Universität Stuttgart

IRIS

Language in Social Context: Bridging NLP and Sociolinguistics

Agnieszka Faleńska
Filip Miletić

Day 2: Demographic Factors (I)

Image credit: DALLE (2024)

Recap: Explaining language variation

- Language use varies across levels of linguistic structure
- Sociolinguists focus on identifying **sociolinguistic variables** and accounting for differences in the use of their **variants**
- They do so relying on a range of **internal** and **external** factors
- Social media is a rich but **challenging** source of data for analyzing sociolinguistic variables
- **NLP methods** allow for analyzing large amounts of data

Recap: Approaches to data collection

- Dialect survey (Dollinger, 2015; Preston, 2002)
- The sociolinguistic interview (Labov, 1992; Tagliamonte, 2006)
- Participant observation (Eckert, 2000)
- Rapid & anonymous survey (Labov, 1972)
- Surreptitious recording (cf. Milroy & Gordon, 2003: 83–83)

tens to hundreds of speakers

highly controlled information

sampling issues

- Social media corpora
- Demographics: self-reported, derived, approximated

thousands of speakers

noisy information

Day 2: Outline

Geographic origin
PFC Project (Detey et al., 2016)
NARVS Project (Boberg, 2005)

Socioeconomic status
Foundational work in the US (Labov, 1972)
Estimating socioeconomic status today (Dodsworth, 2009)

NLP methods for analysing linguistic variations
Logistic regression, SAGE, POS tags, syntactic dependencies, bleaching

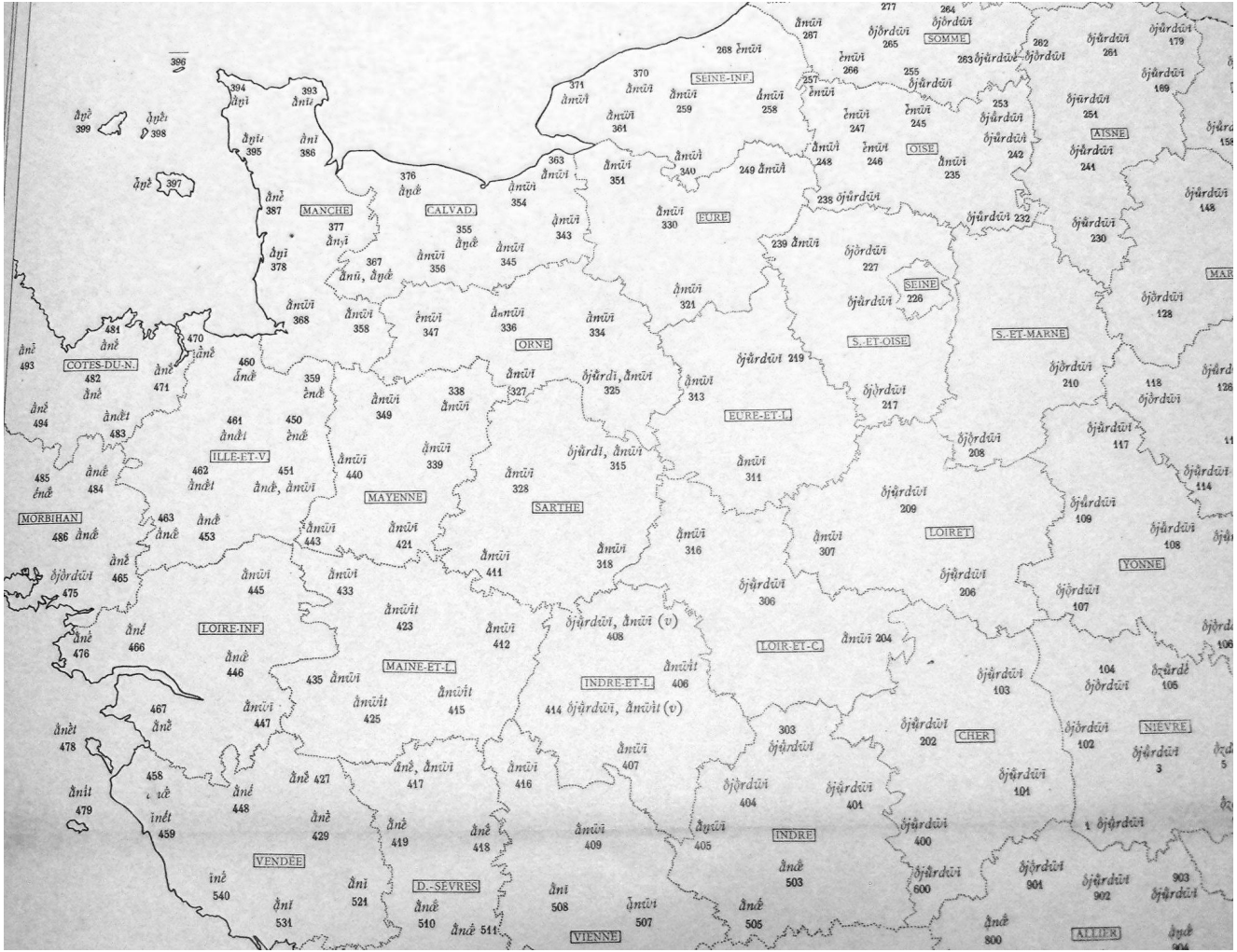
Practical exercise
Exploring regional variation in a large-scale corpus

Geographic origin

Background

- Long tradition of analyzing language variation across regions
  - Georg Wenker’s survey of German (1876–1887)
  - Jules Gilléron & Edmond Edmont’s survey of French (1896–1900)

Chambers (2000); Trudgill (2000); Tagliamonte (2002)



Gilliéron & Edmont (1910)
via Lotusfleurie/Wikimedia Commons

Background

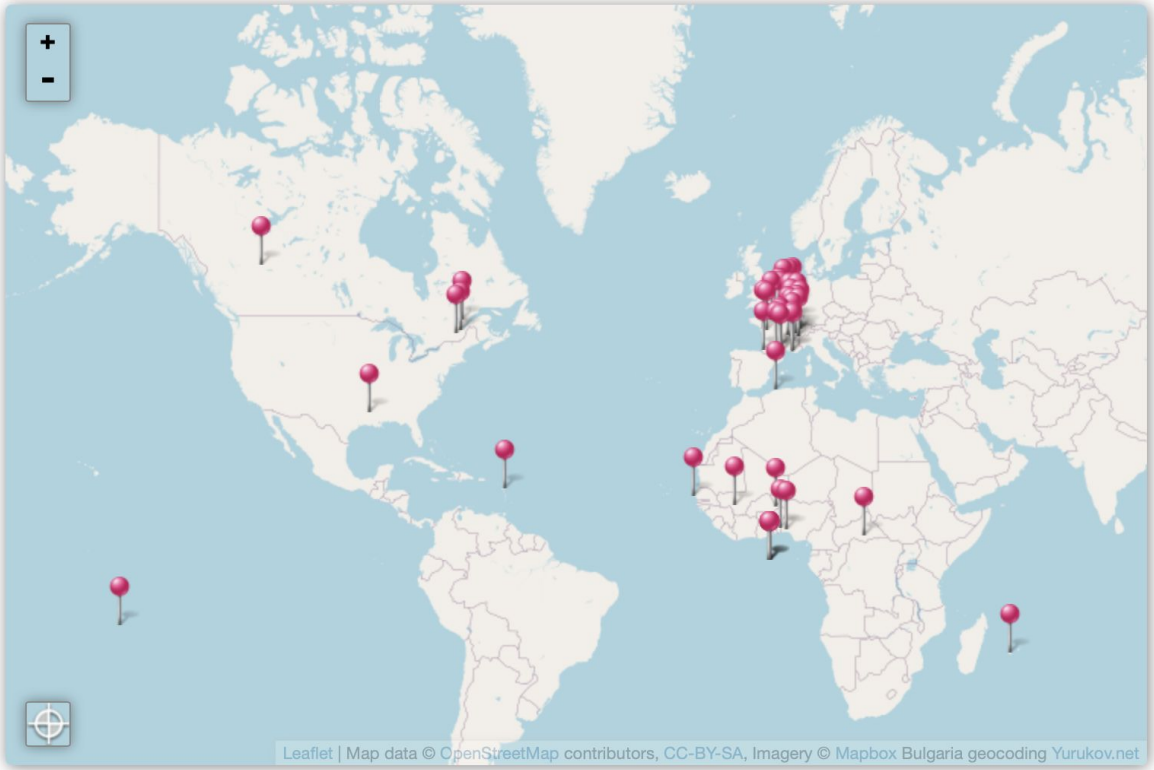
- **Long tradition** of analyzing language variation across regions
  - Georg Wenker's survey of German (1876–1887)
  - Jules Gilléron & Edmond Edmont's survey of French (1896–1900)
- Effect of **barriers & distance**
- Traditional focus on **NORMs** – nonmobile, older, rural males
- **Explanatory power decreased** over the 20th century
  - Increased social mobility ⇒ linguistic homogeneization
  - Interactions with other social factors
- **Ongoing interest** in present-day analyses
  - Target phenomena that are geographic by definition
  - Comparative sociolinguistic method

Chambers (2000); Trudgill (2000); Tagliamonte (2002)

PFC Project: Phonology of Contemporary French

- Large-scale project using **sociolinguistic interviews**
  to examine **variation in pronunciation** across French-speaking regions
  (Detey et al., 2019)
- Aims at “a large reference corpus of spoken French”
- The same protocol across many collaborators & survey points:
  50+ locations, 400+ speakers ⇒ maximize geographic coverage
- https://www.projet-pfc.net/

PFC Project: Phonology of Contemporary French



Map source: https://research.projet-pfc.net

PFC Protocol

- Pragmatic approach to sampling
  - Small cohorts (N=10–15) on network principle
  - Control for binary gender & age
- Labovian method
  - Reading tasks: word list, text
  - Conversation tasks: guided, informal
- Post-processing
  - Orthographic & phonemic transcription
  - Coding of target variables

20–30 hours of work per speaker

Detey et al. (2019)

PFC Protocol

Word list

1. roc
2. rat
3. jeune
4. mal
5. ras
6. fou à lier
7. des jeunets
8. intact
9. nous prendrions
10. fêtard
11. …

Text

Le Premier Ministre ira-t-il à Beaulieu?

Le village de Beaulieu est en grand émoi. Le Premier Ministre a en effet décidé de faire étape dans cette commune au cours de sa tournée de la région en fin d'année. Jusqu'ici les seuls titres de gloire de Beaulieu étaient son vin blanc sec, ses chemises en soie, un champion local de course à pied (Louis Garret), quatrième aux jeux olympiques de Berlin en 1936, et plus récemment, son usine de pâtes italiennes. […]

Guided conversation

Speakers: 44aar2, cqams1

PFC: A speaker from Toulouse

Sociolinguistic profile and recording situation

The speaker we focus on is a man, YR (PFC code: byr), who was 68 years old at the time of the recording, which took place in 2010. He was born in Toulouse in 1942 and has always lived there. The other two participants in this informal conversation are his sister (CM), 65 years old, and his daughter (CR), 38 years old. YR's main qualification is an Industrial Training Certificate (Brevet d'enseignement industriel) and he started his career as a coppersmith before joining the aviation industry. He rose through the ranks of his firm, became a draughtsman, then an industrial designer, and ended up in a managerial role. His father was born in a small village near Toulouse, whereas his mother came from Lille. The only language spoken at home was French and YR has no command of Occitan. His sister (CM), who had retired at the time of the recording, had been a manager in the clothing industry and his daughter (CR) was working as an assistant engineer in a university research centre. Like YR, both CM and CR have always lived in Toulouse or its outskirts.

Courdès-Murphy et al. (2019)

PFC: A speaker from Toulouse

Schwa

- Typical feature of southern French: **presence of schwas**
  in contexts where they are absent or variably present in Reference French
- Words ending in consonant(s) vs. **consonants(s) + schwa** (e.g. *total* vs. *sociale*):
  distinction maintained
- Final schwas in **polysyllabic words** at the **end of a rhythmic group**:
  always maintained (e.g. *à l'époque on appelait ça la promotion sociale*)
- Final schwas **preceding a consonant-initial** word in the **same rhythmic group**:
  ~80% maintained (e.g. *davantag(e) besoin* with *à faire ton métier*)

Courdès-Murphy et al. (2019)

PFC: Variation in Douzens (Aude)

TABLE 32.8. Correlation between the behaviour of final schwas and age differences.

	schwa absent		schwa uncertain		schwa present		total	
	occurrences	%	occurrences	%	occurrences	%	occurrences	%
Total	149	20	14	2	592	78	755	100

Durand & Tarrier (2019). Map source: Superbenjamin/Wikimedia Commons

NARVS: North American Regional Vocabulary Survey

- Large-scale project using **a written dialect survey** to examine **lexical variation** in North American English (Boberg, 2005)
- Aims at “identifying new variables that delineate regional divisions in Canadian English and distinguish Canadian from American English”
- Lecture assignment ⇒ media attention ⇒ broad survey coverage: 1,800 Canadians & 360 Americans (3 years)

NARVS Protocol

- Participants “who grew up entirely in one region and still live in that region”
- Basic demographic information
- 53 lexical variables

the first year of school after kindergarten: *first form / first grade / form one / grade one / year one*

a carbonated, nonalcoholic beverage, like Coke, Pepsi, Sprite, or Mountain Dew: *coke / cola / cold drink / fizzy drink / pop / soda / sodapop / soft drink / tonic*

a small store, open late, that sells milk, newspapers, snacks, etc.: *convenience store / corner store / dep / dépanneur / general store / variety store / commercial name* (e.g., 7-11, Becker’s, Mac’s, etc.)

Net variation

Given a pair of regions,

→ for each variant, calculate absolute difference between the two regions

→ sum all absolute differences

Boberg (2005)

NARVS: Lexical Boundaries in Canada

Boberg (2005)

TABLE 2

Relative Lexical Strength of Regional Divisions in Canada

Division	*Net Variation*
MR-NB	62%
EO-MR	54%
NS-NL	45%
CB-NL	44%
PE-NS	40%
MB-NO	39%
NB-PE	37%
SK-MB	37%
BC-AB	33%
SO-TO	32%
NS-CB	31%
NO-SO	30%
VV-BC	29%
TO-EO	28%
NB-NS	27%
AB-SK	26%

NOTE: Lexical boundary strength is measured by mean net variation per variable on the left and by the number of major isoglosses on the right, in descending order. The column labeled "Sig. less than" shows the results of *t*-tests of the difference between the net variation, with $p = 0.05$ as the criterion for significance.

NARVS: Key Lexical Variables in Canada

TABLE 3
Top Ten Canadian Regional Differentiators by Total Net Variation

	Variable	*Net Variation*
Q45	cabin/cottage / chalet	1,212%
Q48	parking garage/parkade	1,107%
Q12	pizza with all the toppings	1,062%
Q36	internship/stage	1,045%
Q47	convenience/corner store/dépanneur	1,020%
Q32	see-saw/teeter-totter	996%
Q39	notebook/scribbler	960%
Q41	backpack/bookbag/schoolbag	917%
Q10	carbonated beverage	908%
Q13	burger with all the toppings	852%

"*Chalet* definitely, that's the word we use in Quebec."

"Yeah *chalet* is pretty normal I think for Montrealers."

"This is super, super comfortable for me as well. I don't think that this is awkward at all. I think for people outside of Quebec it would be."

Boberg (2005); Miletić et al. (2023)

NARVS: American Influence on Canadian Regions

Variation between Each Canadian Region and the Mean of U.S. Regions

VV	BC	AB	SK	MB	NO	SO	TO	EO	MR	NB	PE	NS	CB	NL
78	87	83	89	80	86	80	76	82	85	83	85	81	83	82
21	26	21	28	20	28	23	21	20	18	24	22	25	20	24
(Seattle)				(Twin Cities)			(Cleveland)			(Boston)				

NOTE: The upper number is the average net variation (%) per question; the lower number is the total number of major isoglosses between each Canadian region and the United States. Canadian regions are shown from west to east. Below the figures are the approximate locations of major U.S. cities.

- Canadian regions are more similar to each other than to the US
- No Canadian region is consistently more or less American than another

Boberg (2005)

Takeaways

- The effect of regional variation remains pronounced
  in individual speakers & across broad geographic regions
- Data collection choices ⇒ trade-offs on multiple dimensions
  - Coverage: number of speakers, geographic areas
  - Target variables: e.g. phonological vs. lexical, number of instances
  - Degree of control: spontaneous communication vs. eliciting target items
- Challenging to move beyond previously established variables

NLP for geographic origin

So far…

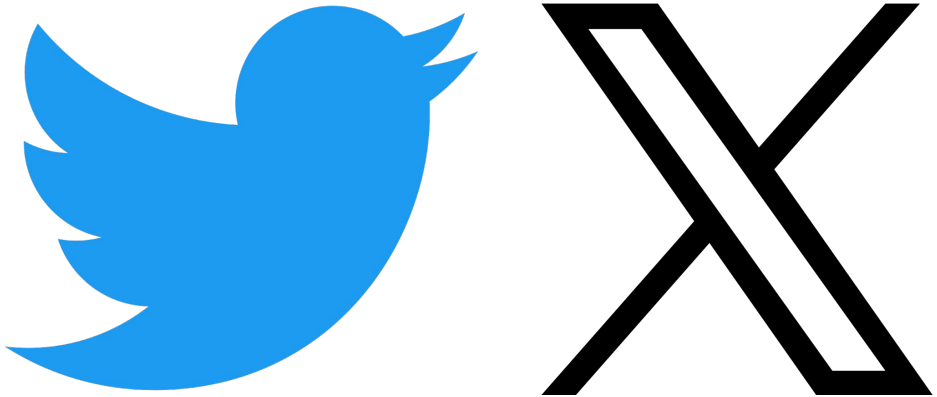
- Small sample of participants or years of collecting data
- Starting from known linguistic variables (coke vs. soda)

- Let’s analyze linguistic variation across regions through social media…
  - Does known regional variability **persist** on social media?
  - How to **discover** new lexical variables?
  - How to select **regions** for the analysis?

… a story that will repeat multiple times during this week

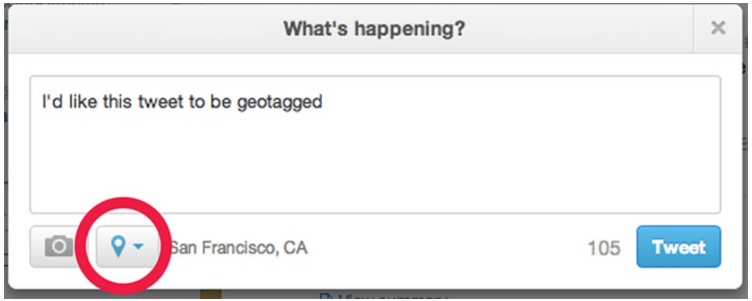
Does known regional variability persist on social media?

Which data (used to) come with geographical tags?



Twitter as a source of geotagged data (until 2019)

- Geotags – coordinates specified by a GPS system
- User-specified coordinates
- User specification of a home location



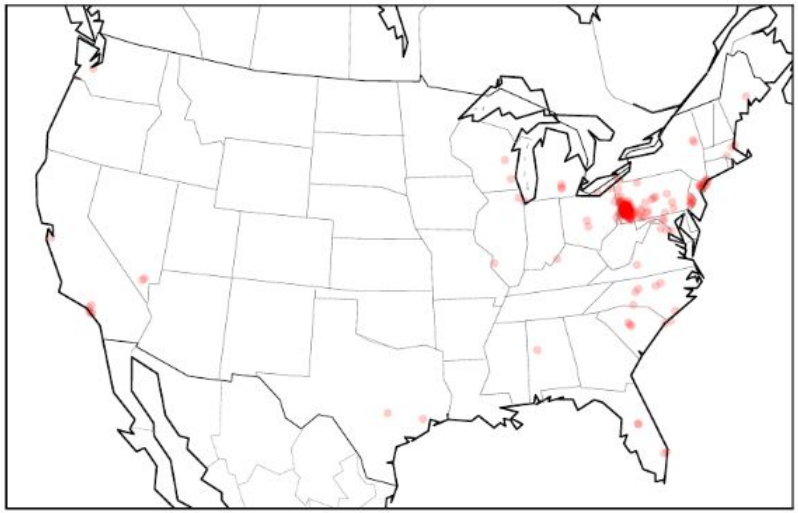
Source: https://blog.x.com/en_us/a/2013/tweet-marks-the-spot

Do regional dialect words from spoken language persist on Twitter?

Eisenstein (2014): 114 million US geotagged messages from 2.77 million different users within **metropolitan statistical areas (MSA)**

Yinz – a form of the second-person pronoun, used around Pittsburg

- rare, appearing in only a few hundred tweets
- the geographical distribution around Pittsburgh



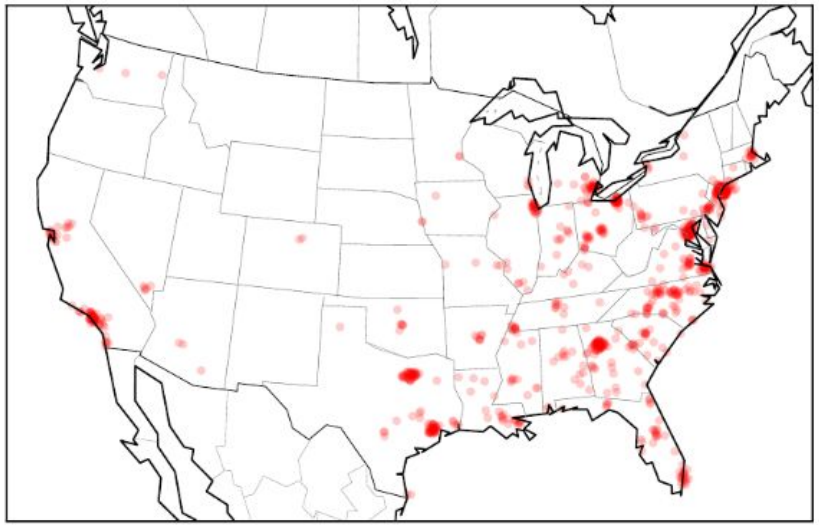
(a) *yinz*

Source: Eisenstein (2014)

Do regional dialect words from spoken language persist on Twitter?

Yall, **y'all** – an alternative form of the second-person pronoun, often associated with the Southeastern United States

- frequent, approximately one per 250 messages
- popular in the Southeast, but also in many other parts of the United States



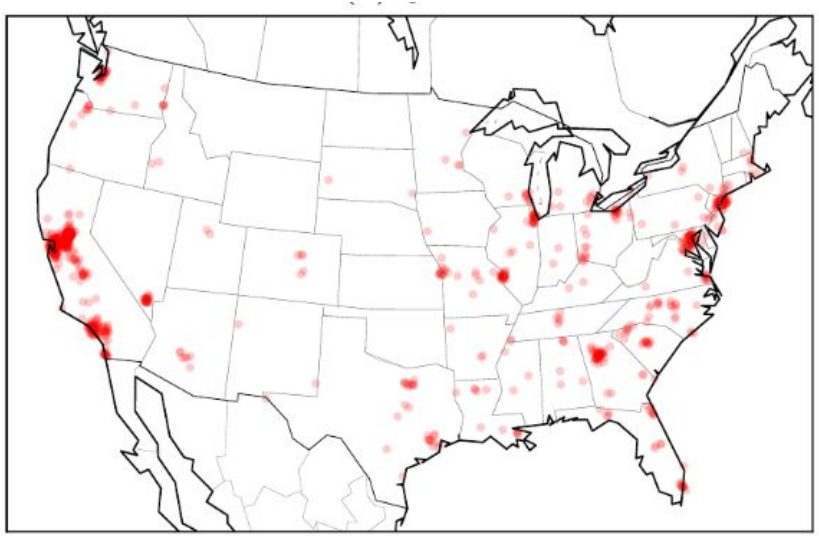
(b) *yall*

Source: Eisenstein (2014)

Do regional dialect words from spoken language persist on Twitter?

Hella – an intensifier popularly associated with Northern California

- common, nearly one out of every thousand messages
- appears in Northern California at a higher-than-average rate, but also throughout the country



(c) *hella*

Source: Eisenstein (2014)

How to **search Twitter** while controlling for the frequency?

- Doyle (2014): *SeeTweet* search tool
- Frequency-based approaches show only positive results (posts with “yinz”)
- Major cities produce more tweets than the rest of the country
- Doyle (2014): Conditional distribution of a dialectal variant given a location

D – data

M – metadata

$$P(D|M) = \frac{P(M|D)\,P(D)}{P(M)}$$

How to calculate?

Bayes’ Rule

How to find “the rest” on Twitter?

How to estimate $P(M)$ *and* $P(D)$?

- Find a baseline query q with distribution independent of the metadata
  - *I, of, the, a*
  - $P(m|q)$ approximately constant for all $m \in M$

- How to estimate $P(M|D)$?
  - switch to the unnormalized distribution

$$P(d|M) \propto \tilde{P}(d|M) = \frac{P(M|d)}{P(M|q)}$$

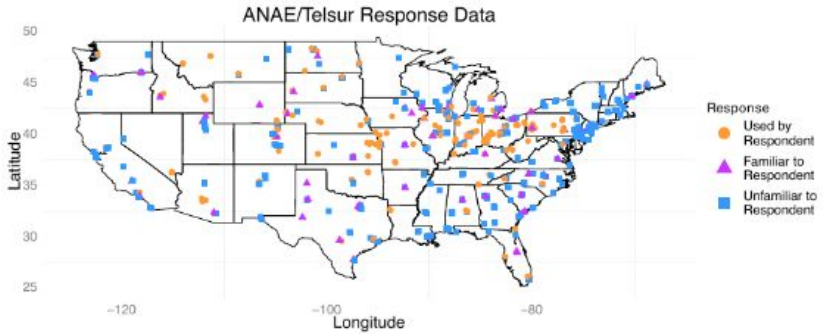
- *SeeTweet* search tool

Source: Doyle (2014)

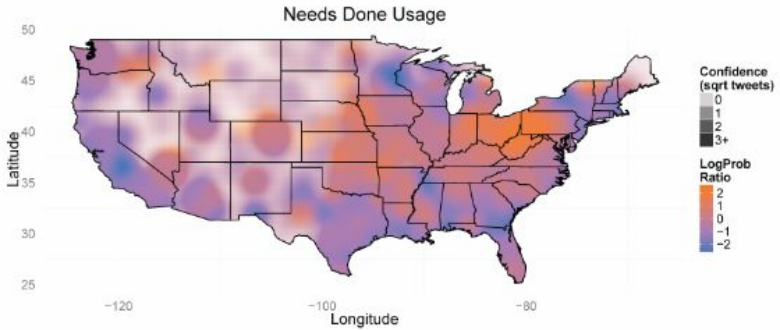
Tweeter vs. gold-standard data

***needs* + past participle** as in *The car needs washed*

The *Atlas of North American English* (577 survey participants)

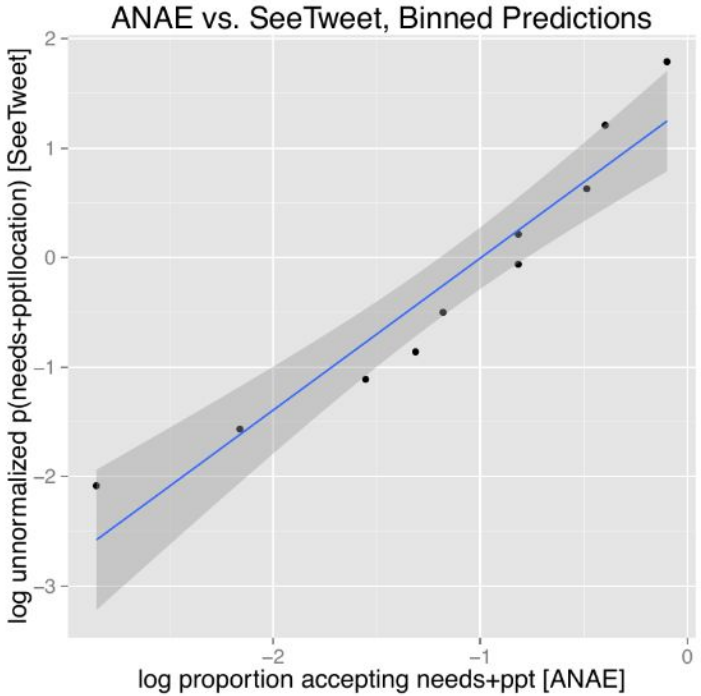


"Needs done" in SeeTweet (480 positive, 32275 baseline tweets)



Source: Doyle (2014)

Twitter vs. gold-standard data



- $R^2 = 0.90$
- SeeTweet can generate data that is tightly correlated with gold-standard data from controlled surveys

Source: Doyle (2014)

How to discover new lexical variables?

How to discover new lexical variables?

- Check all the words from the dictionary
- Raw difference in frequencies will overemphasize very common words (imagine 1% change for *the*)
- SAGE (Sparse Additive Generative Models of Text, Eisenstein et al., 2011)
  - method to compare two corpora
  - identifies terms that are over- or under-represented in **target** compared to **background** dataset

How to discover new lexical variables?

Eisenstein (2014): SAGE on 114 million US geotagged tweets within metropolitan statistical areas (MSA)

r – region

w – word

$$P_r(w) = \frac{exp\left(m_w + \beta_w^{(r)}\right)}{\sum_i exp\left(m_i + \beta_i^{(r)}\right)}$$

log of the empirical frequency across all regions

deviation from the empirical log frequency in region r

Source: Eisenstein (2014)

Top words for largest MSAs

New York: flatbush, baii, brib, bx, staten, mta, odee, soho, deadass, werd

Los Angeles: pasadena, venice, anaheim, dodger, disneyland, angeles, compton, ucla, dodgers, melrose

Chicago: #chicago, lbvs, chicago, blackhawks, #bears, #bulls, mfs, cubs, burbs, bogus *fake*

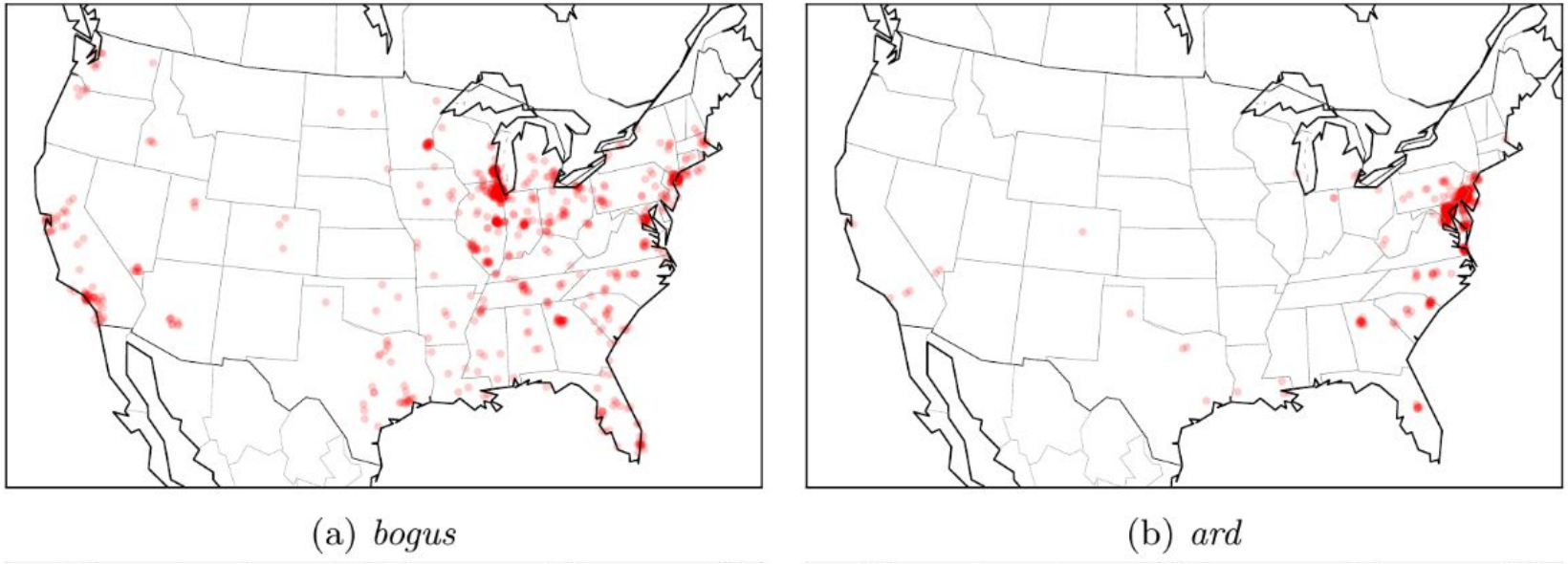
Philadelphia: jawn, ard, #phillies, sixers, phils, wawa, philadelphia, delaware, philly, phillies

alright: @name ard let me know

→ might have gone unnoticed without the use of automated methods

Source: Eisenstein (2014)

Top words for largest MSAs

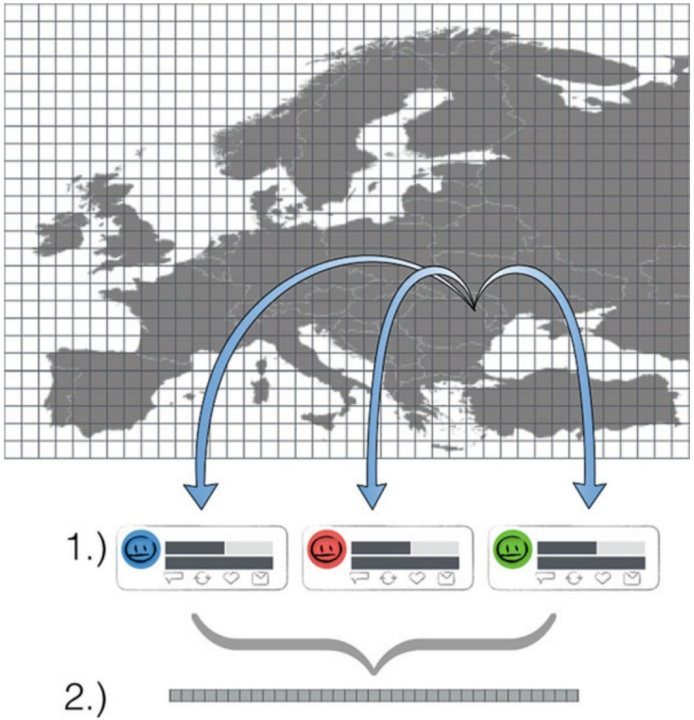


(a) *bogus* (b) *ard*

Source: Eisenstein (2014)

How to select regions for the analysis?

How to select regions for the analysis?



Hovy et al. (2020):

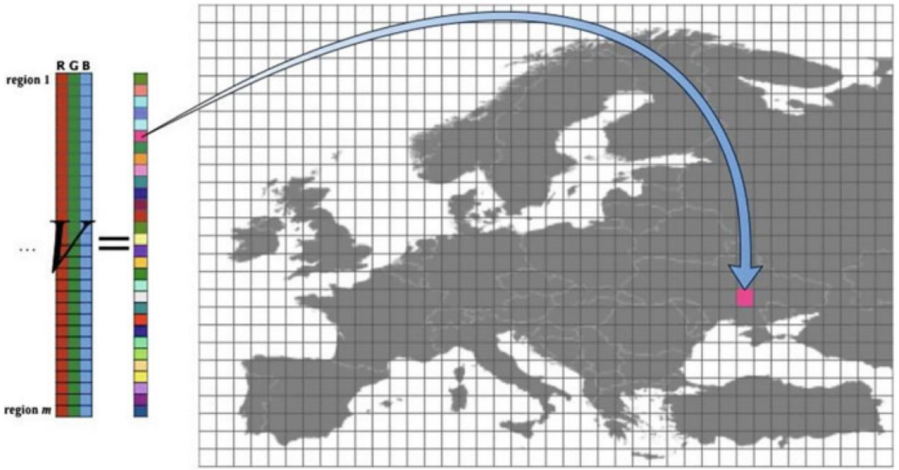
- “Coordinate grid” across Europe
- squares with sides around 11 km
- 18 million cells across Europe

- Linguistic profile for each cell
  - collect all character trigrams
  - run DocToVec to get a vector representation for all cell tweets

Source: Hovy et al.(2020)

Visualizing regional language variation

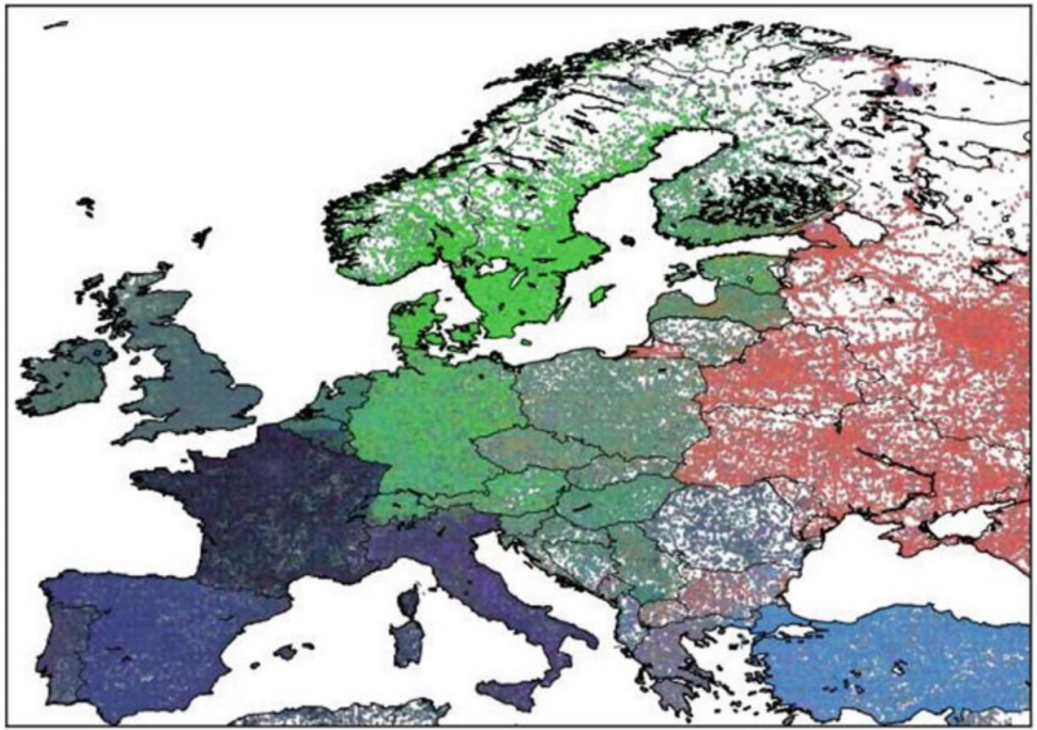
- PCA to reduce cell vectors to only three dimensions
- Convert to RGB



Source: Hovy et al.(2020)

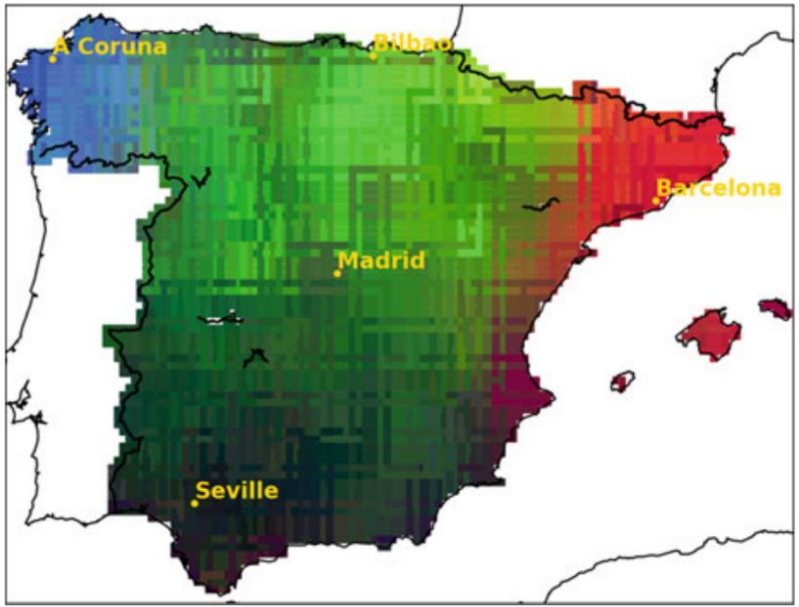
Visualizing regional language variation

- Color division along major language families (Romance vs. Germanic vs. Slavic)
- Russian stands out due to Cyrillic script
- Multilingual regions: Belgium, Switzerland, north of Italy



Source: Hovy et al.(2020)

Example: Spain

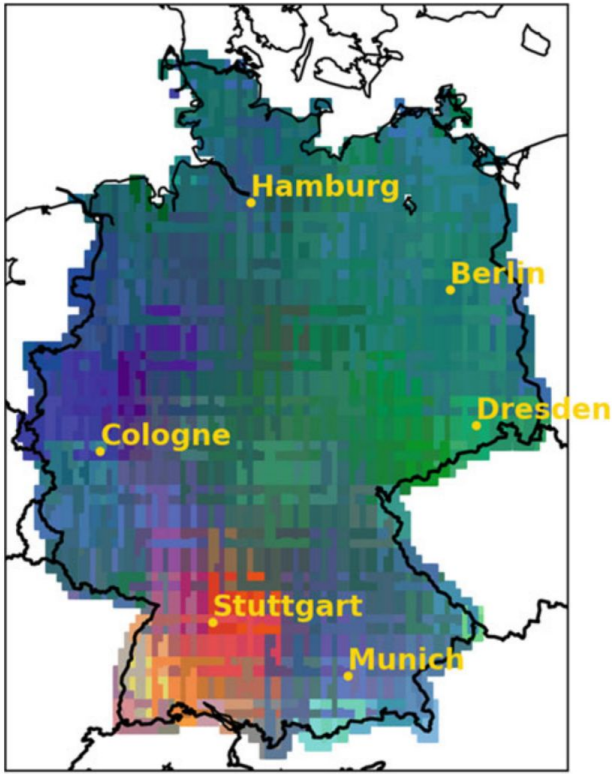


Andalucian influence?

Source: Hovy et al.(2020), https://commons.wikimedia.org/wiki/File:Spain_languages.PNG

Example: Germany

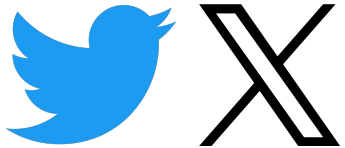
- Clear distinction of the southwestern Swabian varieties (Stuttgart)
- Clear difference between the Swabian varieties in the southwest and Bavarian varieties in the southeast (Munich)



Source: Hovy et al.(2020)

Takeaways

- Twitter as a source of data for analysing regional linguistic varieties
- Do regional dialect words from spoken language persist on Twitter?
  - Yes, examples *yall, hella*
- How to discover new lexical variables?
  - Example: SAGE
- How to select regions for the analysis?
  - metropolitan statistical areas (MSA)
  - automatically



Questions?

Socioeconomic status

Background

- Variation may be linked to social barriers and social distance
- **Social stratification:** “any hierarchical ordering of groups within a society especially in terms of power, wealth and status” (Trudgill, 2000: 25)
- **Social classes:** “aggregates of individuals with similar social and/or economic characteristics” (Trudgill, 2000: 25)
- **Challenges:** differences across societies & social mobility
- **Different types of measures** (Curry et al., 2024)
  - Objective ⇒ education, income, occupation
    ⇒ economic, social, cultural capital
  - Subjective ⇒ self-perception, e.g. Macarthur scale

Labov (1972): Department Stores Study

- Preliminary study using **rapid and anonymous observation** to investigate **(r) as a social differentiator in NYC English**
- **Hypothesis:** “if any two subgroups of New York City speakers are ranked in a scale of social stratification, then they will be ranked in the same order by their differential use of (r)” (p. 169)
- **Operationalization:** speech of salespeople in three department stores: Saks Fifth Avenue > Macy’s > S. Klein

Labov (1972: Ch. 2)

Labov (1972): Department Stores Study

- Casual and anonymous speech event:
  - *Excuse me, where are the women's shoes?*
  - *Fourth floor.* (casual)
  - *Excuse me?*
  - *Fourth floor.* (emphatic)
- 264 informants
- Basic perceived demographics

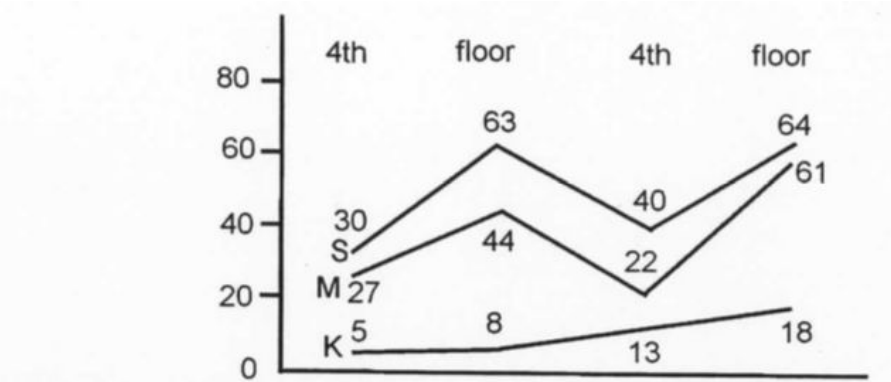


Figure 13.2: Percentage of all (r – 1) by store for four positions (S = Saks, M = Macy's, K = Kleins)

Labov (1972: Ch. 2)

Labov (1972): Interviews in the Lower East Side

- In-depth study using **the sociolinguistic interview**
  to investigate **phonological variables** in New York City English:
  (r), (th), (dh), (oh), (eh)
- Subset of a previously established representative sample
  of the Lower East Side population (N=207)
- Socioeconomic index
  - occupation (of the breadwinner)
  - education (of the respondent)
  - income (of the family)

Labov (1972: Ch. 5)

Labov (1972): Interviews in the Lower East Side

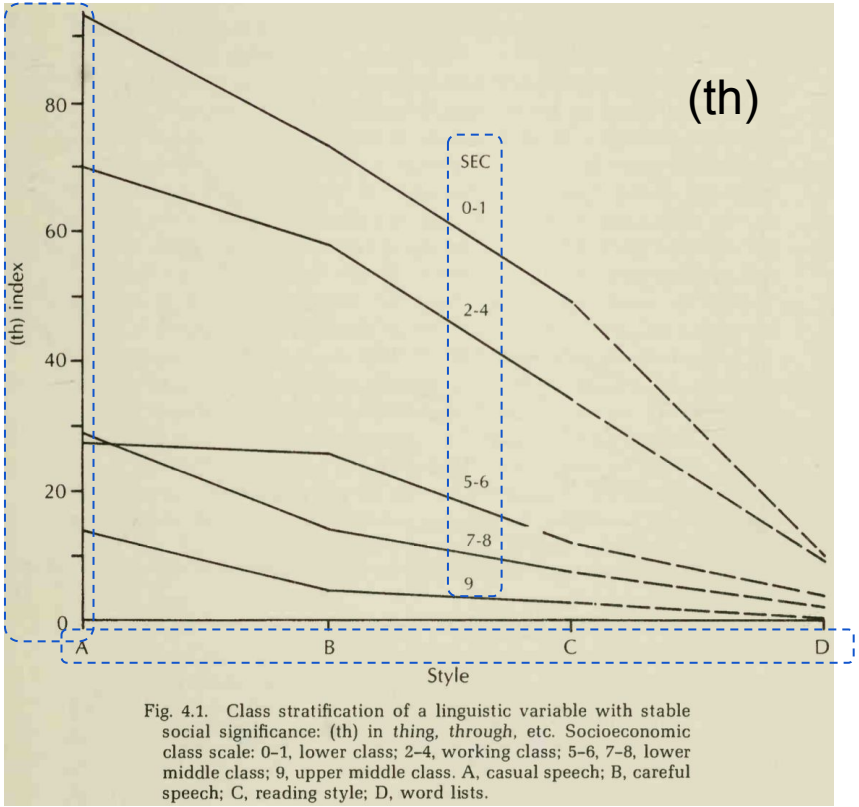


Fig. 4.1. Class stratification of a linguistic variable with stable social significance: (th) in *thing, through*, etc. Socioeconomic class scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists.

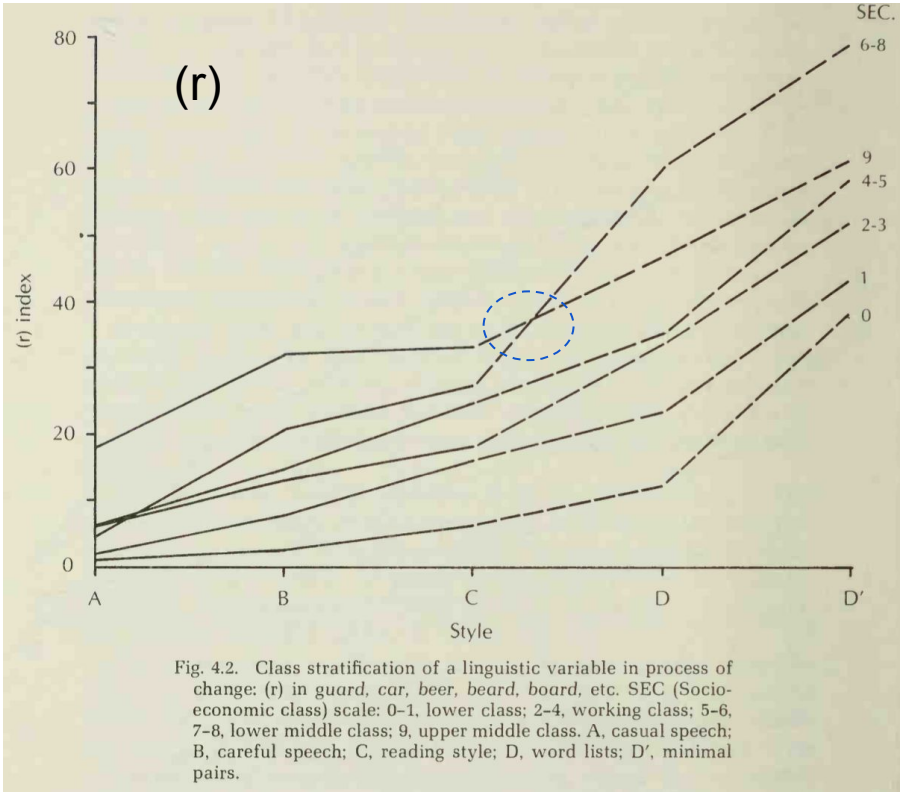


Fig. 4.2. Class stratification of a linguistic variable in process of change: (r) in *guard, car, beer, beard, board*, etc. SEC (Socioeconomic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs.

Labov (1972: Ch. 5)

Labov (1972): Interviews in the Lower East Side

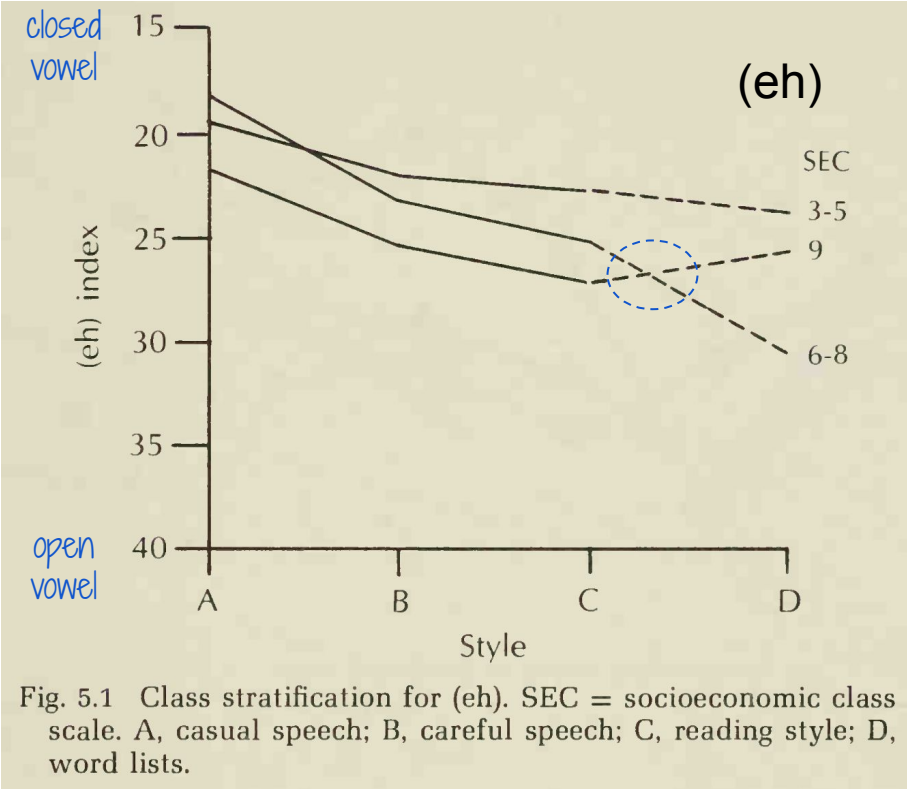


Fig. 5.1 Class stratification for (eh). SEC = socioeconomic class scale. A, casual speech; B, careful speech; C, reading style; D, word lists.

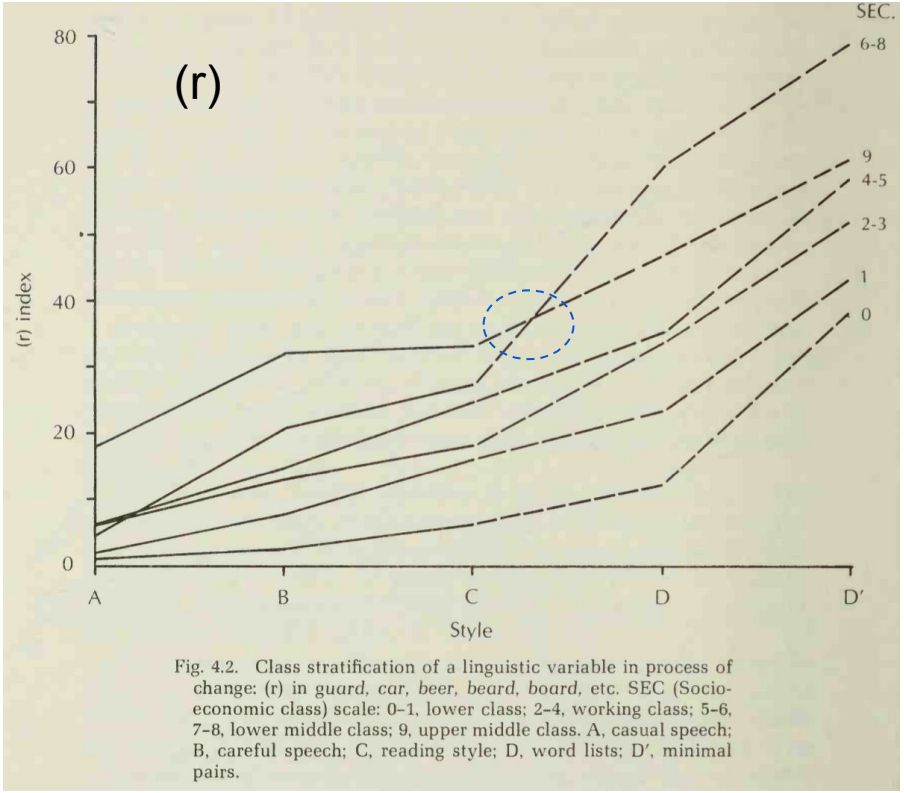


Fig. 4.2. Class stratification of a linguistic variable in process of change: (r) in *guard, car, beer, beard, board,* etc. SEC (Socio-economic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs.

Labov (1972: Ch. 5)

Labov (1972): Interviews in the Lower East Side

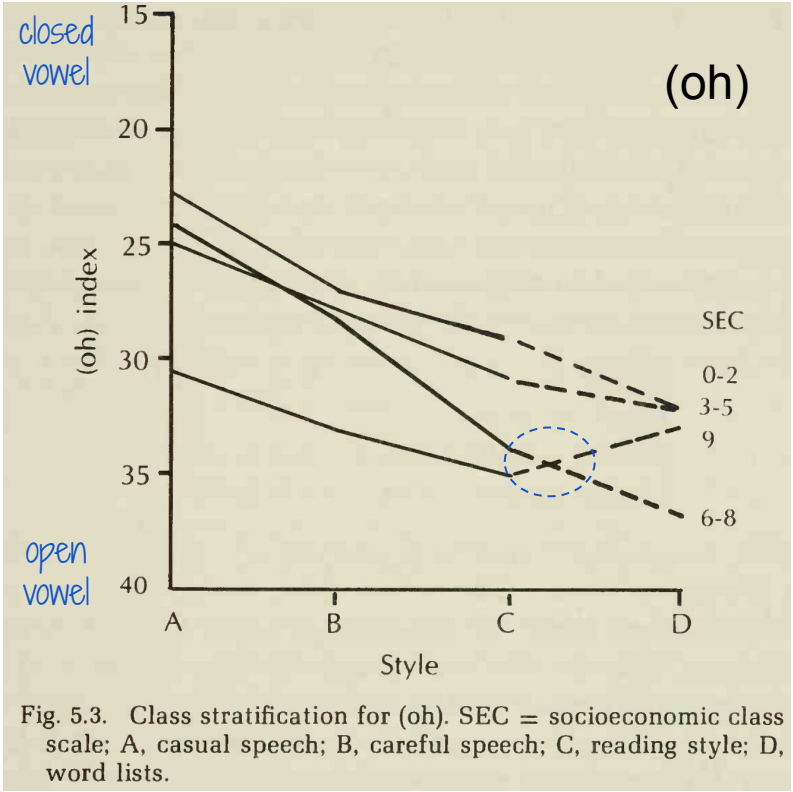


Fig. 5.3. Class stratification for (oh). SEC = socioeconomic class scale; A, casual speech; B, careful speech; C, reading style; D, word lists.

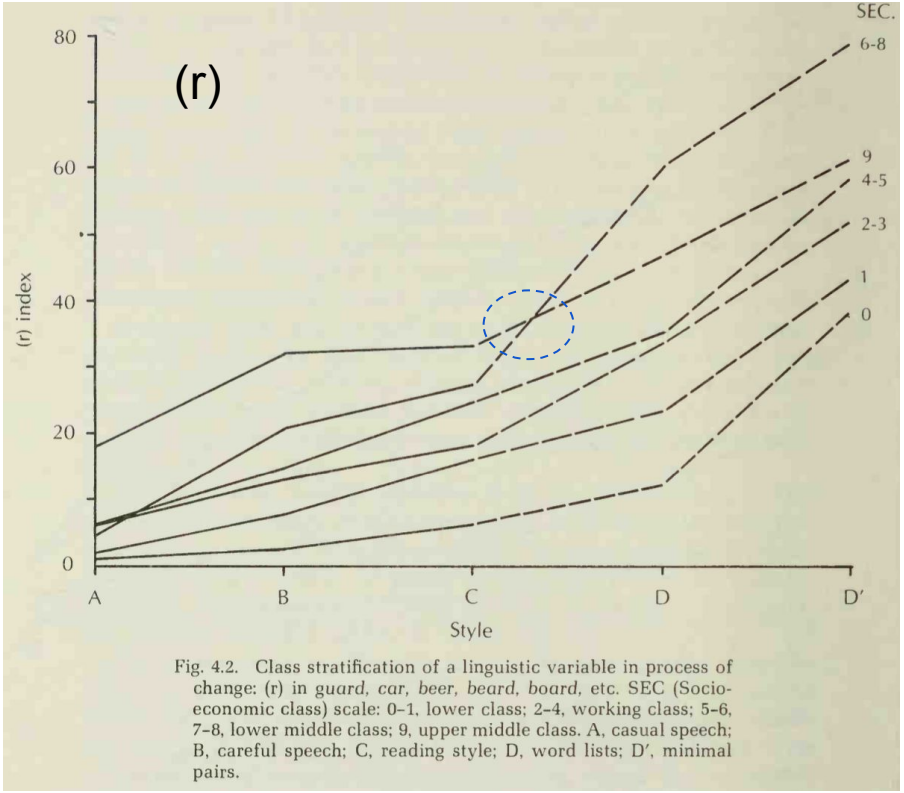


Fig. 4.2. Class stratification of a linguistic variable in process of change: (r) in *guard, car, beer, beard, board,* etc. SEC (Socio-economic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs.

Labov (1972: Ch. 5)

Labov (1972): Interviews in the Lower East Side

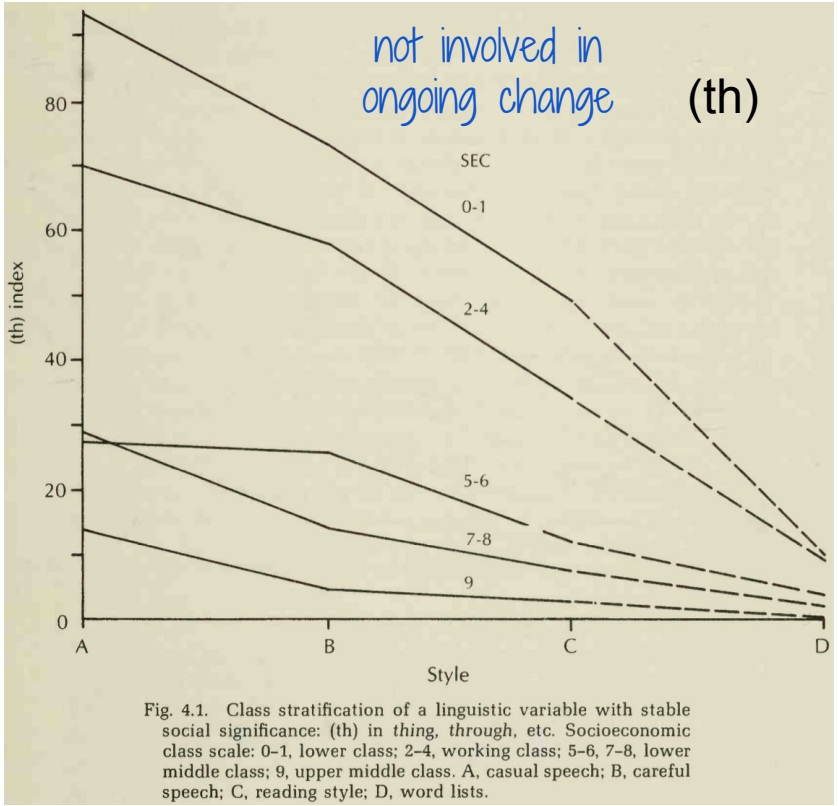


Fig. 4.1. Class stratification of a linguistic variable with stable social significance: (th) in *thing, through*, etc. Socioeconomic class scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists.

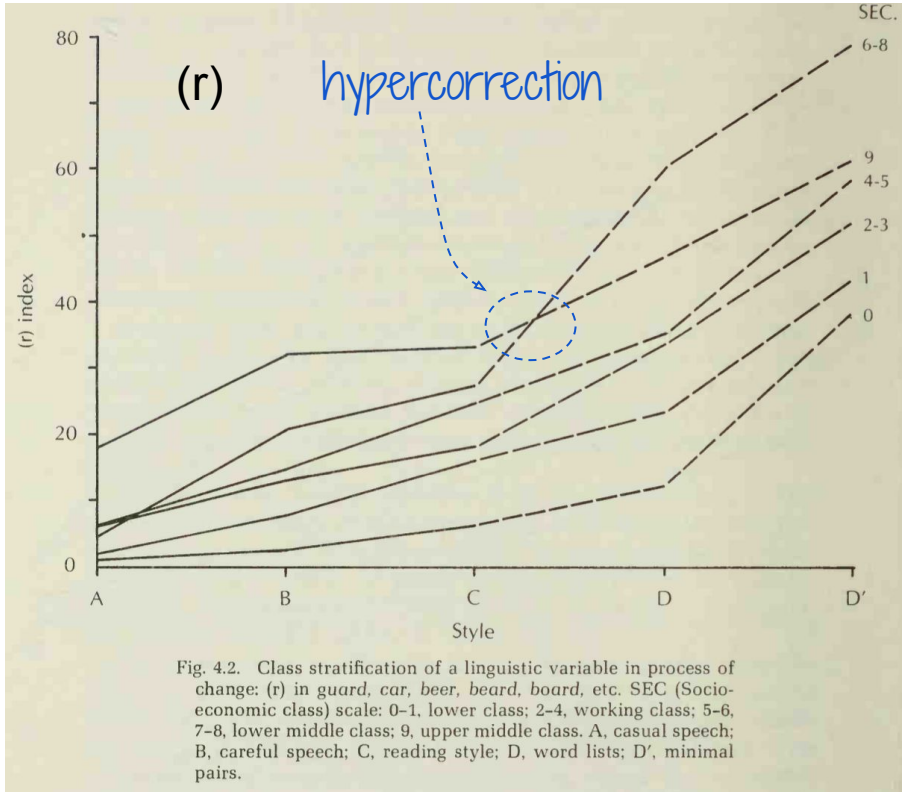


Fig. 4.2. Class stratification of a linguistic variable in process of change: (r) in *guard, car, beer, beard, board*, etc. SEC (Socioeconomic class) scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists; D′, minimal pairs.

Labov (1972: Ch. 5)

Labov (1972): Interviews in the Lower East Side

- Subjective reaction test: given recordings of target variables (+ controls), rate speakers on an occupational suitability scale

realized closed vowels stop

Social class		**(r) pos.**	**(oh) neg.**	**(eh) neg.**	**(th) sens.**
Lower	0–2				
Working	3–5				
Lower middle	6–8				
Upper middle	9				

Labov (1972: Ch. 5)

Modeling socioeconomic class (Dodsworth, 2009)

Early-class models (cf. Labov)

- Artificial boundaries determined by the researcher
- Ignores social mobility
- Women classified in terms of husbands and fathers
- Isolation from other aspects of social identity

Local and identity-based models

- The linguistic market (Bourdieu & Boltanski, 1975; Sankoff & Laberge, 1978)
- Density & multiplexity of social networks (Milroy & Milroy, 1992)
- Communities of practice (Wenger, 1998; Eckert, 2000)
- Relational class (Mallinson, 2007)

Takeaways

- Socioeconomic status has been consistently linked to language variation
  - In stable variation: clearly aligned with level of formality
  - In ongoing change: frequent hypercorrection of lower middle class, “[going] beyond the highest-status group [when using] the forms considered correct and appropriate for formal styles” (Labov, 1972: 244)
- Its operationalization is not straightforward
  - Methodological decisions required, ideally before data collection
  - More recent approaches provide further context for seminal work

NLP for socioeconomic status

So far…

- Small sample of participants (clerks in particular shops in New York)
- Starting from known linguistic variables (pronounced /R/)
- Let’s analyze linguistic variation related to socioeconomic status through social media
    - How to **measure** socioeconomic status of people writing online?
    - How to **discover** new linguistic variables from these texts?

How to measure socioeconomic status of people writing online?

How to measure socioeconomic status?

- Objective
  - education – access to higher salaries and more prestigious occupations
  - income – measure of an individual's access to goods and services
  - occupation – strong indicator of prestige and other formative experiences

- Subjective
  - asking for people's perception of their social class
  - Macarthur scale – people are asked to place themselves on a ladder (the higher, the more privileged)

- Multidimensional concept

Source: Curry et al.(2024)

The lack of (social) class in NLP

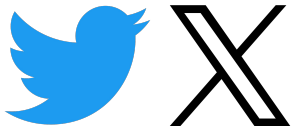
	Measurement	Granularity
Lampos et al. (2014)	Unemployment	Country
Preoţiuc-Pietro et al. (2015)	Occupation	Individual
Flekova et al. (2016)	Income	Individial
Hasanuzzaman et al. (2017)	Income	Individual
Giorgi et al. (2018)	Income, Education	County (census data)
Zamani et al. (2018)	Income, Education, Unemployment	Country-wise
Degaetano-Ortlieb (2018)	Class (high, low)	Individual
Van et al. (2019)	Income, poverty education	State-level (census)
Jawahar and Seddah (2019)	Income, geolocation	Neighbourhood-level
Basile et al. (2019)	Restaurant price	Individual
Ghazouani et al. (2019)	socio-economic status	Mixed
Abraham et al. (2020)	Income, area	Group
Tafreshi et al. (2021)	Income, education	Individual
Abbasi et al. (2021)	Income, education	Individual
Strømberg-Derczynski et al. (2021)	SES (high, mix, unknown)	Aggregated by dataset
van Boven et al. (2022)	Low-income countries	Country
Ngao et al. (2022)	Low-income countries	Country
Grützner-Zahn and Rehm (2022)	GDP	Country
Cole (2022)	Class (high, low)	Individual
Malik et al. (2022)	Caste, occupation	General (bias)
Hržica et al. (2022)	Class (middle)	Group

Source: Curry et al.(2024)

Preoţiuc-Pietro et al. (2015): User **occupational** class through… Twitter content

- Filter Twitter account that self-disclose occupations in their profile information

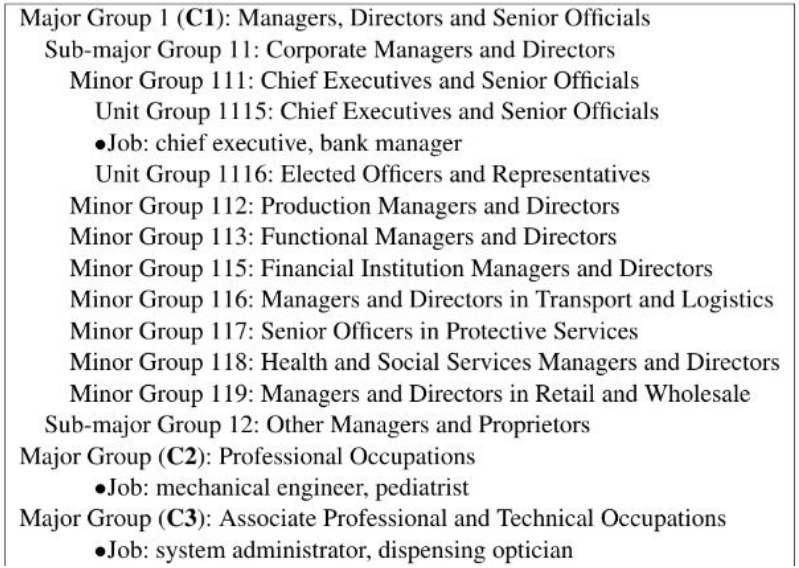
- Validation on a random set of 500 users
  - no description 12.2%
  - random information 22%
  - user information but not occupation related 45.8%
  - job related information 20%



Source: Preoţiuc-Pietro et al. (2015)

User occupational class through Twitter content

- **The Standard Occupational Classification (SOC)** – UK government system developed by the Office of National Statistics for classifying occupations
- Query Twitter API for self-disclosed SOC job titles
- Max 200 accounts for each job title

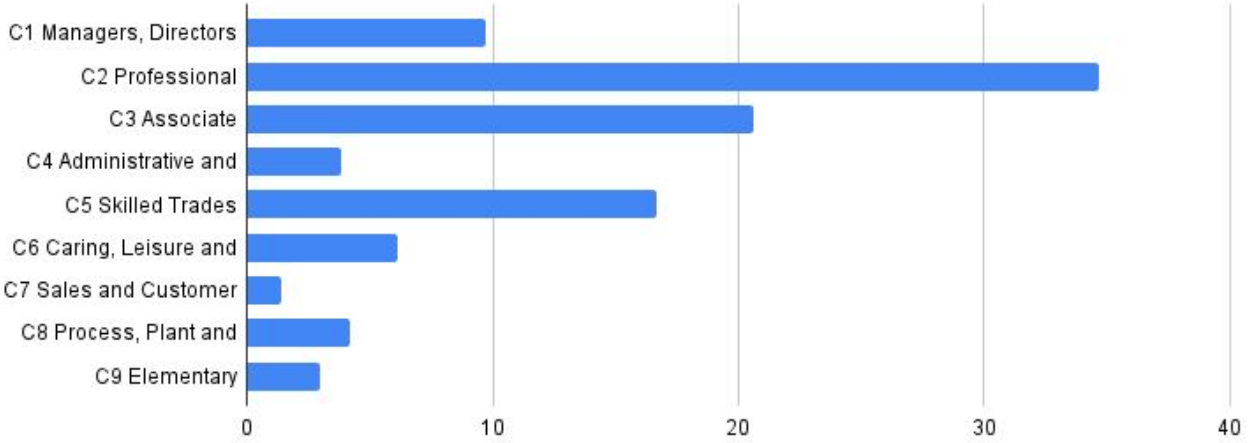


(...) 9 major groups

Source: Preoţiuc-Pietro et al. (2015)

User occupational class through Twitter content

- 50% removed after manual filtering
- Final data: 5,191 users, all 9 major SOC groups, 22 sub-major and 55 minor



Source: Preoţiuc-Pietro et al. (2015)

How to discover new linguistic variables from texts?

How to **discover** new linguistic variables from texts?

Basile et al. (2019): “You Write Like You Eat”

- assumption: socio-economic background of people can be predicted from the restaurants they visit

- data: restaurant reviews from Yelp!

Source: Basile et al. (2019)

Yelp reviews

- Restaurants have price categories $, $$, $$$, $$$
- After filtering: 138 authors for each category

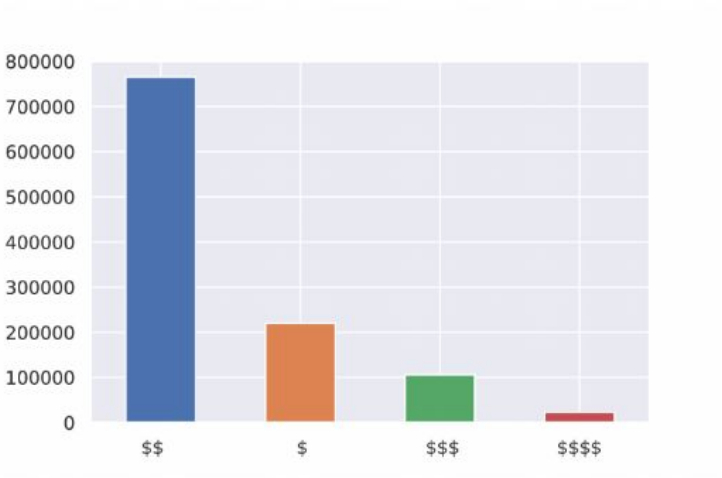


Figure 2: Author distribution before filtering. While users belonging to class $$$$ might visit cheaper places, the same is not true in the opposite direction: this explains the small size of class $$$$.

Source: Basile et al. (2019)

CLASS $	CLASS $$$$
So freaking good. That's all I'm gonna say. Don't believe me? Walk into the place and smell it. [...] Will definitely go back.,Fresh, hand-made pepperoni rolls..... oh yeah. Their cheesy focattia (did I spell that right?) is amazing. Take it home, throw it in the oven, drizzle a little EVOO on top and you're golden. Friendly people there. Parking sucks, but I'm not taking off a point for that! Their marinara is dee-lish,Super tasty!!!	Let me start off saying that 2 years ago my husband and I had a spectacular dinner at L'Atelier by Joel Robuchon and finally got the "Time" to visit Joel Robuchon.We got a limo service and a nice tour inside the mansion of Robuchon which was very memorable and the hostess escorted us to the dining area. Decore: In comparison to L'Atelier this place was much more chic and elegant. However, I still loved the idea to see all the chefs preparing and decorating my plates at L'Atelier.

Figure 3: Sample reviews for classes $ and $$$$.

Source: Basile et al. (2019)

Hypothesis: some classes are typified by writing styles

$ – *hand-made pepperoni rolls.... oh yeah*

$$ – *Their marinara is dee-lish,Super tasty!!!*

$$$ – *When Jet first opened, I loved the place.*

$$$$ – *compared to pierre gagnaire in paris, the food here is way less ambitious*

- Differences in formality and approach to informal speech
- Frequency of interjections, abbreviations and/or emojis
- Extent to which the style approximates speech, e.g., using exclamation marks
- Incorrect spelling, orthography

Source: Basile et al. (2019)

How to find stylistic variation in restaurant reviews?

- Automatically predict restaurant's price range based on a review
- Analyze the classifier and check which features are the most important
- These features are the indicators of the stylistic variation

Source: Basile et al. (2019)

How to find stylistic variation?

- Bag-of-word representations
- 3-6 word and character n-grams
- Logistic regression classifier
- F1 accuracy 0.53

Relying on words captures restaurant aspects

$	$$	$$$	$$$$
fast	tried	at	excellent
kids	happy	clubs	gras
coffee	staff	wynn	we
customer	won	music	las
clean	put	pretty	steak
they	phoenix	night	tasting
order	find	club	foie
came	try	vegas	wine
always	place	buffet	course
pizza	salsa	hotel	vega

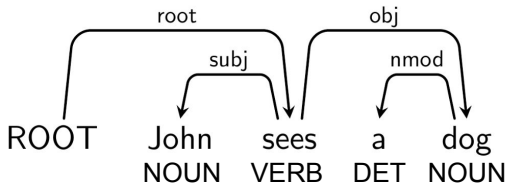
Table 4: The 10 most important word features per class. We omit character-level (ngram) features to facilitate interpretability.

Source: Basile et al. (2019)

Is there any stylistic variation in Yelp reviews?

- Stronger classifier: CNN + multilayer perceptron

- Style-aiming representations:
  - substitute words with their respective POS-tags
  - transform words into triplets (incoming arc, POS tag, POS of head)



John → (subj, NOUN, VERB)

  - bleaching

Source: Basile et al. (2019)

Bleaching

Token	Bleached representation
I	X_01_True_V_2117
love	xxxx_04_True_CVCV_617
pizza	xxxxx_05_True_CVCCV_15
!	!_01_False_!_21

- Number of chars
  - X capitalized
  - x small letters
- Number of chars numerically
- Alphanumeric True/False
- Vowels or Consonants
- Frequency

Source: Basile et al. (2019)

Is there any stylistic variation in Yelp reviews?

model	F1
random baseline	0.25
LR BOW (lexical) baseline	0.53
CNN lexical	0.54
CNN pos tags	0.33
CNN dependency tree	0.52
CNN bleaching	0.46

relying on content words

type of classifier does not matter

no content words, but still above the baseline

Possible to differentiate authors based on stylistic parameters

Source: Basile et al. (2019)

Takeaways

- How to **measure** socioeconomic status of people writing online?
  - Multidimensional concept
  - NLP uses various proxies, such as income, occupation, or life experiences
  - Examples: Twitter and Yelp reviews as sources of data
- How to discover new lexical variables?
  - Example: logistic regression, CNN
  - Examples of features: POS tags, syntactic dependencies, bleaching
- Under-researched topic in NLP

Questions?

Practical exercise

Day 2: Quebec linguistic variation and SAGE

- Today's goal:
  Explore large-scale quantitative estimates of lexical variation across regions
  - Open **Geographical Variability** notebook
  - Follow all the steps and fill in the missing pieces of code (marked with TODO)

Takehomes

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

Tomorrow…

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

Universität Stuttgart

IRIS

Thank you for your attention!

Image credit: DALLE (2024)