THE LANGUAGE OF FOOD

Dan Jurafsky
Linguistics and Computer Science
Stanford University

IPAM, May 23, 2016
America’s national foods

- Hamburger
- Frankfurter
- French fries
- Ketchup
Europe

Hamburg

Frankfurt

French
But what about ketchup?
Fish preservation 2000 years ago

- Mon-Khmer, Tai-Kadai, Hmong-Mien speakers in Southeast China/Southeast Asia

Preserving fish to last through dry season
- Fish, rice, salt, lactic acid fermentation
200 BCE: China expands

Han Dynasty: Emperor Wu invades tribal Guangdong and Fujian.
- Assimilating locals and their fermented seafoods
- Pushing the rest south
Innovations in Japan

700 CE: fermented fish rice ➔ Japan called *sushi* (technically *narezushi*)

18th century: Use vinegar instead of lactic fermentation (and eat the rice)

19th century: eat the fish fresh
Meanwhile...
Southern Chinese sail/emigrate to SE Asia
Propagating fish sauces and fish products.
Also red rice wine and wine lees pastes 红糟
The fish/soy fermentation isogloss!

Ishige (1993)
The Age of European Exploration

1600: British and Dutch sail to Asia for spices, textiles, porcelain

The sailors drink

beer and wine

But both go sour in the tropical heat

(no hops yet)

They arrive in Indonesia to find ethnic Chinese making:
Arrack

The ancestor of rum
Distilled palm wine & red rice
(from Arabic word for ‘sweat’)

1609: British start buying arrack
1704: The “common drink” of Europeans in Asia

Sailors add limes (for scurvy) to make “punch”, first cocktail
While buying arrack, the British picked up barrels of:

Fish sauce
Vietnamese *nuoc mam* or Thai *nam pla*

What was this fish sauce called in the Zhangzhou dialect of Southern Min?

“Ketchup”

*ke* = 鰤 salted fish  
*chiap* = 汁 “sauce”

The British bring it home.
Charles Lockyer’s 1703 trip to Asia

British trader in Asia

Ketchup (fish sauce) now an important commodity
Soy comes in Tubs from Jappan, and the best Ketchup from Tongqueen; yet good of both sorts, are made and sold very cheap in China.

I know not a more profitable Commodity.
Fish sauce transforms
What to do when you have an expensive imported luxury?
Create fakes!
Expensive imported ketchup

To Make KATCH-UP that will keep good Twenty Years.

Take a Gallon of strong stale Beer, one Pound of Anchovies wash'd and clean'd from the Guts, half an Ounce of Mace, half an Ounce of Cloves, a quarter of an Ounce of Pepper, three large Races of Ginger, one Pound of Eschallots, and one Quart of flap Mushrooms ...boil ... strain ...... bottle and stop it very close...This is thought to exceed what is brought from India... 1742
Innovations: Jane Austen’s family recipe for walnut ketchup, 1800

Take green **walnuts** ... **vinegar** with a handful of **salt**. ... boil .... **clove**, **mace**, sliced **ginger**, sliced **nutmeg**, Jamaica **peppercorns**, little **horse radish** with a few **shallots**. ... bottle it up ...
Innovations: tomatoes! 1817

Tomata Catsup

Gather a gallon of fine, red, and full ripe tomatas; mash them with one pound of salt; ...add a quarter of a pound of anchovies, two ounces of shallots, and an ounce of ground black pepper; boil ... strain ......mace... allspice and ginger.... nutmeg, .... coriander ... will keep for seven years
Sugar!

1850: anchovies disappear
1870: especially in the US, lots of sugar added
Hidden in the name of our national sauce: A history of innovation!

- Tai tribal fermented fish and rice
- Japanese sushi
- Vietnamese fish sauce
- Anchovy, mushroom, walnut ketchups
- Modern sweet tomato ketchup

Recipes are just a technology
The ketchup theory of innovation

Innovation happens at interstices, as we borrow and extend the ideas of our neighbors
Other stories of early borrowed innovation reflected in food words

How ice cream came from military technology!

- How Chinese gunpowder and Syrian chemists led to the Italian invention of ice cream and the word “sherbet”)
How about innovations in science?

Most historical accounts of scientific change:
  Internal revolution from one paradigm to another

But what about external influences?
Is there a similar role for borrowing from the neighbors?
From recipes to ideas: Using computational linguistics to study the flow of ideas in science

Ashton Anderson  
Dan McFarland  
David Hall  
Will Hamilton  
Raine Hoover  
David Jurgens  
Minkyoung Kim  
Jure Leskovec  
Chris Manning  
Vinod Prabhakaran
The history of one field: computational linguistics

ACL Anthology Corpus
20,000 papers
15,000 authors
100,000 citations
Most papers in Computational Linguistics
Topic models for measuring language

**Topics**

- gene 0.04
- dna 0.02
- genetic 0.01
- life 0.02
- evolve 0.01
- organism 0.01
- brain 0.04
- neuron 0.02
- nerve 0.01
- data 0.02
- number 0.02
- computer 0.01

**Documents**

**Seeking Life’s Bare (Genetic) Necessities**

*COLD SPRING HARBOR, NEW YORK—* How many genes does an organism need to survive? Last week at the genome meeting here,* two* genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known *genomes,* concluded that today’s *organisms* can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—*but* that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson at Umeå University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are compared, sequenced. “It may be a way of organizing any newly sequenced genome,” explains Ardy Kosmider, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

**Topic proportions and assignments**

Blei 2011
Topic models and other text processing for history of science


(1) Induce 73 topics from distribution over words:

**Anaphora resolution**: pronoun, anaphora, antecedent, pronouns, coreference, anaphoric...

**Parsing**: grammar parse chart context-free edge production CFG symbol terminal left items nonterminal...

**Probability**: probability distribution estimate entropy statistical likelihood parameters smoothing modeling stochastic prior Bayesian...

70 more...

(2) Each CL paper has a distribution over topics:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Anaphora</th>
<th>Parsing</th>
<th>MT</th>
<th>Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.12</td>
<td>0.08</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.22</td>
<td>0.16</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.38</td>
<td>0.04</td>
<td>0.01</td>
</tr>
</tbody>
</table>

ACL anthology
Topic zeitgeist: \[ p(\text{topic}=z \mid \text{year}) \propto \sum_{d_i} p(z \mid d_i) \]
What happened in 1988?

We read all the papers with new statistics.
The Origins of Statistical Modeling

... Speech Researchers with Electrical Engineering background!

Kenneth Church, et al.

Peter Brown, Bob Mercer, Frederick Jelinek,
Implication

Speech researchers brought these ideas to computational linguistics.
  ◦ Computational linguists borrowed these new models
  ◦ And started adapting them to new problems

The **ketchup model** of innovation:
  ◦ We borrowed technology from the neighbors
  ◦ Interdisciplinarity plays a key role
But wait: With topic models alone, no hard evidence of a connection between rise and fall of topics X and Y
But wait: with topic models alone, no hard evidence of a connection between rise and fall of topics X and Y.

By tracking the movements of people over time, we can better understand the causal story.
Studying movement of people across topics

First cluster topics on how people move in and out of them

Then study the flow of people between these 9 clusters:

- **1980-1988**
  - Early Natural Language Understanding
  - Discourse
  - Parsing
  - Finite Automata

- **1989-1994**
  - Government Sponsored “Bakeoff” Period
  - Grantees required to attend these workshops

- **1995-2008**
  - Early Probability Models
  - Supervised Learning of Linguistic Classes
  - Probabilistic Methods
  - Big Data Computational Linguistics
Flow between clusters is the average flow between topics in those clusters


2002–04 — 2005–07
Topic 8: the funnel of the hourglass

US DARPA funding agency “bakeoffs”

Researchers from many fields
  ◦ work on common tasks
  ◦ shared evaluations
  ◦ show up at to annual workshop and present results
  ◦ interdisciplinary participation

Successful innovations were replicated
The field converges: 3 conferences move toward the new empirical

Papers at the main 3 conferences over time, showing their language projected onto 2 dimensions.

Blue is the new statistical conference
The field converges: continued authorship

Overlap between authors in neighboring time windows
Pollination

Of the prolific authors who first published in the database in 1989, 50% were from speech recognition. Most left (returned to EE).

Government-sponsored period led to a large influx of speech recognition researchers

The people returned, but these new ideas stayed.

A new model of spread of innovation

◦ “Pollination” rather than “Colonization”
Summary

Use language to locate a field-redefining innovation

- Trace the flow of people to find causal patterns

We found:

- Innovations came from neighboring field whose researchers pollenated new field
- Midwifed by government funding for collaboration/group work on shared problems
In that spirit, let’s apply computational linguistics to food!
The Linguistics of the Everyday

Words on potato chip packages reveal identity in American food
Expensive chips

Cheap chips
Sample from the corpus:
Bourdieu’s *Distinction*

Survey: French taste in the 1960s correlated with class

- Working class had “popular” tastes
  - the *Blue Danube waltz*
  - heavy starchy meals (cassoulet)
- High status class had “refined” tastes
  - *Well-Tempered Clavier* or Breughel
  - new ethnic or health foods (curry, brown rice)

Bourdieu’s proposal:

- Not about absolute quality
- About distinguishing upper class from lower class
Bourdieu’s Distinction

In matters of taste, more than anywhere else, all determination is negation; and tastes are perhaps first and foremost distastes, disgust provoked by horror or visceral intolerance (“sick-making”) of the tastes of others.

Bourdieu, Distinction
Let’s measure linguistic **distinction**

**Comparison**
- “more”, “less”
- “least”, “best”, “finest”
- “unique”

**Negation**
- “not”, “no”, “never”, “didn’t”
Distinction in expensive chips

5x more frequent in expensive chips

Because of our unique baking process...

in a class of their own

...deliciously different...

best in America...

crunchy bite you won't find in any other chip

less fat than other leading brands...

Every additional negative word adds 4 cents to the price per ounce
Say “no”

nothing fake or phony.
no fake colors, no fake flavors,
no fluorescent orange fingertips,
no wiping your greasy chip hand
on your jeans. no, really.
Expensive Chips: Health

Chips are a health food!

But expensive bags mention health 6 times more than cheap!
Cheap Chips: Traditional authenticity

in the shadow of the Cascade Mountains
made in the great Pacific Northwest
classic American snacks
using an old family recipe
time-tested standard
85-year-old recipe
a time-honored tradition
since 1921
the chips that built our company
Bill and Sally Utz believed
Expensive chips: Natural authenticity

all natural
great taste...naturally
still made with all natural oil
absolutely nothing artificial
only real food ingredients
Yukon Gold potatoes
only the finest potatoes
hand-rake every batch
Language and class/expense

Expensive
  Health
  “Natural” Authenticity
Negation

Cheap
  Traditional Authenticity
How do menus reflect socio-economic differences?

What linguistic differences are reflected in cheap versus expensive menus?

Menus as reflections of attitudes toward socio-economic class

Online menus from 6562 restaurants
- 650,000 menu items
- 5,000,000 words
- allmenus.com, yelp.com
Lots of adjectives in $$ menus

**zesty, rich, golden brown, crispy, creamy**

Crispy white-meat tenders served with a **creamy** Creole sauce
rich, **creamy** spinach artichoke dip
Creamy, homemade fettuccine alfredo
**zesty** chili pepper cream sauce
Lots of vague filler words in $ menus

Delicious, freshly, flavorful

The delicious taste
delicious outdoor grill flavor
flavorful entrées
flavorful ancho-chile
two freshly made sides
freshly steamed broccoli
Why Expensive Menus are short

Why say food is “fresh”?

- Grice’s Maxims of Quantity and Relevance:
  - The hearer needs to know it’s fresh
  - I am trying to communicate the freshness of my food

Why would you need to know?

- Because you don’t already know if it’s fresh
- i.e. you aren’t sure if it’s fresh

Expensive restaurants want you to be sure

- I say nothing so you assume food is fresh
Why Middle Priced Menus are long


Game theoretic model

• Assume 3 class: high, mid, low,

• And assume other noisy cues to quality (location, expensive, décor)

• Mid signals to show isn’t low

• High countersignals (doesn’t signal) to show isn’t mid
Grice in action: *real* on menus

$  
- chocolate chip pancakes served with **real whipped cream**
- home made meatloaf served with **real mashed potatoes**
  chicken cutlet: melted swiss cheese on a roll with lettuce, tomato, russian dressing and **real bacon bits**

$$  
- california roll: **real crab** and avocado
- blueberry whole grain pancakes with **real maple syrup**

no $$$ or $$$$$
Aside: semantics and the history of artificial food

From the New York Public Library’s Buttolph collection:

1990s: real bacon (not Bacon Bits).
1970/80s: real whipped cream (not Cool Whip)
           real sour cream (not Imo).
1960s: real butter (not margarine).
1930/40s: genuine calves liver.
1900: real German beer and real turtle (not ale or mock turtle)
What do fancy restaurants do instead?

Use rare words
- *tonnarelli, bastilla, persillade*

*Use long words:*
- *decaffeinated, accompaniments, complements, exquisitely*

*cheap restaurants:*
- *decaf* not *decaffeinated*, *sides* not *accompaniments*.

What do fancy restaurants do?

Each additional average letter = 18 cents
Expensive restaurant menu:

## Cheap restaurant menu

### Wood-fired Pizza

**$11.99**

*Wood-fired brick oven, made-to-order, fresh ingredients*

**Ingredients:**
- **Dough:** Perfect blend of organic flour, fresh yeast, and pure filtered water
- **Toppings:** Your choice of fresh or dried ingredients

*Available in a variety of sizes and toppings*  

### SUSHI

**$9.95**

*Freshly prepared, traditional sushi rolls*

**Ingredients:**
- Fresh, sustainably sourced seafood
- Organic rice
- Fresh vegetables

*Available in a variety of types and sizes*  

### STEAK
deliciously aged prime cut

**$34.95**

*Grilled to perfection, served with a side of your choice*

**Ingredients:**
- Premium-grade, hand-picked steak
- Assorted sides

*Available in different cuts and styles*  

### WINE

**$19.95**

*Selected from our extensive wine list*

**Ingredients:**
- Fine wines from around the world

*Available by the glass or bottle*  

### JACK DANIELS WINGS

**$19.95**

*Grilled or fried, served with your choice of sauce*

**Ingredients:**
- Premium chicken wings
- Assorted sauces

*Available in different styles and flavors*  

### CAESAR SALAD

**$9.95**

*Classic romaine lettuce with romaine leaves, grated parmesan cheese, and croutons*

**Ingredients:**
- Fresh lettuce
- Shredded cheese
- Croutons

*Available with a variety of dressings*  

###onia’s PHILLY CHEESESTEAK

**$11.95**

*Philadephia-style steak sandwich with sautéed onions and peppers*

**Ingredients:**
- Sliced steak
- Sautéed onions
- Peppers

*Available with a choice of sides*  

### BURRITO BOWL

**$10.95**

*Grilled chicken or beef, served over a bed of rice with beans and vegetables*

**Ingredients:**
- Grilled protein
- Rice
- Beans
- Veggies

*Available with a variety of sauces*  

### MEATLOAF

**$14.95**

*Classic comfort food, served with a side of mashed potatoes*

**Ingredients:**
- Ground beef
- Tomatoes
- Onions

*Available with a choice of sides*  

### BAKED BACON W/ CHEESE & MAPLE SYRUP

**$12.95**

*Brunch favorite, wrapped in bacon and served with a drizzle of maple syrup*

**Ingredients:**
- Bacon
- Cheese
- Maple syrup

*Available with a choice of sides*  

### DEEP-FRIED RIBS

**$22.95**

*Slow-cooked, fall-off-the-bone ribs, served with a side of coleslaw*

**Ingredients:**
- Slow-cooked ribs
- Coleslaw

*Available with a choice of sides*  

### BBQ RIBS

**$16.95**

*Smoked and slow-cooked, served with a side of your choice*

**Ingredients:**
- Pork ribs
- Assorted sides

*Available with a choice of sides*  

### LEMONADE

**$3.95**

*Refreshing, homemade lemonade*

**Ingredients:**
- Fresh lemons

**Note:** Available in different flavors*  

---

**Terms and Conditions:**
- All prices subject to change without notice.
- No substitutions allowed.
- Specials and offers cannot be combined.
Linguistics and economics

Expensive menus are linguistically modest

- Modest advertising is a way of displaying luxury status
Traditional authenticity in $$ restaurants:

fresh **homemade** guacamole and chip

**old fashioned** beef stew

annie’s famous pot roast **homemade just like mom’s**

**grandma minnie’s** fried chicken salad
Expensive restaurants use natural authenticity

HERB ROASTED ELYSIAN FIELDS FARMS LAMB
Eggplant Porridge, Cherry Peppers,
Greenmarket Cucumbers and Pine Nut Jus

GRASS FED ANGUS BEEF CARPACCIO
Pan Roasted King Trumpet Mushrooms
Dirty Girl Farm Romano Bean Tempura
Persillade, Extra Virgin Olive Oil

BISON BURGER
8 oz. blue star farms, grass fed & pasture raised,
melted gorgonzola, grilled vegetables
More on social aspects of word meaning: The Tiki Lounge effect

exotic, oriental, spices, spicy

exotic five spices
exotic blend of indian spices
island spices you crave: An exotic, delicious sauce.
oriental vinaigrette
Thai curry herbs and spices
kick of southwestern spice
spicy Santa Fe sauce
spicy garlic & lime grilled shrimp

Dishes with these words cost 3 cents more per word
Framing in restaurant reviews


900,000 online reviews

The bartender... absolutely horrible... we waited 10 min before we even got her attention... and then we had to wait 45 - FORTY FIVE! - minutes for our entrees... stalk the waitress to get the cheque... she didn't make eye contact or even break her stride to wait for a response
Log odds ratio

**Log likelihood ratio:** does “horrible” occur more % in corpus A or B?

\[
\log P_A \left( \text{"horrible"} \right) - \log P_B \left( \text{"horrible"} \right)
\]

\[
= \log \left( \frac{\text{count}_A \left( \text{"horrible"} \right)}{\sum_{\text{word in } A} \text{count(\text{word})}} \right) - \log \left( \frac{\text{count}_B \left( \text{"horrible"} \right)}{\sum_{\text{word in } B} \text{count(\text{word})}} \right)
\]

**Log odds ratio:** does “horrible” have higher odds in A or B?

\[
\log \left( \frac{\text{count}_A \left( \text{"horrible"} \right)}{\frac{NA}{1 - \text{count}_A \left( \text{"horrible"} \right)}} \right) - \log \left( \frac{\text{count}_B \left( \text{"horrible"} \right)}{\frac{NB}{1 - \text{count}_B \left( \text{"horrible"} \right)}} \right)
\]

\[
\log \left( \frac{\text{count}_A \left( \text{"horrible"} \right)}{\text{NA} - \text{count}_A \left( \text{"horrible"} \right)} \right) - \log \left( \frac{\text{count}_B \left( \text{"horrible"} \right)}{\text{NB} - \text{count}_B \left( \text{"horrible"} \right)} \right)
\]
Log odds ratio with a prior

Log odds ratio:

$$\log \left( \frac{f_A(\text{"horrible"})}{N_A - f_A(\text{"horrible"})} \right) - \log \left( \frac{f_B(\text{"horrible"})}{N_B - f_B(\text{"horrible"})} \right)$$

With a prior:

$$\log \left( \frac{f_A(\text{"horrible"}) + f_{\text{prior}}(\text{"horrible"})}{N_A + N_{\text{prior}} - (f_A(\text{"horrible"}) + f_{\text{prior}}(\text{"horrible"}))} \right) - \log \left( \frac{f_B(\text{"horrible"}) + f_{\text{prior}}(\text{"horrible"})}{N_B + N_{\text{prior}} - (f_B(\text{"horrible"}) + f_{\text{prior}}(\text{"horrible"}))} \right)$$
Log odds ratio informative Dirichlet prior

Monroe, Colaresi and Quinn (2008)

Find words that are statistically overrepresented in a particular category of review compared to another

\[
\delta_w^{(i-j)} = \log \left( \frac{y_w^i + \alpha_w}{n_i^i + \alpha_0 - (y_w^i + \alpha_w)} \right) - \log \left( \frac{y_w^j + \alpha_w}{n_j^j + \alpha_0 - (y_w^j + \alpha_w)} \right)
\]

\(n_i^i\) is the size of corpus \(i\), \(n_j^j\) is the size of corpus \(j\), \(y_w^i\) is the count of word \(w\) in corpus \(i\), \(y_w^j\) is the count of word \(w\) in corpus \(j\), \(\alpha_0\) is the size of the background corpus, and \(\alpha_w\) is the count of word \(w\) in the background corpus.\)

\[
\sigma^2 \left( \delta_w^{(i-j)} \right) \approx \frac{1}{y_w^i + \alpha_w} + \frac{1}{y_w^j + \alpha_w}
\]

Final statistic for a word: z-score of its log-odds-ratio:

\[
\frac{\delta_w^{(i-j)}}{\sqrt{\sigma^2 \left( \delta_w^{(i-j)} \right)}
\]
Top 50 words associated with one-* reviews by Monroe, *et al.* (2008) method

<table>
<thead>
<tr>
<th>Linguistic Class</th>
<th>Words in Class</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Negative sentiment</strong></td>
<td>worst, rude, terrible, horrible, bad, awful, disgusting, bland, tasteless, gross, mediocre, overpriced, worse, poor</td>
</tr>
<tr>
<td><strong>Linguistic negation</strong></td>
<td>no, not</td>
</tr>
<tr>
<td><strong>1 pl pronouns</strong></td>
<td>we, us, our</td>
</tr>
<tr>
<td><strong>3rd pronouns</strong></td>
<td>she, he, her, him</td>
</tr>
<tr>
<td><strong>Past tense verbs</strong></td>
<td>was, were, asked, told, said, did, charged, waited, left, took</td>
</tr>
<tr>
<td><strong>Narrative sequencers</strong></td>
<td>after, then</td>
</tr>
<tr>
<td><strong>Common nouns</strong></td>
<td>manager, waitress, waiter, customer, customers, attitude, waste, poisoning, money, bill, minutes</td>
</tr>
<tr>
<td><strong>Irrealis modals</strong></td>
<td>would, should</td>
</tr>
<tr>
<td><strong>Complementizers</strong></td>
<td>to, that</td>
</tr>
</tbody>
</table>
Language of bad reviews?

Negative sentiment language
  horrible awful terrible bad disgusting

Past narratives about people (Biber’s factor)
  waited, didn’t, was
  he, she, his, her,
  manager, customer, waitress, waiter

Frequent mentions of we and us
  ... we were ignored until we flagged down one
  waiter to go get our waitress ...
Other narratives with this language

A genre using:

Past tense, we/us, negative, people narratives

Texts written by people suffering trauma

- Chat group discussions after Princess Diana’s death
- Blog posts after September 11, 2001
- Student newspaper reports after a campus tragedy

Why? Pennebaker’s social stage model of coping

- people feel a need to tell stories expressing their negative emotion,
- “we/us/our” suggests we are seeking comfort in community
- Past tense used to distance ourselves from the traumatic event
Implications

1-star reviews are not descriptions of bad food. They are trauma narratives!

The lesson of bad reviews:

We are very sensitive to personal interaction
Positive Reviews?

- addicted to wings
- the fries are like crack
- ....crave... cupcakes
- orgasmic pastry
- seductively seared...
- very naughty pork belly

![Bar chart for Drugs and Sex](image)
Why the addiction narrative?

<table>
<thead>
<tr>
<th>Meaty, fatty foods</th>
<th>Starchy comfort food</th>
<th>Sweet food</th>
<th>Small ethnic dishes</th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>burgers</td>
<td>pizza</td>
<td>sweets</td>
<td>sushi</td>
<td>comfort</td>
</tr>
<tr>
<td>barbecue</td>
<td>mac and cheese</td>
<td>pancakes, breakfast</td>
<td>dim sum</td>
<td>fried, greasy</td>
</tr>
<tr>
<td>chicken wings</td>
<td>pasta/noodles</td>
<td>sugar</td>
<td>tacos, burritos</td>
<td>unhealthy</td>
</tr>
<tr>
<td>french fries</td>
<td>soups</td>
<td>chocolate</td>
<td>spam, musubi</td>
<td>hearty, satisfying</td>
</tr>
<tr>
<td></td>
<td>sandwiches</td>
<td>beignets</td>
<td>dumplings</td>
<td>junk</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>falafel</td>
<td>authentic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>tapas</td>
<td>cheap</td>
</tr>
</tbody>
</table>

Craved foods aren’t vegetables, or main courses like meatloaf or fish or even side dishes like mashed potatoes. They are **junk foods** or at least non-normative foods:

- Assuage the guilt
- It’s not my fault, I had no control, the cupcake made me eat it
Gender and the addiction metaphor?

Are women more likely to use the metaphor of drugs than men?

- women are more pressured to conform to healthy eating
- women more likely to mention food cravings
  - Rozin, *et al.* (1991): females are significantly more likely to express cravings for chocolate than males.

**How to test:**

- Figure out the gender of Yelp users from first names
  - Using Social Security Baby Name database
  - See if women use the drug metaphor more.
Are women more likely to use the metaphor of drugs than men?

- Women significantly more likely than men to talk about food as a drug ($p=0.000832$).
- But we don’t know the cause:
  - women might be more likely than men to have these cravings
  - women might be more comfortable than men in admitting to these cravings
  - women might have identical desires but be more likely than men to use this particular linguistic metaphor
- By the way
  - Men were more likely to use the language of trauma
The linguistics of food requests


Question: What language do people use when making successful requests?

Problem: it depends on the request

Solution: control for the request

21,000 requests for pizza on “Random Acts of Pizza” Reddit.
“My gf and I have hit some hard times with her losing her job and then unemployment as well for being physically unable to perform her job due to various hand injuries as a server in a restaurant. She is currently petitioning to have unemployment reinstated due to medical reasons for being unable to perform her job, but until then things are really tight and ANYTHING would help us out right now.

I [...] would certainly return the favor again when I am able to reciprocate.”

**Evidentiality:** Urgent requests are met more frequently than non-urgent requests (Yinon and Dovrat 1987; Shotland and Stebbins 1983; Colaizzi, Williams, and Kayson 1984; Gore, Tobiasen, and Kayson 1997)

**Length:** Long requests demonstrate extra effort and can provide more evidence (Lettice et al. 2012)

**Reciprocity:** Promises to return the favor (Wilke and Lanzetta 1970; Willer et al. 2013; Gray, Ward, and Norton 2012; Plickert, Côté, and Wellman 2007)

**Status:** People of high status (e.g. occupation or wealth) receive help more often. (Solomon and Herman 1977; Goodman and Gareis 1993)
Who gets pizza?
Predictable from the language!
Most important factor: **Pro-social behavior**
- Generalized Reciprocity
  Promising to Pay it Forward
- Saying thank you
- Having karma in community
Computing the Language of Food

The language of menus and ads
- economics of “modest” advertising
- Health, authenticity, and class

The language of reviews and requests
- social psychology of language
- the linguistic signs of trauma
- our complex attitudes toward food and body
- the rewards of pro-social behavior/language
Some conclusions from the Language of Food

- Innovation happens at interstices, as we borrow and extend the ideas of our neighbors.
- People suffer when you are mean to them, and are generous when you are kind.
- You can tell a lot about psychology, economics, even evolution if you just look very carefully at the language of food.
Other stuff going on in our lab

With Will Hamilton and Jure Leskovec

Computational historical linguistics

- Test linguistic theories of meaning change
- Using 200 years of online corpora

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.

2. Diachronic embedding methods

The following sections outline how we construct diachronic (historical) word embeddings, by first constructing embeddings in each time-period and then aligning them over time, and the metrics that we use to quantify semantic change.

3.2.1 Embedding algorithms

We use three methods to construct word embeddings within each time-period: PPMI, SVD, and SGNS (i.e., word2vec).

These distributional methods represent each word $w_i$ by a vector $\mathbf{w}_i$ that captures information about its co-occurrence statistics. These methods operationalize the 'distributional hypothesis' that word semantics are implicit in co-occurrence relationships (Harris, 1954; Firth, 1957). The semantic similarity/distance between two words is approximated by the cosine similarity/distance between their vectors (Turney and Pantel, 2010).

2.1.1 PPMI

In the PPMI representations, the vector embedding for word $w_i$ contains the positive point-wise mutual information (PPMI) values between $w_i$ and a large set of pre-specified 'context' words. The word vectors correspond to the rows of the matrix $\mathbf{M}_{PPMI} \in \mathbb{R}^{|V| \times |V_C|}$ with entries given by

$$M_{PPMI}^{i,j} = \max(\log(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w_i) \hat{p}(c_j)}), 0),$$

where $c_j \in V_C$ is a context word and $\gamma > 0$ is a negative prior, which provides a smoothing bias (Levy et al., 2015). The $\hat{p}$ correspond to the smoothed empirical probabilities of word (co-)occurrences within fixed-size sliding windows of text. Clipping the PPMI values above zero ensures they remain finite and has been shown to dramatically improve results (Bullinaria and Levy, 2007; Levy et al., 2015); intuitively, this clipping ensures that the representations emphasize positive word-word correlations over negative ones.

4. Synchronic applications of these three methods are reviewed in detail in Levy et al. (2015).
Other stuff going on in our lab

With Jennifer Eberhardt, Nick Camp, Camilla Griffiths, Rob Voigt, Vinod Prabhakaran, David Jurgens, Will Hamilton

Police language and race

- Use Oakland police department footage
- Can we extend Cristian’s work on politeness to measure procedural justice
  - Respect, neutrality, etc. of police officers
- Do police interactions with whites and blacks differ in aspects of procedural justice as measured from language?