How math helps us better understand language

Linguistics Studies Lecture Series
College of Charleston
February 18, 2021

Lisa S. Pearl
Professor
Department of Language Science
SSPB 2219
University of California, Irvine
lpearl@uci.edu

Computation of Language Laboratory
UC Irvine
Language is a uniquely human ability.
We can investigate many things about language, including what we know when we know language.
About language

We can investigate many things about language, including how language knowledge develops.
About language

We can investigate many things about language, including how we use language to communicate different information.
So what kinds of things do we **know** when we know language?
About language

how to identify words in fluent speech (speech segmentation)

what a pretty kitty!
About language

what a pretty kitty!

how to pronounce words (phonology)

speech segmentation

✔️ Kitty

❌ ki TTY
About language

what a pretty kitty!

Noun

penguin  kitty  owl

certain words behave like other words (syntactic categorization)

speech segmentation

phonology
About language

how to interpret words in context (syntax, semantics)

“Oh look — a pretty kitty!”
“Look — there’s another one!”
how to put words together to ask questions (syntax)

This kitty was bought as a present for someone.

Lily thinks this kitty is pretty.

“Who does Lily think the kitty for is pretty?”
how to identify the right interpretation in context
(pragmatics)

“Every kitty didn’t sit on the stairs”

No kitties sat on the stairs.

Not all kitties sat on the stairs.
About language

What a pretty kitty!

Speech segmentation

Phonology

Kitty

Noun

Syntactic categorization

Penguin

Owl

Syntax, semantics

What we know

“Who does Lily think the kitty for is pretty?”

“Every kitty didn’t sit on the stairs”

Not all kitties sat on the stairs.
About language

speech segmentation

what a pretty kitty!

phonology

syntactic categorization

speech segmentation

“Every kitty didn’t sit on the stairs”

Not all kitties sat on the stairs.

“Who does Lily think the kitty for is pretty?”

syntax

syntactic categorization

syntax, semantics

How did we develop that knowledge?

“Oh look — a pretty kitty!” “Look — there’s another one!”

syntax, semantics
Children are amazing at learning language.
Much of the linguistic system is already known by age 4.
Also, children figure language out mostly without explicit instruction.
About language

What they’re doing: Extracting patterns and making generalizations from the surrounding data mostly just by hearing examples of what’s allowed in the language.

What’s so hard about that?
There are often many ways to generalize beyond the input, and most of them aren’t right.
There are often many ways to generalize beyond the input, and most of them aren’t right.

“What a pretty birdie!”
About language

There are often many ways to generalize beyond the input, and most of them aren’t right.

“birdie”

“Look - a birdie!”
There are often many ways to generalize beyond the input, and most of them aren’t right.

“Look at that birdie!”
There are often many ways to generalize beyond the input, and most of them aren’t right.

What generalization to make?
About language

There are often many ways to generalize beyond the input, and most of them aren’t right.

???

“birdie”

= blue creature
There are often many ways to generalize beyond the input, and most of them aren’t right.

“birdie”

= creature on branch
About language

There are often many ways to generalize beyond the input, and most of them aren’t right.

“birdie”

= [whatever makes something a bird]
These kinds of induction problems are everywhere in cognitive development, including language development.
These kinds of induction problems are everywhere in cognitive development, including language development.

Language development = Solving a lot of induction problems.
About language

It’s amazing how good kids are at this.

Language development = Solving a lot of induction problems.
About language

How do we *use* language knowledge to communicate?
About language

How do we use language knowledge to communicate?

And can we help machines do it too?
About language

“C’mon — don’t you think this is awesome?”
“C’mon — don’t you think this is awesome?”

- addressee
- mental state
- good++++
"C’mon — don’t you think this is awesome?"

something salient in the context
About language

“C’mon — don’t you think this is awesome?”
But more subtle information is communicated too.

“C’mon — don’t you think this is awesome?”
The speaker likely has a persuasive intention.

“C’mon — don’t you think this is awesome?”
“C’mon — don’t you think this is awesome?”

If the speaker actually doesn’t like penguins, he could be **intending to ingratiate** himself with the addressee (using deception).
“C’mon — don’t you think this is awesome?”

At face value, the speaker seems to have a good feeling about penguins (positive sentiment).
“C’mon — don’t you think this is awesome?”

The casual style of speaking suggests familiarity with the addressee, and may indicate something about the speaker’s identity.
So, our knowledge of language use involves communicating both core and subtle information.
About language

This is what we’d like machines to be able to do just as well as humans.
About language

My research focuses on understanding development and use.

And I use math to do it!
About **math**

Quantitative techniques = techniques that rely on **math**

One main part: **Counting** things
About math

Quantitative techniques = counting things

(sometimes we count a lot of things)
About math

Quantitative techniques = counting things

Another part: principled reasoning based on those counts

Bayesian inference

\[ p(\text{Generalization} \mid \text{Data}) \propto p(\text{Generalization}) \cdot p(\text{Data} \mid \text{Generalization}) \]

Tolerance & Sufficiency Principles

\[ \text{exceptions} < = \frac{\# \text{ items}}{\ln(\# \text{ items})} \]
About math

Quantitative techniques = counting things + principled reasoning

Then we use these quantitative techniques to help us understand how little humans develop language knowledge and how adult humans and machines use language knowledge
About math

Quantitative techniques = counting things + principled reasoning

But what do we count and reason over? How do we connect that information to language development and use?
Quantitative techniques for language development

To understand language development, we’re typically using computational cognitive modeling to encode a child’s learning process very precisely.
We think the child is learning by *counting* different parts of her input and *reasoning* over those counts in a sensible way.

So, the model will *count* those same things and learn about language by doing principled *reasoning* over those counts.
Quantitative techniques for language development

Let’s think about this for speech segmentation.

whataprettykitty!
Quantitative techniques for language development

Let’s think about this for speech segmentation.

what a pretty kitty!
Quantitative techniques for language development

These are the kinds of utterances infants hear.

- what a pretty kitty!
- what a cute penguin!
- look at the pretty birdie!
- the kitty is very cute!
- do you see the kitty?
Quantitative techniques for language development

But they’re more like this before infants know where the words are.

what a pretty kitty!
what a cute penguin!
look at the pretty birdie!
the kitty is very cute!
do you see the kitty?
Quantitative techniques for language development

But they’re more like this before infants know where the words are.

whataprettykitty
whatacutepenguin
lookattheprettybirdie
thekittyisverycute
doyouseethekitty
Quantitative techniques for language development

One idea is that children initially perceive syllables.

whataprettykitty
whatacutepenguin
lookattheprettybirdie
thekittyisverycute
doyouseethekitty
Quantitative techniques for language development

One idea is that children initially perceive syllables.

what a pretty kitty
what a cute penguin
look at the pretty bird
the kitty is very cute
do you see the kitty
Quantitative techniques for language development

One learning theory is that infants count syllables and where they occur.

what a pretty kitty
what a cute penguin
look at the pretty bird
the kitty is very cute
do you see the kitty
Quantitative techniques for language development

what a pretty kitty
what a cute penguin
look at the pretty bird
doe you see the kitty

Then, infants can reason about those counts to figure out where words are.
Quantitative techniques for language development

Then, infants can reason about those counts to figure out where words are.

what a pretty kitty
what a cute penguin
look at the pretty bird
the kitty is very cute
do you see the kitty

Then, infants can reason about those counts to figure out where words are.
Quantitative techniques for language development

So this is what our computational cognitive model can do.

what a pretty kitty
what a cute penguin
look at the pretty bird
the kitty is very cute
do you see the kitty
Quantitative techniques for language development

Example model from my research group: Use Bayesian inference to reason about counts of syllables in a child’s input.

what a pretty kitty
what a cute penguin
look at the pretty birdie
the kitty is very cute
do you see the kitty

Quantitative techniques for language development

This learning strategy involves the child imagining what collection of words (a lexicon) could be used to create the utterances she hears.

what a pretty kitty
what a cute penguin
look at the pretty bird
the kitty is very cute
do you see the kitty

Bayesian inference

Quantitative techniques for language development


Some possible **lexicons** a child might consider for the first utterance:

**what a pretty kitty**

whataprettykitty
Quantitative techniques for language development


Bayesian inference

Each possible lexicon could be used to generate the observed utterance.

what a pretty kitty
Quantitative techniques for language development


Bayesian inference

Each possible lexicon could be used to generate the observed utterance.

what a pretty kitty

whata pretty kitty

what a pretty kitty

what a pretty kitty
Quantitative techniques for language development


Bayesian inference

Each possible lexicon could be used to generate the observed utterance.

what a pretty kitty
whata pretty kitty
Quantitative techniques for language development


Bayesian inference

Each possible lexicon could be used to generate the observed utterance.

what a pretty kitty
whata pretty kitty
Quantitative techniques for language development


Bayesian inference

Each possible lexicon could be used to generate the observed utterance.

what a pretty kitty

whatapretty kitty
Quantitative techniques for language development

The child tries to identify the most probable one.

what a pretty kitty

The modeled child in the computational cognitive model chooses the most probable one using Bayesian inference.
Quantitative techniques for language development


Bayesian inference

\[ p(\text{lexicon} | \text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances} | \text{lexicon}) \]

We want to figure out the lexicon with the highest probability, given the utterances the child encounters (and the syllables in them).
Quantitative techniques for language development


what a pretty kitty

Bayesian inference

\[ p(\text{lexicon} | \text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances} | \text{lexicon}) \]

The model reasons about this by considering two other probabilities.
Quantitative techniques for language development


what a pretty kitty

Bayesian inference

\[ p(\text{lexicon}|\text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances}|\text{lexicon}) \]

The prior probability of the lexicon captures any preferences the modeled child has (which we think real children would have too).
Quantitative techniques for language development


what a pretty kitty

\[ p(\text{lexicon} \mid \text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances} \mid \text{lexicon}) \]

One preference: Prefer lexicons with fewer words.
Quantitative techniques for language development


**Bayesian inference**

\[ p(\text{lexicon}|\text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances}|\text{lexicon}) \]

Another preference: Prefer lexicons with shorter words.
Quantitative techniques for language development


Bayesian inference

\[ p(\text{lexicon} | \text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances} | \text{lexicon}) \]

These preferences compete with each other.

shorter words fewer words
Quantitative techniques for language development


Bayesian inference

$p(\text{lexicon}|\text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances}|\text{lexicon})$

The modeled child uses Bayesian inference to reason about the lexicon that’s the best balance between these preferences…
Quantitative techniques for language development

Bayesian inference

\[ p(\text{lexicon} | \text{utterances}) \propto p(\text{lexicon}) \cdot p(\text{utterances} | \text{lexicon}) \]

…that also has a high probability of generating the observed utterances = a high likelihood.
Quantitative techniques for language development


Bayesian inference

We can then see if the lexicon the modeled child identifies has the right kind of things in it (like real words).
Quantitative techniques for language development


Bayesian inference

If so, then the computational cognitive model has captured (using this math) how a child could identify words in fluent speech.
Quantitative techniques for language development


Bayesian inference

what a pretty kitty

It turns out this learning strategy for speech segmentation is useful.
Quantitative techniques for language development


Bayesian inference

It can segment realistic English input to children fairly well...though the inferred lexicons aren’t perfect.

useful

what a pretty kitty

✓
But it turns out these imperfect lexicons are very useful for subsequent stages of language development, like learning what a word form refers to.
Quantitative techniques for language development


Bayesian inference

This was true for idealized modeled children, with perfect memory and perfect processing abilities.
Quantitative techniques for language development


But what about modeled children with more realistic constraints on their memory and processing abilities?

Is this segmentation strategy useable by children, who have these kinds of cognitive limitations?
Yes! Modeled children with cognitive constraints on their memory and processing abilities can still use this strategy to segment English quite well.
Quantitative techniques for language development


Bayesian inference

Does it work for different languages (besides English)?
Quantitative techniques for language development


Bayesian inference

Yes! It segments well for languages with different properties: Spanish, Italian, German, Hungarian, Japanese, Farsi

useful

useable

different languages
Quantitative techniques for language development


Bayesian inference

We were able to discover how good this segmentation strategy is by using computational cognitive modeling.
Quantitative techniques for language development

Some general questions about language development that my research has tried to answer this way
Which learning strategies could children be using?

Quantitative techniques for language development

Which learning strategies could children be using?

Which learning biases are necessary?

Quantitative techniques for language development

Which learning strategies could children be using?
Which learning biases are necessary?

Which knowledge representations are learnable — and which aren’t?

(Pearl & Sprouse in press, Bates & Pearl 2019, Pearl, Ho, & Detrano 2017, 2014; Pearl 2017, Pearl 2011, Pearl 2009)
Quantitative techniques for language development

Which learning strategies could children be using?
Which learning biases are necessary?
Which knowledge representations are learnable — and which aren’t?

When do children learn different aspects of the linguistic system?

Quantitative techniques for language development

Which learning strategies could children be using?
Which learning biases are necessary?
Which knowledge representations are learnable — and which aren’t?
When do children learn different aspects of the linguistic system?

What factors affect children’s observable behavior?

What about using math to help us understand how humans use language and how machines could learn to do the same thing?
Let’s focus on subtle information that can be expressed in language.
Math is again at the heart of the techniques researchers use.
So, we’re counting things and reasoning about those counts in principled ways.
Quantitative techniques for language use

But what are we counting?
Quantitative techniques for language use

precise psychological and linguistic theoretical constructs
Quantitative techniques for language use

It's just two minutes away from Hyde Park.
Quantitative techniques for language use

It's just two minutes away from Hyde Park.

Psychological: “specific details” in a description
Quantitative techniques for language use

Linguistic: prepositional phrases

log \left( \frac{p(f_i|T)}{p(f_i|D)} \right)

It's just two minutes away from Hyde Park.
Quantitative techniques for language use

The counts of these features are used in mathematical equations

$$\log\left(\frac{p(f_v | T)}{p(f_v | D)}\right)$$
Quantitative techniques for language use

The counts of these features are used in mathematical equations that underlie machine learning techniques.

\[ \log \left( \frac{p(f_i|T)}{p(f_i|D)} \right) \]
Quantitative techniques for language use

The counts of these features are used in mathematical equations that underlie machine learning techniques available in common computational tools.
Quantitative techniques for language use

Some discoveries from CoLaLab about subtle information in language text alone

electronic (more conversational)

written text
Quantitative techniques for language use

Pearl & Steyvers 2013, Pearl & Enverga 2015: Detecting emotions, attitudes, and intentions in short messages
Quantitative techniques for language use

Pearl & Steyvers 2013, Pearl & Enverga 2015: Detecting emotions, attitudes, and intentions in short messages

What features we counted:

n-grams (strings of n units) that abstracted across linguistic constructs

the+best
the+brightest
the+most+fantastic
the+most+fun

the+POSITIVE-ADJECTIVE-IN-THE-SUPERLATIVE
Quantitative techniques for language use

Pearl & Steyvers 2013, Pearl & Enverga 2015: Detecting emotions, attitudes, and intentions in short messages

We then used standard computational tools to achieve better accuracy than previous approaches.
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?

Answer: Yes and no. The features the author manipulated (which did create several fairly distinct characters) weren’t the ones that signified his own style. His own style features were still present.
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?

What features we counted:
character-level (like punctuation)

“I HAVE both your letters at once. It is very unhappy, my dear, … And why? Shall I venture to tell you? – Because they are nearer…”
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?

What features we counted: word-level (like total words)

“I HAVE both your letters at once. It is very unhappy, my dear, ... And why? Shall I venture to tell you?— Because they are nearer...”
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people? ✓ X

What features we counted: syntactic (like first person pronouns)

“I HAVE both your letters at once. It is very unhappy, my dear, … -And why? Shall I venture to tell you? - Because they are nearer...”
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?

✓ X

What features we counted: semantic (like endearments)

“It HAVE both your letters at once. It is very unhappy, my dear, … And why? Shall I venture to tell you? - Because they are nearer...”
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people? ✓ ✗

What features we counted: formatting (like all capitals)

“I HAVE both your letters at once. It is very unhappy, my dear, … — And why? Shall I venture to tell you? - Because they are nearer...”
Quantitative techniques for language use

Pearl, Lu, & Haghighi 2016: Authorship in epistolary novels — can one person (the author) really write in the style of multiple other people?

✓

We then used standard computational tools to determine the stylistic components that were distinct from the author’s own style vs. those that were alike.

“I HAVE both your letters at once. It is very unhappy, my dear, ... And why? Shall I venture to tell you? - Because they are nearer...”
Vogler & Pearl 2019: Can we more accurately detect **deception** across different content domains, like online product reviews, short opinion essays, and transcripts of job interviews?
Quantitative techniques for language use

Vogler & Pearl 2019: Can we more accurately detect deception across different content domains, like online product reviews, short opinion essays, and transcripts of job interviews?

Answer: Yes.

When the content (and form) of the language text changes a lot from sample to sample, we can do much better if we use features that are linguistically-defined and also capture the psychological idea of “specific details”.

\[
\log \left( \frac{p(f_i | T)}{p(f_i | D)} \right)
\]
Quantitative techniques for language use

What features we counted:
Specific details expressed with linguistic constructions like exact numbers and prepositional phrases.

Vogler & Pearl 2019: Can we more accurately detect deception across different content domains, like online product reviews, short opinion essays, and transcripts of job interviews?

It's just two minutes away from Hyde Park.
Quantitative techniques for language use

Vogler & Pearl 2019: Can we more accurately detect deception across different content domains, like online product reviews, short opinion essays, and transcripts of job interviews?

We then used standard computational tools to achieve better detection than previous approaches.

It's just two minutes away from Hyde Park.
How **math** helps us better understand **language**

Math is at the heart of **quantitative** techniques, where we **count** things and **reason** about those counts.
How math helps us better understand language

We can use quantitative techniques to better understand many different questions about the utterly human capacity of language.
How **math** helps us better understand **language**

In language **development**, quantitative techniques are used in **computational cognitive modeling**.
How **math** helps us better understand **language**

In language **use**, quantitative techniques underlie common **computational tools** that are used in combination with insights from psychology and linguistics.
So let’s keep using math to help us better understand language!
Thank you!