BC COMS 2710: Computational Text Analysis

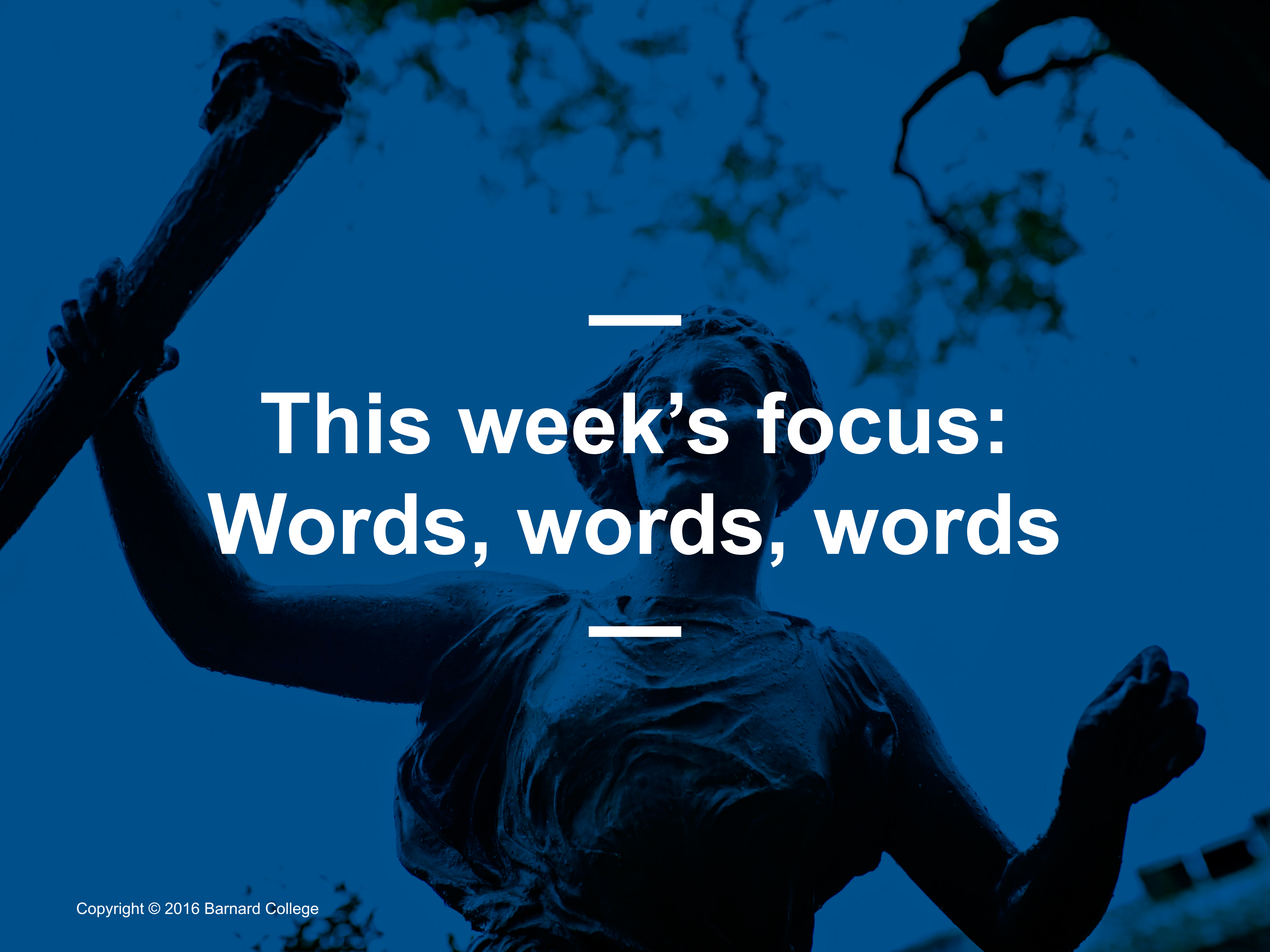
Lecture 5 – What’s in a word

Announcements – Assignments

- Homework 01
  - Due tomorrow night
- Readings:
  - Reading 01 – due last night, get it in ASAP if haven't
  - Reading 02 – link course site, due Sunday
- Week 2 Tutorials:
  - 2.1 – Tokenization, lemmatization, stopwords, etc
    - Based on today's lecture
  - 2.2 – Exploring dictionary-based methods
    - Based on Wednesday's lecture

Scheduling Announcement
Monday 05/17 & Tuesday 05/18

- Gauri will be lecturing on Regular Expressions and holding open hours during course time

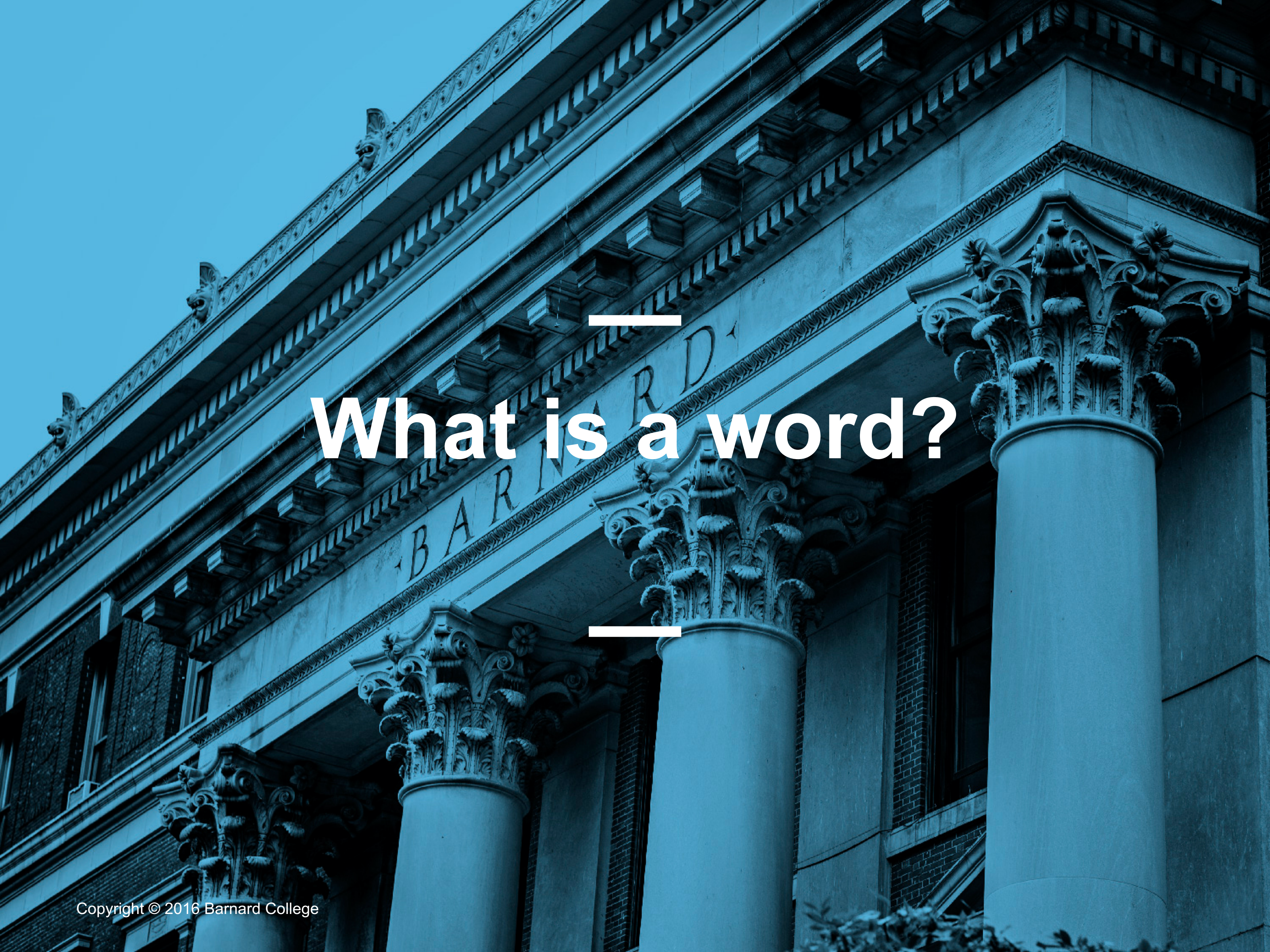


This week's focus: Words, words, words

Why focus on words?

- Words suggest meaning
- If we can identify words, we can count them
- If we we can count words, we can quantify (aspects of) a text that contains those words.
- If we can quantify a text, we can compute with it.
  - Answer quantitative questions about text
- Caveat:
  - Quantifying a text isn't the same thing as being *correct* about what that text means, nor is meaning solely a function of word counts(!).

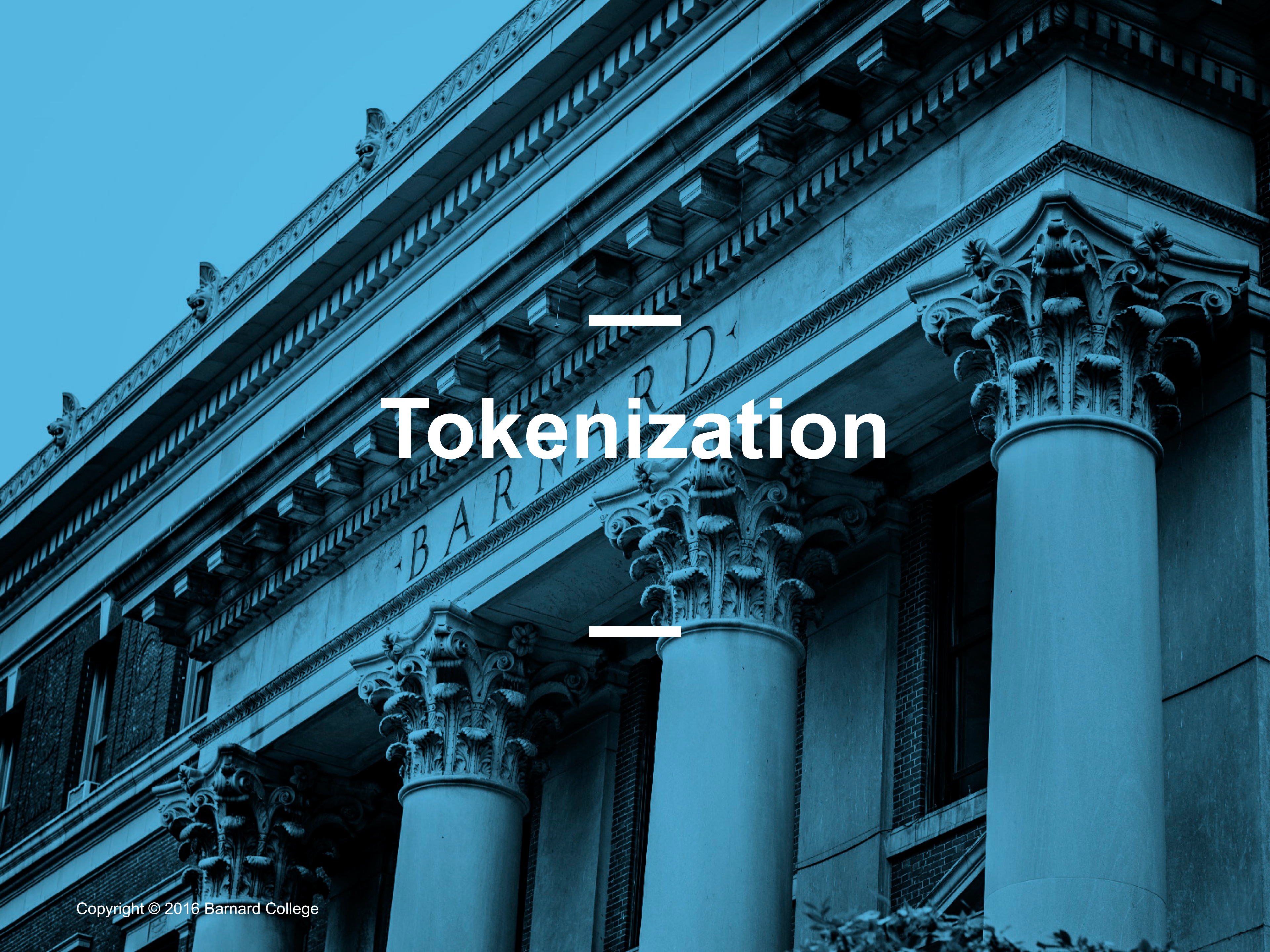
Matthew Wilkens - https://mattwilkens.com/



What is a word?

Outline

- Tokenization
- Lemmatization
- Stemming
- Stopwords
- Part of Speech
- Dependency Parsing
- Named Entities



Tokenization

Tokenization

“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”

Chapter 5
Basic Text Processing In: Text Mining:
A Guidebook for the Social Sciences

Tokenization - Example

“Mr. Smith doesn’t like apples.”

How many tokens are in the sentence?

Tokenization - Example

“Mr. Smith doesn’t like apples.”

> *“The process of identifying the words in the input sequence of characters, mainly by **separating the punctuation marks** but also by identifying contractions, abbreviations, and so forth”*

Tokenization - Example

“Mr. Smith doesn’t like apples.”

*“The process of identifying the words in the input sequence of characters, mainly by **separating the punctuation marks** but also by identifying contractions, abbreviations, and so forth”*

Tokenization - Example

“Mr. Smith doesn’t like apples.”

“The process of identifying the words in the input sequence of characters, mainly by separating the punctuation marks but also by identifying contractions, abbreviations, and so forth”

Tokenization - Example

“Mr. Smith doesn’t like apples.”

Mr.

Smith

does

n’t

like

apples

.

Type vs Token

- **Type**: An element of the vocabulary

- **Token**: an instance of a type in the text

- ***N*** = number of tokens
- ***V*** = vocabulary, i.e. set of tokens
- |***V***| = size of Vocabulary

Type vs Token

- **Type**: An element of the vocabulary

- **Token**: an instance of a type in the text

“We refuse to believe that there are insufficient funds in the great vaults of opportunity of this nation. And so we've come to cash this check, a check that will give us upon demand the riches of freedom and the security of justice”

- Q: How many types, tokens?

Lemmatization & Stemming

Lemmatization

"reduces the inflectional forms of a word to its root form"

Chapter 5

Basic Text Processing In: Text Mining:

A Guidebook for the Social Sciences

boys ->

children ->

am, are, is ->

Lemmatization - example

I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression will be ***transformed*** *into an oasis of freedom and justice.*

With this faith we will be able to ***transform*** *the jangling discords of our nation into a beautiful symphony of brotherhood.*

Stemming

“applies a set of rules to an input word to remove suffixes and prefixes and obtain its stem, which will now be shared with other related words.”

Chapter 5

Basic Text Processing In: Text Mining:

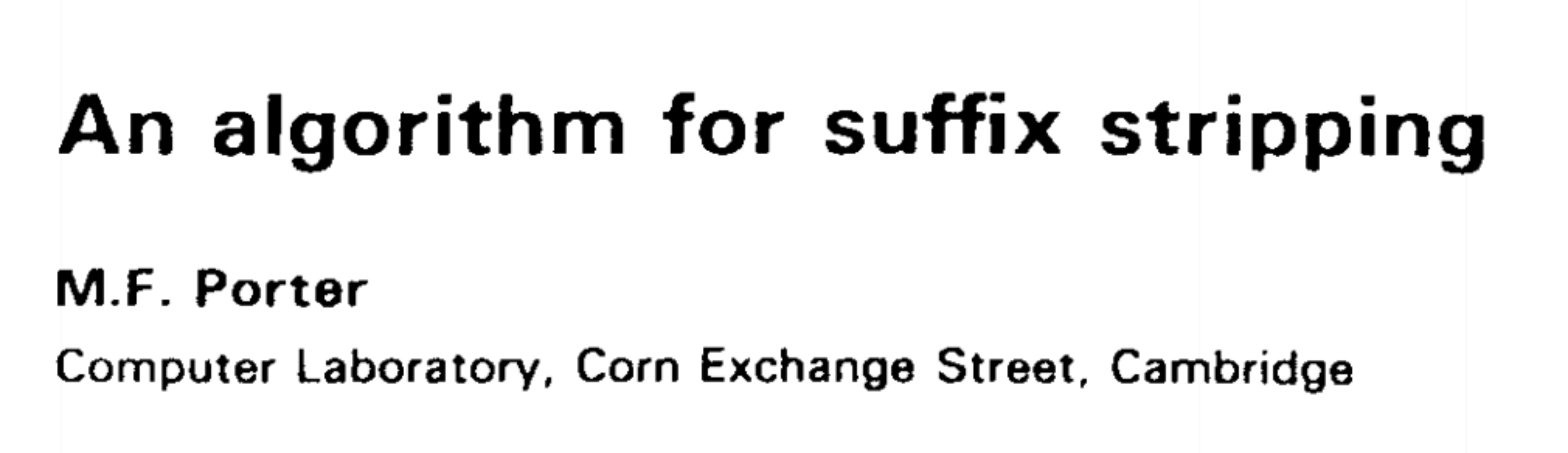
A Guidebook for the Social Sciences

“more radical way to reduce variation”

Chapter 2

Dirk Hovy textbook

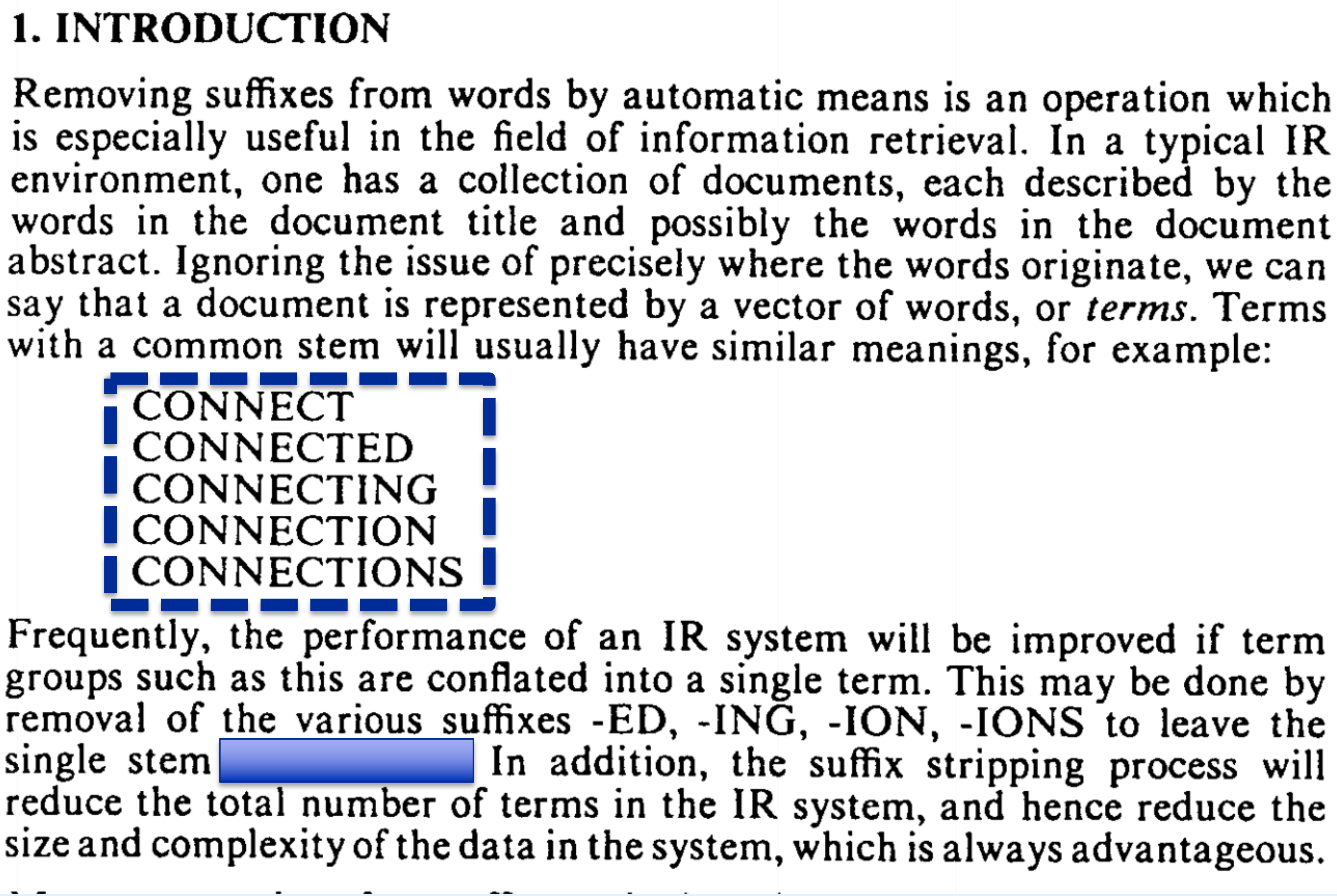
Porter Algorithm for Stemming



An algorithm for suffix stripping

M.F. Porter

Computer Laboratory, Corn Exchange Street, Cambridge



1. INTRODUCTION

Removing suffixes from words by automatic means is an operation which is especially useful in the field of information retrieval. In a typical IR environment, one has a collection of documents, each described by the words in the document title and possibly the words in the document abstract. Ignoring the issue of precisely where the words originate, we can say that a document is represented by a vector of words, or *terms*. Terms with a common stem will usually have similar meanings, for example:

CONNECT
CONNECTED
CONNECTING
CONNECTION
CONNECTIONS

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single stem [redacted] In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous.

Porter Stemming Explained

"For each language, it defines a number of suffixes (i.e., word endings) and the order in which they should be removed or replaced. By repeatedly applying these actions, we reduce all words to their stems."

Chapter 2
Dirk Hovy textbook

https://www.cs.toronto.edu/~frank/csc2501/Readings/R2_Porter/Porter-1980.pdf

Stemming Example

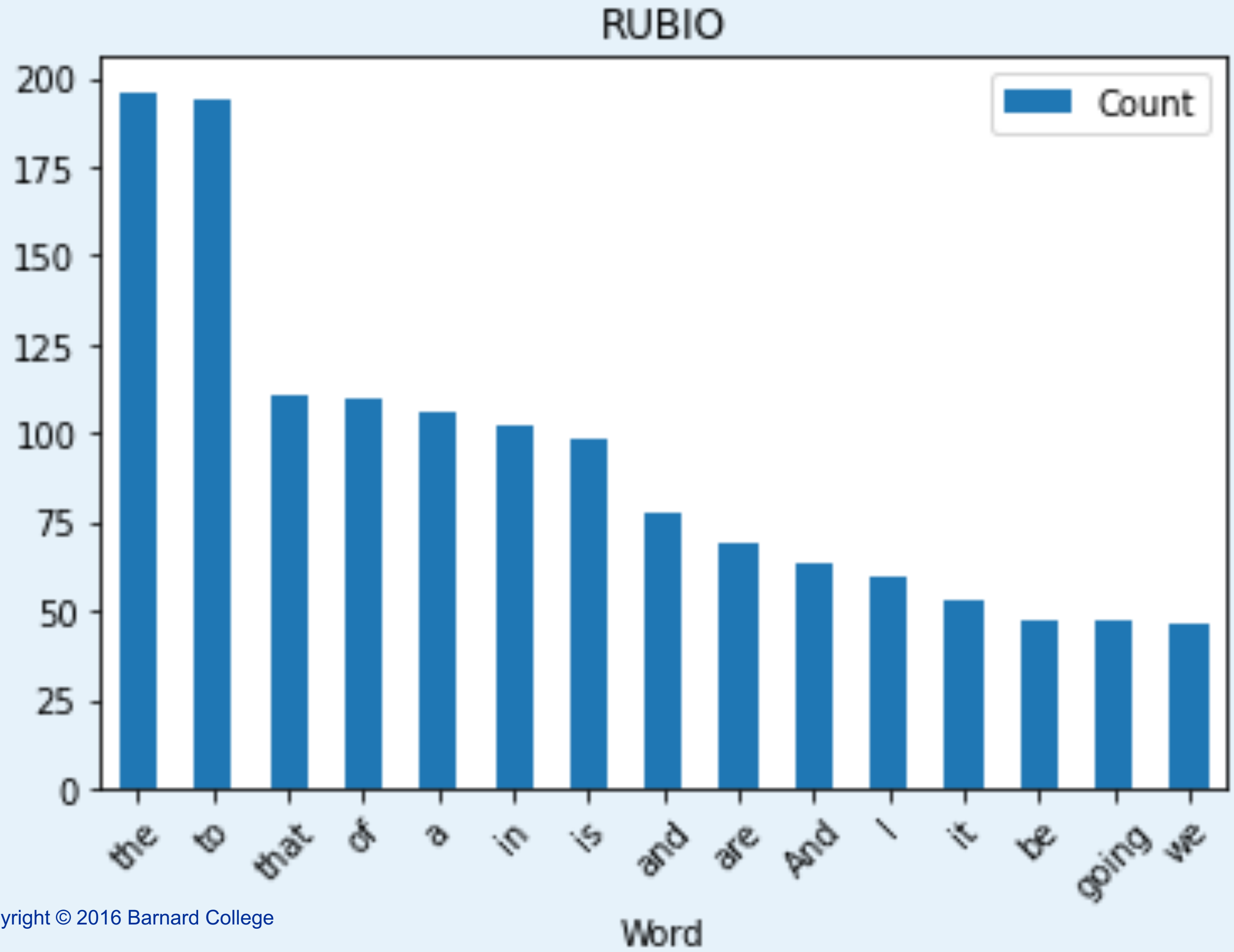
This was not the map we found in Billy Bones's chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.

Thi wa not the map we found in Billi Bone s chest but an accur copi complet in all thing name and height and sound with the singl except of the red cross and the written note .

Example from:
https://web.stanford.edu/~jurafsky/slp3/slides/2_TextProc_Mar_25_2021.pdf

Stop Words

Frequency of Rubio’s terms in 2016 Miami debate

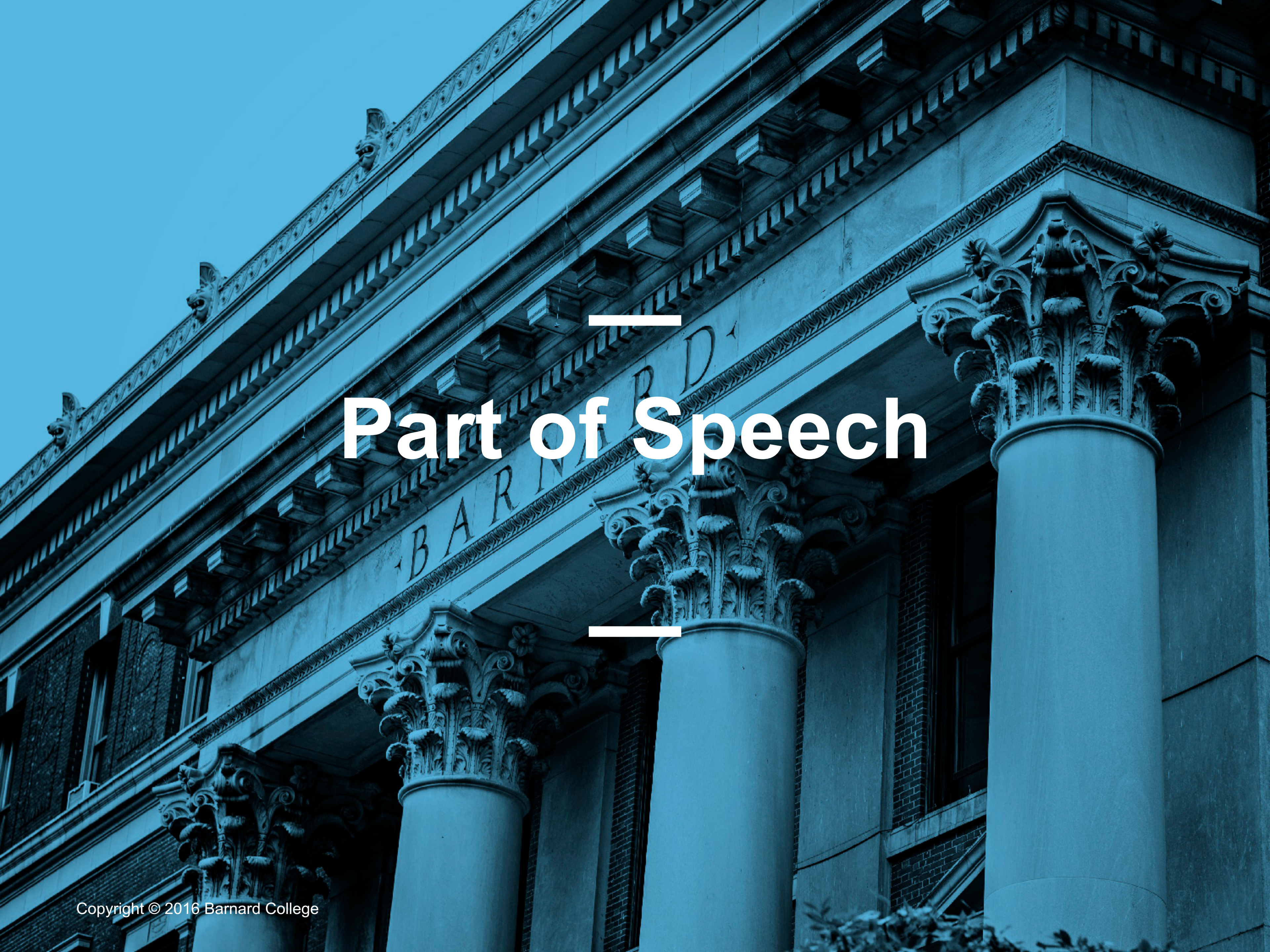


Stopwords

“set of ignorable words that occur often, but not contribute much to our task, so it can be beneficial to remove.”

Chapter 2

Dirk Hovy textbook



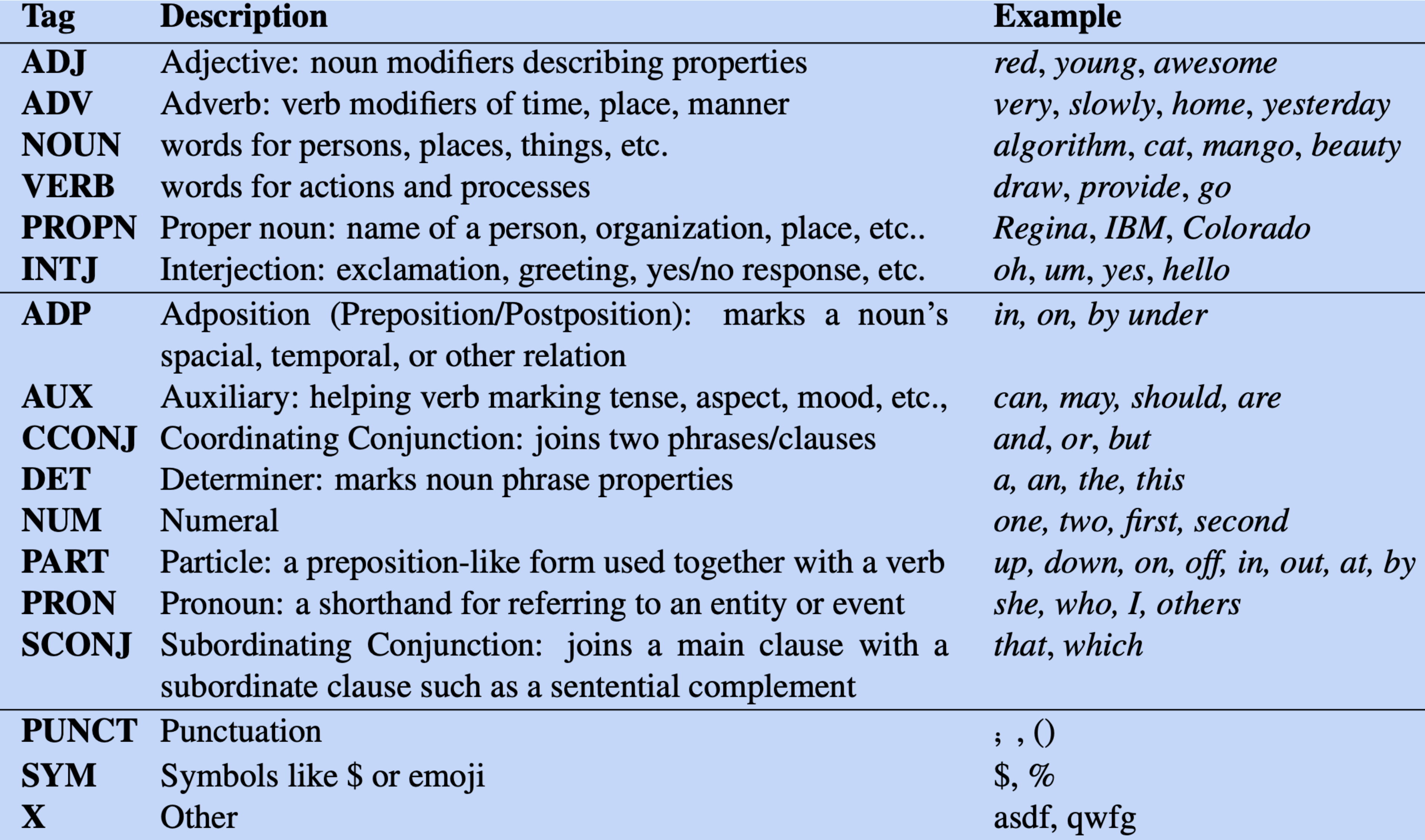
Part of Speech

Part of Speech

- Categorize words based on their grammatical properties

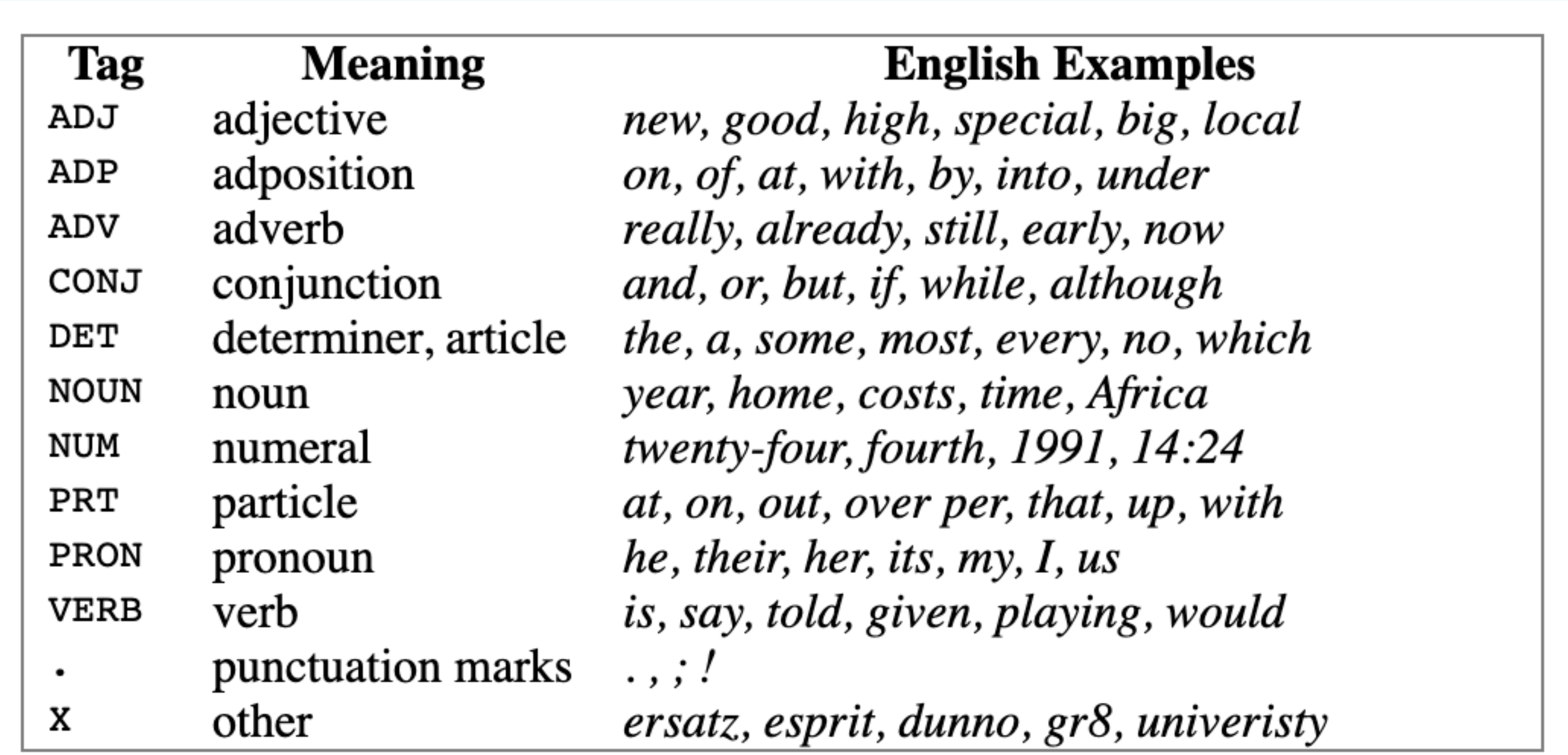
- Part-of-speech tagging:
  - Process of identifying the grammatical category of tokens in a corpus

Universal Tag Set



Tag	Description	Example
ADJ	Adjective: noun modifiers describing properties	*red*, *young*, *awesome*
ADV	Adverb: verb modifiers of time, place, manner	*very*, *slowly*, *home*, *yesterday*
NOUN	words for persons, places, things, etc.	*algorithm*, *cat*, *mango*, *beauty*
VERB	words for actions and processes	*draw*, *provide*, *go*
PROPN	Proper noun: name of a person, organization, place, etc..	*Regina*, *IBM*, *Colorado*
INTJ	Interjection: exclamation, greeting, yes/no response, etc.	*oh*, *um*, *yes*, *hello*
ADP	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	*in, on, by under*
AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	*can, may, should, are*
CCONJ	Coordinating Conjunction: joins two phrases/clauses	*and*, *or*, *but*
DET	Determiner: marks noun phrase properties	*a, an, the, this*
NUM	Numeral	*one, two, first, second*
PART	Particle: a preposition-like form used together with a verb	*up, down, on, off, in, out, at, by*
PRON	Pronoun: a shorthand for referring to an entity or event	*she, who, I, others*
SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	*that*, *which*
PUNCT	Punctuation	; , ()
SYM	Symbols like $ or emoji	$, %
X	Other	asdf, qwfg

Simplified Tag set



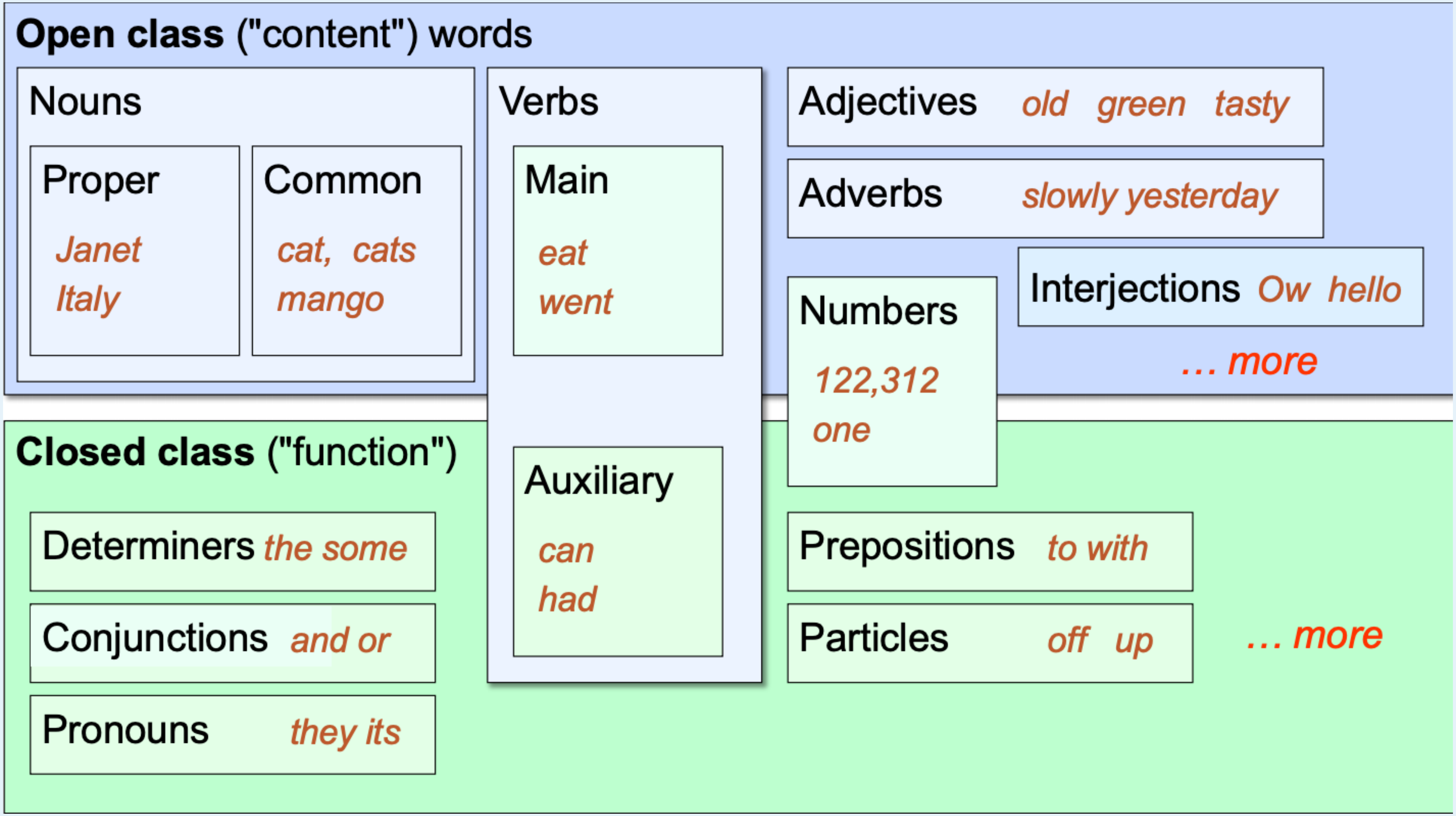
Tag	Meaning	English Examples
ADJ	adjective	*new, good, high, special, big, local*
ADP	adposition	*on, of, at, with, by, into, under*
ADV	adverb	*really, already, still, early, now*
CONJ	conjunction	*and, or, but, if, while, although*
DET	determiner, article	*the, a, some, most, every, no, which*
NOUN	noun	*year, home, costs, time, Africa*
NUM	numeral	*twenty-four, fourth, 1991, 14:24*
PRT	particle	*at, on, out, over per, that, up, with*
PRON	pronoun	*he, their, her, its, my, I, us*
VERB	verb	*is, say, told, given, playing, would*
.	punctuation marks	*. , ; !*
X	other	*ersatz, esprit, dunno, gr8, univeristy*

Word Classes: Open vs Closed

- Closed class words
  - Relatively fixed membership
  - Usually **function** words: short, frequent words with grammatical function
    - determiners: *a, an, the*
    - pronouns: *she, he, I*
    - prepositions: *on, under, over, near, by, …*
- Open class words
  - Usually **content** words: Nouns, Verbs, Adjectives, Adverbs
    - Plus interjections: oh, ouch, uh-huh, yes, hello
  - New nouns and verbs like iPhone or to fax

Slide from https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pdf

Word Classes Graphic



Open class ("content") words

Nouns

Proper *Janet* *Italy*

Common *cat, cats* *mango*

Verbs

Main *eat* *went*

Auxiliary *can* *had*

Adjectives *old green tasty*

Adverbs *slowly yesterday*

Numbers *122,312* *one*

Interjections *Ow hello*

… more

Closed class ("function")

Determiners *the some*

Conjunctions *and or*

Pronouns *they its*

Prepositions *to with*

Particles *off up*

… more

Slide from https://web.stanford.edu/~jurafsky/slp3/slides/8_POSNER_intro_May_6_2021.pdf

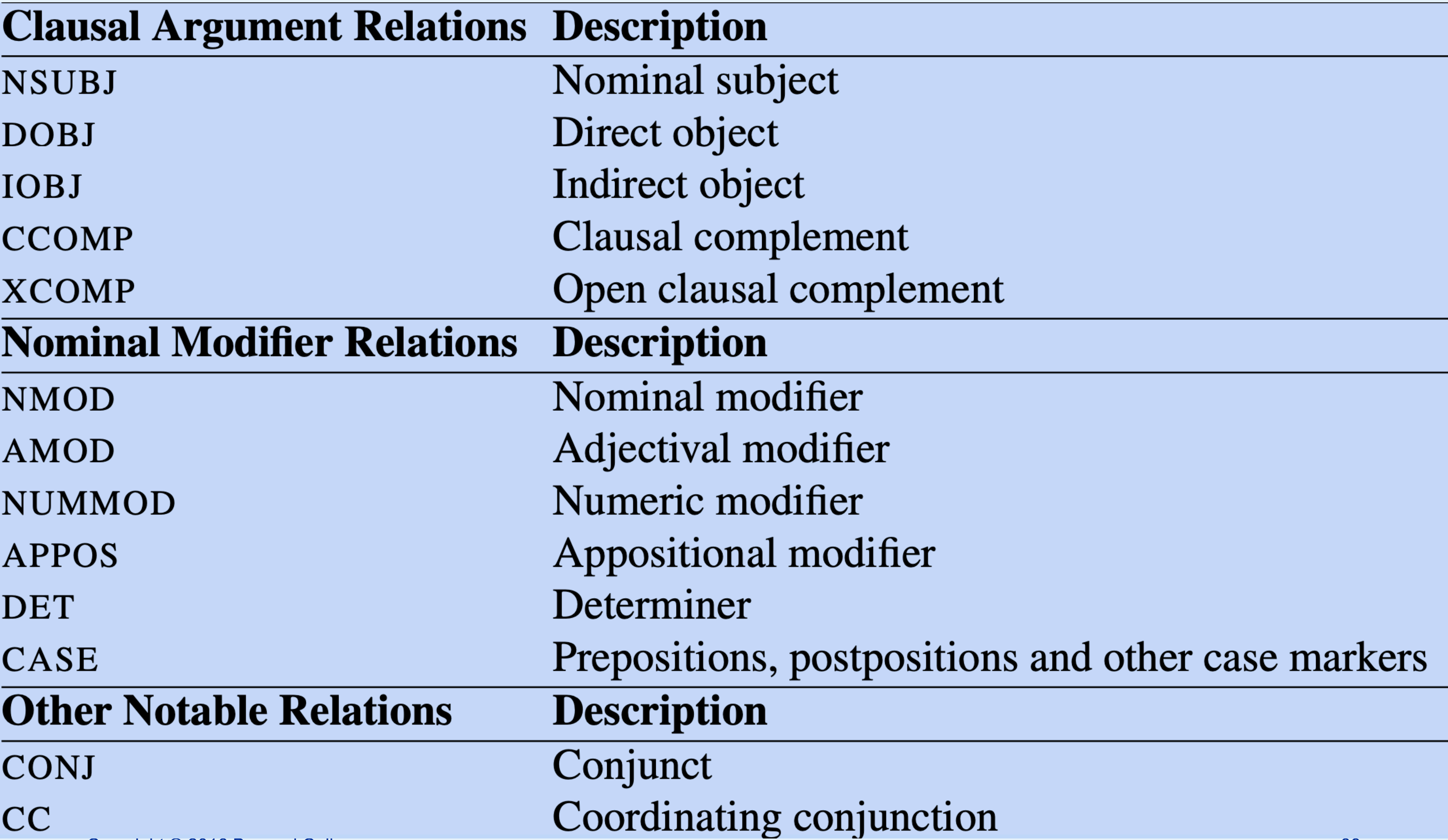
Dependency Parsing

Dependency Parsing - Idea

The idea in dependency grammar is that the sentence “hangs” off the main verb like a mobile. The links between words describe how the words are connected.

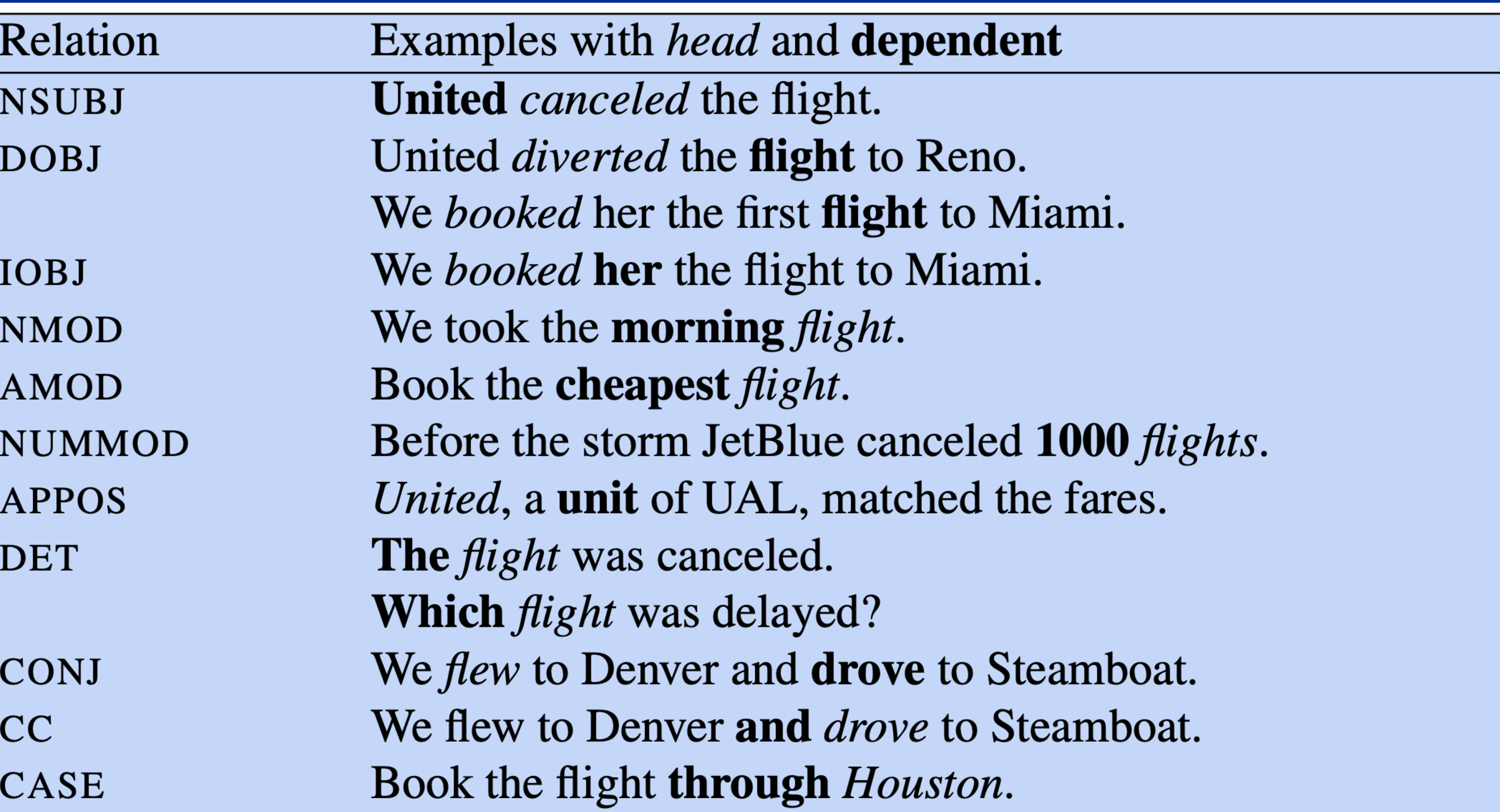
Chapter 2
Dirk Hovy textbook

Universal DP Tags



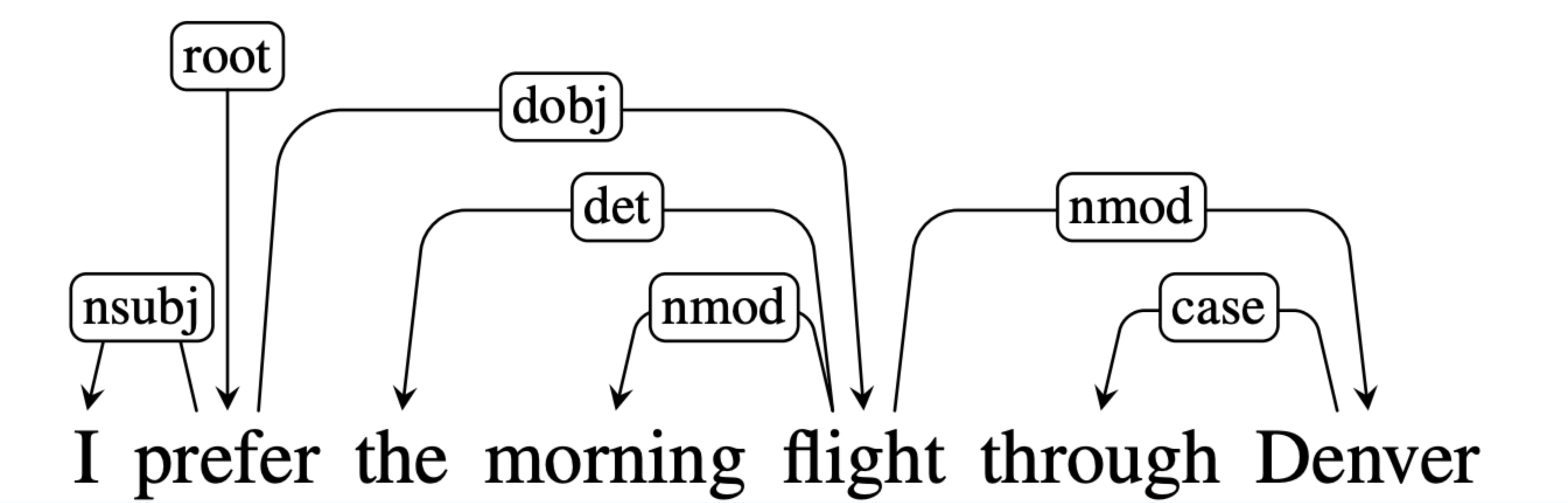
Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

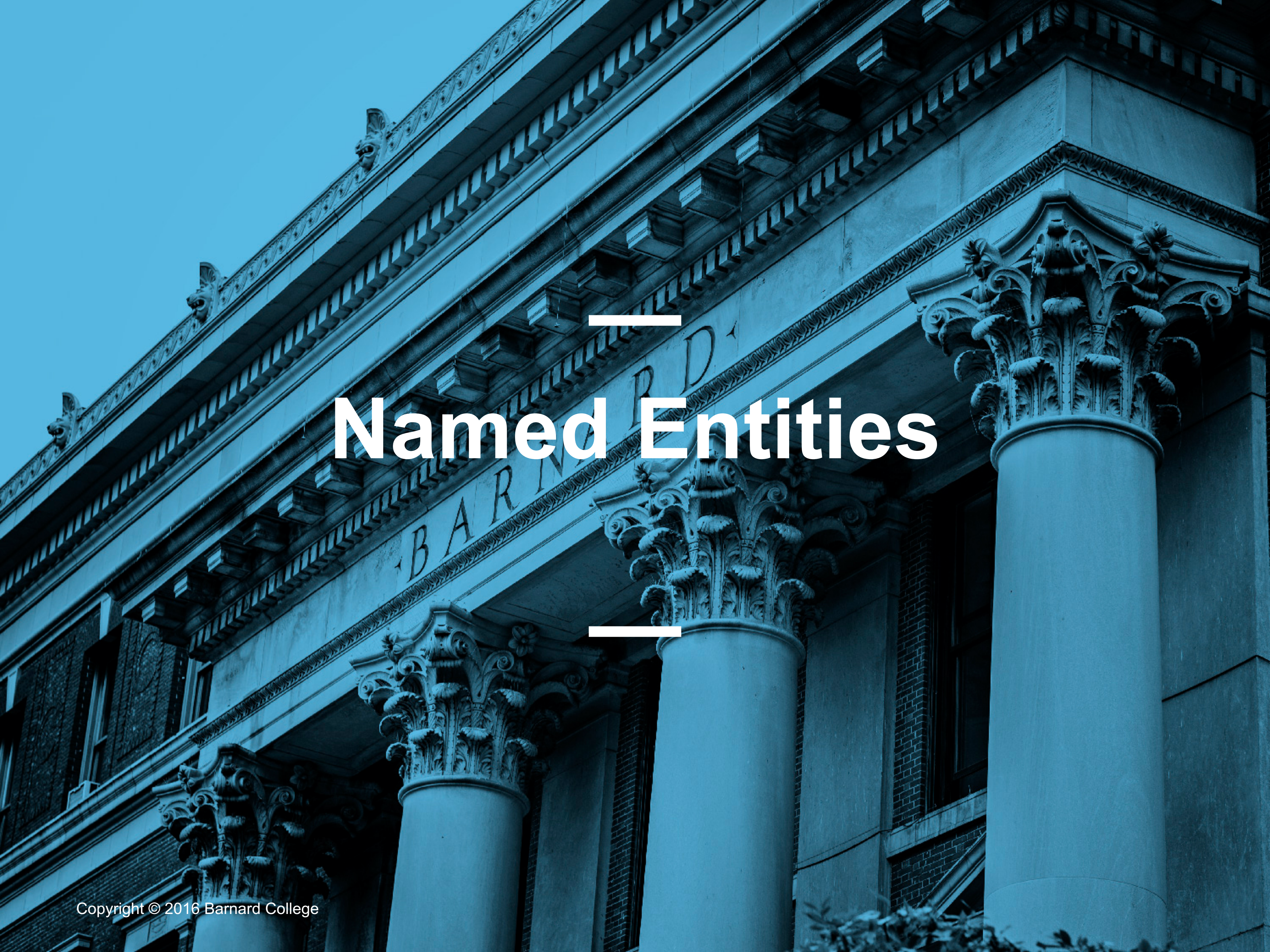
Examples of tags



Relation	Examples with *head* and **dependent**
NSUBJ	**United** *canceled* the flight.
DOBJ	United *diverted* the **flight** to Reno.
	We *booked* her the first **flight** to Miami.
IOBJ	We *booked* **her** the flight to Miami.
NMOD	We took the **morning** *flight*.
AMOD	Book the **cheapest** *flight*.
NUMMOD	Before the storm JetBlue canceled **1000** *flights*.
APPOS	*United*, a **unit** of UAL, matched the fares.
DET	**The** *flight* was canceled.
	Which *flight* was delayed?
CONJ	We *flew* to Denver and **drove** to Steamboat.
CC	We flew to Denver **and** *drove* to Steamboat.
CASE	Book the flight **through** *Houston*.

Dependency Parsing - Example





Named Entities

Named Entity Recognition

- Classify words into predefined categories:
  - persons
  - organizations
  - locations
  - expressions of times
  - quantities
  - monetary values
  - percentages

Slide from Federico Nanni

Named Entity Recognition

- Classify words into predefined categories:
  - persons
  - organizations
  - locations
  - expressions of times
  - quantities
  - monetary values
  - percentages

Monday, October 30, Hillary Clinton will present her book in Chicago at the University of Chicago.

Slide from Federico Nanni

Named Entity Recognition

- Classify words into predefined categories:
  - persons
  - organizations
  - locations
  - expressions of times
  - quantities
  - monetary values
  - percentages

Monday, October 30, Hillary Clinton will present her book in Chicago at the University of Chicago.

Slide from Federico Nanni

Approaches for NER

- regular expression to extract:
- Gazetteers
- Patters
- Machine Learning

Slide from Federico Nanni

Approaches for NER – Regular Expressions

- Extract:
  - telephone numbers
  - E-mails
  - Dates
  - Prices
  - Locations (e.g., word + “river” indicates a river -> Hudson river)

Slide from Federico Nanni

Approaches for NER - Gazetteers

- Dictionaries or list of proper names of:
  - Person
  - Location
  - Organization

Slide from Federico Nanni

Approaches for NER – Context Patterns

- context patterns, such as:
  - [Person] earns [Money]
  - [PERSON] joined [ORGANIZATION]
  - [PERSON] fly to [LOCATION]

Slide from Federico Nanni