WORDS

WORD COUNTING

- How many unique words are in the passage?

- Which words appear the most?

- How many words only appear once in the text?

WHAT IS A WORD?

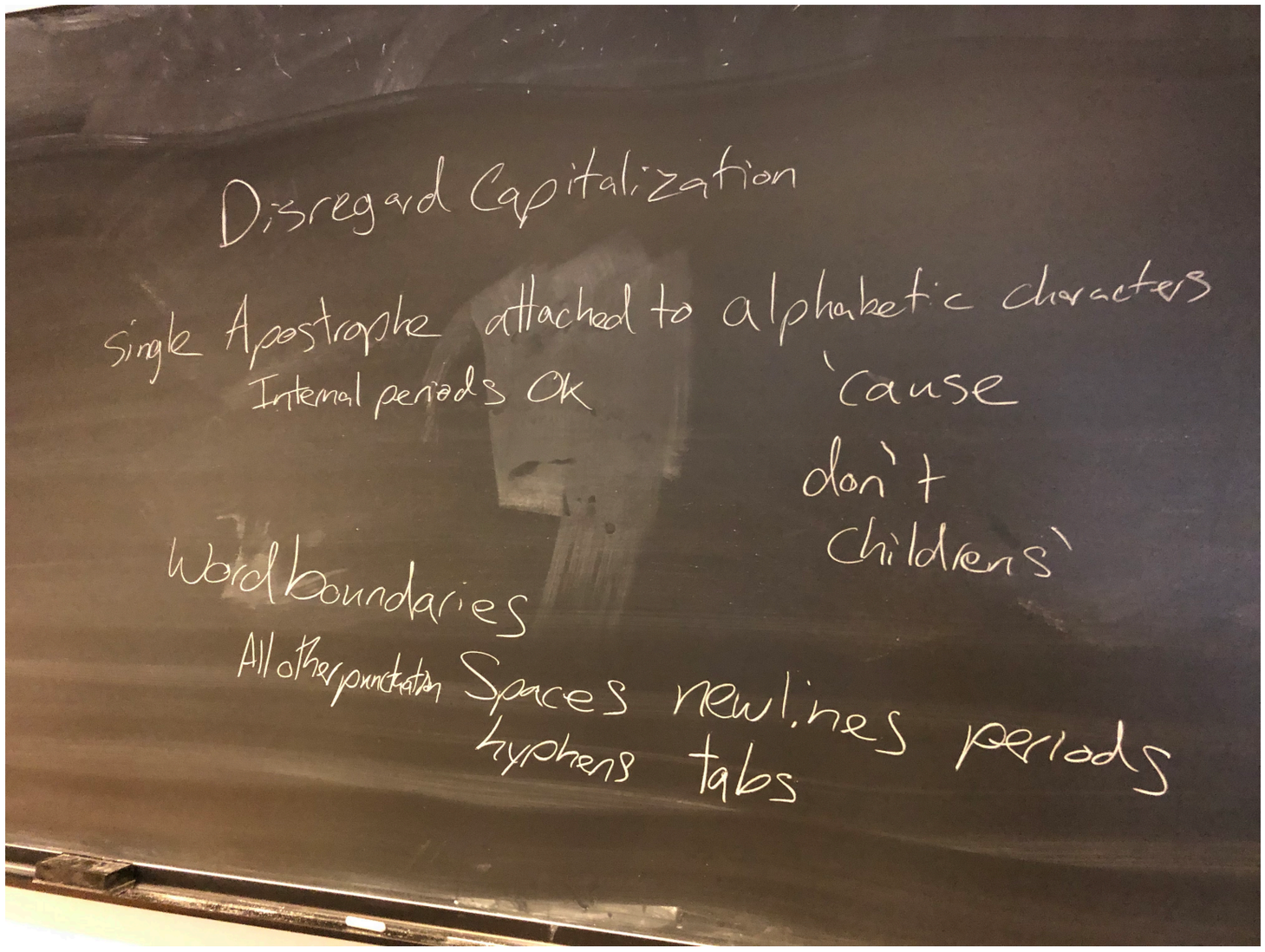
- sharp-edged
  - 1 word or 2?
- don’t
  - 1 word or 2?
- occasion / occasions
  - Are they the same word?

TO WHAT EXTEND DOES REPRESENTATION DEFINE A WORD?

- **Bill** always picks up the **bill** when we eat out.
  - Are these the same word?
- **Bill** me now or **bill** me later
  - What about these?
- **I** remember Mrs. Hubble as a little curly sharp-edged person in sky-blue, who held a conventionally juvenile position, because she had married Mr. Hubble,—**I** don't know at what remote period,--when she was much younger than he.
  - Are these the same word?

CLASS DEFINITION OF A WORD

- Does capitalization matter?
- What (if any) punctuation should we keep?
- What do we define as a word boundary?



MORPHEMES

- A **morpheme** is the smallest unit of meaning in a word.
- Morphemes are classified into two main types:
  - Free morphemes can stand alone
  - Bound morphemes must appear with other morphemes, usually as a prefix, suffix, or infix

prefix	stem	suffix
	dog	
	wait	ed
un	kind	
de	forest	ation

LEMMAS

- A **lexeme** is a set of words that share the same meaning and base morpheme.
- The **lemma** (or **headword**) is the form that represents the entire lexeme. This is usually the form found in the dictionary.

Lemma	Lexeme			
fly	fly	flys	flew	flying
run	run	runs	ran	running
quiet	quiet	quiets	quieted	quietly
go	go	goes	went	going

WORD STEMS

➤ A **word stem** is the part of the word that doesn't change, even as the form changes

Word	Lemma	Word stem
running	run	run
flying	fly	fli
produced	produce	produc
go	go	go
went	go	went

WHY??

Suppletion, the use of one word as the inflected form of another, when the two words have different backgrounds.
Go -> from Old English *gan*

Went -> from Old English *wenden* (origin of 'to wend')

LEMMATIZATION

- Lemmatization is the process of automatically converting words into their lemma.
- This process is hard!
  - Need to determine part of speech/meaning through context first

I walked out of the last meeting feeling like it went better than I expected

I walk out of the last meeting feel like it go good than I expect

STEMMING

- Stemming is the process of converting words to their stems automatically
- This process is way easier!

I walked out of the last meeting feeling like it went better than I expected

I walk out of the last meet feel like it went better than I expect

STEMMING

we dine on these occas in the kitchen, and adjourn, for the nut and orang and appl to the parlor; which wa a chang veri like joe's chang from hi working-cloth to hi sunday dress. my sister wa uncommonli live on the present occas, and inde wa gener more graciou in the societi of mrs. hubbl than in other compani. i rememb mrs. hubbl as a littl curli sharp-edg person in sky-blu, who held a convent juvenil posit, becaus she had marri mr. hubbl, -- i don't know at what remot period, -- when she wa much younger than he . i rememb mr hubbl as a tough, high-should, stoop old man, of a sawdusti fragranc, with hi leg extraordinarili wide apart : so that in my short day i alway saw some mile of open countri between them when i met him come up the lane. i walk out of the last meet feel like it went better than i expect

PORTER STEMMER

Written as a series of steps, where each word can match up to 1 rule per step.

Step 1a

SSES -> SS	caresses -> caress
IES -> I	ponies -> poni
	ties -> ti
SS -> SS	caress -> caress
S ->	cats -> cat

PORTER STEMMER

Written as a series of steps, where each word can match up to 1 rule per step.

Step 1b

(C?(VC)+V?) EED ->	EE	feed -> feed
		agreed -> agree
(*v*) ED ->		plastered -> plaster
		bled -> bled
(*v*) ING ->		motoring -> motor
		sing -> sing