



# Applied Natural Language Processing

Info 256

Lecture 21: Multiword expressions (April 11, 2019)

David Bamman, UC Berkeley

# Words

- One morning I shot an elephant in my pajamas
- I didn't shoot an elephant
- **Imma** let you finish but Beyonce had one of the best videos of all time
- I do uh main- mainly business data processing
- 一天早上我穿着睡衣射了一只大象

- The White House pledged to cut down the red tape for access to public documents.

# Multiword expressions

- The **White House** pledged to **cut down** the **red tape** for access to public documents

# Multiword expressions

The

White\_House

pledged

to

cut\_down

the

red\_tape

for

access

to

public

documents

# Multiword expressions

type	examples
MW compounds	red tape, motion picture, daddy longlegs, hot air balloon, trash talk
verb-particle	pick up, dry out, take over, cut short, hold hostage, take seriously
verb-noun	pay attention (to), go bananas, lose it, break a leg, make the most of
support verbs	make decisions, take breaks, take pictures, have fun, perform surgery
coordination	cut and dried/dry, more or less, up and leave
connective	as well as, let alone, in spite of, on the face of it/on its face
fixed phrase	easy as pie, scared to death, go to hell, bring home the bacon
proverbs	Beggars can't be choosers. The early bird gets the worm.

# Multiword expressions

- Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical **idiomaticity**

# Predictability

- The meaning and behavior of multiword expressions is typically not predictable from the individual words that comprise it.

- “dog”

- “top”

- “days”

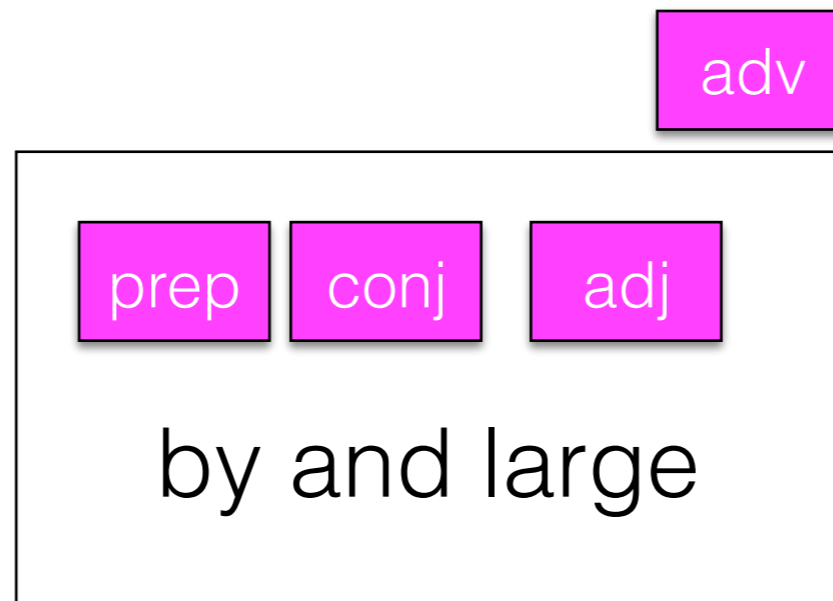
- “dog days”

- “top dog”



# Syntactic idiomaticity

- The syntax of the MWE isn't predictable from its components



# Semantic idiomatcity

- The meaning of a MWE is not predictable from its components

kick



- **S: (v)** kick (drive or propel with the foot)
- **S: (v)** kick (thrash about or strike out with the feet)
- **S: (v)** kick (strike with the foot) *"The boy kicked the dog"; "Kick the door down"*
- **S: (v)** kick (kick a leg up)
- **S: (v)** **kick back**, **recoil**, kick (spring back, as from a forceful thrust) *"The gun kicked back into my shoulder"*
- **S: (v)** kick, **give up** (stop consuming) *"kick a habit"; "give up alcohol"*
- **S: (v)** kick (make a goal) *"He kicked the extra point after touchdown"*
- **S: (v)** **complain**, **kick**, **plain**, **sound off**, **quetch**, **kvetch** (express complaints, discontent, displeasure, or unhappiness) *"My mother complains all day"; "She has a lot to kick about"*

bucket



- **S: (n)** **bucket**, **pail** (a roughly cylindrical vessel that is open at the top)
- **S: (n)** **bucket**, **bucketful** (the quantity contained in a bucket)

kick the bucket



- **S: (v)** **die**, **decease**, **perish**, **go**, **exit**, **pass away**, **expire**, **pass**, **kick the bucket**, **cash in one's chips**, **buy the farm**, **conk**, **give-up the ghost**, **drop dead**, **pop off**, **choke**, **croak**, **snuff it** (pass from physical life and lose all bodily attributes and functions necessary to sustain life) *"She died from cancer"; "The children perished in the fire"; "The patient went peacefully"; "The old guy kicked the bucket at the age of 102"*

# Pragmatic idiomaticity

- An MWE is associated “with a fixed set of situations or particular context”

good morning!	Fixed greeting used at same time of day
all aboard!	used in specific situation of boarding a train/ship
shock and awe	fixed phrased associated with specific moment in Iraq War

# Lexical idiomaticity

- At least one component of the MWE doesn't appear in the vocabulary on its own.

ad hoc	“created or done for a particular purpose as necessary”
--------	---

- Neither “ad” nor “hoc” are English words on their own.

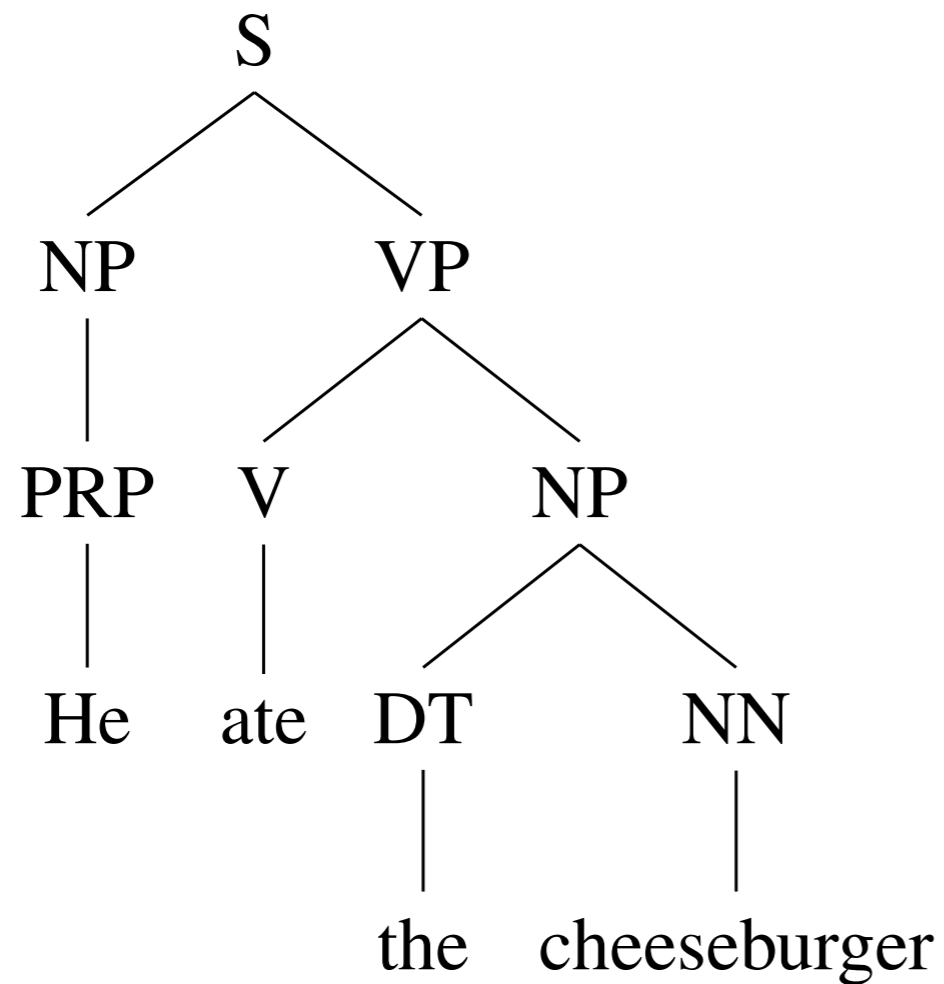
# Statistical idiomaticity

- The words in a MWE occur unusually frequently together compared to their individual frequency.

	flawless	immaculate	impeccable	spotless
condition	+	−	+	+
credentials	−	−	+	−
hair	−	+	?	−
house	?	+	?	+
logic	+	−	+	−
timing	?	+	+	−

*Note:* “+” = strong lexical affinity, “?” = neutral lexical affinity, “−” = negative lexical affinity.

# Compositionality

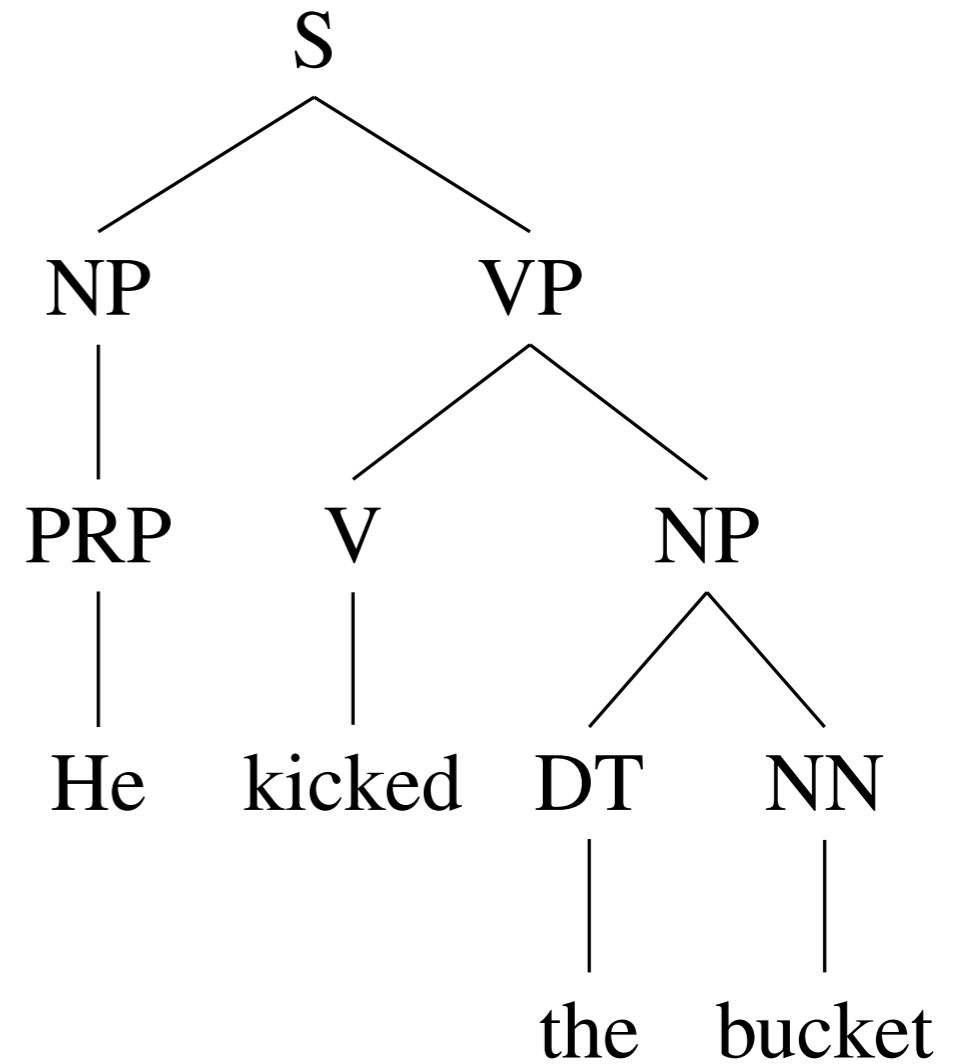
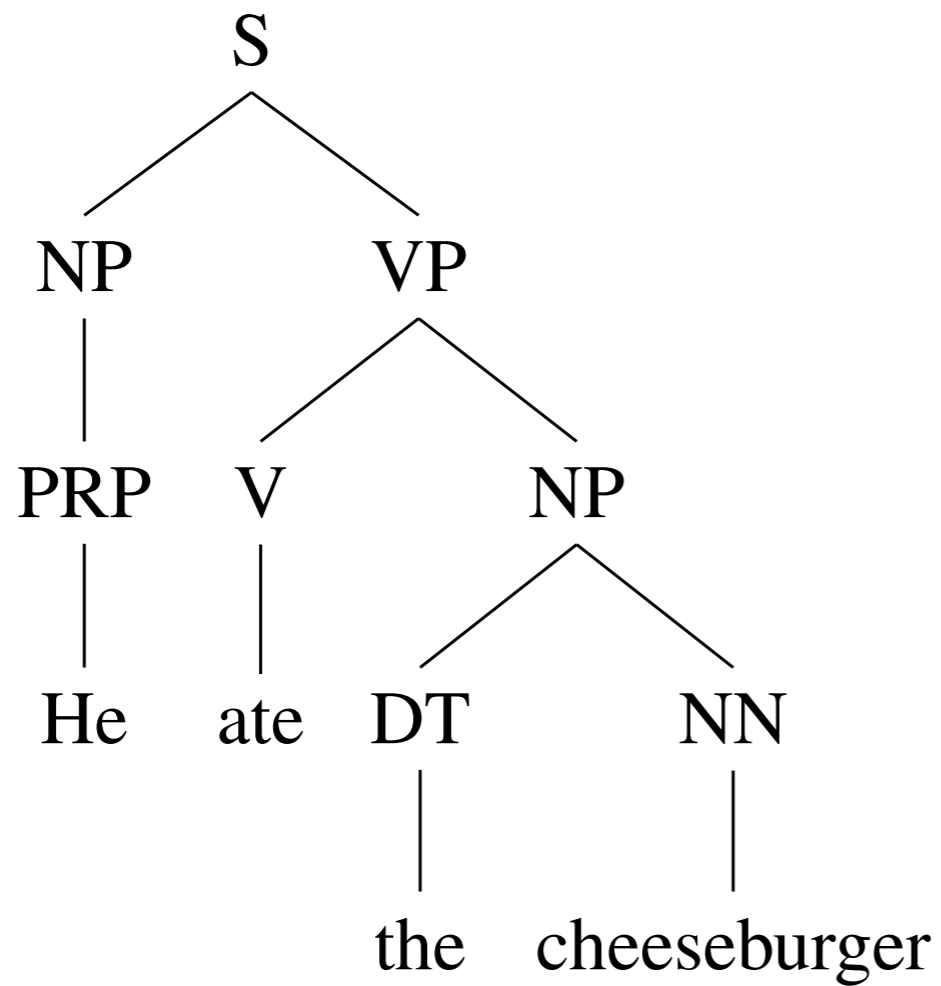


We can build up the meaning of a sequence by the combination of its parts

He  
the cheeseburger  
ate  
ate the cheeseburger  
He ate the cheeseburger

# Compositionality

idioms resist  
compositionality



# MWE dictionaries

Random sample of WordNet MWEs:

- WordNet contains multiword entries

arctic willow  
blade apple  
cardiac valve  
de bakey  
glycerol tristearate  
line of descent  
madagascar cat  
vaginal discharge  
western red cedar  
works program





Wiktionary  
The free dictionary

Main Page  
Community portal  
Preferences  
Requested entries  
Recent changes  
Random entry  
Help  
Glossary  
Donations  
Contact us

Tools

What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Cite this page

Visibility

Show translations  
Show quotations

In other languages

Not logged in [Talk](#) [Contributions](#) [Preferences](#) [Create account](#) [Log in](#)

Entry

[Discussion](#)

[Citations](#)

Read

[Edit](#)

[History](#)

# red tape

## Contents [\[hide\]](#)

### 1 English

#### 1.1 Etymology

#### 1.2 Noun

##### 1.2.1 Synonyms

##### 1.2.2 Translations

#### 1.3 See also

##### 1.3.1 Usage notes

#### 1.4 Anagrams

## English [\[edit\]](#)

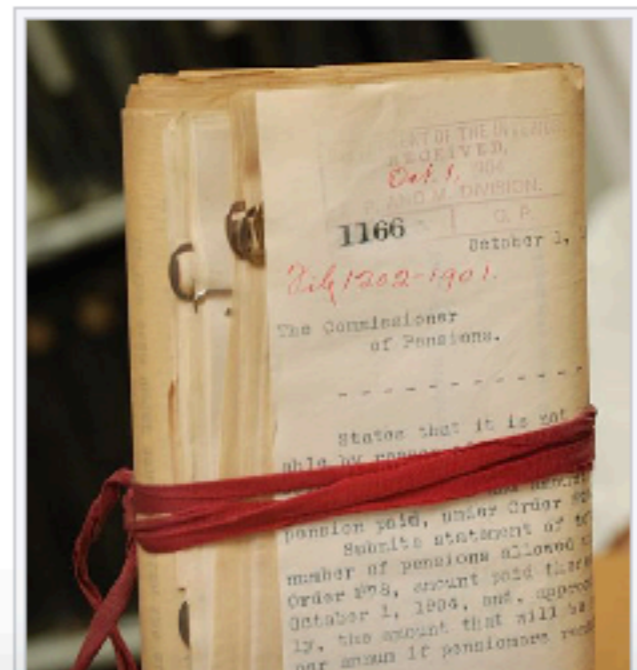
### Etymology [\[edit\]](#)

- Thought to allude to the former practice of binding government documents in red-coloured tape

### Noun [\[edit\]](#)

**red tape** (*uncountable*)

- The binding tape once used for holding important documents together. [\[quotations ▼\]](#)
- (metonymically, idiomatic)* Time-consuming regulations or bureaucratic procedures.



# WikiMWE

- 350,000 MWEs (2-4grams) of technical terminology and collocations

- Anchor Text (Internal Links):  
[[target|term\_candidate]]
- Section Headers:  
===\* term\_candidate ===\*
- Phrases in Boldface: "**term\_candidate**"
- Phrases in Italics: "*term\_candidate*"

His administration filed briefs that urged the [[Supreme Court of the United States|**Supreme Court**]] to strike down [[Same-sex marriage in the United States|**same-sex marriage**]] bans as unconstitutional (''[[United States v. Windsor]]'' and ''[[Obergefell v. Hodges]]'');



Search for phrases...

SEARCH

# A B C D E F G H I J K L M N O P Q R S T U V W X Y Z NEW RANDOM

The Web's Largest Resource for Phrases, Verbs & Idioms

A MEMBER OF THE STANDS4 NETWORK



Like 496

# Welcome to Phrases.com

Phrases.com is a large collection of common phrases, casual expressions and idioms – collaboratively assembled by our contributing editors. Entries can be easily browsed, searched for, rated, heard and translated to several common and not so common languages.

Improve your English and writing skills by navigating our comprehensive phrases dictionary **alphabetically**, or simply search by **keywords**.

## Hot

Our most popular phrases



ma vaillante fille 263

I'll see you and ... 117

ain't no mountain... 106

you can't educate 105

Rate it: ★★☆☆☆ (3.00 / 1 vote)

# stand for

To tolerate.

2 Views

COLLECTION EDIT

# MWE Extraction

- In many cases, existing MWE lexical don't cover the specific MWE present in a new domain.
- Several methods for extracting MWE from a corpus.

# Collocations

“An arbitrary and recurrent word combination”

[Benson 1990; Baldwin and Kim 2010]

$$\chi^2$$

- $\chi^2$  (chi-square) is a statistical test of dependence—here, dependence between the two variables of word 1 identity and word 2 identity.
- For assessing the difference in two datasets, this test assumes a 2x2 contingency table:

	word 1	$\neg$ word 1
word 2	7	104023
$\neg$ word 2	104	251093

$$\chi^2$$

To test whether “white house” is a meaningful collocation, we can ask: does the word *house* occur **significantly** more frequently after *white*?

	$w_1 = \text{white}$	$w_1 = \neg \text{white}$	
$w_2 = \text{house}$	104	1004	“red house”, “my house”
$w_2 = \neg \text{house}$	2	13402	“red car”, “my dog”

“white dog”, “white truck”

$$\chi^2$$

For each cell in contingency table, sum the squared difference between observed value in cell and the expected value assuming independence.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$



	$w_1 = \text{white}$	$w_1 = \neg \text{white}$	sum	frequency
$w_2 = \text{house}$	104	1004	1108	0.076
$w_2 = \neg \text{house}$	2	13402	13404	0.924
sum	106	14406		
frequency	0.007	0.993		

Assuming independence:

$$P(w_1 = \text{white}, w_2 = \text{house}) = P(w_1 = \text{white}) \times P(w_2 = \text{house}) \\ = 0.007 \times 0.076 = 0.00053$$

Among 14512 words, we would expect to see 7.69 occurrences of *white house*.

	$w_1 = \text{white}$	$w_1 = \neg \text{white}$		
$w_2 = \text{house}$	7.69	1095.2	$P(w_2 = \text{house})$	0.076
$w_2 = \neg \text{house}$	93.9	13315.2	$P(w_2 = \neg \text{house})$	0.924

$P(w_1 = \text{white})$   $P(w_1 = \neg \text{white})$

0.007      0.993

$$\chi^2$$

- What  $\chi^2$  is asking is: how different are the observed counts different from the counts we would expect given complete independence?

	$w_1 = \text{white}$	$w_1 = \neg \text{white}$
$w_2 = \text{house}$	104	1004
$w_2 = \neg \text{house}$	2	13402

	$w_1 = \text{white}$	$w_1 = \neg \text{white}$
$w_2 = \text{house}$	7.69	1095.2
$w_2 = \neg \text{house}$	93.9	13315.2

$$\chi^2$$

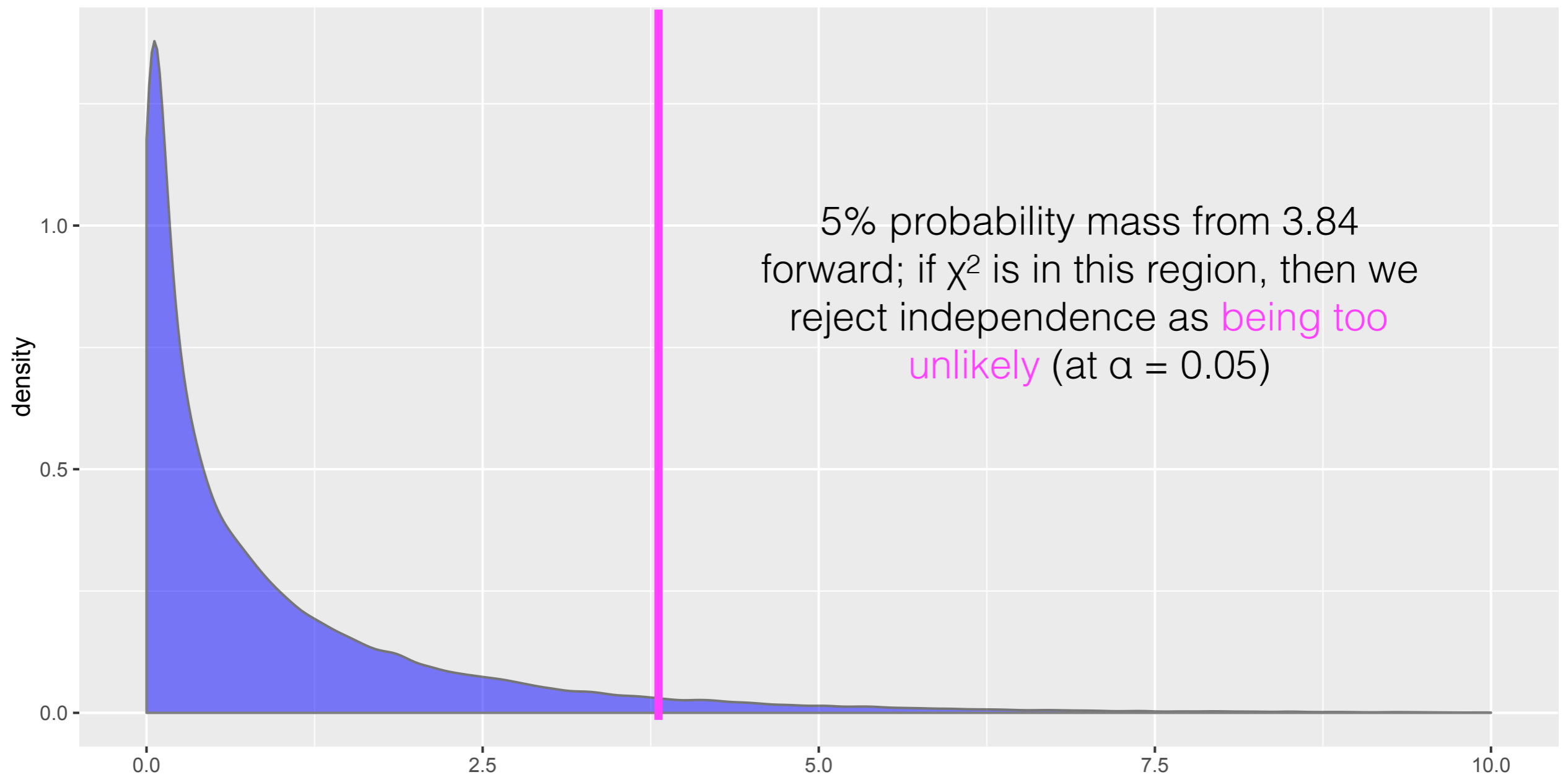
- With algebraic manipulation, simpler form for 2x2 table  $O$  (cf. Manning and Schütze 1999)

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

$$\chi^2$$

- The  $\chi^2$  value is a statistic of dependence with a probability governed by a  $\chi^2$  distribution; if this value has low enough probability in that measure, we can reject the null hypothesis of the independence between the two variables.

$\chi^2$



$$\chi^2$$

- Chi-square is ubiquitous in corpus linguistics (and in NLP as a measure of collocations).
- A few caveats for its use:
  - Each cell should have an *expected* count of at least 5
  - Each observation is independent

Why is part of speech tagging useful?



# POS indicative of MWE

at least one adjective/noun or noun phrase

and definitely  
one noun

$$((A | N)^+ | ((A | N)^*(NP))(A | N)^*)N$$

*AN*: linear function; lexical ambiguity; mobile phase

*NN*: regression coefficients; word sense; surface area

*AAN*: Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase

*ANN*: cumulative distribution function; lexical ambiguity resolution; accessible surface area

*NAN*: mean squared error; domain independent set; silica based packing

*NNN*: class probability function; text analysis system; gradient elution chromatography

*NPN*: degrees of freedom; [*no example*]; energy of adsorption

# MWE prediction

- Many phrases are ambiguous about whether they are a MWE in context.
  - **the white house** pledged to reduce red tape
  - he lives in **the white house** on the corner
  - Kim **made a face** at the policeman.
  - Kim **made a face** in pottery class.

# MWE prediction

- Data: 55,000 tokens of web reviews annotated for MWE *in context*.

I googled restaurants in the area and Fuji\_Sushi came\_up and  
reviews were great so I made\_ a carry\_out \_order

<https://github.com/nert-nlp/streusle/>

# BIO notation

Standard BIO entity notation

**no gaps,** *he was willing to budge a little on the price which means a lot to me .*  $(0|BI^+)^+$   
**1-level** 0 0 0 0 0 B I 0 0 0 0 B I I I I 0

Expanded BIO to accomodate one layer of nesting

**gappy,** *he was willing to budge a little on the price which means a lot to me .*  $(0|B(o|bi^+|I)^*I^+)^+$   
**1-level** 0 0 0 0 B b i I 0 0 0 B I I I I 0

# MWE prediction

<i>preexisting lexicons</i>	entries	max gap length	LOOKUP				SUPERVISED MODEL			
			$\bar{P}$	$\bar{R}$	$\bar{F}_1$	$\sigma$	$\bar{P}$	$\bar{R}$	$\bar{F}_1$	$\sigma$
none	0						74.39	44.43	55.57	2.19
WordNet + SemCor	71k	0	<u>46.15</u>	28.41	35.10	2.44	74.51	45.79	56.64	1.90
6 lexicons	420k	0	35.05	46.76	<u>40.00</u>	2.88	<u>76.08</u>	<u>52.39</u>	<u>61.95</u>	1.67
10 lexicons	437k	0	33.98	<u>47.29</u>	39.48	2.88	75.95	51.39	61.17	2.30
best configuration with in-domain lexicon		1	<b>46.66</b>	<b>47.90</b>	<b>47.18</b>	2.31	<b>76.64</b>	51.91	61.84	1.65
			2 lexicons + $MWtypes(train)_{\geq 1}$				6 lexicons + $MWtypes(train)_{\geq 2}$			

Schneider et al. (2014), "Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut" (TACL)

# Activity

- [13.mwe/JustesonKatz95.ipynb](#)