

Through the looking glass

What NLP can reveal about
sociolinguistic variation



Ian Stewart

Georgia Tech NLP Seminar

25 September 2020

Language and society

Human language is not monolithic but varies through time and space (Labov 2001).

- Dialects, slang, language varieties

People use language to achieve social goals and construct **social meaning** (Eckert 2008) in interpersonal interactions.



Language variation

When communicating, people have to choose between different **linguistic forms**.

- <I'm going> vs. <I'm goin'>

While typical linguistics investigates **cognitive** systems behind choices, sociolinguistics explains the **social systems** that govern language (Bell 1984).

- What makes Southern American English different from other **dialects**?
- What makes a word socially **acceptable** to say?
- How do **multilingual** people choose between languages?

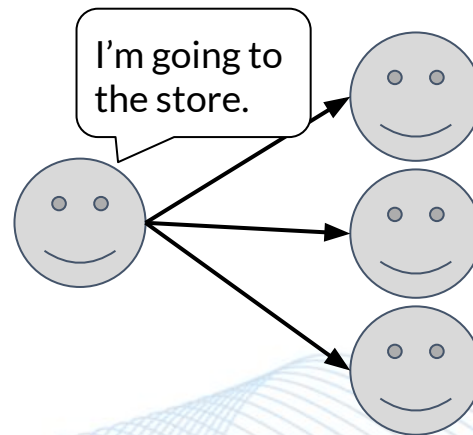
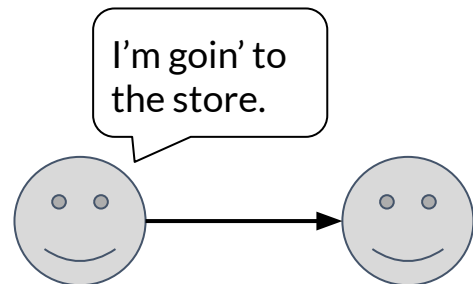
Language variation

When communicating, people have to choose between different **linguistic forms**.

- <I'm going> vs. <I'm goin'>

While typical linguistics investigates **cognitive** systems behind choices, sociolinguistics explains the **social systems** that govern language (Bell 1984).

- What makes Southern American English different from other **dialects**?
- What makes a word socially **acceptable** to say?
- How do **multilingual** people choose between languages?



Language variation: social change

“Like the rest of language, variation does not simply reflect the social, but enacts it, and in the course of this enactment, it participates in **social change**.”

- Penny Eckert



0:00-0:10, 0:36-0:45

Sociolinguistics: data collection

Traditional sociolinguistics (spoken)



I don't like egg I don't like er mayonnaise I was going to make a coffee
about an hour ago
(pause)
dear Sir with reference to your forthcoming holiday with us to Ostend an
explanatory letter that we have today received from H. Limited we
sincerely apologize for the inconvenience this may cause

Coupland (1980)

Computational sociolinguistics (written)



Justin Trudeau @JustinTrudeau · Aug 2
Happy Pride, Vancouver! Enjoy the last day of #VanVirtualPride week.



Justin Trudeau @JustinTrudeau · Aug 2
Bonne Fierté, Vancouver! Profitez bien de la dernière journée de la semaine de la #FiertéVirtuelle de Vancouver.

Twitter (2020)

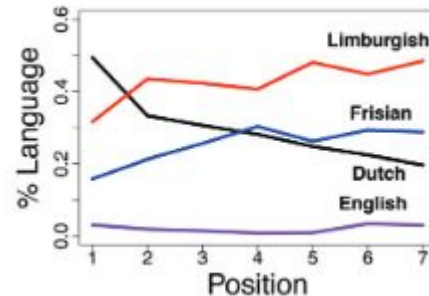
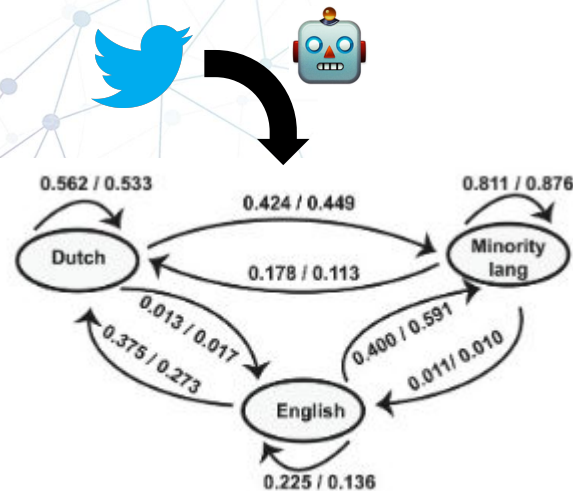
← [- FORMAL] [+ FORMAL] →

Computational sociolinguistics

Computational methods can readily **process** and **categorize** language patterns in online discussions.

NLP unlocks fundamentally different questions for sociolinguists! (Nguyen et al. 2016)

- How often do people **code-switch** with friends vs. strangers?
- What **linguistic factors** contribute to a new word's success?



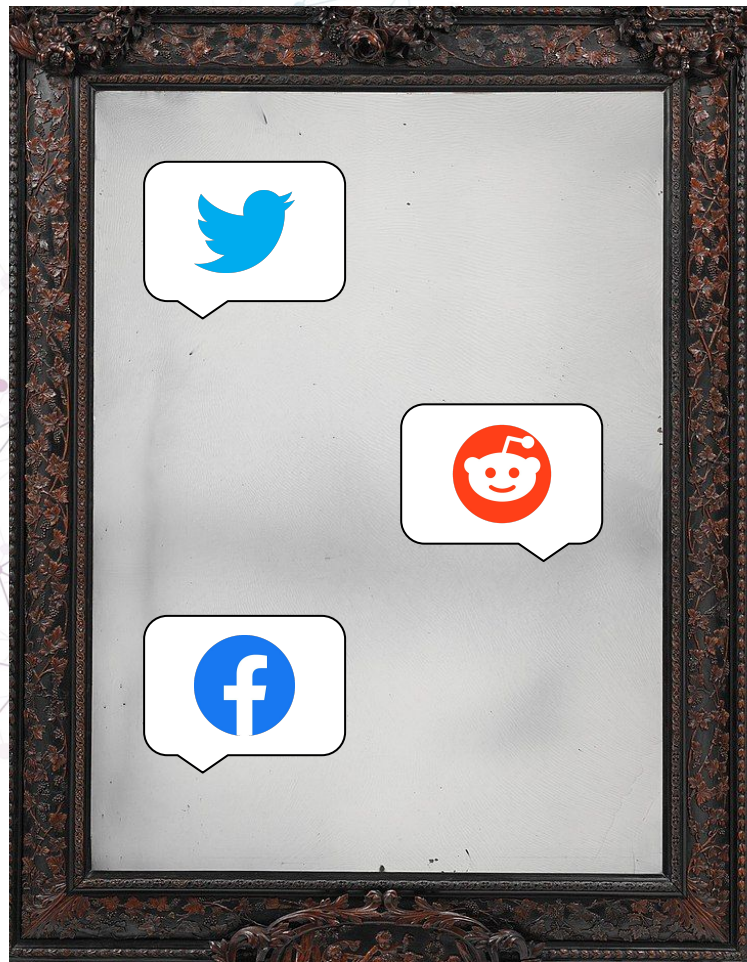
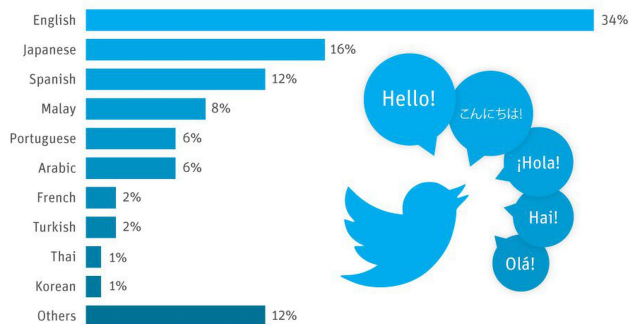
Nguyen et al. (2015)

Social media: a mirror for society?

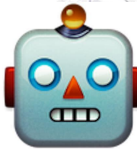
Social media is not a perfect representation of society (Ruths and Pfeffer 2014), but it fosters healthy **diversity of language patterns**.

Only 34% of All Tweets Are in English

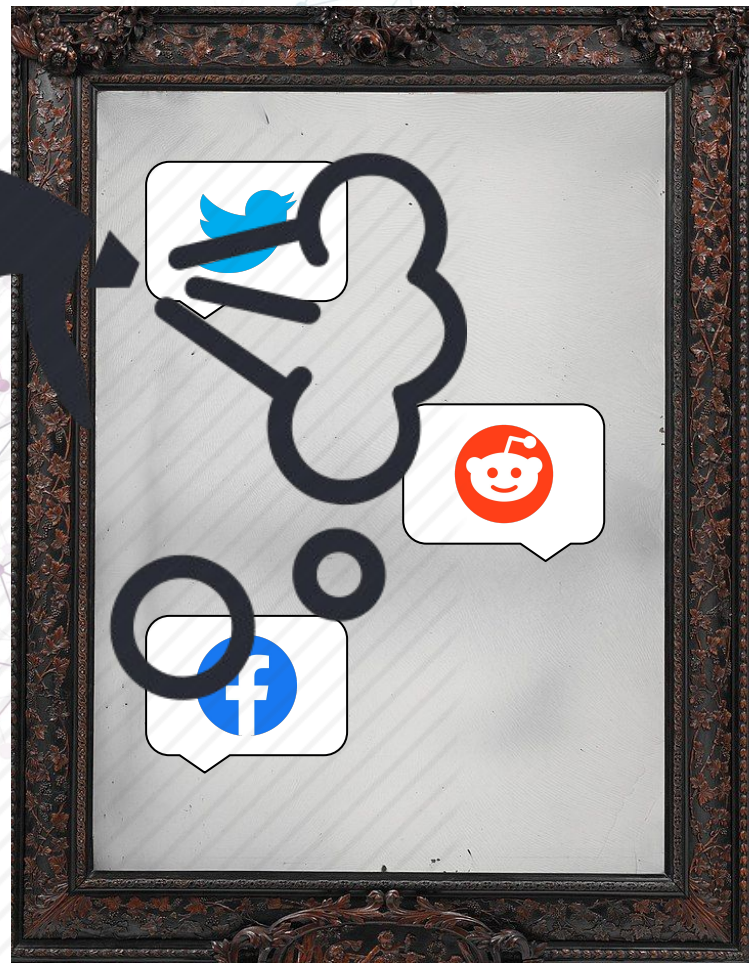
Distribution of languages used in Tweets around the world (September 2013)



Social media: a mirror for society?



dependency parsing
language identification
verb conjugation



Social media: a mirror for society?

NLP gives access to **rare** patterns that take a long time to identify manually.

NLP reveals the full **variety** of language choices that otherwise appear limited.

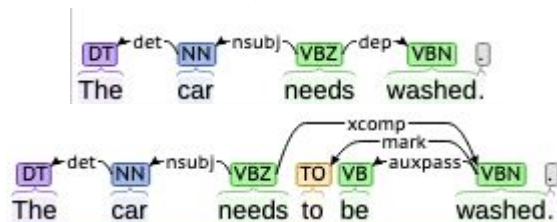
NLP helps linguists capture **complex** patterns that would require difficult judgment.

Yo lo **googleo**.

Van a **googlear** eso?

Si, se **puede**!

Yes we **can**!



Talk outline

Why sociolinguistics?

Benefits of NLP for sociolinguistics

- Rare patterns
- Variety of patterns
- Complex patterns

NLP for social science, and vice versa

Talk outline

Why sociolinguistics?

Benefits of NLP for sociolinguistics

- Rare patterns
- Variety of patterns
- Complex patterns

NLP for social science, and vice versa

Sociolinguistics: rare patterns

Many language choices do not occur all the time but only a few times in a given conversation (code-switching).

Sociolinguistics requires large data sample to test hypotheses about variation, e.g. **consistent audience effects**.

NLP can identify such patterns automatically in large data to uncover robust sample for analysis.

Sociolinguistics: rare patterns

Many language choices do not occur all the time but only a few times in a given conversation (code-switching).

Sociolinguistics requires large data sample to test hypotheses about variation, e.g. **consistent audience effects**.

NLP can identify such patterns automatically in large data to uncover robust sample for analysis.



↑ English



↓ French

Rare patterns: multilingual choices

Speakers have to choose between languages based on their social goals (Myers-Scotton 1995).

Code-switching is rarer in interviews but common on social media (Shoemark et al. 2017; Rijhwani et al. 2017) and **readily detected** with language identification algorithms (Lui and Baldwin 2012).

Identifying multilingual behavior at scale helps test subtle questions around speaker **attitude**.

Rare patterns: multilingual choices

Speakers have to choose between languages based on their social goals (Myers-Scotton 1995).

Code-switching is rarer in interviews but common on social media (Shoemark et al. 2017; Rijhwani et al. 2017) and **readily detected** with language identification algorithms (Lui and Baldwin 2012).

Identifying multilingual behavior at scale helps test subtle questions around speaker **attitude**.



República
Catalana ara!

(Catalonian Republic now!)

República
Catalana ahora!



Sí o no, ¿què penses?

Catalonian Independence and Linguistic Identity on Social Media

Ian Stewart, Diyi Yang, Jacob Eisenstein
NAACL 2018

Language choice: political attitudes



<http://www.noticiasdegipuzkoa.eus/2018/03/28/politica/intensa-jornada-de-protestas-para-defender-la-republica-catalana>



<https://www.heraldo.es/noticias/nacional/2017/09/22/un-centenar-personas-manifiesta-frente-sede-anc-por-unidad-espana-1198045-305.html>

How does local language use correlate with political attitudes, within an independence movement?

Pro-independence

██████████ · 23 Apr 2017
A punt pel Sí a la [#republicacatalana](#) [#viuSantJordi](#) [#Valls](#) [#esquerra_valls](#)

██████████ · 3 Oct 2017
Entiende [@marianorajoy](#) Si tu limitada capacidad de deja. [#CatalunaLibre](#) te manda este mensaje!!

██████████ · 30 Sep 2017
Piolín, no ens oblidem de tu! [#EscolesObertes](#) [#FreePiolin](#) [#NoPassaran](#)

Anti-independence

██████████ · 9 Oct 2017
Replying to [@el_pais](#)
[@AdaColau](#) está cagadita, al igual que todos los independentistas.
[#EspanaEnPie](#) [#EspanaUnida](#)

██████████ · 4 Oct 2017
[@JoanTarda](#) Ya en tu foto de perfil se ve que eres un descerebrado iluminado. [#CatalunaEsEspana](#) Y si te molesta, vete a Francia.

██████████ · 4 Oct 2017
Me explica alguien eso de la "resistencia pacífica"? [#CatalanReferendum](#)
[#Constitucion](#) [#10Oct](#) [#EspanaNoSeRompe](#)

Pro-independence

██████████ · 23 Apr 2017
A punt pel Sí a la [#republicacatalana](#) [#viuSantJordi](#) [#Valls](#) [#esquerra_valls](#)

Catalan



██████████ · 3 Oct 2017
Entiende [@marianorajoy](#) Si tu limitada capacidad de deja. [#CatalunaLibre](#) te manda este mensaje!!

Spanish



██████████ · 9 Oct 2017
Replying to [@el_pais](#)
[@AdaColau](#) está cagadita, al igual que todos los independentistas.
[#EspanaEnPie](#) [#EspanaUnida](#)

Spanish



██████████ · 4 Oct 2017
[@JoanTarda](#) Ya en tu foto de perfil se ve que eres un iluminado. [#CatalunaEsEspana](#) Y si te molesta, vete a Francia.

Spanish



██████████ · 30 Sep 2017
Piolín, no ens oblidem de tu! [#EscolesObertes](#) [#FreePiolin](#) [#NoPassaran](#)

Catalan



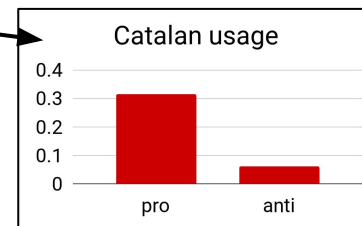
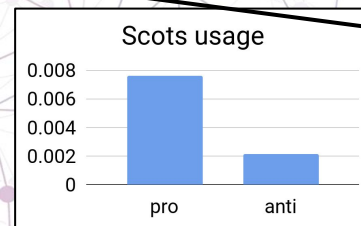
██████████ · 4 Oct 2017
Me explica alguien eso de la "resistencia pacífica"? [#CatalanReferendum](#)
[#Constitucion](#) [#10Oct](#) [#EspanaNoSeRompe](#)

Spanish



Language choice: results

In discussions of Catalonia referendum, **pro-independence** speakers chose minority language more often than anti-independence speakers.

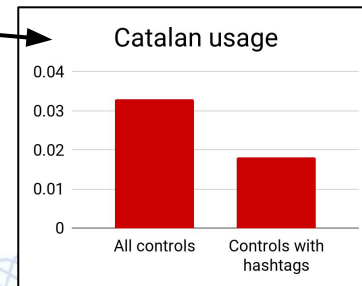
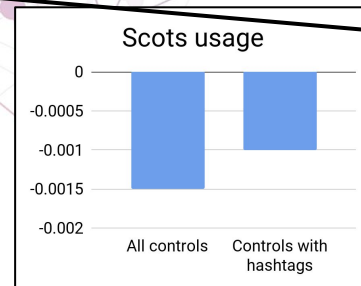
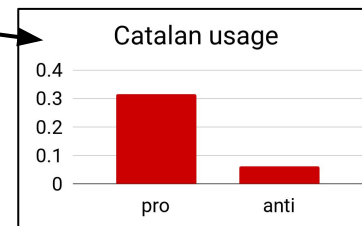
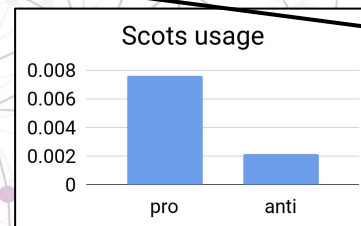


Language choice: results

In discussions of Catalonia referendum, **pro-independence** speakers chose minority language more often than anti-independence speakers.

All speakers chose minority language more often in **any referendum discussion** than in control discussion.

Political value of minority language as identity marker is **stronger** than similar Scottish scenario.



Language choice: cultural attitudes

For multilingual people, language choices may reflect **cultural attitudes** (Auer 2013).

Speakers may adopt more English **loanwords** if they feel aligned to US/UK culture.

Loanwords are **rare in spoken conversation** (Poplack et al. 1988): NLP can help uncover large sample with language identification.

Language choice: cultural attitudes

For multilingual people, language choices may reflect **cultural attitudes** (Auer 2013).

Speakers may adopt more English **loanwords** if they feel aligned to US/UK culture.

Loanwords are **rare in spoken conversation** (Poplack et al. 1988): NLP can help uncover large sample with language identification.



Diccionario de la lengua española

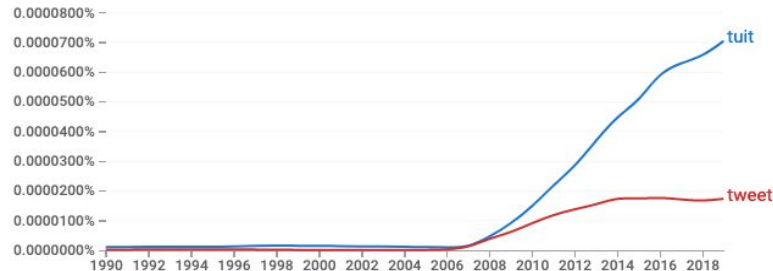
Edición del Tricentenario

Actualización 2019

tuit

Del ingl. *tweet*.

1. m. Mensaje digital que se envía a través de la red social Twitter® y que no puede rebasar un número limitado de caracteres.



The background features a complex network graph with nodes and edges in various colors (purple, blue, green, orange). In the top-left and bottom-right corners, there are decorative wavy lines in orange and blue respectively. The main text is centered over the network graph.

Tuiteamos o pongo un tweet?

Characterizing the social constraints of English loanword integration in Spanish social media

Ian Stewart, Diyi Yang, Jacob Eisenstein
in preparation

Loanword integration

Speakers do not always adopt loanwords “as-is” but often *integrate* them to align to their native language grammar (Kang 2011).

Speakers tend to use *integrated* forms if they are more linguistically conservative (Poplack 1988) - does this apply to cultural attitudes as well?

How do **cultural attitudes** affect integration of loanwords?

less integrated

poner un tweet

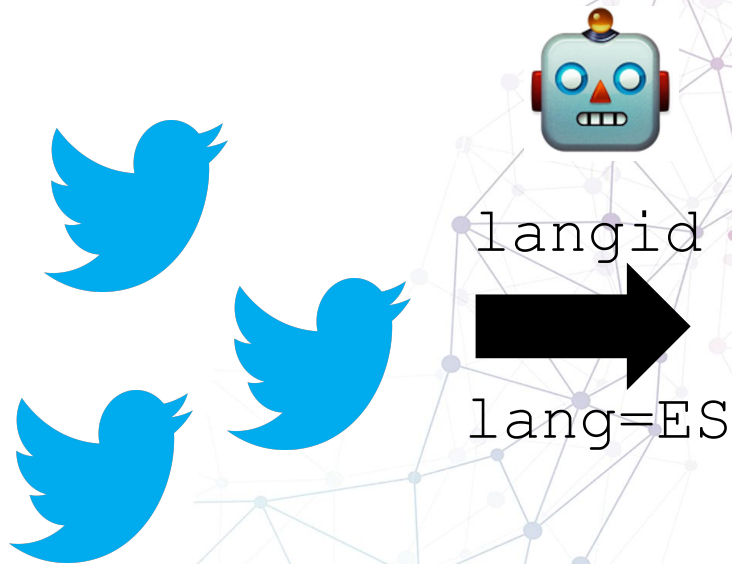
(send a tweet)

tweetear

(tweet)

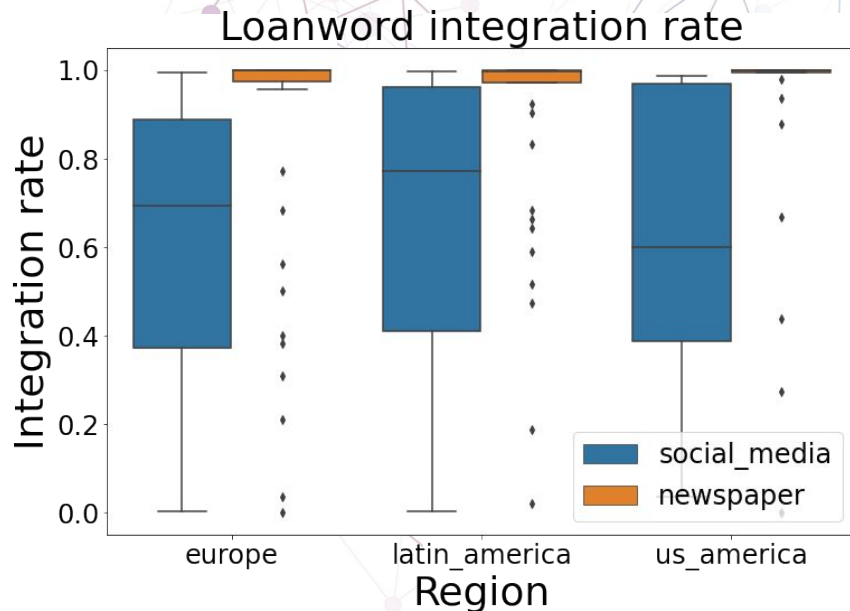
more integrated

Loanword integration: data



Loanword	Verbs	Count
Like	<i>likear, dar un like</i>	13,154
Connect	<i>conectar, hacer un conexión</i>	7857
Flip	<i>flippear, hacer flip</i>	6904
Stalk	<i>stalkear, ser un stalker</i>	5508
Tweet	<i>tweetear, poner un tweet</i>	5294

Loanword integration: domain results



Newspapers use integrated verbs at **higher rate** than social media: integration relates to **formality** (conservative cultural attitude?).

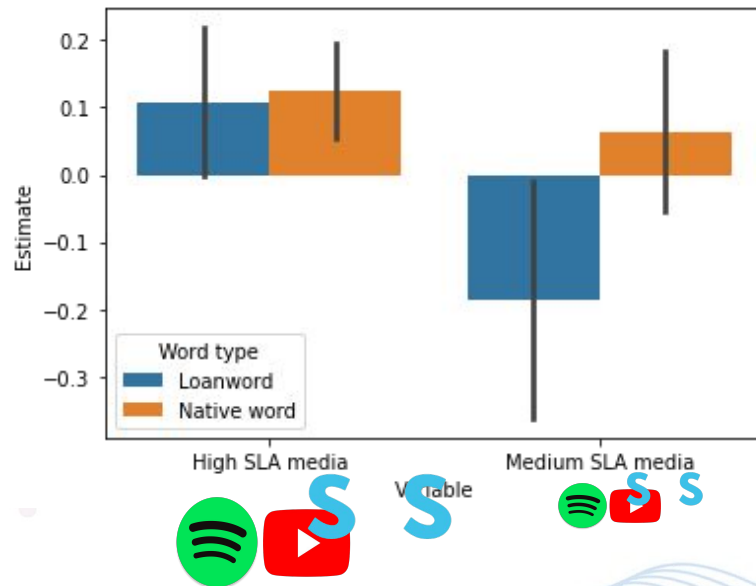
Loanword integration: speaker results

Q: which types of speakers on social media tend to use the integrated forms?

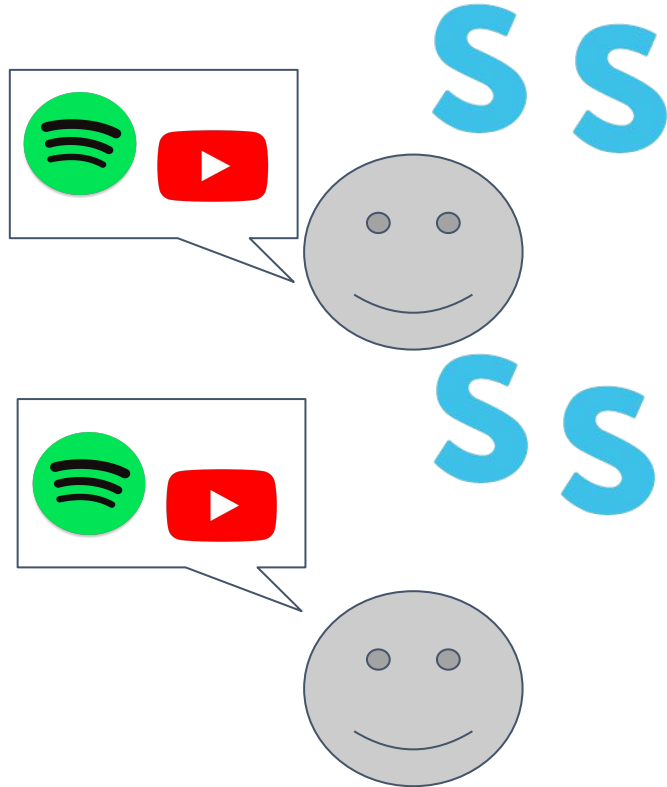
Cultural attitude does not explain loanword integration ($\beta=0.108, p > 0.05$) but does explain **native verb integration**

($\beta=0.126, p < 0.01$).

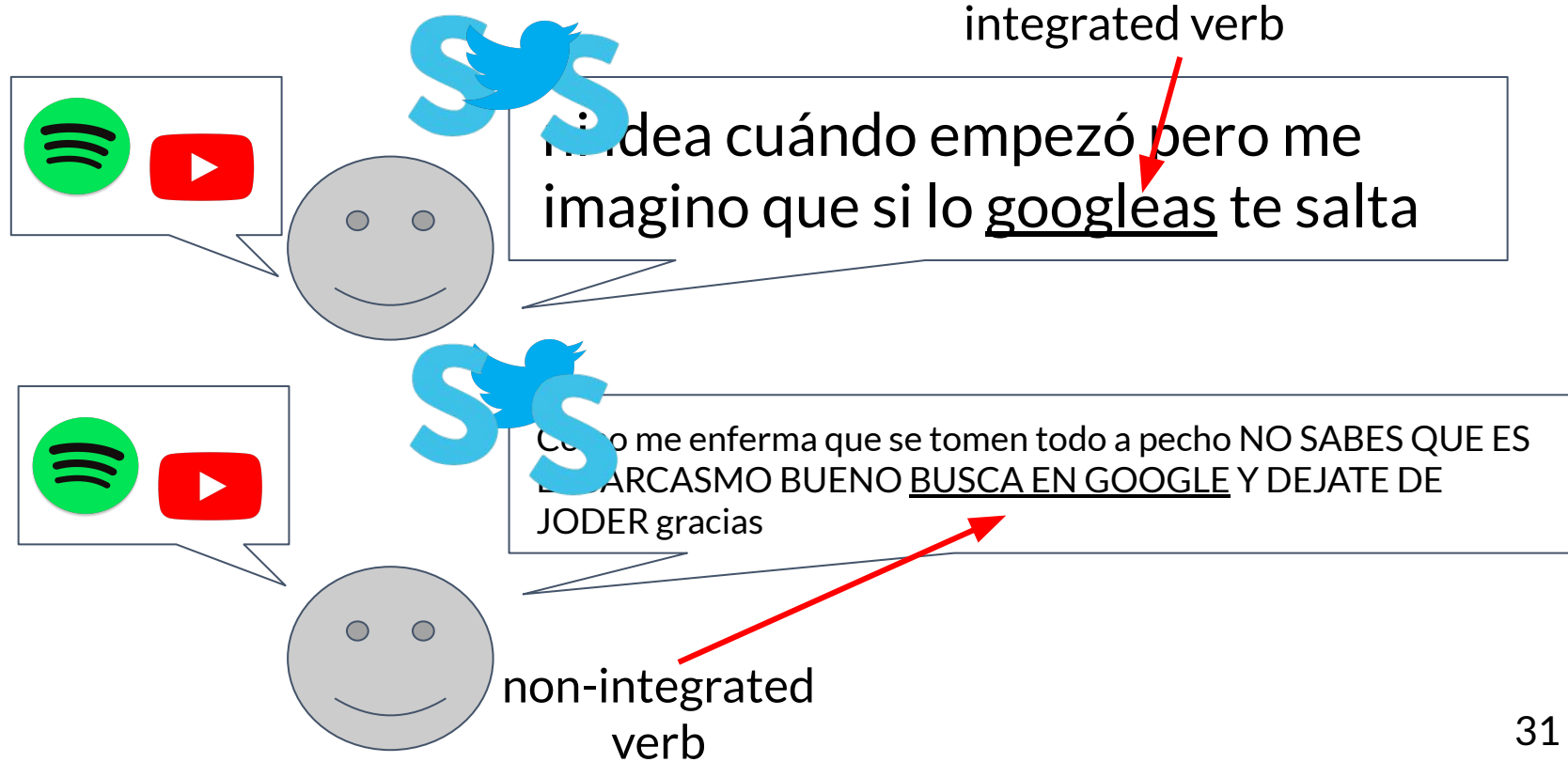
- Native verbs are more well-known among speakers, may have **stronger connection** to language norms and latent social attitudes.



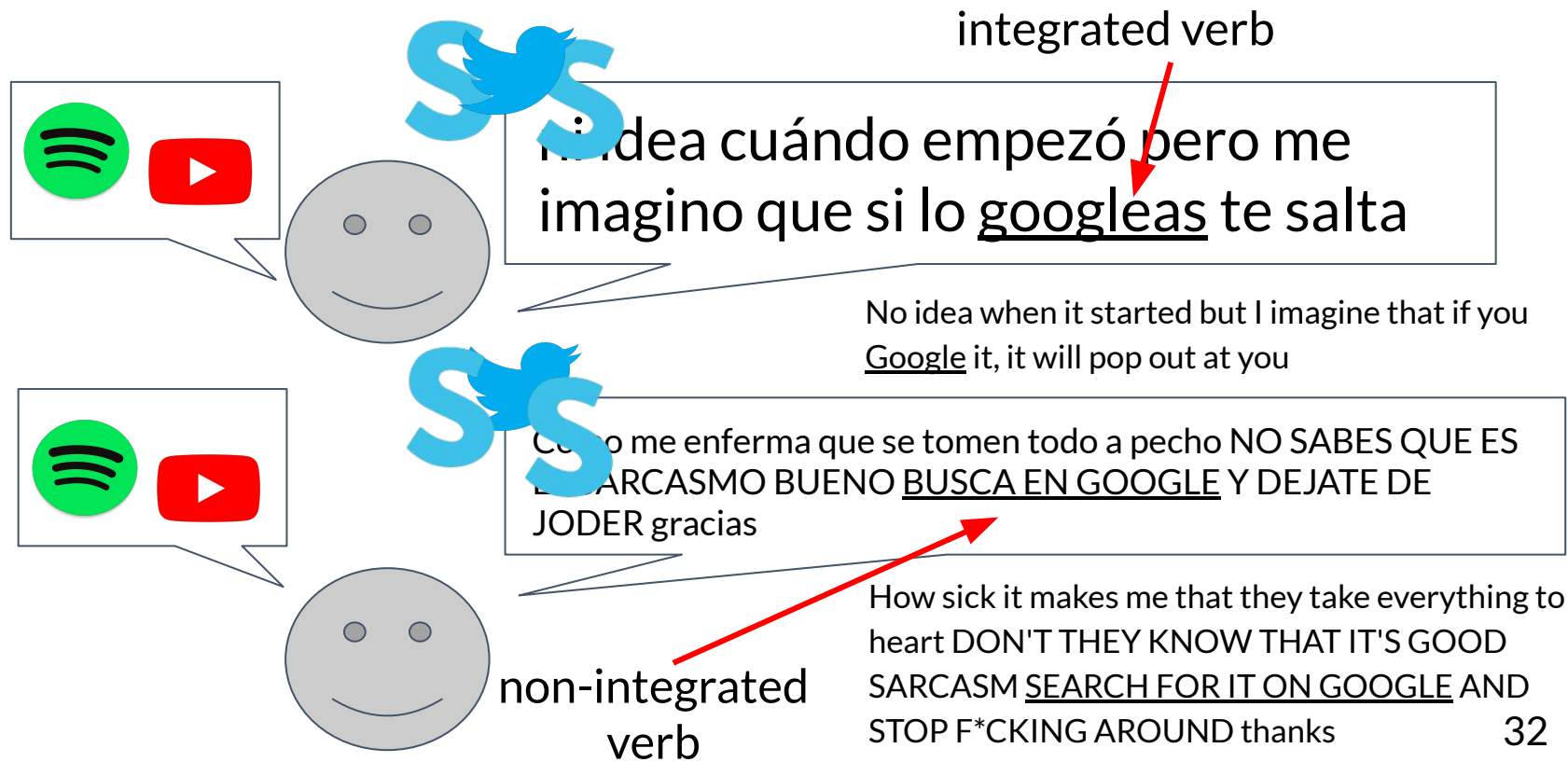
Loanword integration: attitude result



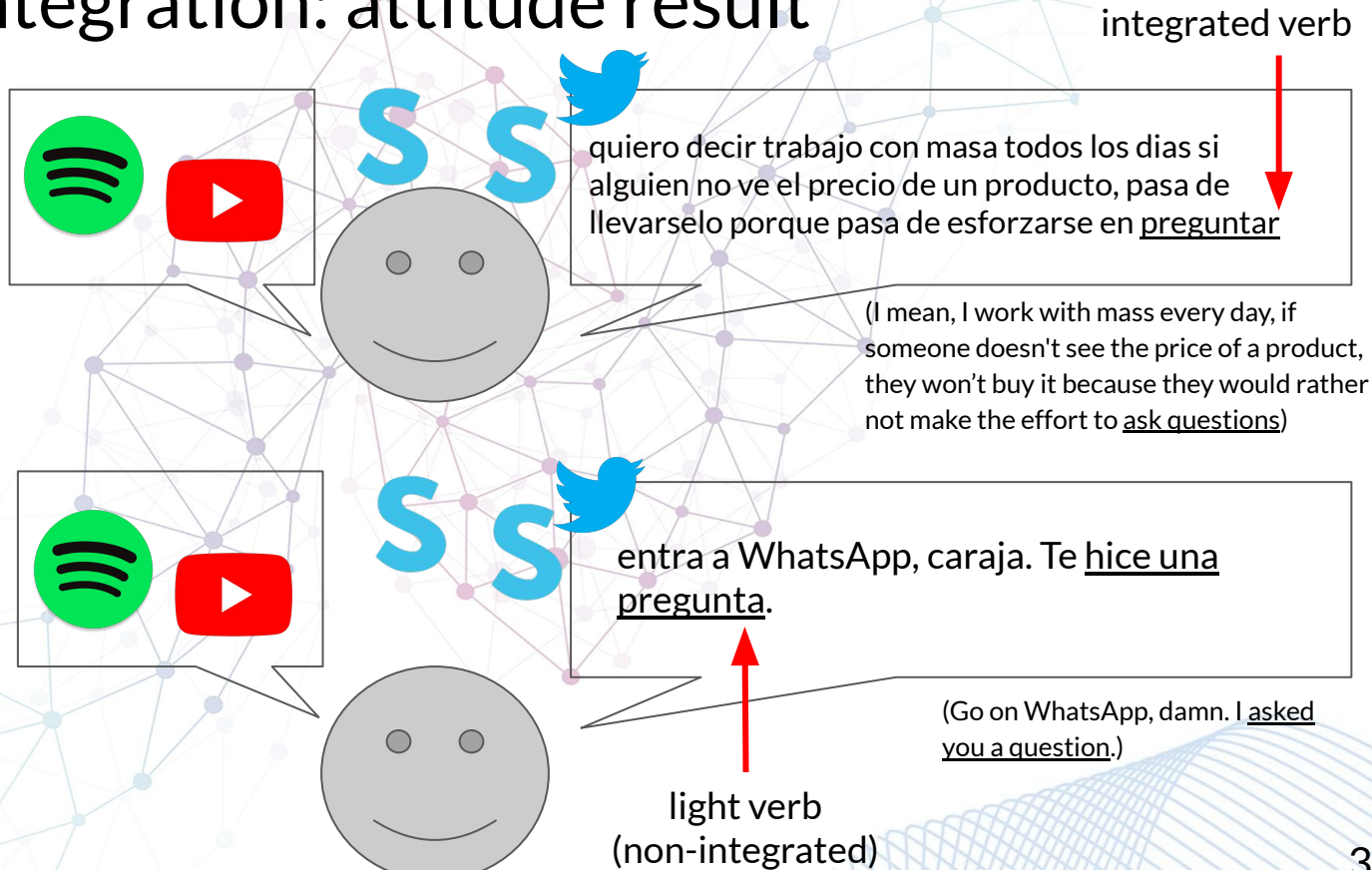
Loanword integration: attitude result



Loanword integration: attitude result



Loanword integration: attitude result



Loanword integration: summary

Verb integration is likely considered more **formal** → social factors matter for **grammatical decision!**

Authors who share more Spanish/Latin American music are not more likely to use integrated loanwords: **cultural attitude** less important than e.g. political attitude.

NLP helps to **magnify rare multilingual phenomena** in social media.

Talk outline

Why sociolinguistics?

Benefits of NLP for sociolinguistics

- Rare patterns
- Variety of patterns
- Complex patterns

NLP for social science, and vice versa

Sociolinguistics: variety of patterns

NAME + **DESCRIPTION**



Atlanta, **GA**



Atlanta, the
capital of **Georgia**

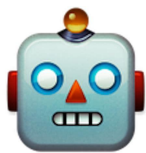


Atlanta and Georgia

Some language patterns are relatively common but difficult to observe in their full **variety**.

- Spelling (cooooooool)
- Syntax

NLP can capture the full variety of a pattern with **generic representations**, e.g. dependency parses.

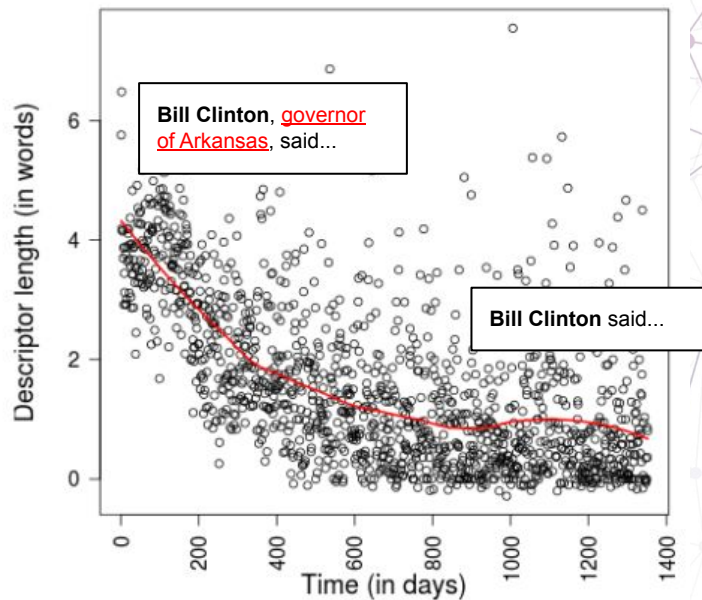




Characterizing collective attention via descriptor context: A case study of public discussions of crisis events

Ian Stewart, Diyi Yang, Jacob Eisenstein
ICWSM 2020

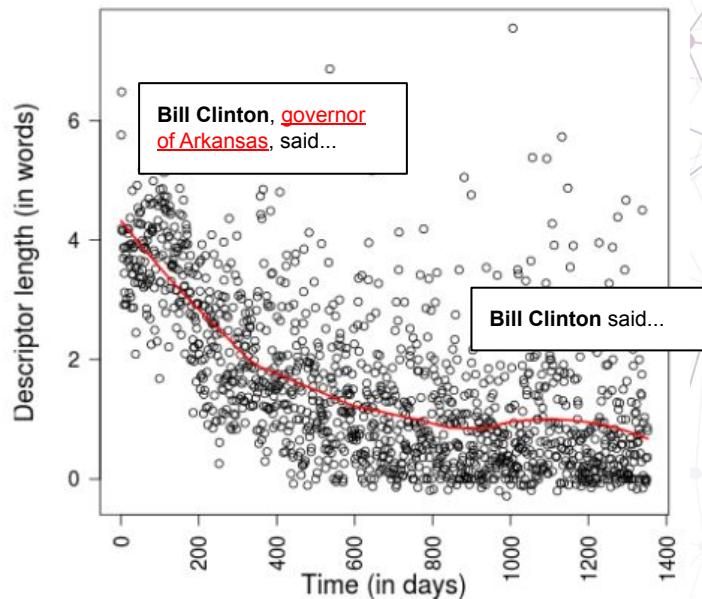
Collective attention: tracking expectations



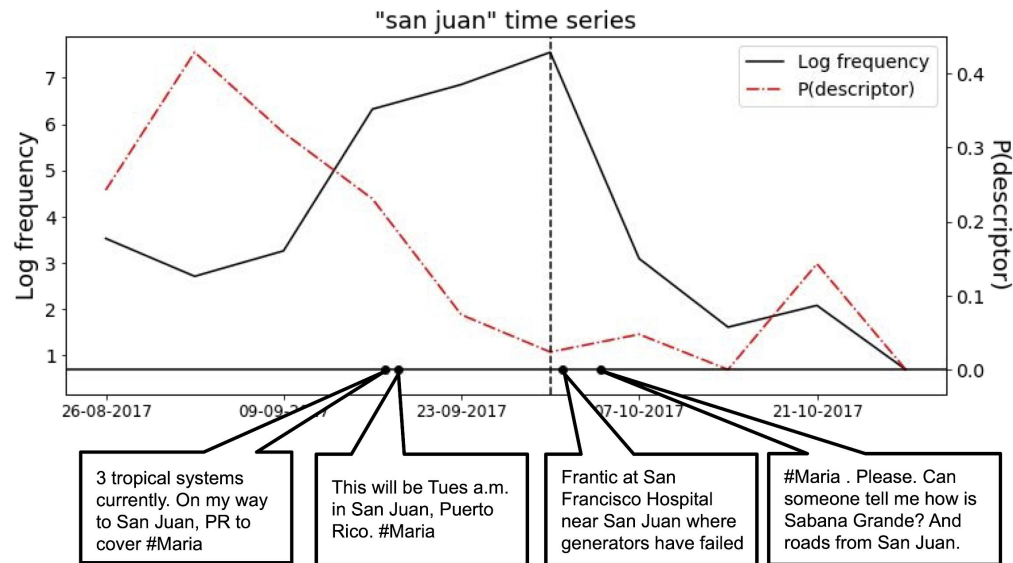
Staliunaite et al. (2018)

Descriptive information use reveals expectations about **audience** (Prince 1992) and robust to data sampling error.

Collective attention: tracking expectations

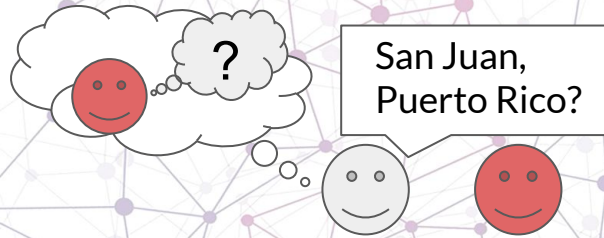


Staliunaite et al. (2018)



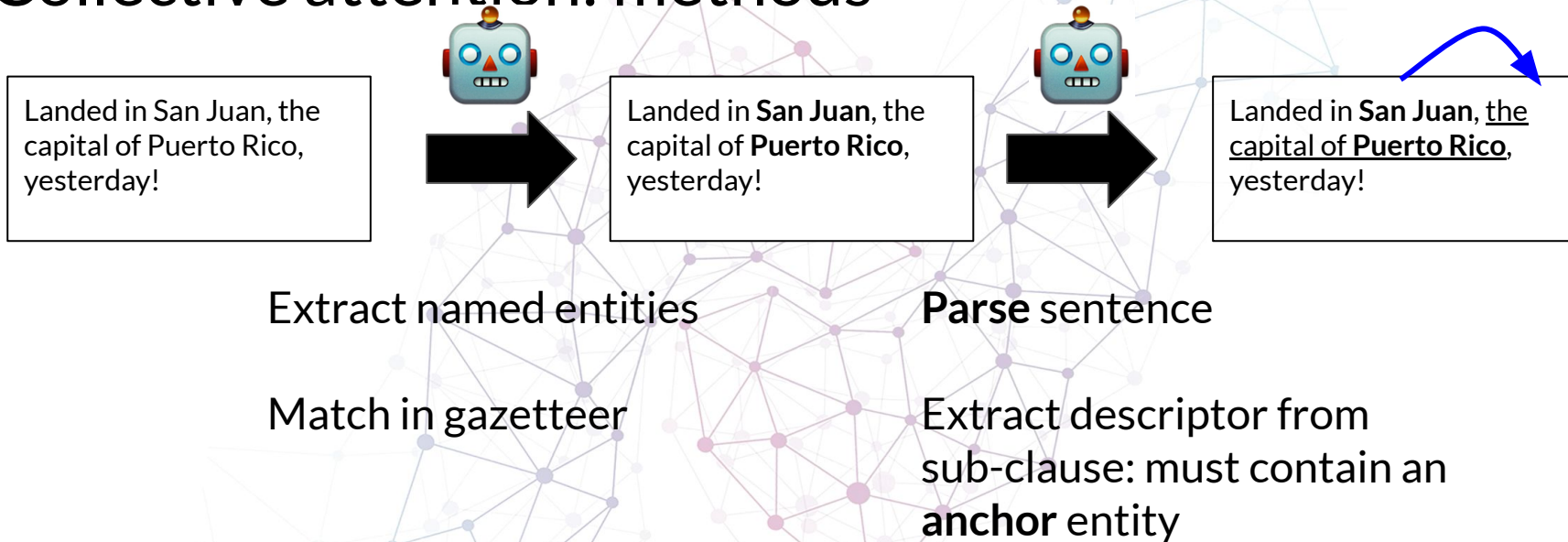
Descriptive information use reveals expectations about **audience** (Prince 1992) and robust to data sampling error.

Collective attention: study goal



How do people change their use of description information in reaction to crisis events?

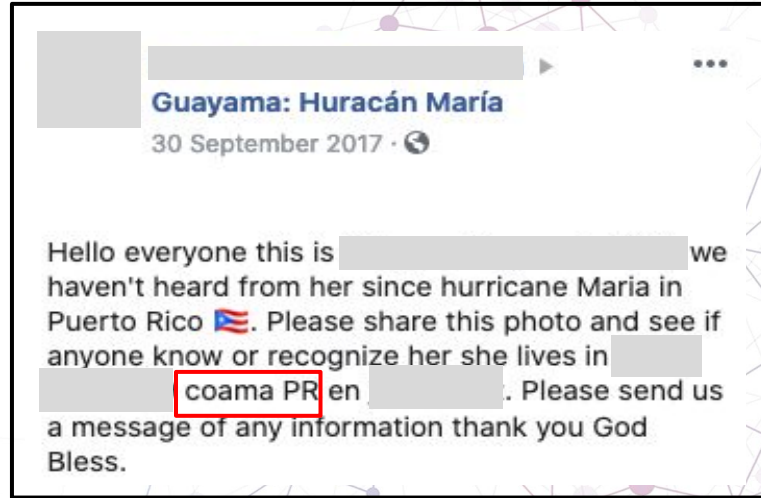
Collective attention: methods



Collective attention: data



60 public groups
(Hurricane Maria)

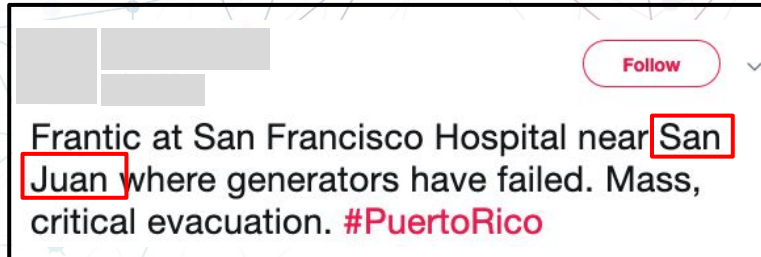


English/Spanish

30,000
posts



Hashtags for 5 hurricanes
(Florence, Harvey, Irma,
Maria, Michael)



English/Spanish

2,000,000
tweets

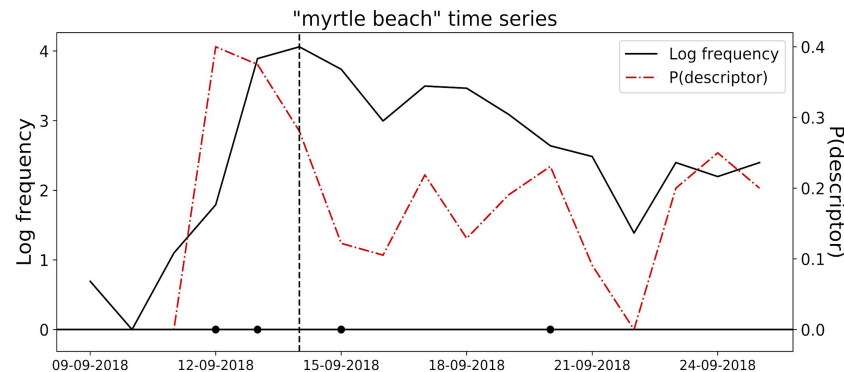
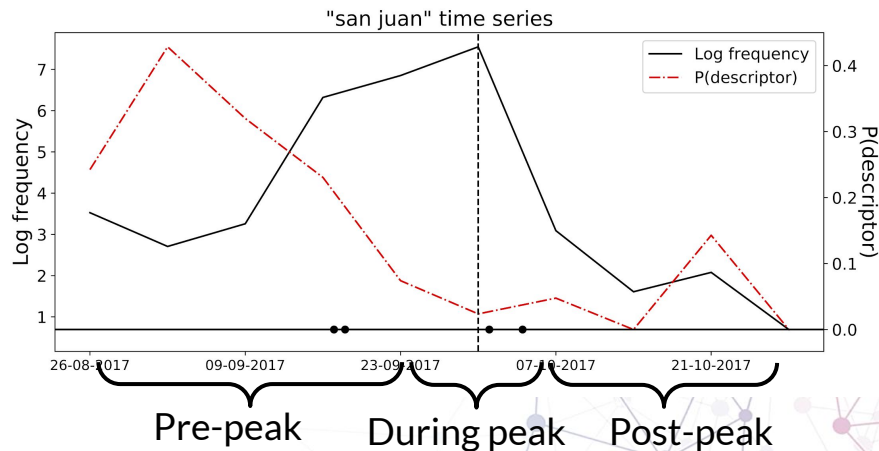
Collective attention: results



Authors add descriptors for **non-local** and **less popular** locations.

($\beta=0.623, p < 0.05$; $\beta=0.172, p < 0.05$)

Collective attention: results



Authors drop descriptors **after** the peak in post volume. ($\beta = -0.127$, $p < 0.05$)

Collective attention: results

CBS Evening News
@CBSEveningNews

"No access to major hospitals" once winds reach a sustained 40 mph @JerickaDuncan reports near Jacksonville, Florida



6:57 PM · Sep 10, 2017 · SnapStream TV Search

CBS Evening News
@CBSEveningNews

Jacksonville Sheriff's office hopes the 356 people they rescued "will take evacuation orders seriously" in future
[cbsn.ws/2xj2LVx](https://www.cbsn.ws/2xj2LVx)



1:55 PM · Sep 12, 2017 · Sprinklr

Highly active authors drop descriptors after prior mentions.

($\beta = -0.237$, $p < 0.05$)

Collective attention: results



Highly active authors drop descriptors after more social engagement. ($\beta=0.292, p < 0.05$)

Collective attention: takeaways

Authors modulate their use of descriptors on social media to **accommodate variable communication needs.**

NLP helps address **variety** of descriptor forms that relate to the same underlying pattern of variation.

Talk outline

Why sociolinguistics?

Benefits of NLP for sociolinguistics

- Rare patterns
- Variety of patterns
- Complex patterns

NLP for social science, and vice versa

Sociolinguistics: complex patterns

Some aspects of language variation are easy to intuitively understand but hard to quantify.

The **linguistic utility** of a new word can play a key role in its acceptance, but “utility” is complex.

NLP provides statistical methods to capture “squishy” linguistic constructs that relate to variation.

fetch



that's so **fetch**

this **fetch** purse

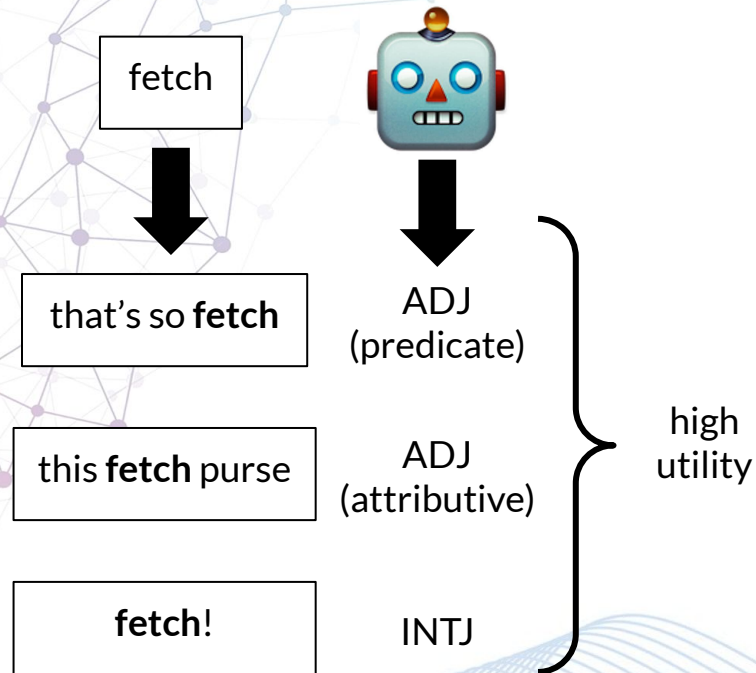
fetch!

Sociolinguistics: complex patterns

Some aspects of language variation are easy to intuitively understand but hard to quantify.

The **linguistic utility** of a new word can play a key role in its acceptance, but “utility” is complex.

NLP provides statistical methods to capture “squishy” linguistic constructs that relate to variation.



The background features a complex network graph with nodes and edges in various colors (purple, blue, green, orange). Overlaid on this are several wavy, concentric lines in orange and blue, creating a layered, abstract effect.

Making “fetch” happen: the influence of social and linguistic context on nonstandard word growth and decline

Ian Stewart, Jacob Eisenstein
EMNLP 2018

Word adoption


New words come and go constantly in online communities, due to turnover in membership (Danescu-Niculescu-Mizil et al. 2013) and need for new conventions to fill a role (Kooti et al. 2016).

Not all words last! What enables a word to outlast its competitors?

af 

Stands for 'as f  . The SI unit of everything.

asf 

A wrong way of abbreviating the words "as f  the right one being "af" (or "a.f./AF/A.F").

Word adoption: what factor is most important?



Does the **social** context of a word influence its adoption more than its **linguistic** context?

Word adoption: social dissemination

Observed **social count** normalized by expected count. (Altmann et al. 2011)

Compute for: users (\mathbf{D}^U), subreddits (\mathbf{D}^S), threads (\mathbf{D}^T).

user dissemination

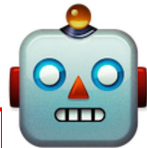
$$D_{(w)}^U = \log \frac{U_{(w)}^{\text{observed}}}{\tilde{U}_{(w)}^{\text{expected}}}$$

Word adoption: linguistic dissemination

Observed count of **trigram contexts** normalized by expected count.

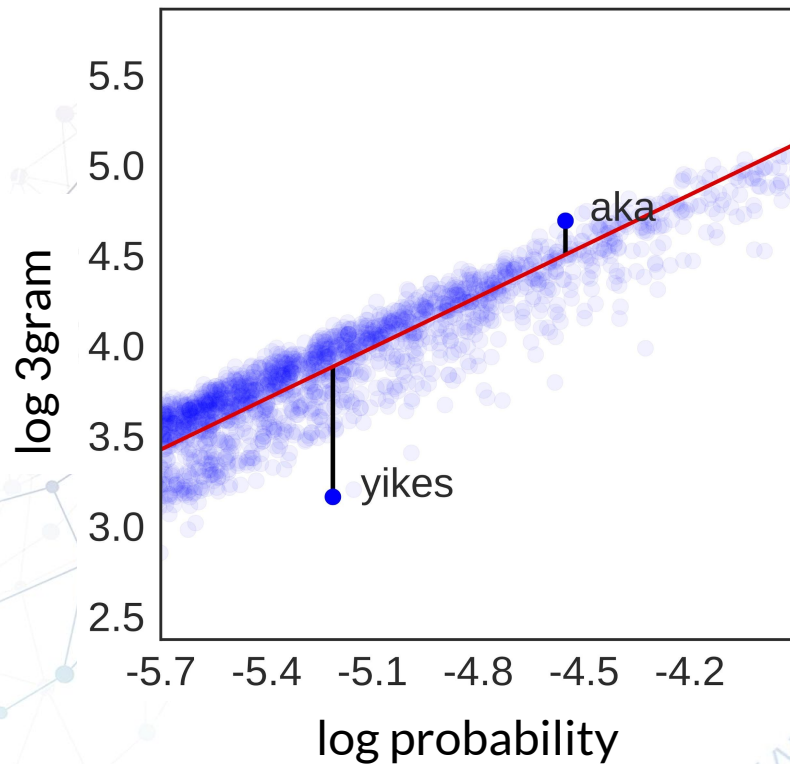
Scalable to 1000s of words without memory problems.

Similar to prior “dispersion” metrics. (Chesley and Baayen 2010)

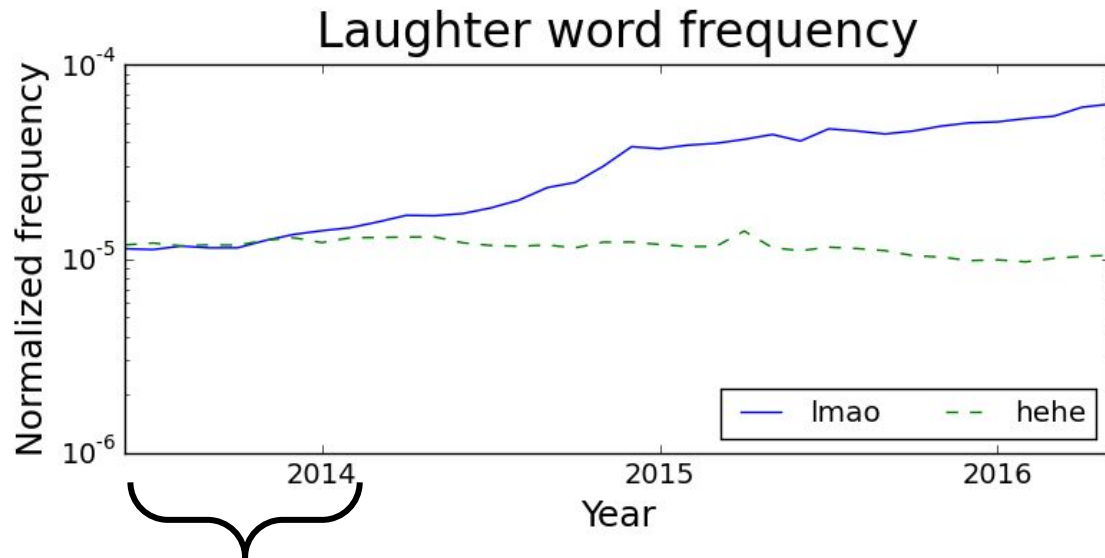


$$D_{(w)}^L = \log \frac{C_{(w)}^3}{\tilde{C}_{(w)}^3}$$

Word adoption: linguistic dissemination

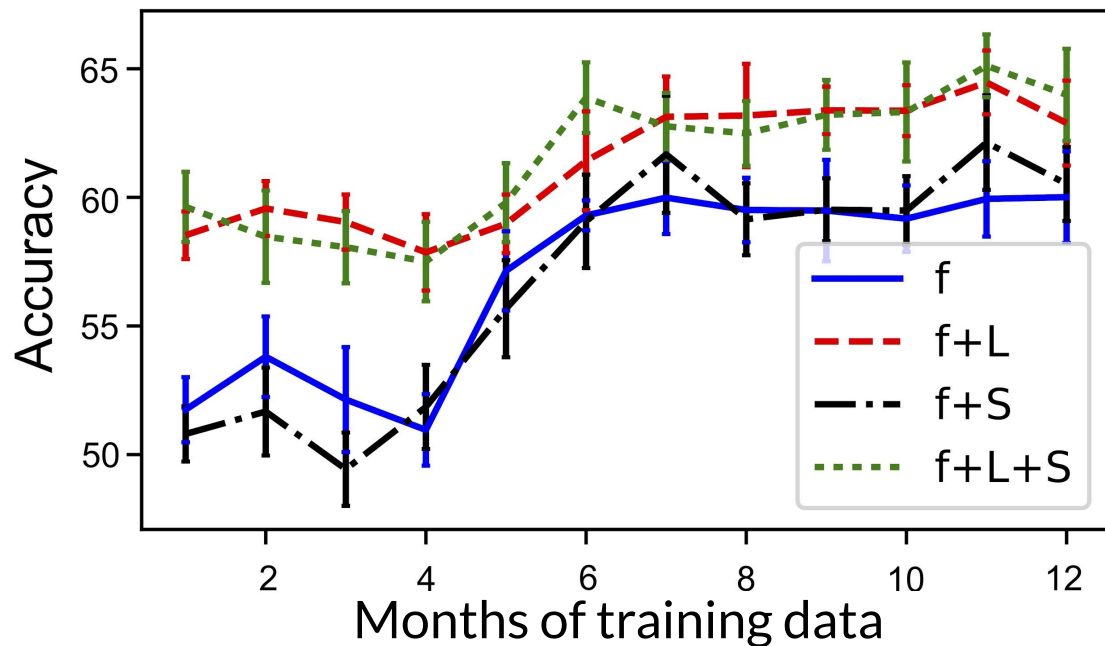


Word adoption: prediction task



Predict whether a nonstandard word will **grow or decline**, based on the word's early levels of dissemination.

Word adoption: prediction results



Models:

f = frequency

L = linguistic dissemination

S = social dissemination

Linguistic dissemination is a strong predictor of word adoption, even in early stages!

Word adoption: takeaways

Linguistic dissemination plays an important role in word adoption and abandonment in online communities (Metcalf 2003).

NLP helps to quantify **complex concept** of linguistic utility for better analysis of language change.

Social media: a mirror for society?

NLP gives access to **rare** patterns that take a long time to identify manually.

NLP reveals the full **variety** of language choices that otherwise appear limited.

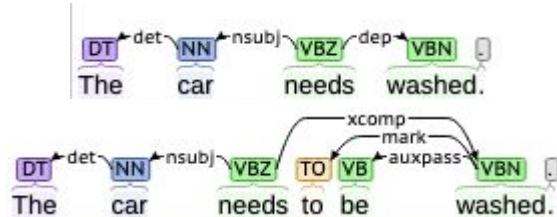
NLP helps linguists capture **complex** patterns that would require difficult judgment.

Yo lo **googleo**.

Van a **googlear** eso?

Si, se **puede**!

Yes we **can**!



Talk outline

Why sociolinguistics?

Benefits of NLP for sociolinguistics

- Rare patterns
- Variety of patterns
- Complex patterns

NLP for social science, and vice versa

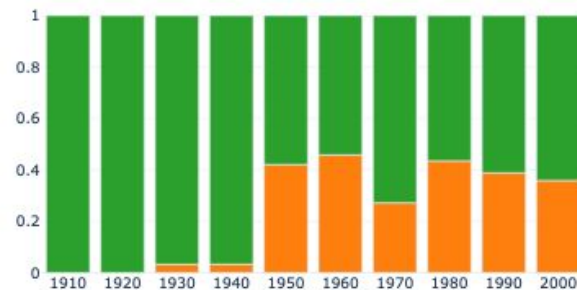
What do sociolinguists need from NLP?

More **accurate** tools for more subtle phenomena like semantics (Shoemark et al. 2020).

- Which senses of *lol* are the most socially prominent (e.g. sarcastic vs. literal)?

Better coverage of non-English languages.

- Lots of variety even within English (AAE) that many tools consider “noisy” and therefore outside the scope of detection (Blodgett et al. 2016).



■ the polished *disk* // a *disk* on a rigid backing
■ floppy and hard-*disk* drives // portable *disk*-radio
(d) *disk*

Giulianelli et al. (2020)

What do social scientists need from NLP?

Good social science requires accurate measurements of **complicated perceptual** constructs:

- Attitudes, beliefs (Augenstein et al. 2016)
- Relationships (Chang et al. 2020)
- Deception (Niculae et al. 2015)

Surveys and interviews can address these constructs, but not **at scale** and not in **rapidly changing environments** like social media.

With the right data and assumptions, NLP can extract social constructs from text data to boost social science research.

- Estimates of **uncertainty** (90% chance that X has negative sentiment)
- **Explanations** of language use (X is negative because of “messed up”)

What do social scientists need from NLP?

Good social science requires accurate measurements of **complicated perceptual constructs**:

- Attitudes, beliefs (Augenstein et al. 2016)
- Relationships (Chang et al. 2020)
- Deception (Niculae et al. 2015)

Surveys and interviews can address these constructs, but not **at scale** and not in **rapidly changing environments** like social media.

With the right data and assumptions, NLP can extract social constructs from text data to boost social science research.

- Estimates of **uncertainty** (90% chance that X has negative sentiment)
- **Explanations** of language use (X is negative because of “messed up”)



actually, COVID-19 has really messed up my life



90% negative → COVID
50% willingness to
disclose information

What does NLP need from sociolinguists?

Traditional NLP has focused on representing language models using **text alone**, assumes that language is “generated” without social context.

Recent work incorporates **demographics** into NLP models for better performance (Hovy 2015; Garimella et al. 2017) .

Sociolinguistics provides social **constructs** that can inform language modeling!

- Audience, relationships, community norms...

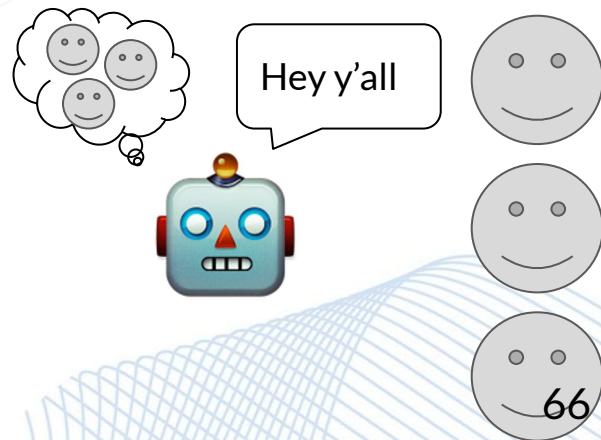
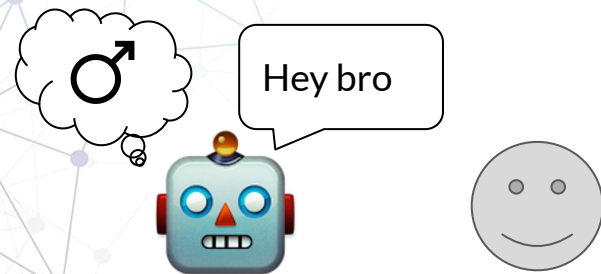
What does NLP need from sociolinguists?

Traditional NLP has focused on representing language models using **text alone**, assumes that language is “generated” without social context.

Recent work incorporates **demographics** into NLP models for better performance (Hovy 2015; Garimella et al. 2017) .

Sociolinguistics provides social **constructs** that can inform language modeling!

- Audience, relationships, community norms...



Thanks!

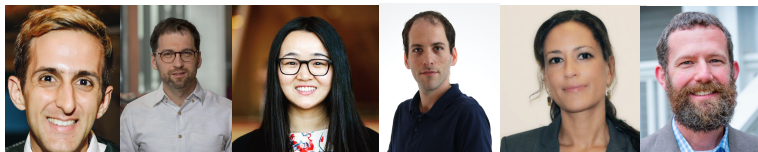


ianbstew@umich.edu
ianbstewart.github.io



Computational sociolinguistics:

<https://www.frontiersin.org/research-topics/9580/computational-sociolinguistics>



Methods questions

1. How do we verify that methods work?
 - a. Intrinsic evaluation, testing against known distributions, etc.
2. What do NLP methods get us that regular word counting does not?
 - a. For “fetch” study, 3-gram counting was good enough, may not even need **semantic representation**.
3. How do we address linguistic biases in social media?
 - a. Forget about population mismatch - need to focus on whether linguistic patterns are **representative of language as whole** or unique to online writing.
4. What **different questions** do sociolinguists get to ask with social media data?
 - a. Methods may expose entirely new patterns (e.g. complicated syntax) but when do we really need **new patterns** (vs. validating known patterns)?