PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 4: Parts of Speech and Phrases

Today

- Lecture: key points from readings
- Reading discussion
- Theory assignment questions
- Lab: partsofspeech.R
- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Parts of Speech

- Grammatical Categories for words ←→ parts of speech.

- Nouns: typically refer to people, animals, concepts and things.
  - Cat, dog, watermelon, car, boat, chair, Matt, Susan, space, time.
- Verbs: typically express the action in a sentence.
  - Run, throw, treat, show, write, walk, climb
- Adjectives: Describe the properties of nouns:
  - Fast, green, interesting, small, hard, frequent

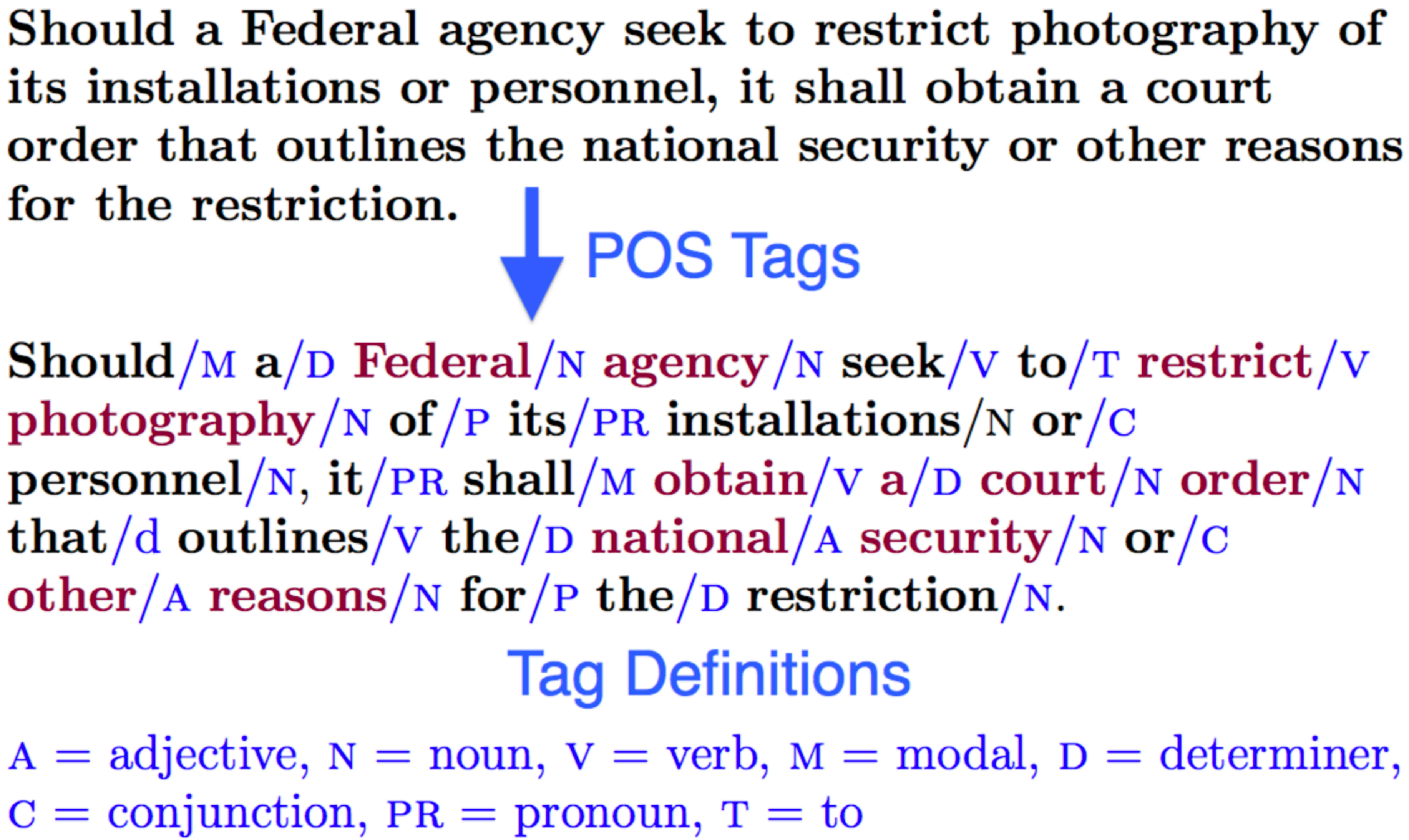
- “Substitution test” – determine which words belong in the same class.

Coarse tag	PTB tags	Examples
N: Nouns	NN: Common nouns (singular)	paragraph, adoption, member, extension, exploration
	NNS: Common nouns (plural)	barriers, additions, objects, policies, negotiations
	NNP: Proper nouns (singular)	Advanced, Assessment, Notice, Contents, Injury
	NNPS: Proper nouns (plural)	AUTHORIZATIONS, Limitations, Indians, Presidents
	FW: Foreign words	novo, parte, pima, de, tempore, officio, pro, bona
	CD: Numbers	48, 632, 1834, 2009, 1129, 1302, 381, 586, 810
A: Adjectives	JJ: Adjectives and ordinals	renewable, following, other, scientific, subsequent, last
	JJR: Adjectives (comparative)	younger, higher, Higher, More, less, smaller, earlier
	JJS: Adjectives (superlative)	latest, highest, greatest, Best, least, largest
	VBG: Gerunds, present participles	setting, resulting, being, working, operating, beginning
	RB: Adverbs	respectively, generally, forth, previously, so, no, fully
D: Determiners	DT	this, either, any (*Most common:* the, a, this)
P: Prepositions	IN: Most prepositions	*Most common words:* of, in, for, by, under, as, with
	TO: The word "to"	to, To, TO
V: Verbs	VB: Verbs (base form)	itemize, supply, TERMINATE, guarantee, concentrate
	VBD: Past tense	overestimated, trained, expired, GENERATED, switched
	VBN: Past participle	eliminated, intercepted, owed, advertised, Incorporated
	VBP: Present tense (non-3rd sing)	mitigate, nullify, Benefit, insert, fulfill, produce, seize
	VBZ: Present tense (3rd sing)	distributes, announces, directs, respects, upholds, uses
M: Verb Modifiers	RB: Adverbs (base form)	extremely, hard, rapidly, after, now
	RBR: Comparative adverbs	better, faster, slower, easier, shorter
	RBS: Superlative adverbs	best, worst, fastest, slowest, easiest
	RP: Particle adverbs	about, off, on, up
	MD: Modal auxiliary verbs	can, should, might, musn't
C: Coord. Conj.	CC: Coordinating conjunctions	and, or, but

Part of Speech Tagging

- Process of assigning a part of speech (POS) “tag” to each word in a document.
- Canonical training set of POS tags (in English) is the Penn Treebank (1999) which is maintained by the Linguistic Data Consortium:
  - https://catalog.ldc.upenn.edu/LDC99T42
- In practice, this means employing either hand coding, a heuristic approach, a maximum likelihood model, a neural net, or some combination to assign POS tags to terms.
- Quality of POS taggers is dependent on training data
  - Challenges in other languages, Twitter, etc.

Part of Speech Tagging



Should a Federal agency seek to restrict photography of its installations or personnel, it shall obtain a court order that outlines the national security or other reasons for the restriction.

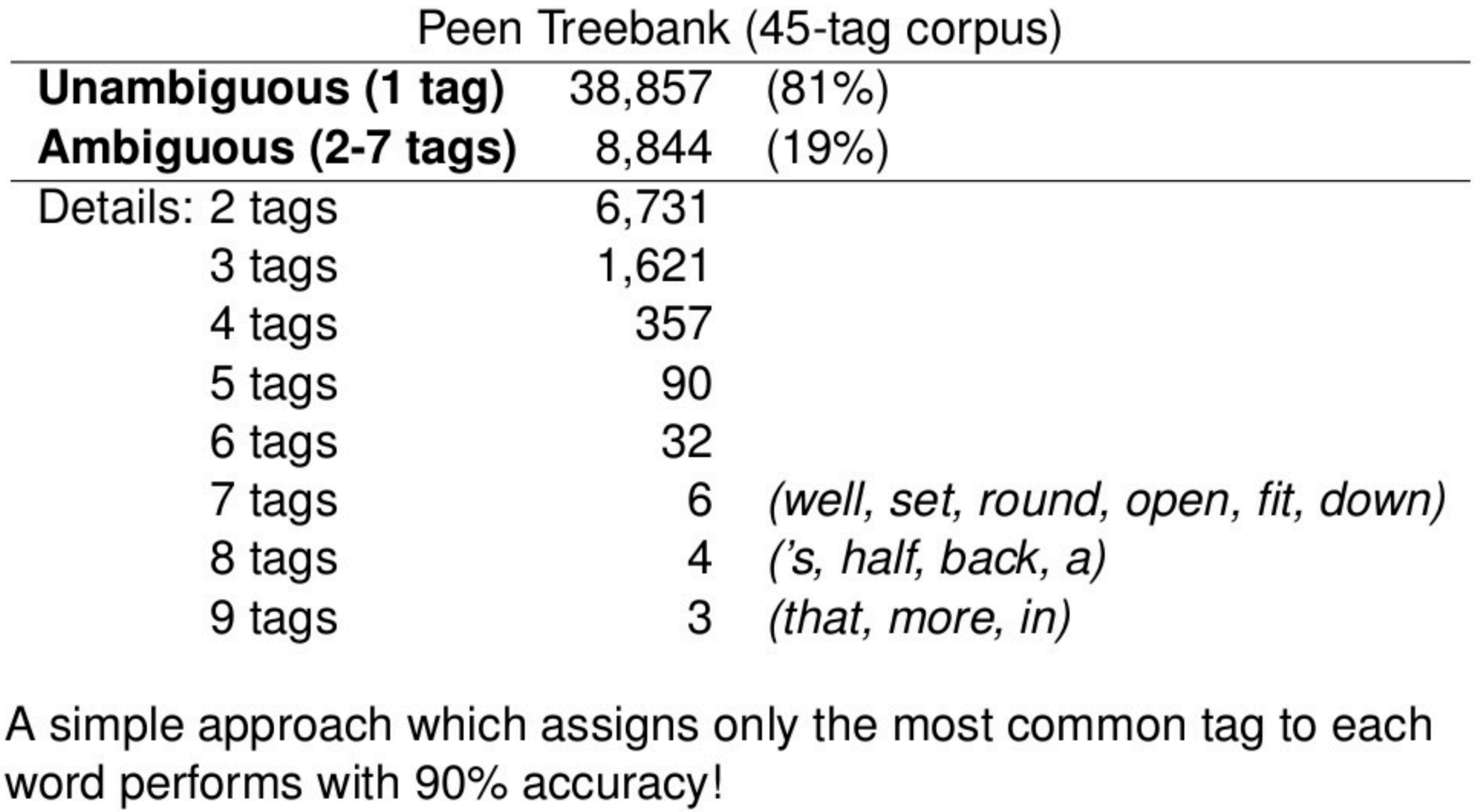
POS Tags

Should/M a/D Federal/N agency/N seek/V to/T restrict/V photography/N of/P its/PR installations/N or/C personnel/N, it/PR shall/M obtain/V a/D court/N order/N that/d outlines/V the/D national/A security/N or/C other/A reasons/N for/P the/D restriction/N.

Tag Definitions

A = adjective, N = noun, V = verb, M = modal, D = determiner, C = conjunction, PR = pronoun, T = to

Part of Speech Tagging



Peen Treebank (45-tag corpus)

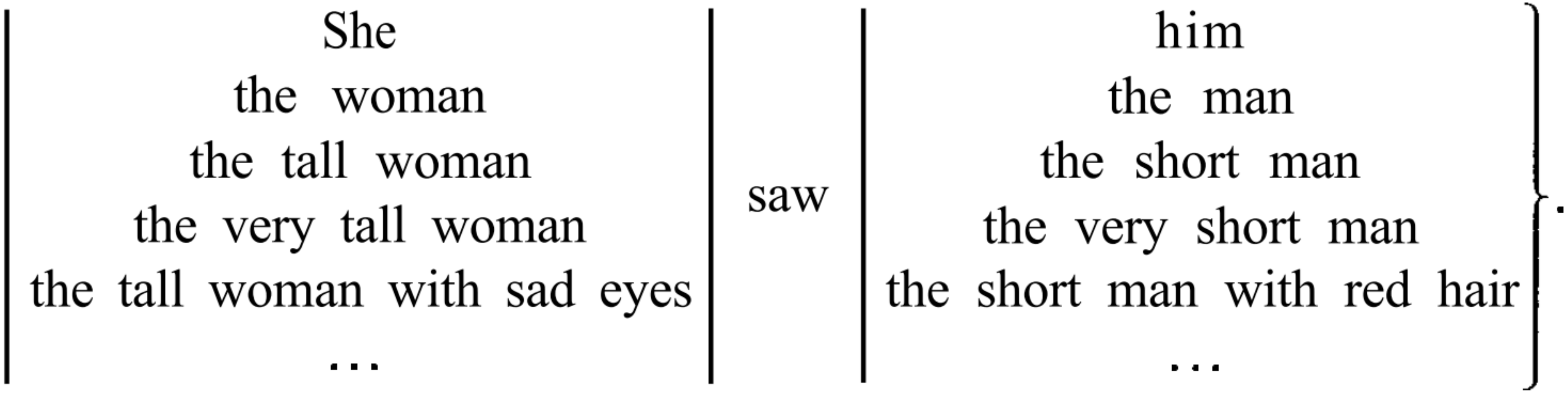
Unambiguous (1 tag)	38,857	(81%)
Ambiguous (2-7 tags)	8,844	(19%)
Details: 2 tags	6,731	
3 tags	1,621	
4 tags	357	
5 tags	90	
6 tags	32	
7 tags	6	*(well, set, round, open, fit, down)*
8 tags	4	*('s, half, back, a)*
9 tags	3	*(that, more, in)*

A simple approach which assigns only the most common tag to each word performs with 90% accuracy!

https://www.slideshare.net/franciscozamoraceu/a-connectionist-approach-to-partofspeech-tagging

Syntax → Phrases

- **Syntax** is the study of the regularities and constraints of word order and phrase structure.
- **Phrases** are *syntactically coherent* groupings of words.



She		him
the woman		the man
the tall woman		the short man
the very tall woman	saw	the very short man
the tall woman with sad eyes		the short man with red hair
…		…

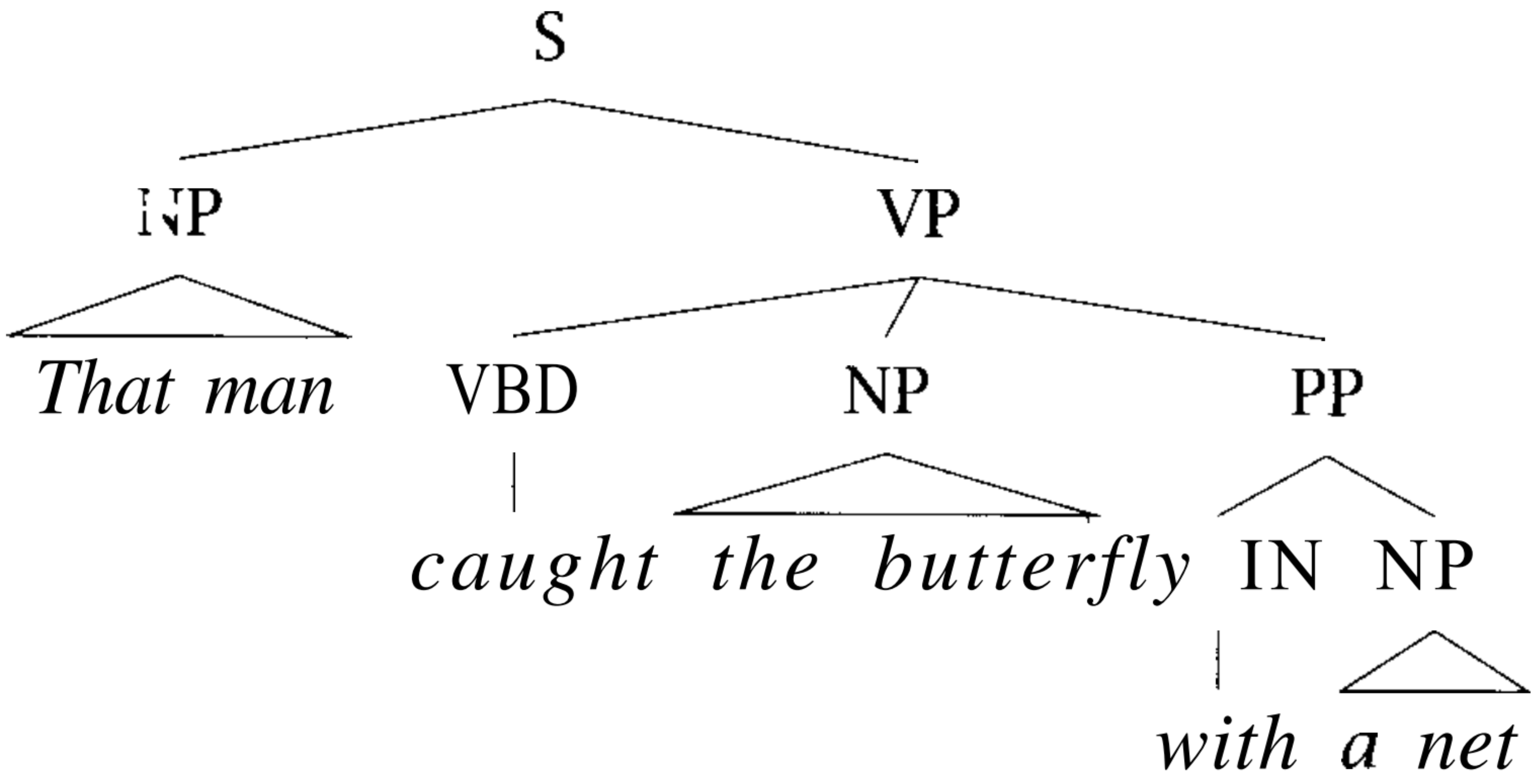
Manning and Schutze (1999, p.94)

Phrases

- **Noun phrases.** A noun is usually embedded in a phrase a syntactic unit of the sentence in which information about the noun is gathered. The noun is the head of the noun phrase, the central constituent that determines the syntactic character of the phrase. Noun phrases are usually the arguments of verbs, the participants in the action, activity or state described by the verb.
- **Verb phrases.** Analogous to the way nouns head noun phrases, the verb is the head of the verb phrase (VP). In general, the verb phrase organizes all elements of the sentence that depend syntactically on the verb

Manning and Schutze (1999, p.95)

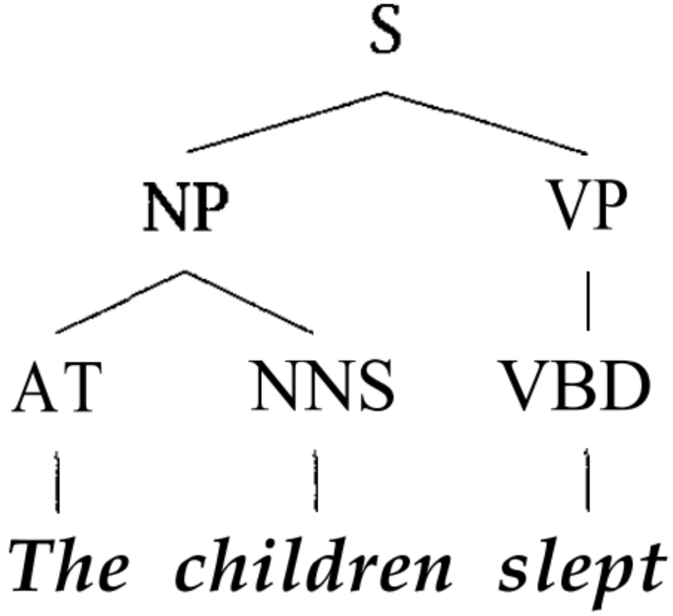
Tree Structure of Sentences



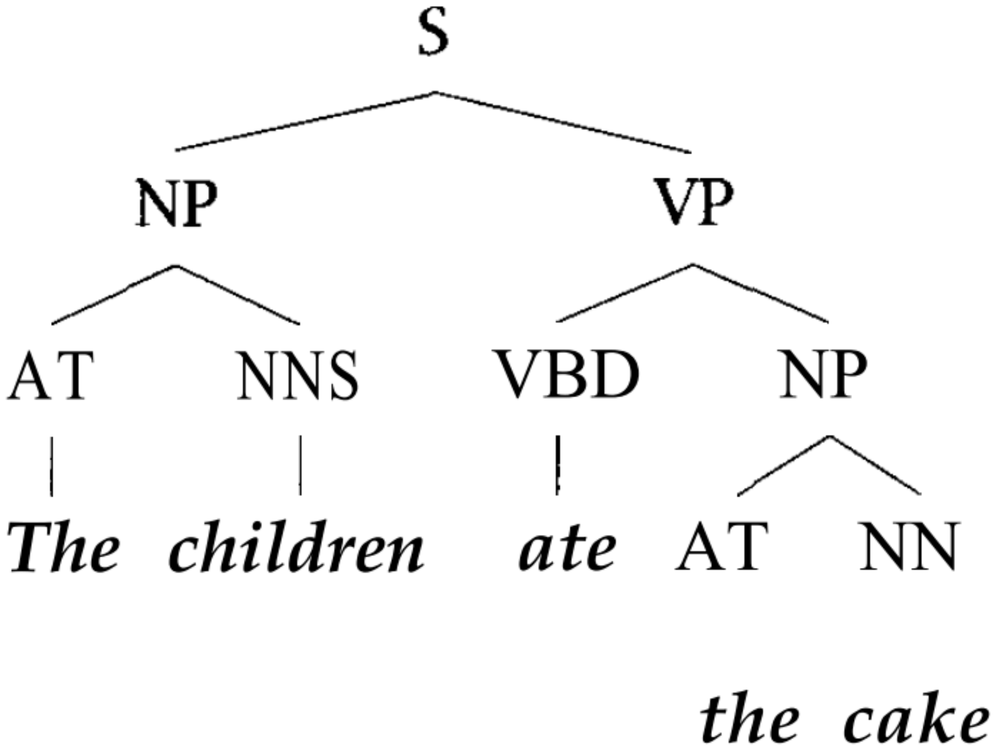
Manning and Schutze (1999, p.95)

Tree Structure of Sentences

(3.42)

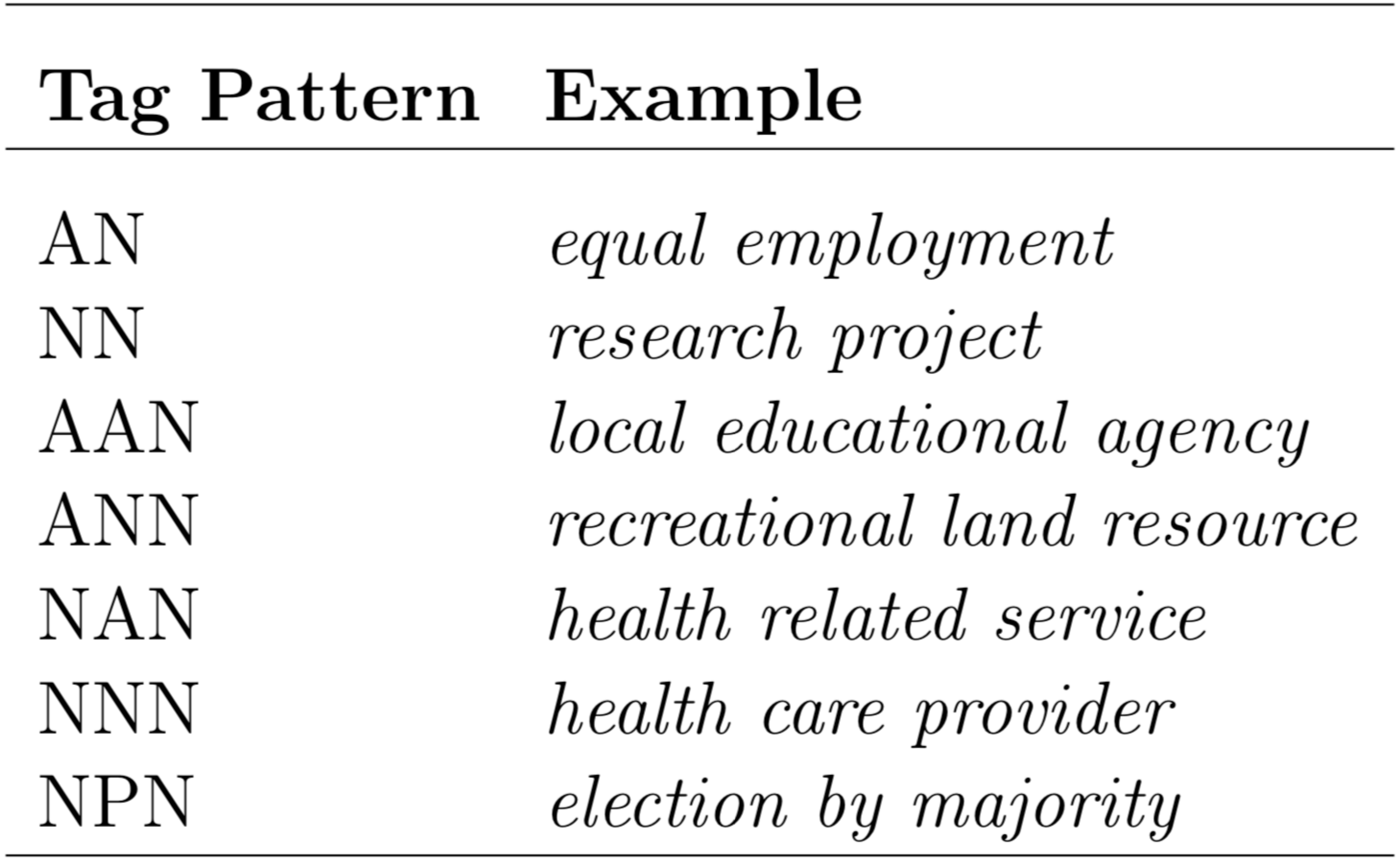


(3.43)



Manning and Schutze (1999, p.95)

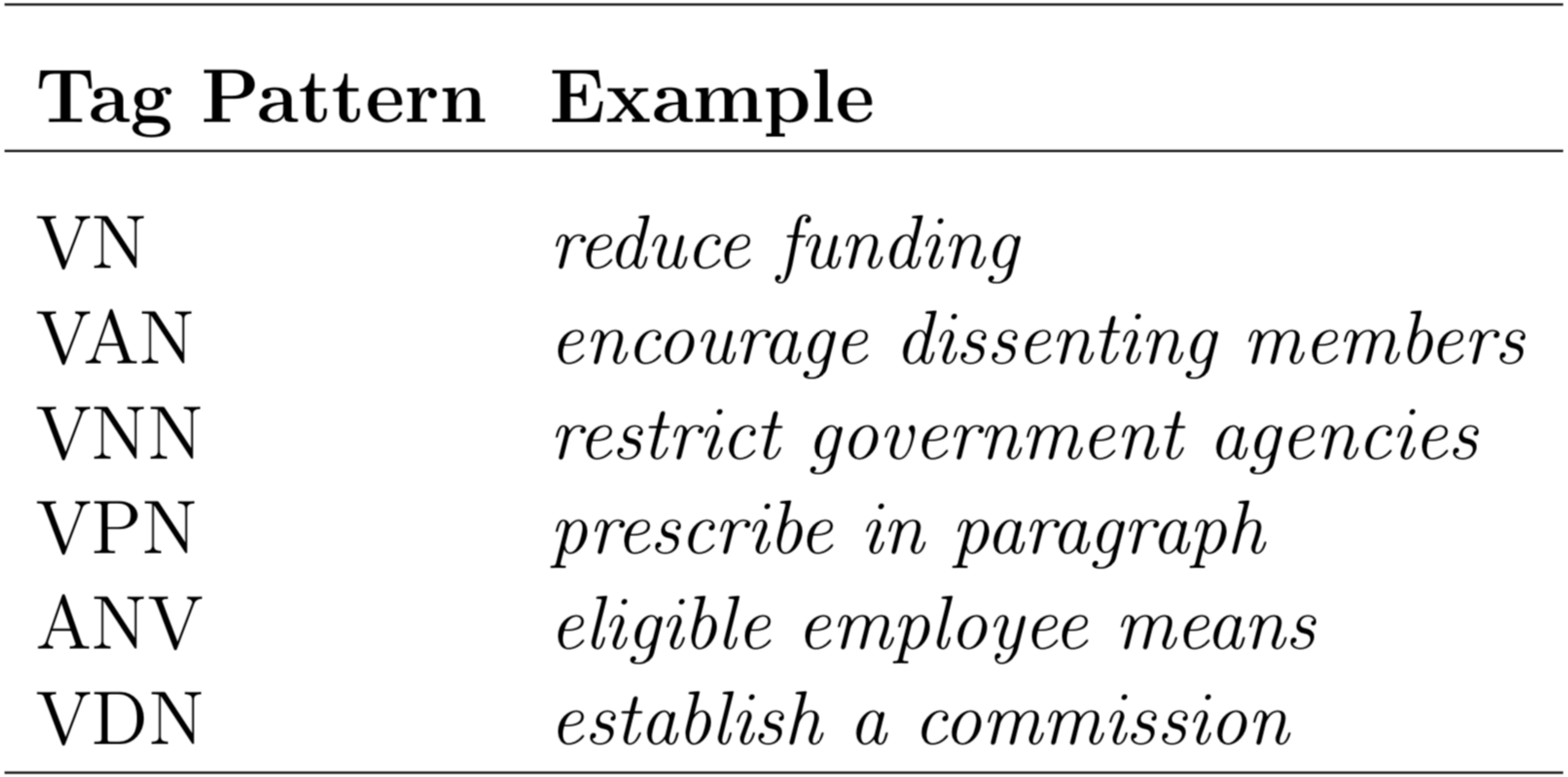
POS Tag Patterns → Noun Phrases



Tag Pattern	**Example**
AN	*equal employment*
NN	*research project*
AAN	*local educational agency*
ANN	*recreational land resource*
NAN	*health related service*
NNN	*health care provider*
NPN	*election by majority*

(Justeson and Katz, 1995)

POS Tag Patterns → Verb Phrases



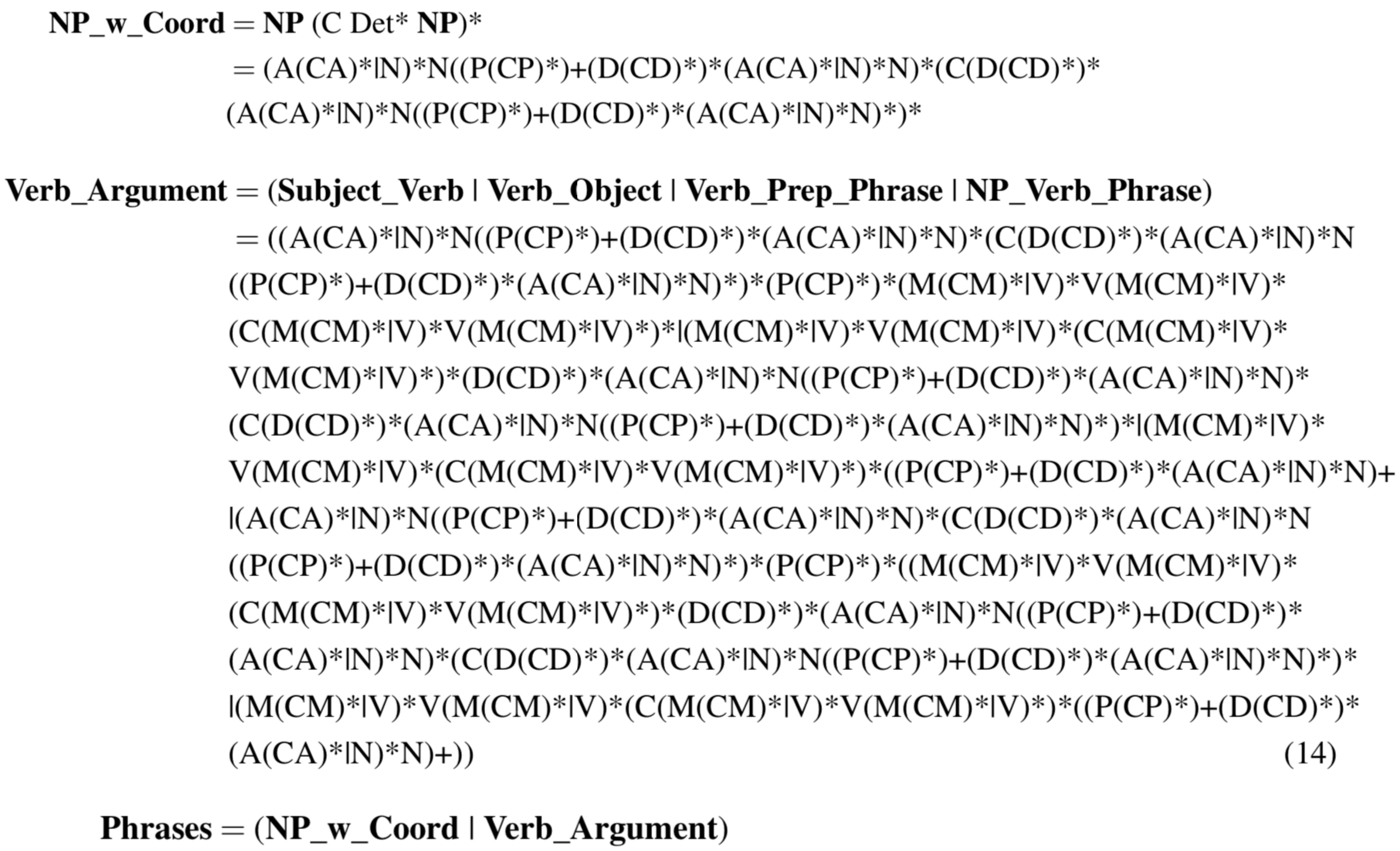
Tag Pattern	**Example**
VN	*reduce funding*
VAN	*encourage dissenting members*
VNN	*restrict government agencies*
VPN	*prescribe in paragraph*
ANV	*eligible employee means*
VDN	*establish a commission*

Regular Expressions over POS Tags

- **Noun Phrases**: (A|N)*N(PD*(A|N)*N)*
  - Zero or more adjectives or nouns, followed by a noun, followed (optionally) by zero or more groups of terms containing a preposition and zero or more determiners, then zero or more adjectives or nouns, and ending in a noun.

- **Verb Groups**: (M(CM)*|V)*V(M(CM)*|V)*
  - Modifier followed by zero or more coordinating conjunction-modifier pairs, or a verb, all repeated zero or more times, then a verb, then a modifier followed by zero or more coordinating conjunction-modifier pairs, or a verb, all repeated zero or more times.

Full Generalization



NP_w_Coord = **NP** (C Det* **NP**)*
= (A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*(C(D(CD)*)*
(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*)*

Verb_Argument = (**Subject_Verb** | **Verb_Object** | **Verb_Prep_Phrase** | **NP_Verb_Phrase**)
= ((A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*(C(D(CD)*)*(A(CA)*|N)*N
((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*)*(P(CP)*)*(M(CM)*|V)*V(M(CM)*|V)*
(C(M(CM)*|V)*V(M(CM)*|V)*)*|(M(CM)*|V)*V(M(CM)*|V)*(C(M(CM)*|V)*
V(M(CM)*|V)*)*(D(CD)*)*(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*
(C(D(CD)*)*(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*)*|(M(CM)*|V)*
V(M(CM)*|V)*(C(M(CM)*|V)*V(M(CM)*|V)*)*((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)+
|(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*(C(D(CD)*)*(A(CA)*|N)*N
((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*)*(P(CP)*)*((M(CM)*|V)*V(M(CM)*|V)*
(C(M(CM)*|V)*V(M(CM)*|V)*)*(D(CD)*)*(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*
(A(CA)*|N)*N)*(C(D(CD)*)*(A(CA)*|N)*N((P(CP)*)+(D(CD)*)*(A(CA)*|N)*N)*)*
|(M(CM)*|V)*V(M(CM)*|V)*(C(M(CM)*|V)*V(M(CM)*|V)*)*((P(CP)*)+(D(CD)*)*
(A(CA)*|N)*N)+)) (14)

Phrases = (**NP_w_Coord** | **Verb_Argument**)

The Readings This Week

- Manning & Schtze, H. (1999). Foundations of Statistical Natural Language Processing. Chapter 3.
- Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(01).
- Handler, A., Denny, M. J., Wallach, H., & OConnor, B. (2016). Bag of What? Simple Noun Phrase Extraction for Text Analysis. EMNLP + CSS