Universität Stuttgart

IRIS

Language in Social Context: Bridging NLP and Sociolinguistics

Agnieszka Faleńska
Filip Miletić

Day 3: Demographic Factors (II)

Image credit: DALLE (2024)

Recap: Demographic Factors of Language Variation

- Language variation associated with speakers' **geographic origin** is due to distance and barriers, but also interacts with other factors
- Language variation is also linked to **socioeconomic status**, which has been extensively theorized but also difficult to operationalize
- Both have been addressed by sociolinguists & NLP practitioners
  - Different sources of data and types of analyses
- Complementary perspectives

Day 3: Outline

Gender

Sex differentiation (Trudgill, 1972), Gender Paradox (Labov, 2001)

Gender variation (Bucholtz, 2002), beyond the binary (Zimman, 2013)

Age

Language variation specific to age groups (Tagliamonte, 2016)

Age as a reflection of language change (Chambers & Heisler, 1999)

Example NLP methods for analysing age- and gender-related variables

Gender

Background

- Major social variables generally include **sex** (binary male/female), mostly analyzed in relation to social class & style differences
- Subsequent studies shift to **gender**
- Only very recent inclusion of **non-binary** views

- Sex ⇒ biological differences between male and female individuals
- Gender ⇒ social, psychological and cultural construct
  ⇒ not necessarily defined by biological sex
  ⇒ not necessarily aligned with a heteronormative binary view

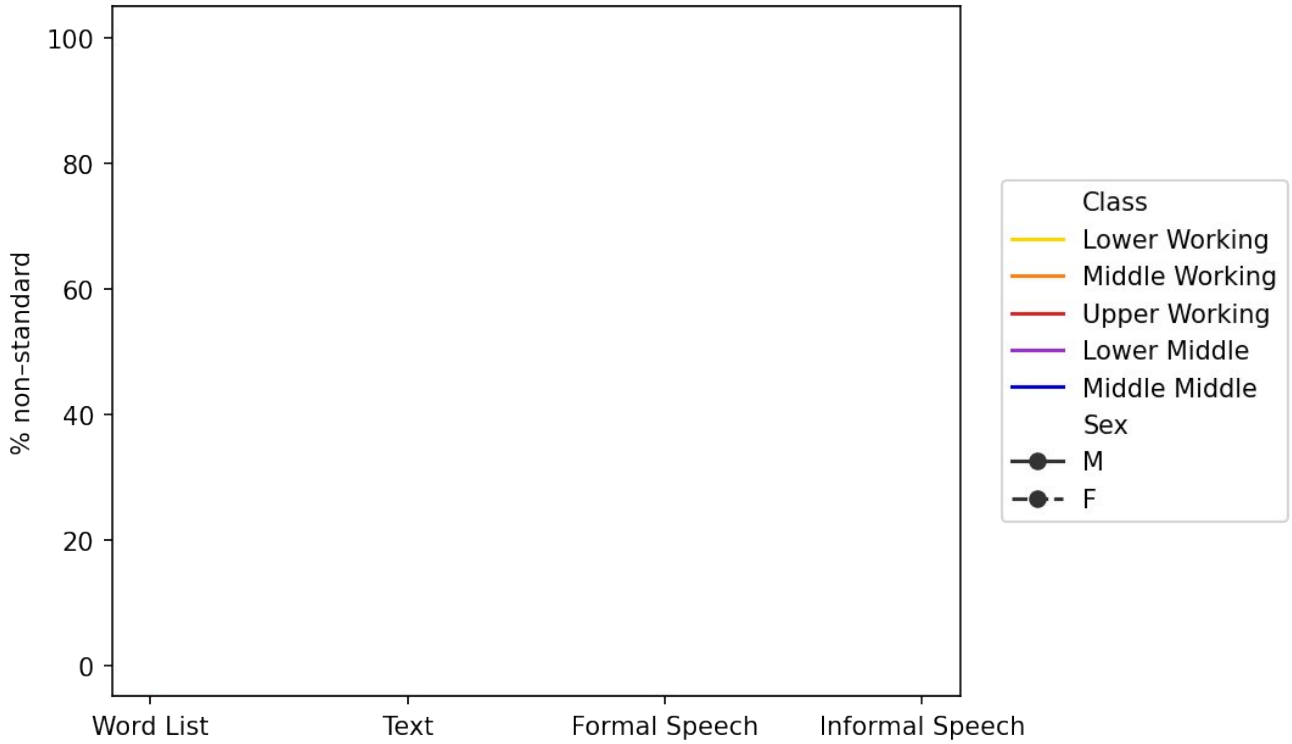
Romaine (2003); Council of Europe <https://www.coe.int/en/web/gender-matters/sex-and-gender>

Trudgill (1972): Social differentiation in Norwich

- Sociolinguistic interviews in Norwich in 1968
- Same protocol as Labov (1966),
  eliciting four styles of different formality
- Random sample (N=60)
- Variable: **(ing)** in *walking*, *laughing*, …
- Explanatory factors:
  age, social class, sex, stylistic context

Map credit: Nilfanion/Wikimedia Commons

Trudgill (1972): Social differentiation in Norwich



Labov (2001): Gender Paradox

- **Stable variables:**
  women use less stigmatized & more prestige variants — (ing) in Norwich
- **Change from above:**
  women adopt prestige forms more — (r) in NYC
- **Change from below:**
  women use innovative forms more — Northern Cities Shift

- **Gender Paradox:**
  Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not.

Romaine (2003): Limitations

Traditional views

- Explanations generally sought for the behavior of women – not men
- Supposed use of language to achieve status that is otherwise out of reach

Issues

- Limited consideration of access to institutions where language-related prestige matters (e.g. education)
- Alternative plausible explanations (e.g. social networks)
- Most early work conducted by men

Bucholtz (2002): From “sex differences” to gender variation

- Style as a social practice:
  “All sociolinguists must understand gender not as a variable that transcends particular situations but as a complex and context-specific system for producing identities and ideologies” (p. 33)

- “A good deal of sociolinguistic work must be done on [transgender] speakers, in all their diversity, if we are to understand the linguistic consequences of stepping outside the sex/gender system.” (p. 35)

Nerd girls			*Cool girls*		
Speaker	**(uw)**	**(ow)**	**Speaker**	**(uw)**	**(ow)**
Bob	88	46	Claudia	115	97
Christine	71	39	Erin	136	129
Claire	56	52	Iris	142	76
Fred	113	78	Josie	138	116
Kate	80	33	Lumiere	130	114
Loden	82	56	Paige	144	111
			Rebecca	148	120
			Sophia	135	138
			Zoe	137	122

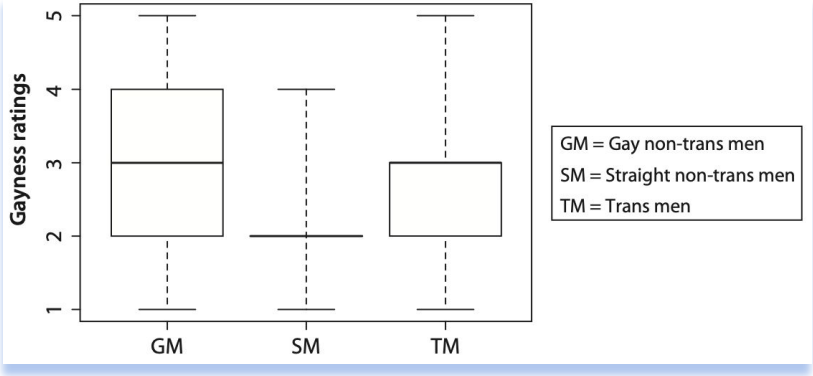
Table 1. Fronting scores for (uw) and (ow) for 15 European American girls at Bay City High School, by cultural style

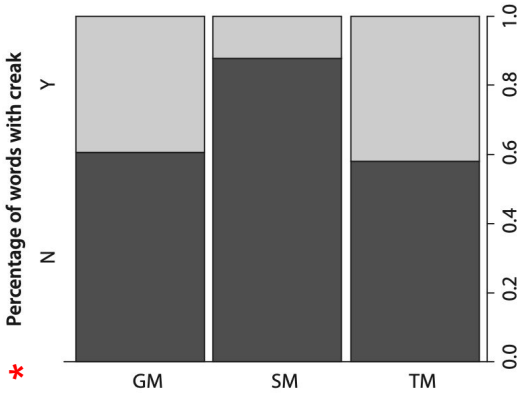
Recordings source: https://web.stanford.edu/~eckert/vowels.html

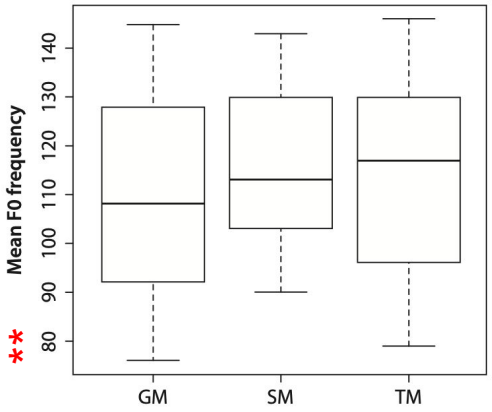
Zimman (2013): Beyond the heteronormative binary

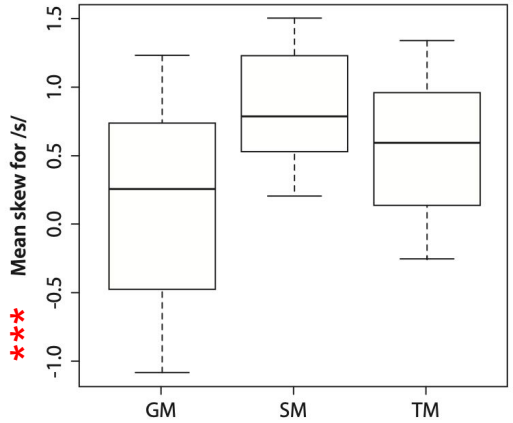
- An **acoustic analysis** to investigate the **perceived sexuality** of men
- Sample of 15 men: trans, cis gay, cis straight
- Hypothesis: multiple phonetic styles may index the same sexuality
- Focus on experience of trans men
  - Testosterone therapy ⇒ drop in vocal pitch
  - Speaking styles that nevertheless often differ from heteronormative men's voices
  - Potentially primarily perceived as gay rather than trans

Zimman (2013): Beyond the heteronormative binary









Takeaways

- Variability related to gender is well-established & with strong claims,
  e.g. women conforming to overt norms but diverging from covert ones
- Longstanding findings are questioned by biases,
  e.g. binary, biological view isolated from other factors
- More recent work demonstrates the relevance of
  - Gender as a social practice
  - Non-binary identities

NLP for gender

So far…

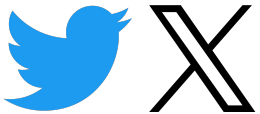
- Using speakers' gender to explain variation in language use via carefully selected, small sets of speakers…
  - … traditionally observed correlations & linking language to identity construction

- Let's analyze linguistic variation related to gender through data…
  - Does gender linguistic variability **persist** on social media?
  - Will we **uncover** new lexical variables?

Does **gender** linguistic variability persist on social media?

Bamman et al. (2014): "Gender **identity** and lexical variation in social media"

- Selection of 14k English Twitter users and 9M tweets (scale!)
- Gender derived from first names
- Modeling:
  - 10,000 most frequent lexical items as features
  - logistic regression + regularization
  - accuracy 88%



Source: Bamman et al.(2014)

Terms significantly associated with one gender

Word class	Previous literature	Our analysis
Pronouns	F	F
Emotion terms	F	F
Kinship terms	F	mixed
CMC words (*lol, omg*)	F	F
Conjunctions	F	ns
Clitics	F	ns
Articles	M	ns
Numbers	M	M
Quantifiers	M	ns
Technology words	M	M
Prepositions	mixed	ns
Swear words	mixed	M
Assent	mixed	F
Negation	mixed	mixed
Emoticons	mixed	F
Hesitation	mixed	F

Women:

- pronouns, including alternative spellings: *u, ur*
- emotion terms: *sad, love, glad,* etc.
- computer-mediated communication: *lol, omg*
  - expressive lengthening: *cooooooool*
  - backchannel sounds: *ah, hmmm, ugh*, and *grr*

Source: Bamman et al.(2014)

Terms significantly associated with one gender

Word class	Previous literature	Our analysis
Pronouns	F	F
Emotion terms	F	F
Kinship terms	F	mixed
CMC words (*lol*, *omg*)	F	F
Conjunctions	F	ns
Clitics	F	ns
Articles	M	ns
Numbers	M	M
Quantifiers	M	ns
Technology words	M	M
Prepositions	mixed	ns
Swear words	mixed	M
Assent	mixed	F
Negation	mixed	mixed
Emoticons	mixed	F
Hesitation	mixed	F

Men:

- numbers
- technology words

Source: Bamman et al.(2014)

Terms significantly associated with one gender

Word class	Previous literature	Our analysis
Pronouns	F	F
Emotion terms	F	F
Kinship terms	F	mixed
CMC words (*lol, omg*)	F	F
Conjunctions	F	ns
Clitics	F	ns
Articles	M	ns
Numbers	M	M
Quantifiers	M	ns
Technology words	M	M
Prepositions	mixed	ns
Swear words	mixed	M
Assent	mixed	F
Negation	mixed	mixed
Emoticons	mixed	F
Hesitation	mixed	F

Men:

- swear words → but *darn* a female marker in the previous literature

Women:

- assent: *okay, yes, yess, yesss, yessss* → bu *yessir* a male marker
- emoticons → but *:) :D and* ;) male markers
- hesitation *um* and *umm*

Source: Bamman et al.(2014)

Word categories

All differences statistically significant

Word class	F (%)	M (%)
Standard dictionary	74.20	74.90
Punctuation	14.60	14.20
Non-standard, not pronounceable (e.g. :) *lmao*)	4.28	2.99
Non-standard, pronounceable (e.g. *luv*)	3.55	3.35
Named entities	1.94	2.51
Numbers	0.83	0.99
Taboo	0.47	0.69
Hashtags	0.16	0.18

Standard dictionary: words found in a standard dictionary and not listed as 'slang', 'vulgar', as proper nouns, or as acronyms

Not pronounceable: *omg, ;) api*

Pronounceable: *nah, haha, lol*

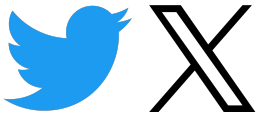
Source: Bamman et al.(2014)

Does **gender** linguistic variability persist on social media?

Bamman et al. (2014): "Gender identity and lexical variation in social media"

- Examples of linguistic variables corroborating sociolinguistic results

- *Informativeness vs. involvement* (Argamon et al. 2003)
  - Women: emotions, pronouns, communication-supporting words
  - Men: numbers, named entities, hashtags



Source: Bamman et al.(2014)

Age

Background

- Age as a factor of variation in its own right
  - Example variety: youth language
  - **Age grading:** changes to the way of speaking over the lifetime
- Age as a reflection of language change over time
  - **Apparent time hypothesis:** once linguistic features are learned, they remain stable ⇒ reflection of the linguistic system at that time
  - This stability depends on type of feature (e.g. phonological vs. lexical)

Bailey et al. (1991); Wagner (2012)

Tagliamonte (2016): Language of youth on the internet

- Youth language assumed to suffer from computer mediated communication but “virtually all [...] is based on anecdote, hearsay, and self-reports” (p. 7)
- Toronto Internet Corpus (TIC)
  - First-year university students asked to provide own samples of CMC with another interlocutor
  - Registers: email, instant messaging, SMS + academic essays
- Variables: short forms, intensifiers, future reference

Tagliamonte (2016): Language of youth on the internet

Laughter variants

friend: umm we never actually made it newhere til we ditched and wei
friend: this girl jst passed out like 7 timed
t: **lmao**.
friend: fell out of elevators, cars, so on
t: omg! no way! **ahahaha**
t: was he ok?
t: *she
friend: nope.took him home with a puke bag
t: **lol**.. her
friend: ya…her…
t: wow thats intense tho. **hhaha** aw poor grl
friend: it was her bday too! shes not gonna member a thing
t: **hahah** your not suppose to! lol
friend: you wanna member ur bday! its the day after to forget (t, IM, 2010)

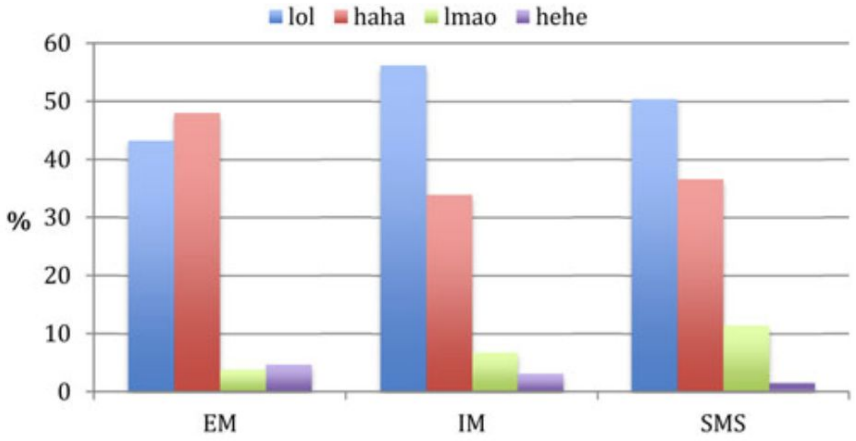


FIGURE 2. Distribution of major laughter variants across registers.

Tagliamonte (2016): Language of youth on the internet

Intensifiers

(22) a. I'm **so** sorry
b. all the **really** hot people were left out

(23) a. its **reallllyyy** hard for me :(
b. loool but ya, **pretty** pointless

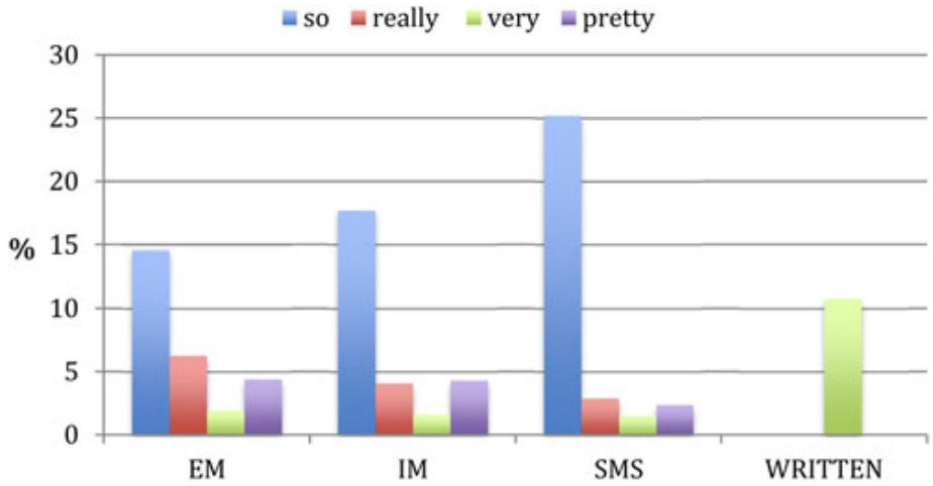


FIGURE 6. Distribution of major intensifiers by register.

Tagliamonte (2016): Language of youth on the internet

Future reference

y: okay **ill** get you a dog
h: im **gona** go brush my teeth
w: so is she **gonna** go bac tmr ?
d: i swear man **ima** go off

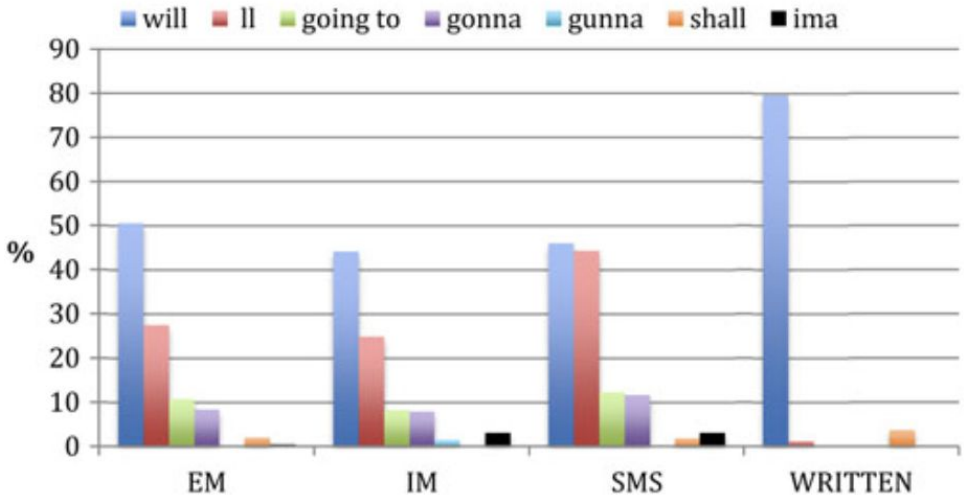
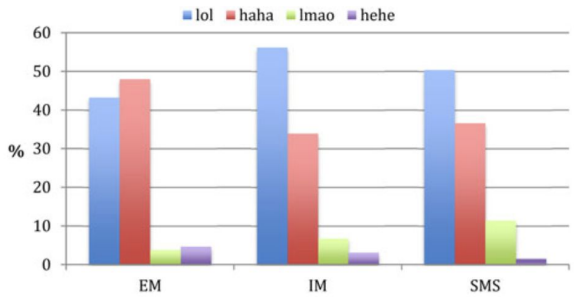
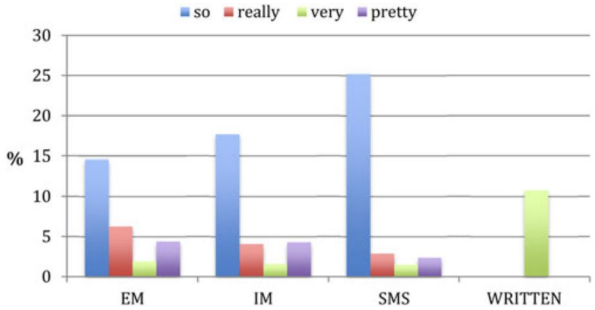


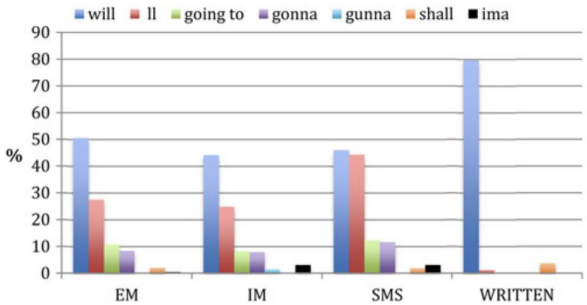
FIGURE 8. Distribution of future temporal reference variants by register.

Tagliamonte (2016): Language of youth on the internet

- Standard language is intact
- Clear differences across CMC registers
- “young people are fluidly navigating a complex range of new written registers and are using conventions that are particular to each one” (p. 28)

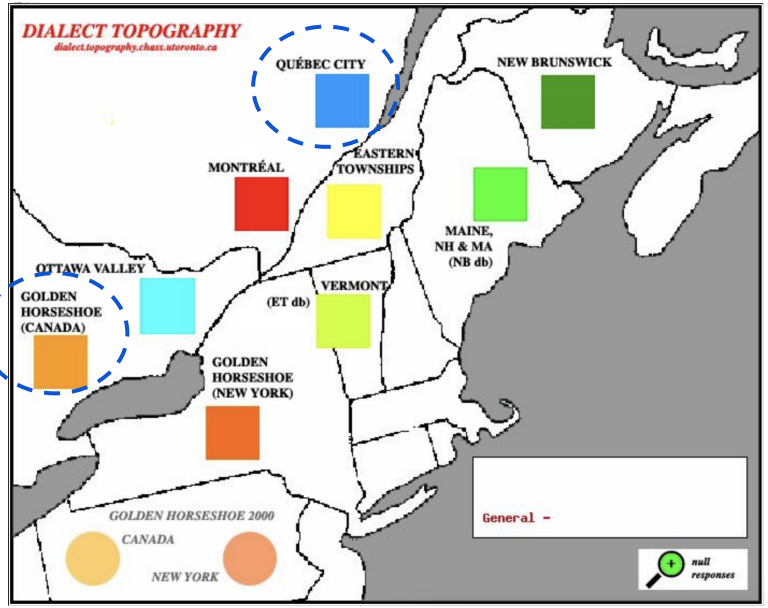






Chambers & Heisler (1999): Quebec City English

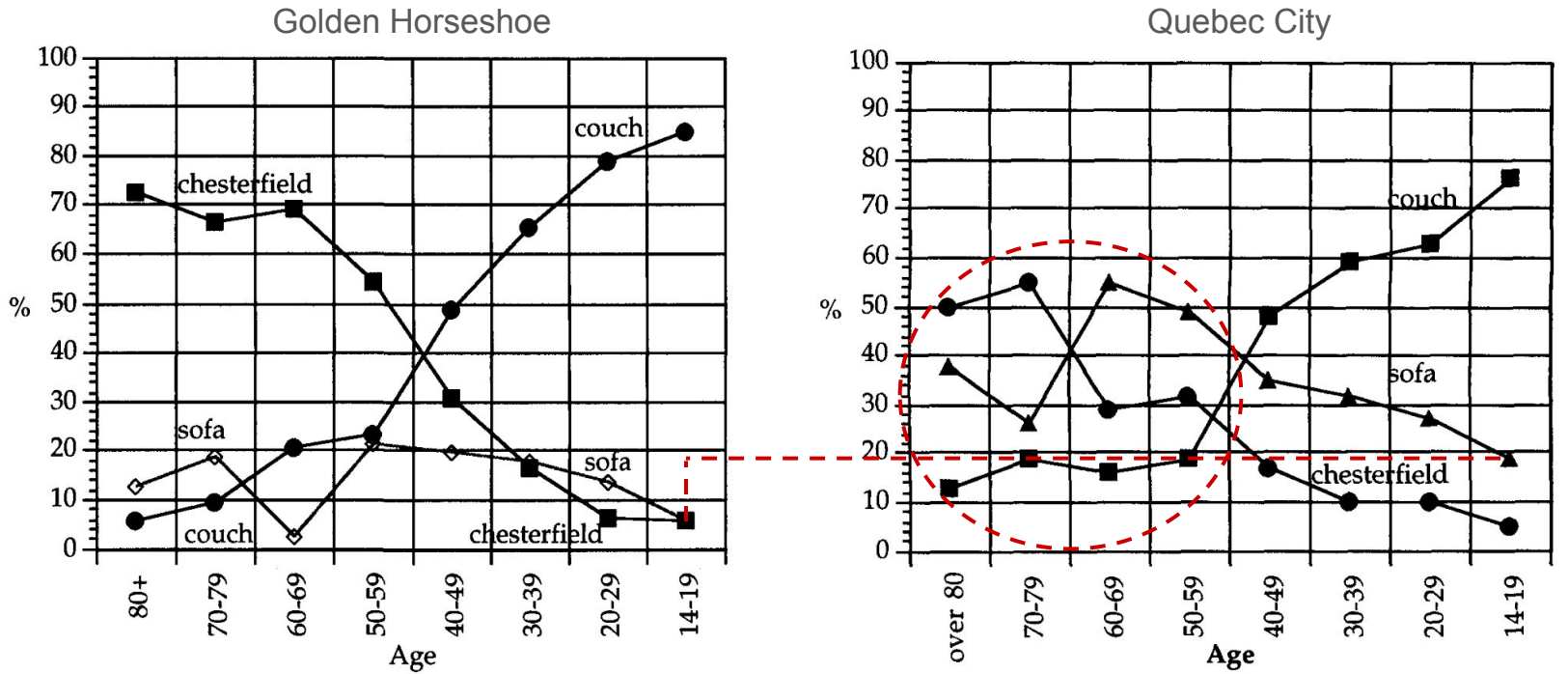
- Dialect Topography Project ⇒ **written dialect survey** to investigate (primarily) **lexical variation** in Canada & bordering US
- 89 linguistic + 12 demographic questions
- Longstanding English-speaking community within a French-language-majority



Map source: https://dialect.topography.artsci.utoronto.ca/atlasgrapher.php

Chambers & Heisler (1999): Quebec City English

What do you call the upholstered piece of furniture that two or three people sit on in the living room?



Chambers & Heisler (1999): Quebec City English

- Quebec City English shows patterns similar to other areas, but less decisive
- Language change: slower adoption, different pathways, more resistance (cf. distance & barriers)
- Possible direct effect of contact in French in further cases and regions – Montreal (Boberg, 2004), Gaspé (Boberg & Hutton, 2015)

Takeaways

- Age is useful in describing age-specific varieties
  as well as ongoing change
- It is vital to distinguish
  age grading vs. change in apparent time
- Need for complementary perspectives,
  e.g. real-time evidence

NLP for age

So far…

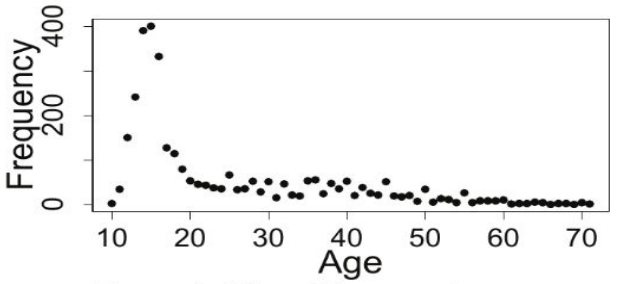
- Using speakers' age to explain variation in language use
  via carefully selected, small sets of speakers…
  - … within a specific age bracket & indirect observations of language change over time

- Let's analyze linguistic variation related to age through data…
  - Does age linguistic variability **persist** on social media?
  - Will we **uncover** new lexical variables?

Does **age** linguistic variability persist on social media?

Nguyen et al. (2013): "How old do you think I am?"

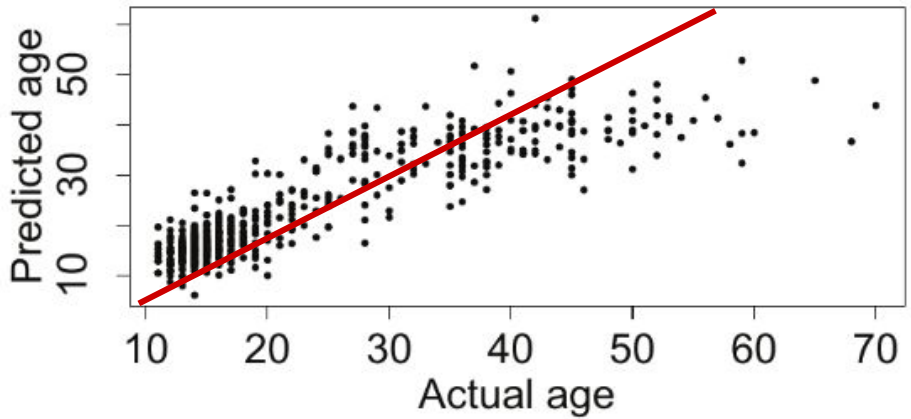
- Random selection of Dutch Twitter users
- Annotations derived from profiles and Internet
- Modeling:
  - words as features
  - linear regression + L2 regularization



Source: Nguyen et al.(2013)

How difficult is to predict exact age of an author?

Accuracy: 0.47



For 30+, the system underpredicts the age

- the language changes less as people get older
- little training data in the older age ranges

Source: Nguyen et al.(2013)

Top predictive features

Young, 20-

Dutch	English	Weight
school	school	-0.081
ik	I	-0.073
:)	:)	-0.071
werkgroep	work group	-0.069
stages	internships	-0.069
oke	okay	-0.067
xd	xd	-0.066
ben	am	-0.066
haha	haha	-0.064
als	if	-0.064

Old, 40+

Dutch	English	Weight
verdomd	damn	0.119
dochter	daugther	0.112
wens	wish	0.112
zoon	son	0.111
mooie	beautiful	0.111
geniet	enjoy	0.110
dank	thanks	0.108
goedemorgen	good morning	0.107
evalueren	evaluate	0.105
sterkte	take care	0.102

Content or style?

Source: Nguyen et al.(2013)

Statistical analysis of variables

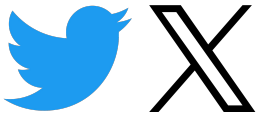
Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

n = 1247, Bonferroni correction, ** $p \leq 0.001$

Source: Nguyen et al.(2013)

Young people:

- capitalized words: *HAHA, LOL*
- alphabetical lengthening: *niiiiiice*
- intensifiers enhance the emotional meaning of words: *so, really, awful*



Statistical analysis of variables

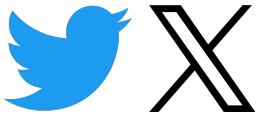
Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

n = 1247, Bonferroni correction, ** $p \leq 0.001$

Source: Nguyen et al.(2013)

Old people:

- prepositions such as *for, by* and *on*
- average word length
- average tweet length



Statistical analysis of variables

Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

n = 1247, Bonferroni correction, ** $p \leq 0.001$

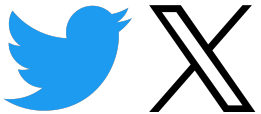
Source: Nguyen et al.(2013)

Young people: *I, you* (interpersonal involvement)

Old people: *we*

Previous research:

- Pennebaker and Stone (2003): as people get older, they make fewer self-references
- Barbieri 2008: older people more often use first-person plurals (e.g. *we*)



Statistical analysis of variables

Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

n = 1247, Bonferroni correction, ** $p \leq 0.001$

Source: Nguyen et al.(2013)

Old women:

- Replies – proportion of tweets that are a reply or mention a user

Statistical analysis of variables

Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

n = 1247, Bonferroni correction, ** $p \leq 0.001$

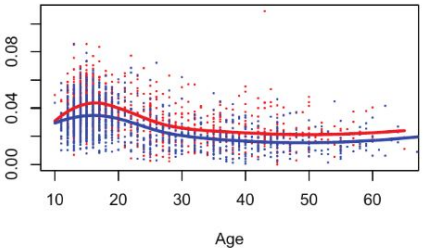
Old people: links and hashtags

Young people: retweets

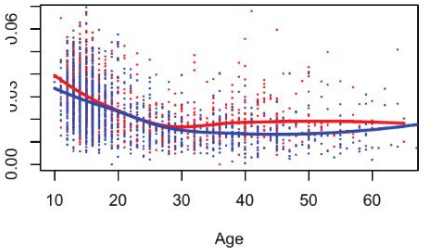
Source: Nguyen et al.(2013)

Linguistic variability across age

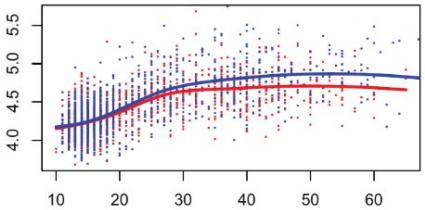
Proportion *I*



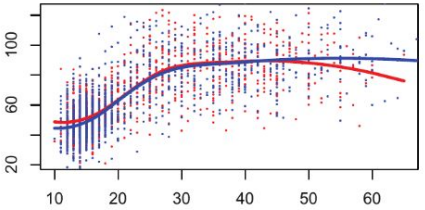
Proportion *you*



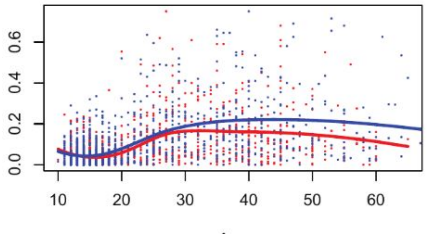
Word length



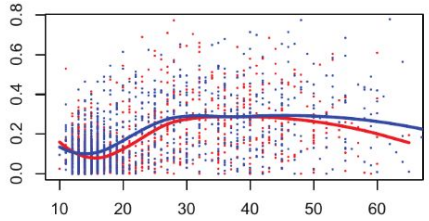
Tweet length



Links



Hashtags

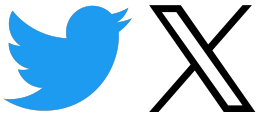


Source: Nguyen et al.(2013)

Does **age** linguistic variability persist on social media?

Nguyen et al. (2013): "How old do you think I am?"

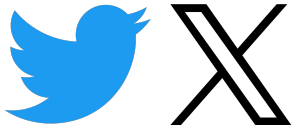
- Examples of linguistic variables corroborating sociolinguistic results
- Most changes occur when people are young; for 30+ studied variables show little change
- Strong influence of gender on how language is displayed



Source: Nguyen et al.(2013)

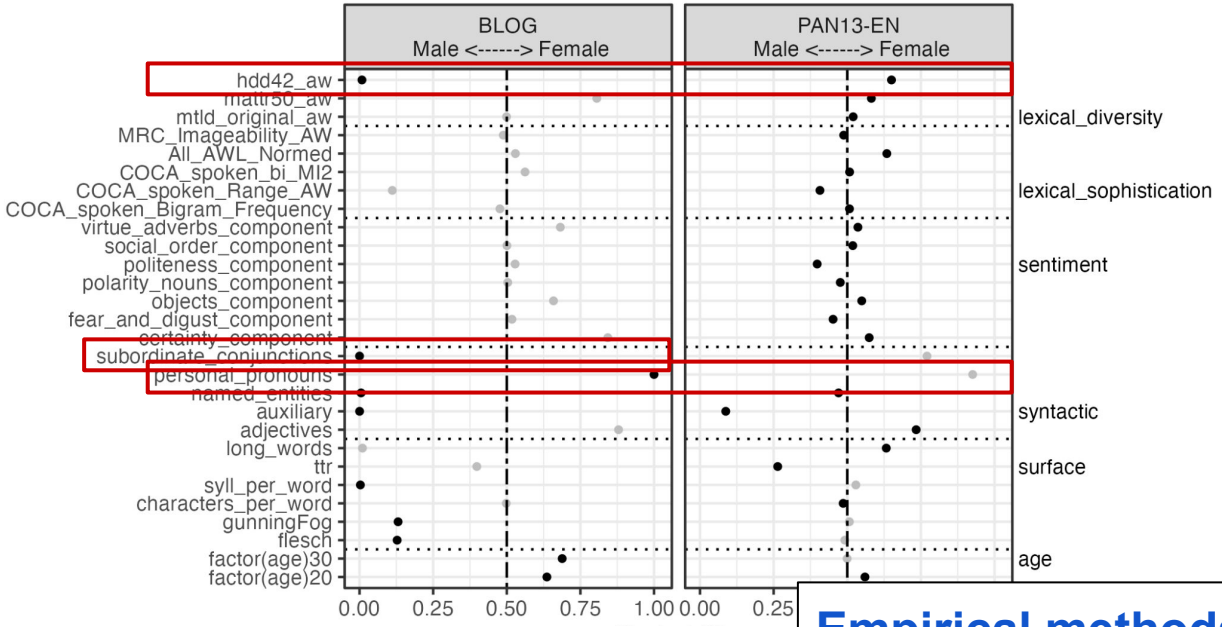
Why don't all the previous findings generalize?

Why don't all the previous findings generalize?



- Different datasets and domains

Gender and age markers across datasets



Empirical methods are based on corpus statistics

Source: Chen et al.(2024)

Why don't all the previous findings generalize?

- Different datasets and domains
  - **Empirical methods are based on corpus statistics**

- Different languages

Gender/age linguistic variability across languages

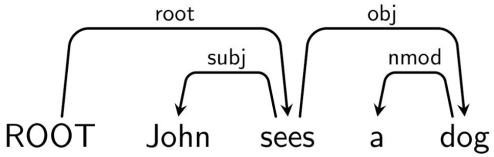
Johannsen et. al, (2015): “Cross-lingual syntactic variation over age and gender”

- Reviews with **self-reported** user meta-data
- Analyze 6/13 of available languages, 11/24 different countries on Trustpilot
  - English: *US, UK, Canada, Australia*
  - French: *France, Belgium*
  - German: *Germany, Austria*
  - Spanish: *Spain*
  - Italian: *Italy*
  - Swedish: *Sweden*

	Users	Age	Gender	Place	All
UK	1,424k	7%	62%	5%	4%
France	741k	3%	53%	2%	1%
Denmark	671k	23%	87%	17%	16%
US	648k	8%	59%	7%	4%
Netherlands	592k	9%	39%	7%	5%
Germany	329k	8%	47%	6%	4%
Sweden	170k	5%	64%	4%	3%
Italy	132k	10%	61%	8%	6%
Spain	56k	6%	37%	5%	3%
Norway	51k	5%	50%	4%	3%
Belgium	36k	13%	42%	11%	8%
Australia	31k	8%	36%	7%	5%
Finland	16k	6%	36%	5%	3%
Austria	15k	10%	43%	7%	5%
Switzerland	14k	8%	41%	7%	4%
Canada	12k	10%	19%	9%	4%
Ireland	12k	8%	30%	7%	4%

Source: (Johannsen et al. 2015)

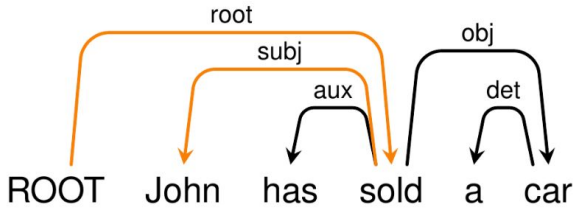
Universal dependencies

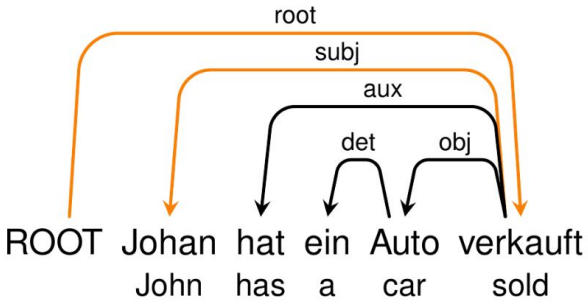


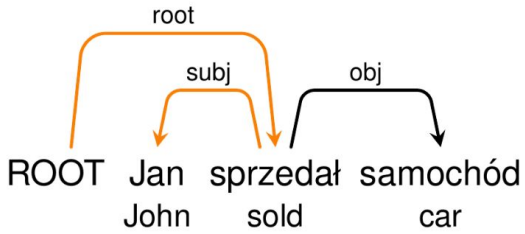
- Ongoing effort to create syntactic annotations for many languages that
  - adhere to the same annotation scheme
  - are publicly available
- Not just dependency trees, also part-of-speech and morphology tags
- Currently more than 200 treebanks in over 150 languages

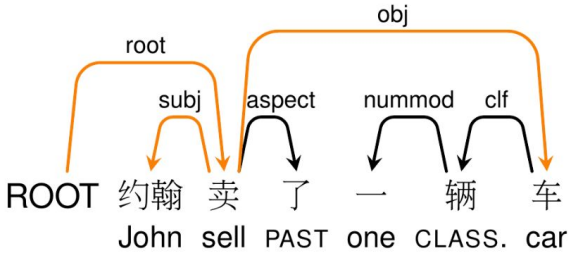
Source: (Nivre et al. 2017)

Universal dependencies









Source: (Nivre et al. 2017)

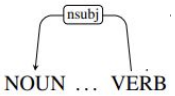
Representation of texts

- Represent texts as max three-token treelets

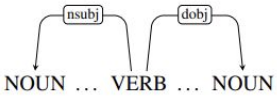
1 token:

NOUN

2 tokens:



3 tokens:



- Logistic regression + regularization + various filtering heuristics

Source: (Johannsen et al. 2015)

Gender linguistic variability across languages

Signif. in # lang.	Rank	Feature	Effect High	By
11	1	NUM	M	32 %
	2	PRON	F	11 %
	3	NOUN	M	6 %
10	4	VERB —ACOMP→ ADJ	F	22 %
	5	VERB	F	6 %
9	6	ADJ ←ACOMP— VERB —CONJ→ VERB	F	36 %
	7	VERB —ACOMP→ ADJ —ADVMOD→ ADV	F	35 %
	8	NOUN —COMPMOD→ NOUN	M	22 %
	9	VERB —NSUBJ→ PRON	F	14 %
8	10	VERB —CONJ→ VERB —ACOMP→ ADJ	F	40 %
	11	VERB —ACOMP→ ADJ —CONJ→ ADJ	F	36 %
	12	ADJ ←ACOMP— VERB —CC→ CONJ	F	28 %
	13	CONJ ←CC— VERB —CONJ→ VERB	F	16 %
	14	VERB —CONJ→ VERB	F	14 %
	15	ADP ←ADPMOD— VERB —NSUBJ→ NOUN	M	14 %
	16	NOUN —ADPMOD→ ADP —ADPOBJ→ NOUN	M	13 %
	17	NOUN —ADPMOD→ ADP	M	13 %
	18	VERB —AUX→ VERB	F	10 %
7	19	ADP —ADPOBJ→ NUM	M	43 %
	20	ADJ ←ACOMP— VERB —NSUBJ→ PRON	F	41 %

Men: numerals and nouns
→ information emphasis (Schler et al., 2006)
→ topics of reviews?

Women: pronouns, verbs

Source: (Johannsen et al. 2015)

Gender linguistic variability across languages

Signif. in # lang.	Rank	Feature	Effect High	Effect By
11	1	NUM	M	32 %
	2	PRON	F	11 %
	3	NOUN	M	6 %
10	4	VERB —ACOMP→ ADJ	F	22 %
	5	VERB	F	6 %
9	6	ADJ ←ACOMP— VERB —CONJ→ VERB	F	36 %
	7	VERB —ACOMP→ ADJ —ADVMOD→ ADV	F	35 %
	8	NOUN —COMPMOD→ NOUN	M	22 %
	9	VERB —NSUBJ→ PRON	F	14 %
8	10	VERB —CONJ→ VERB —ACOMP→ ADJ	F	40 %
	11	VERB —ACOMP→ ADJ —CONJ→ ADJ	F	36 %
	12	ADJ ←ACOMP— VERB —CC→ CONJ	F	28 %
	13	CONJ ←CC— VERB —CONJ→ VERB	F	16 %
	14	VERB —CONJ→ VERB	F	14 %
	15	ADP ←ADPMOD— VERB —NSUBJ→ NOUN	M	14 %
	16	NOUN —ADPMOD→ ADP —ADPOBJ→ NOUN	M	13 %
	17	NOUN —ADPMOD→ ADP	M	13 %
	18	VERB —AUX→ VERB	F	10 %
7	19	ADP —ADPOBJ→ NUM	M	43 %
	20	ADJ ←ACOMP— VERB —NSUBJ→ PRON	F	41 %

Source: (Johannsen et al. 2015)

Patterns beyond single words

Men: noun compounds
→ named entities, company names

Women: VP coordinations
→ English: *is/was/were great/quick/easy*
→ German: *bin/war zufrieden*, *würde bestellen*

… but only 3 patterns appear in all languages

Age linguistic variability across languages

Signif. in # lang.	Rank	Feature	Effect High	Effect By
8	1	NOUN	>45	5 %
7	2	ADP ←ADPOBJ— NOUN —ADPMOD→ ADP	>45	20 %
	3	NOUN —ADPMOD→ ADP —ADPOBJ→ NOUN	>45	14 %
	4	VERB —ADVMOD→ ADV	<35	12 %
	5	ADP —ADPOBJ→ NOUN	>45	8 %
6	6	ADV ←ADVMOD— VERB —CONJ→ VERB	<35	34 %
	7	VERB ←ADVCL— VERB —ADVMOD→ ADV	<35	27 %
	8	VERB —CC→ CONJ	<35	15 %
	9	NOUN —ADPMOD→ ADP	>45	12 %
	10	PRON	<35	10 %
5	11	ADP ←ADPMOD— NOUN —COMPMOD→ NOUN	>45	40 %
	12	VERB —CONJ→ VERB —NSUBJ→ PRON	<35	32 %
	13	ADV ←ADVMOD— VERB —CC→ CONJ	<35	25 %
	14	ADP —ADPOBJ→ NOUN —COMPMOD→ NOUN	>45	23 %
	15	CONJ ←CC— VERB —NSUBJ→ PRON	<35	21 %
	16	CONJ ←CC— VERB —CONJ→ VERB	<35	20 %
	17	ADV ←ADVMOD— VERB —NSUBJ→ PRON	<35	19 %
	18	NOUN —COMPMOD→ NOUN	>45	17 %
	19	VERB —CONJ→ VERB	<35	16 %
	20	VERB —ADVCL→ VERB	<35	11 %

Source: (Johannsen et al. 2015)

Young people: [10] pronouns

Old people: [1] nouns

Old people: [2] in English for **temporal relations**, in German for **comparisons**

- *in time for, within a couple days/hours*
- *im Wert von*

Only few age/gender markers generalize across languages

Why don't all the previous findings generalize?

- Different datasets and domains
  - **Empirical methods are based on corpus statistics**

- Different languages
  - **Only few age/gender markers generalize across languages**

- A diversity of styles to enact age and gender

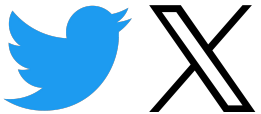
Source: https://www.responsivetranslation.com/blog/what-languages-do-they-speak/

Diversity of styles to enact age and gender

Bamman et al. (2014): "Gender identity and lexical variation in social media"

Word class	F (%)	M (%)
Standard dictionary	74.20	74.90
Punctuation	14.60	14.20
Non-standard, not pronounceable (e.g. :) *lmao*)	4.28	2.99
Non-standard, pronounceable (e.g. *luv*)	3.55	3.35
Named entities	1.94	2.51
Numbers	0.83	0.99
Taboo	0.47	0.69
Hashtags	0.16	0.18

- Cluster authors based on their language
  - represent texts as words
  - modeling: expectation maximization
  - no gender information used



Source: Bamman et al.(2014)

Linguistically similar groups of authors

- Most clusters with strong gender orientation

Skewed cluster	Size	% female	Word class (%) Dict M	Punc F	UnPron F	Pron F	NE M	Num M	Taboo M	Hash M	Top ten words
c14	1,345	89.60%	75.58%	16.44%	3.27%	1.93%	1.66%	0.85%	0.14%	0.13%	hubs blogged bloggers giveaway @klout recipe fabric recipes blogging tweetup

age?

topic?

Dict generally for female: 74.20

Pron generally for female: 4.28

Source: Bamman et al.(2014)

Gender can be enacted through a diversity of styles

Skewed cluster	Size	% female	Word class (%) Dict M	Punc F	UnPron F	Pron F	NE M	Num M	Taboo M	Hash M	Top ten words
c4	1,376	63.00	77.09	15.81	1.84	1.82	2.02	0.78	0.52	0.12	&& hipster #idol #photo #lessambitiousmovies hipsters #americanidol #oscars totes #goldenglobes
c11	432	26.20	68.97	8.32	5.95	11.1	2.01	0.88	2.32	0.38	niggas wyd nigga finna shyt lls ctfu #oomf lmaoo lmaooo
c10	1,865	14.60	77.72	16.17	1.51	1.27	2.03	0.89	0.34	0.06	/cc api ios ui portal developer e3 apple's plugin developers

race?

topic?

No cluster is 100% male or female

Source: Bamman et al.(2014)

Why don't all the previous findings generalize?

- Different datasets and domains
  - **Empirical methods are based on corpus statistics**

- Different languages
  - **Only few age/gender markers generalize across languages**

- A diversity of styles to enact age and gender
  - **Study based on binary gender will only yield results that support this opposition**
  - **Social categories cannot be separated from other aspects of identity**

- Watch out for **essentialism** while interpreting qualitative results

  Koolen and van Cranenburgh, (2017)

Takeaways

- Does gender/age linguistic variability **persist** on social media?
  - Yes, but it varies across domains, topics, and languages

- How to discover new lexical variables?
  - Example: clustering
  - Examples of features: universal dependencies




Practical exercise

Day 3: Gender lexical variables

- Today's goal: find linguistic variables that differ male and female style
  - Open **Gender Variability** notebook
  - Follow all the steps and fill in the missing pieces of code (marked with TODO)

  - Which features are the strongest for both of the models? Can you design a different model that would find different variables?

Takehomes

Recap

- **Gender** has clear links to variation (e.g. women conforming to overt norms)
- Interaction-based and more inclusive analyses ⇒ finer-grained accounts

- **Age** may reflect community-specific patterns and diachronic changes
- Distinguishing between age grading and apparent time remains challenging

- **Social media data** captures patterns similar to real-life communication, but with variability across domains, topics, and languages

Tomorrow…

Adapting to interlocutors

Accommodation theory (Giles, 1973; Bourhis et al., 2007)

Audience design theory (Bell, 1984; Hay et al., 1999; Renn & Terry, 2008)

Social meaning of variation

Indexicality (Eckert, 2008; Beaton & Washington, 2015)

Personae (D'Onofrio, 2020; Podesva, 2011)

Example NLP methods for analysing linguistic adaptation and online communities

Universität Stuttgart

IRIS

Thank you for your attention!

Image credit: DALLE (2024)