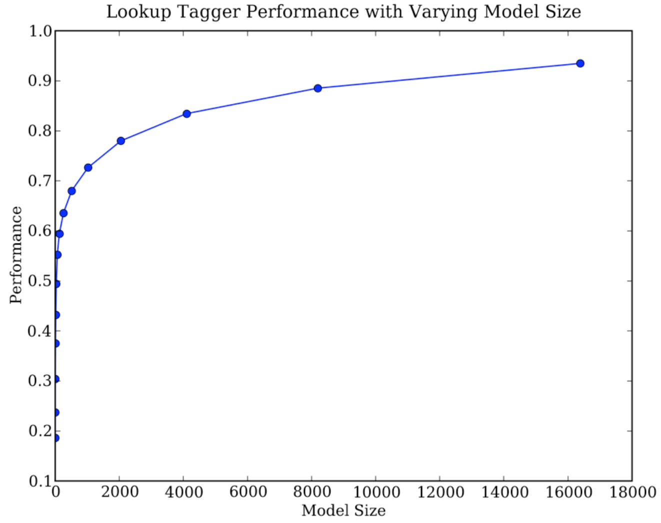
- How would you answer this question computationally?

- How would we find words that had more than one lexical category?

- How might we assign lexical categories automatically?

Detecting the category of a word

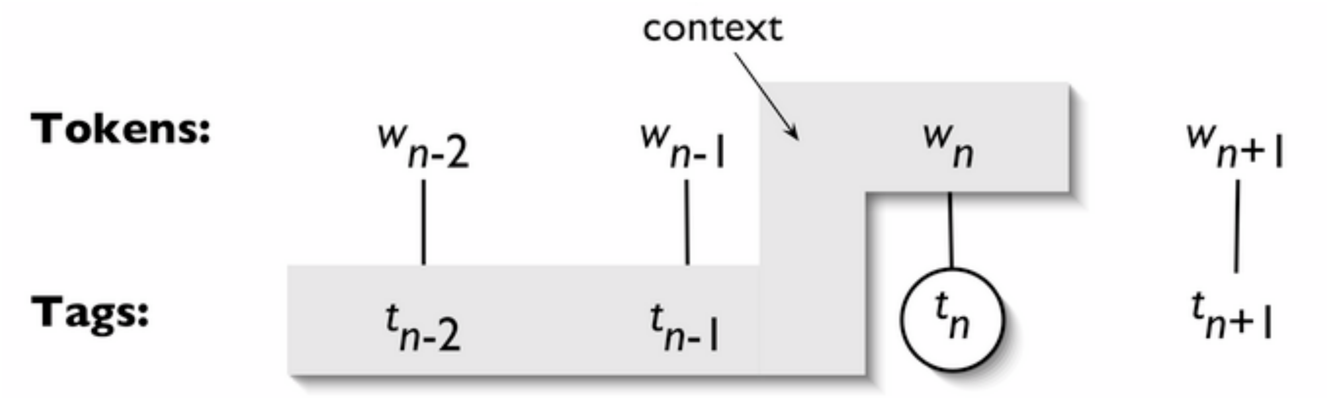
- guess that everything is a noun
- use word-internal clues (what would these be?)
- memorize the most probable tag for the most likely *n* words
  - memory/performance tradeoff
  - open vs closed classes



- exploit context, e.g.:
  - the *watch* vs to *watch* (context is word on the left)
  - old *watch* vs silently *watch* (context is the *category* of the word on the left)
- how much context?
  - they went to *school* vs they want to *school*

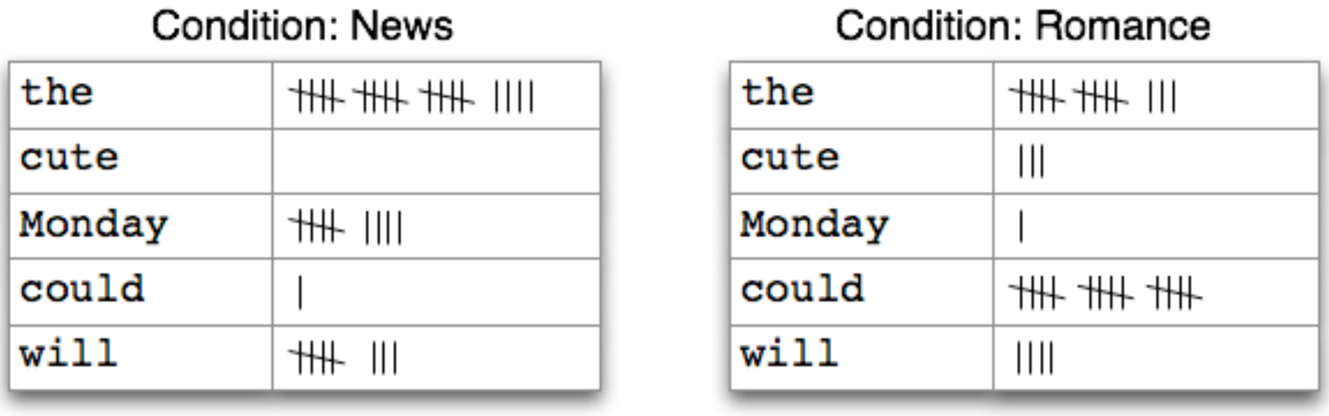
N-Gram Tagging

- Look at current word and n-1 previous tags
- what linguistic intuition is this capturing?



Storage and Training

- What information must a tagger store?
- Simple case: unigram tagger
- Harder case: bigram tagger
- Saving space: backoff
- How do we create this data?



Condition: News

the	𝍸 𝍸 𝍸 ||||
cute	
Monday	𝍸 ||||
could	|
will	𝍸 |||

Condition: Romance

the	𝍸 𝍸 |||
cute	|||
Monday	|
could	𝍸 𝍸 𝍸
will	||||