Using statistics to explore human language

Emily Tucker Prud'hommeaux
Assistant Professor, College of Liberal Arts, RIT
June 30, 2015
Agenda

• About me
• Word statistics in the world
• Word statistics in the classroom
• Word statistics in medicine
• Resources
About me

• Now: assistant professor at RIT
  • computational linguistics
• Before that: grad school
  • computer science
• Before grad school: software engineer
  • speech recognition used by Siri
• First job out of college: secondary school teacher
  • Latin, musical theater
I'm not a mathematician, but...

my brother

Thomas J. Tucker
University of Rochester
Department of Mathematics
I'm not a mathematician, but...

my dad

Thomas W. Tucker
Colgate University
Department of Mathematics
I'm not a mathematician, but...

my grandfather

Albert W. Tucker
Princeton University
Department of Mathematics
I'm not a mathematician, but...

my grandmother

Alice Curtiss Tucker Beckenbach
Princeton (NJ) High School
Geometry Teacher
I'm not a mathematician, but...

my great-grandfather

David Raymond Curtiss
Northwestern University
Department of Mathematics
My math journey

1987  Course I
1988  Course II
1989  Course III
1990  Precalculus
1991  Calculus

2003-2006  Calc 1-3, discrete math, linear algebra, statistics
Why did I return to math?

• I wanted to get a degree in computer science.
• Turns out most algorithms involve math!
• And computational linguistics requires knowledge of calculus, linear algebra, statistics, and probability.
What is computational linguistics?

Using computer algorithms to analyze, understand, and produce human language and speech.
Word statistics

• Central idea in computational linguistics.

• It's often important and helpful to know:
  • how frequent a word is in general
  • how frequent a word is in a particular context

• Also really helpful:
  • how frequent a sequence of words is
    • one word, two words, three words...
  • word sequences are called n-grams
    • unigram, bigram, trigram...
Word statistics in our brains

How would you fill in this sentence?

I'd like to reserve a ______.

- table
- room
- seat
- car

something else?
Word statistics in our brains

• As a speaker of English, you have intuitions.
  • real-world knowledge
  • (unconscious) knowledge of word frequencies and word sequence frequencies
When people don't use context...

Word statistics: brains vs. computers

- You have intuitions about what words are likely to appear together.
- Computers model this knowledge by keeping track of n-grams in very large bodies of text.
- These n-gram models are used in many computational linguistics applications.
What are n-grams used for?

Web search

Google

why is math j
why is math important
why is math so hard
why is math so important
why is math important in life
why is mathilde unsatisfied at the beginning of the story
why is math important to learn
why is math and science important
why is math your favorite subject
why is math required in college
why is math fun

Google Search  I'm Feeling Lucky
What are n-grams used for?

Spelling correction
What are n-grams used for?

Speech recognition
What are n-grams used for?

Plagiarism detection
What are n-grams used for?

Authorship attribution
What are n-grams used for?

Forensic linguistics

"cool-headed logicians"

"eat your cake and have it too"
What are n-grams used for?

Predictive typing
What are n-grams used for?

Word clouds

Obama

Bush
What are n-grams used for?

Personality analysis

Analysis of tweets from katyperry (773 most recent words - 9th March, 2015)

- **Emotional Style**
  - Upbeat (Low): 32
  - Worried (Low): 23
  - Angry (Average): 41
  - Depressed (High): 74

- **Social Style**
  - Plugged In (Low): 40
  - Personable (Low): 28
  - Arrogant/Distant (Low): 33
  - Spacy/Valley girl (Low): 37

- **Thinking Style**
  - Analytic (Low): 30
  - Sensory (Very high): 84
  - In-the-moment (Average): 47

Analysis of tweets from taylorswift13 (842 most recent words - 9th March, 2015)

- **Emotional Style**
  - Upbeat (Very high): 66
  - Worried (Average): 46
  - Angry (Low): 39
  - Depressed (Average): 42

- **Social Style**
  - Plugged in (Average): 55
  - Personable (Average): 54
  - Arrogant/Distant (Average): 51
  - Spacy/Valley girl (Average): 54

- **Thinking Style**
  - Analytic (Average): 51
  - Sensory (Average): 52
  - In-the-moment (Average): 45

Katy Perry

Taylor Swift
You can use n-grams, too!

https://books.google.com/ngrams
Why is this fun?

learn when words go in and out of use
Why is this fun?

learn when words go in and out of use
Why is this fun?

find anachronisms on Mad Men
Why is this fun?

philosophize about the world
Why is this fun?

philosophize about the world

definitive proof that global warming is real
Why is this educational?

• Hundreds of thousands of words in English so the likelihood of any one word or phrase is very small.

• Possible exercises with students:
  
  • comparing small numbers and using visualizations to help reinforce these comparisons
  
  • practice with scientific notation
  
  • converting to log notation to understand why logs are easier to deal with for small numbers
Why is this educational?

small!
Why is this educational?

• Relatively easy-to-understand probability:
  • get lots and lots of books
  • count how many times each word appears
  • divide by total word count

• Possible exercises:
  • converting fractions to percents
  • converting between percents and decimals
  • compiling word statistics using Word and Excel
Why is this educational?

• Teaching that correlation does not equal causation.
• Critical thinking about misinterpreting statistics.
• Understanding how people manipulate statistics to further their particular agenda.
Manipulating Google Ngram

As a society, we have become more tolerant. Even the word "prejudice" is becoming less and less common!
Manipulating Google Ngram

Hey, wait a minute!
Word statistics in medicine

• People with disorders of the mind (e.g., autism, Alzheimer's disease, schizophrenia) often use language in a different way.

• Comparing the way they use words to the way that neurotypical people use words can reveal interesting patterns.

• Such patterns can provide diagnostic hints and information about the deficits specific to those disorders.
Language in autism spectrum disorder

“seemingly nonsensical and irrelevant”
“peculiar and out of place in ordinary conversation”
-Kanner 1946

“a lack of ease in the use of words”
-Rutter 1965

“stereotyped”, “metaphorical”, “inappropriate”
-Bartak et al. 1975

“the use of standard, familiar words or phrases in idiosyncratic but meaningful ways”
-Volden & Lord 1991
Measuring "atypical language"

• How do you measure how inappropriate or idiosyncratic or irrelevant a word or phrase is?

• Idea:
  • ask kids with and without autism to listen to and retell the same story
  • compare what they say to see if the kids with autism are saying something different
Jim was a boy whose best friend was Pepper. Pepper was a big black dog. Jim liked to walk in the woods and climb the trees. Near Jim’s house was a very tall oak tree with branches so high that he couldn’t reach them. Jim always wanted to climb that tree, so one day he took a ladder from home and carried it to the oak tree. He climbed up, sat on a branch, and looked out over his neighborhood. When he started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. Jim held onto a branch so he didn’t fall, but he couldn’t get down. Pepper sat below the tree and barked. Suddenly Pepper took Jim’s shoe in his mouth and ran away. Jim felt sad. Didn’t his friend want to stay with him when he was in trouble? Pepper took the shoe to Anna, Jim’s sister. He barked and barked. Finally, Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. Anna put the ladder up and rescued Jim. Wasn’t Pepper a smart dog?
Jim liked to do stuff in the forest like climb trees. And he wanted to climb up the oak tree so he got the ladder and did. But then when he wanted to get down, he slipped. His shoe fell off, and the ladder fell down so Jim holded on to the branch so he wouldn't fall off. And Pepper was on the ground. He got his shoe and he went. Finally she, Jim's sister, understood what Pepper was saying. She went, she put the ladder up and rescued Jim.

The boy got stuck and someone rescued him and Pepper was a really smart dog. Pepper dogs have a great sense of smell too, like T-rex. T-rex could smell things that were really far away. Well T-rex could be over here and the meat could be way back there under the couch, way back in the couches. Well that guy got stuck on the tree and then his shoe fell out of the tree and Anna rescued it. Pepper brought his shoe back, and Anna rescued them.
The flight attendants told the passengers that they would arrive in Paris in six hours. In the air, Claire became antsy and bored. So, she explored the cockpit where the pilot, who had wings on his shirt, let her look out the wide front window. She also played with the life-vest that she found under her seat. Nanny told her to put it away and go to sleep but she was too excited.
When they were on the plane, the flight attendant told the passengers that they would arrive in Paris in 6 hours. Claire was bored or got bored and went up to the cockpit. And the pilot, who was wearing wings on his shirt, he let her look out of the big front window in the cockpit. And she played with the lifevest under her seat that she found. And nanny was upset with her and told her that she should just go to sleep but Claire was too antsy.

So um, she’s on a plane, and she’s bored. So she goes to the cockpit, where the flight, the employee, the pilot, who apparently lets the girl come in, which which is like the most stupid thing ever, because you know, considering the post-9/11 world, but let’s not get into that. And so she looks around this big wide window that they have to use. So she goes back to her seat and starts playing with a life jacket to which her nanny says don’t do that because that’s a terrible idea, which it is.
Comparing words

• How much does each kid overlap with every other kid in the words that they use?
  • low overlap might indicate that someone is veering off topic and talking about something else

• How many words does each kid use that no other kids use?
  • many "unique" words might indicate that a child is using words that mean the right thing but are unusual or that the words are inappropriate
Results: Word overlap

Kids with typical development tend to use the same sorts of words that other kids use when telling the same story.
Results: Unique word rate

Kids with autism tend to use more words that no other kids are using.
Atypical language in autism

- If you know kids on the spectrum, you've probably noticed these sorts of things.
- And you might guess that it's probably making social interactions more difficult for them.
- But sometimes it's hard to say exactly what is atypical about their language.
- Being able to identify what is atypical can help with therapy and remediation.
Fun with word statistics

• Google Ngram Viewer:
  https://books.google.com/ngrams

• Sentiment analysis:
  http://www.liwc.net/tryonline.php

• Personality analysis:
  http://www.analyzewords.com/

• Word clouds:
  http://www.wordle.net
Thank you!