Part of Speech Tagging

Instructor: Wei Xu

Facebook.com/BizarroComics
©DAN PIRARO. 6·5·12 w/ANDY COWAN
ORANGE JUICE CONCENTRATE
ORANGE JUICE CONCENTRATE
LENNY ALWAYS TREATS "CONCENTRATE" AS A VERB
BIZARROCOMICS.COM
Dist. by King Features

Some slides adapted from Brendan O'Connor, Chris Manning and Yejin Choi

Where are we going with this?

- Text classification: bags of words

- Language Modeling: n-grams

- Sequence tagging:
    - Parts of Speech
    - Named Entity Recognition
    - Other areas: bioinformatics (gene prediction), etc…

What’s a part-of-speech (POS)?

- Syntax = how words compose to form larger meaning bearing units

- POS = syntactic categories for words (a.k.a word class)
    - You could substitute words within a class and have a syntactically valid sentence

I saw the **dog**

I saw the **cat**

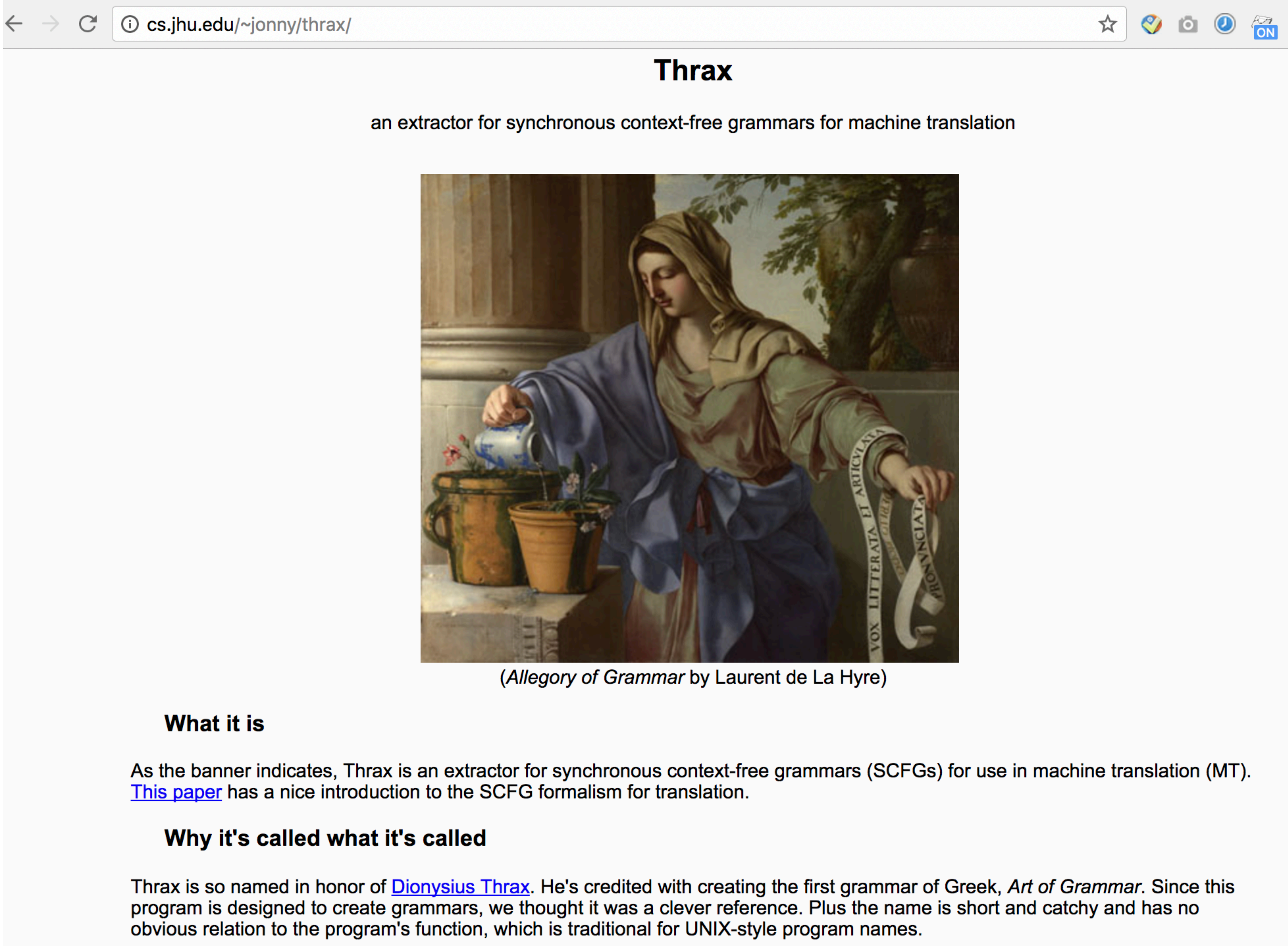
I saw the ___

- Gives information how words combine into larger phrases

Parts of Speech is an old idea

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
- Also, Dionysius Thrax of Alexandria (c. 100 BCE)
- 8 main POS: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection
- Many more fine grained possibilities

https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7



cs.jhu.edu/~jonny/thrax/

Thrax

an extractor for synchronous context-free grammars for machine translation

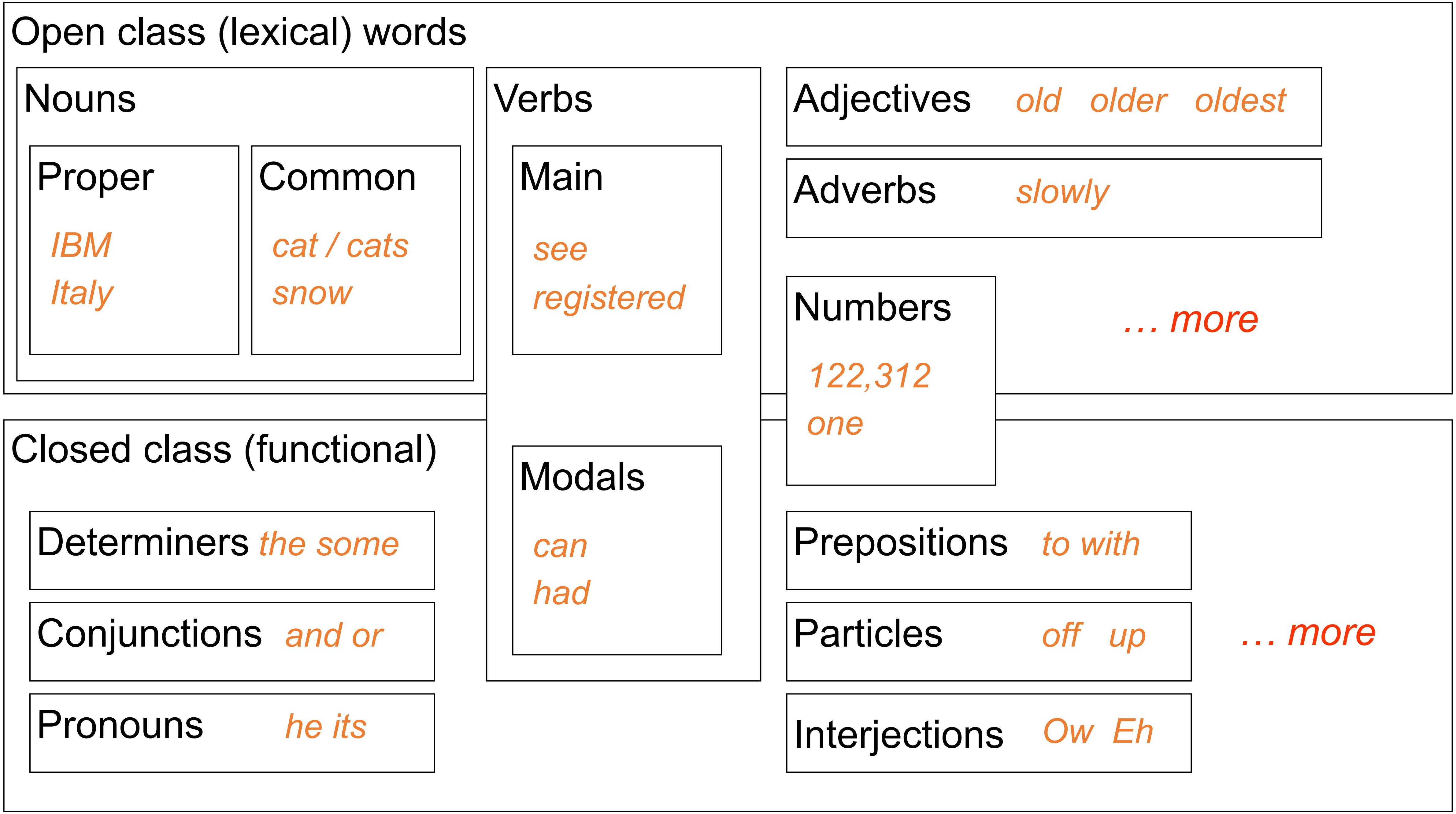
(*Allegory of Grammar* by Laurent de La Hyre)

What it is

As the banner indicates, Thrax is an extractor for synchronous context-free grammars (SCFGs) for use in machine translation (MT). This paper has a nice introduction to the SCFG formalism for translation.

Why it's called what it's called

Thrax is so named in honor of Dionysius Thrax. He's credited with creating the first grammar of Greek, *Art of Grammar*. Since this program is designed to create grammars, we thought it was a clever reference. Plus the name is short and catchy and has no obvious relation to the program's function, which is traditional for UNIX-style program names.



Open class (lexical) words

Nouns

Proper *IBM* *Italy*

Common *cat / cats* *snow*

Verbs

Main *see* *registered*

Modals *can* *had*

Adjectives *old older oldest*

Adverbs *slowly*

Numbers *122,312* *one*

… more

Closed class (functional)

Determiners *the some*

Conjunctions *and or*

Pronouns *he its*

Prepositions *to with*

Particles *off up*

Interjections *Ow Eh*

… more

Open vs. Closed classes

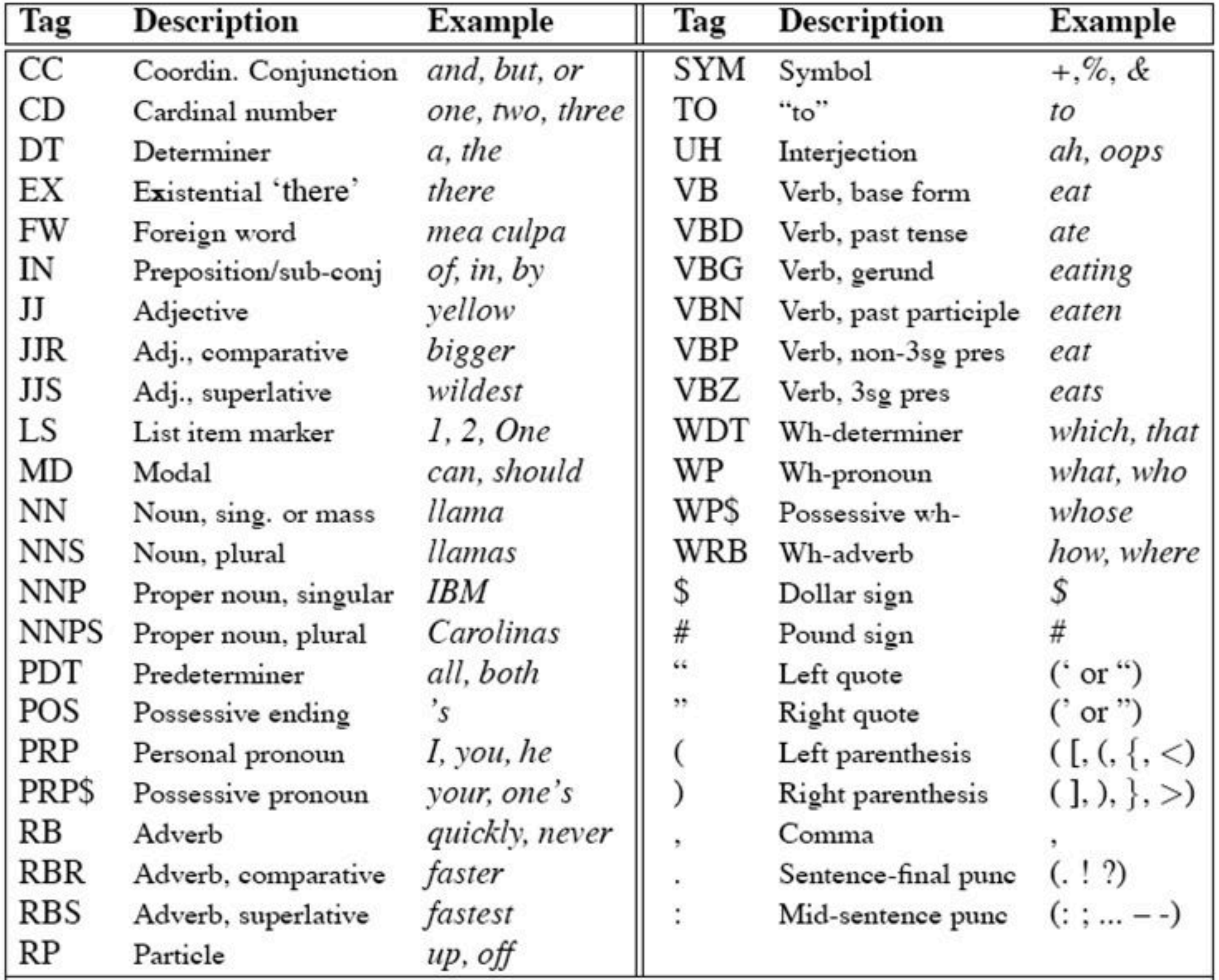
- Open vs. Closed classes
  - Closed:
    - determiners: ***a*, *an*, *the***
    - pronouns: ***she*, *he*, *I***
    - prepositions: ***on*, *under*, *over*, *near*, *by*, …**
    - Q: why called “closed”?
  - Open:
    - Nouns, Verbs, Adjectives, Adverbs.

Many Tagging Standards

- Penn Treebank (45 tags) … this is the most common one
- Brown corpus (85 tags)
- Coarse tagsets
  - Universal POS tags (Petrov et. al. https://github.com/slavpetrov/universal-pos-tags)
  - Motivation: cross-linguistic regularities

Penn Treebank POS

- 45 possible tags
- 34 pages of tagging guidelines



Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*'s*	”	Right quote	(’ or ”)
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	Possessive pronoun	*your, one's*	)	Right parenthesis	(],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

Ambiguity in POS Tagging

- Words often have more than one POS: *back*
  - The *back* door = JJ
  - On my *back* = NN
  - Win the voters *back* = RB
  - Promised to *back* the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

Exercise

POS Tagging

- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS

Penn Treebank POS tags

POS Tagging Performance

- How many tags are correct? (Tag Accuracy)
    - About 97% currently
    - But baseline is already 90%
        - Baseline is performance of stupidest possible method
            - Tag every word with its most frequent tag
            - Tag unknown words as nouns
    - Partly easy because
        - Many words are unambiguous
        - You get points for them (*the*, *a*, etc.) and for punctuation marks!

Deciding on the correct part of speech can be difficult even for people

- “Around” can be a preposition, particle, or adverb

Particle

Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG

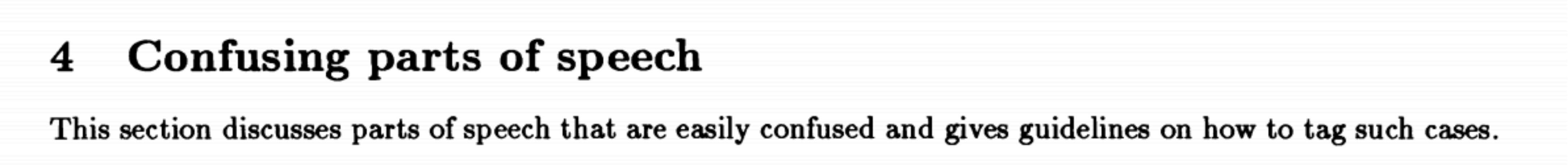
Preposition

All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN

Adverb

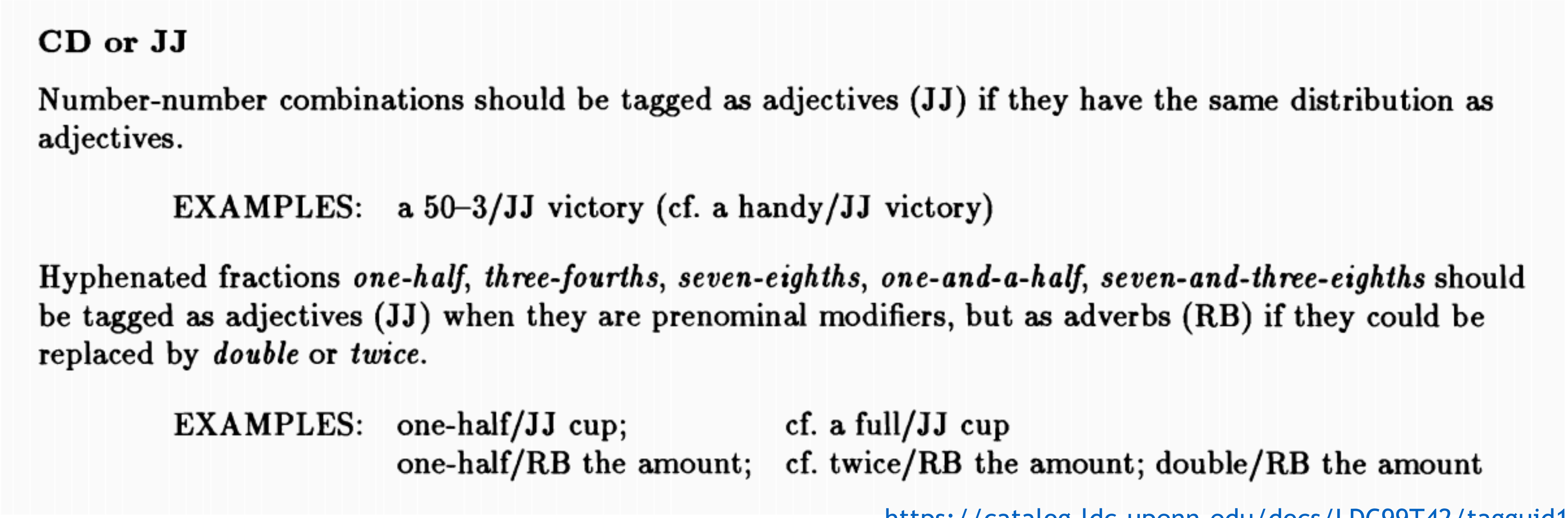
Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

It’s hard for linguists too!



4 Confusing parts of speech

This section discusses parts of speech that are easily confused and gives guidelines on how to tag such cases.



CD or JJ

Number-number combinations should be tagged as adjectives (JJ) if they have the same distribution as adjectives.

EXAMPLES: a 50–3/JJ victory (cf. a handy/JJ victory)

Hyphenated fractions *one-half*, *three-fourths*, *seven-eighths*, *one-and-a-half*, *seven-and-three-eighths* should be tagged as adjectives (JJ) when they are prenominal modifiers, but as adverbs (RB) if they could be replaced by *double* or *twice*.

EXAMPLES: one-half/JJ cup; cf. a full/JJ cup
one-half/RB the amount; cf. twice/RB the amount; double/RB the amount

https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf

How difficult is POS tagging?

- About 11% of the **word types** in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., *that*
  - I know *that* he is honest = IN
  - Yes, *that* play was nice = DT
  - You can’t go *that* far = RB
- 40% of the **word tokens** are ambiguous

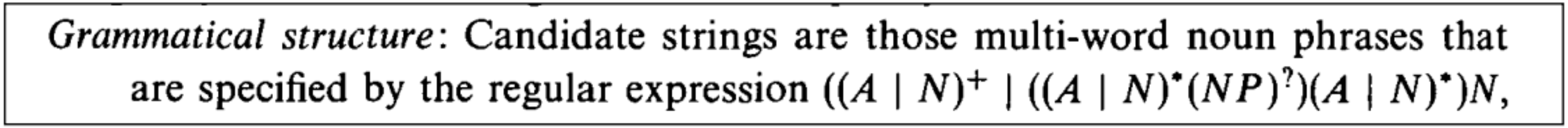
Token vs. Type

Token is instance or individual occurrence of a type.

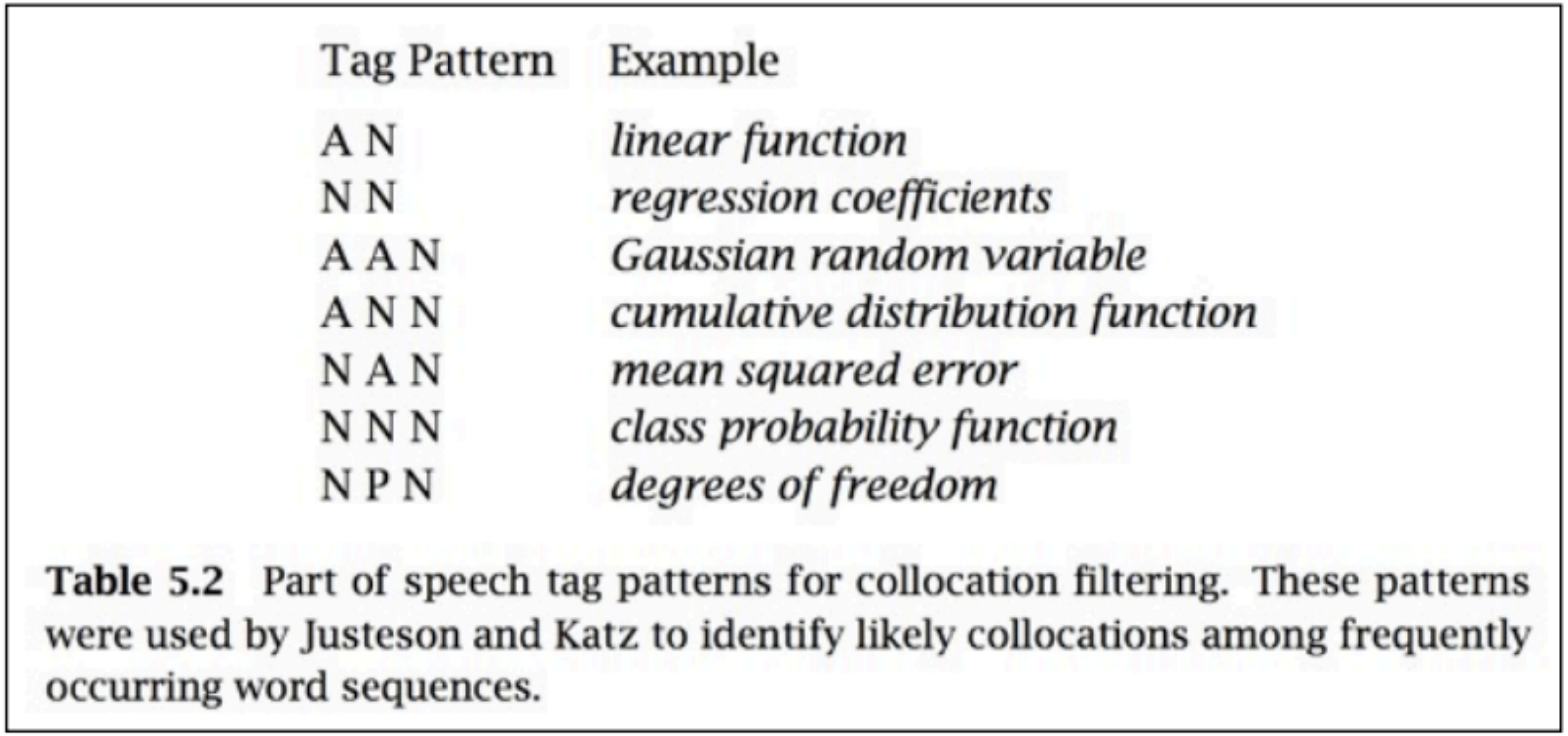
Why POS Tagging?

- Useful in and of itself (more than you'd think)
  - Text-to-speech: record, lead
  - Lemmatization: saw[v] → see, saw[a] → saw
  - Quick-and-dirty NP-chunk detection: grep {JJ | NN}* {NN | NNS}

Quick-and-Dirty Noun Phrase Identification



Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A \mid N)^+ \mid ((A \mid N)^*(NP)^?)(A \mid N)^*)N$,



Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

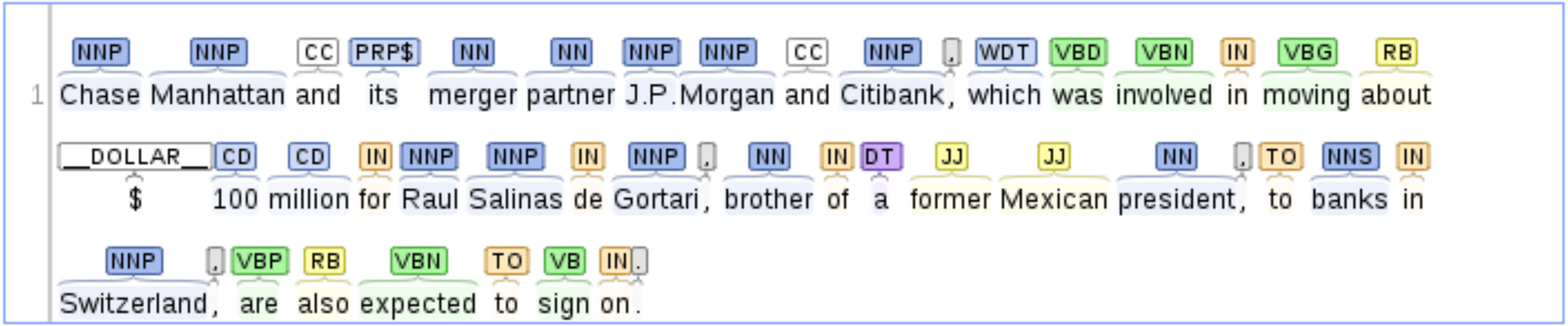
Why POS Tagging?

- Useful in and of itself (more than you'd think)
  - Text-to-speech: record, lead
  - Lemmatization: saw[v] → see, saw[a] → saw
  - Quick-and-dirty NP-chunk detection: grep {JJ | NN}* {NN | NNS}

- Useful for higher-level NLP tasks:
  - Chunking
  - Named Entity Recognition
  - Information Extraction
  - Parsing

Stanford CoreNLP Toolkit

Part-of-Speech:



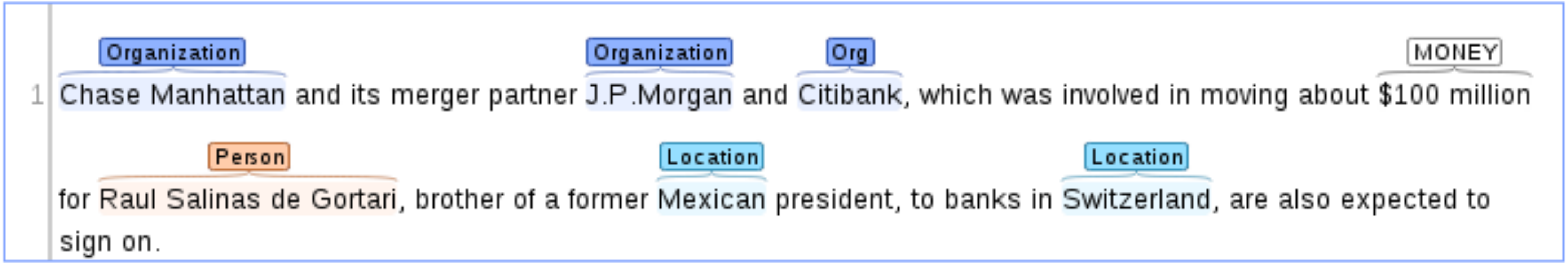
1

NNP	NNP	CC	PRP$	NN	NN	NNP	NNP	CC	NNP	,	WDT	VBD	VBN	IN	VBG	RB
Chase	Manhattan	and	its	merger	partner	J.P.	Morgan	and	Citibank	,	which	was	involved	in	moving	about

__DOLLAR__	CD	CD	IN	NNP	NNP	IN	NNP	,	NN	IN	DT	JJ	JJ	NN	,	TO	NNS	IN
$	100	million	for	Raul	Salinas	de	Gortari	,	brother	of	a	former	Mexican	president	,	to	banks	in

NNP	,	VBP	RB	VBN	TO	VB	IN	.
Switzerland	,	are	also	expected	to	sign	on	.

Named Entity Recognition:

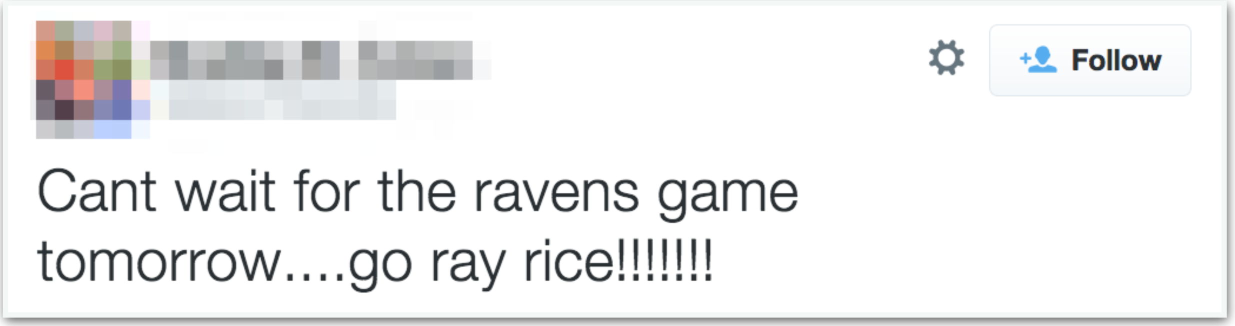


1 Chase Manhattan and its merger partner J.P.Morgan and Citibank, which was involved in moving about $100 million
for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland, are also expected to
sign on.

Organization	Chase Manhattan
Organization	J.P.Morgan
Org	Citibank
MONEY	$100 million
Person	Raul Salinas de Gortari
Location	Mexican
Location	Switzerland

Twitter NLP toolkit (Ritter et al.)

Cant	MD
wait	VB
for	IN
the	DT
ravens	NNP
game	NN
tomorrow	NN
…	:
go	VB
ray	NNP
rice	NNP
!!!!!!!	.



Cant wait for the ravens game tomorrow....go ray rice!!!!!!!