Universität Stuttgart

IRIS

Language in Social Context: Bridging NLP and Sociolinguistics

Agnieszka Faleńska
Filip Miletić

Day 1: Introduction

Image credit: DALLE (2024)

Who are we?

Agnieszka Faleńska

NLP (Natural Language Processing) – subfield of AI that makes computers understand and generate language

- Research group leader: Diversity-Aware NLP Systems

- Computational modeling of subtle linguistic biases
- Cross-lingual structure prediction
- Interpretability and analysis of NLP models

Filip Miletić

Variationist & computational sociolinguistics – identifying and explaining variation in language use

- Postdoctoral researcher in SemRel group (IMS Stuttgart)

- Computational and sociolinguistic modeling of semantic variation
- Analysis of NLP models driven by social and linguistic phenomena
- Resources in interdisciplinary settings

Pema Gurung

- Student assistant
- Help with all the practical exercises

Who are you?

- Students
  - Bachelor?
  - Master?
  - PhD?
  - anybody else?

- Topic
  - linguistics?
  - computational linguistics?
  - other field with “linguistics” in the name?
  - anybody else?

- Who feels confident in…
  - statistics?
  - Python?
  - machine learning?

What is this course about?

What can language tell us about speakers?

Listen to the recording and then fill out the survey.

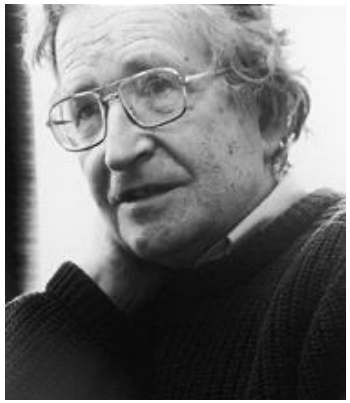
What can language tell us about speakers?

- As speakers, we routinely make choices between competing linguistic forms
- The resulting variability is structured in nature (“orderly heterogeneity”)
  - internal (linguistic) and external (social) factors
  - social structures and patterns of interaction
  - social meaning conveyed by linguistic choices
- These perspectives constitute the focus of sociolinguistics

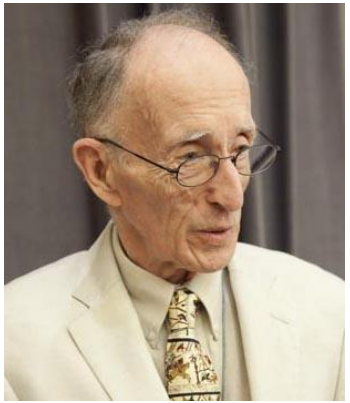
Why study language & society?

- Structural linguistics focuses on *langue* rather than *parole*
- Generativism similarly distinguishes *competence* and *performance*
- Sociolinguistics flips the perspective by investigating variability in language

Ferdinand de Saussure
(1857–1913)



Noam Chomsky
(1928–)



William Labov
(1927–)

“I have resisted the term sociolinguistics for many years, since it implies that there can be a successful linguistic theory or practice which is not social.”
(Labov, 1972: xiii)

Photo credits:
F. Jullien via Wugapodes/Wikimedia
J. Soares via Verdy_p/Wikimedia
Universitat Pompeu Fabra/Flickr

Why study language & society?

Traditional contributions

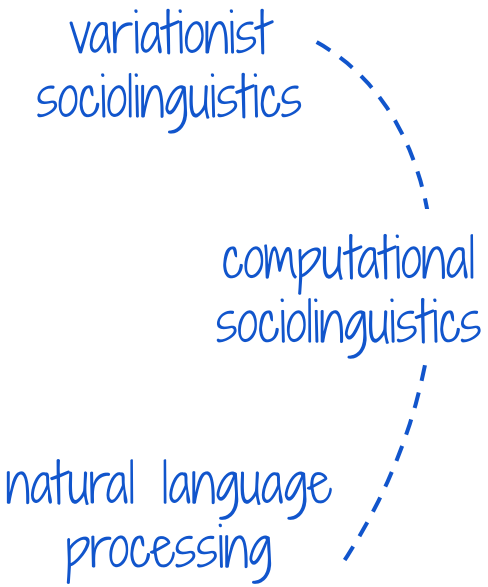
- Explanations of language variation and change
- Insights into social structure

Deployment of new computational methods

- Scaling up sociolinguistic analysis
- Investigating new phenomena

Analysis of emergent language technologies

- Effects of language variation in training data
- Understanding new types of interactions



Course goals

- Get familiar with the central concepts behind **language variation**
- Have a general overview of **computational approaches** to analyze linguistic variants
- Understand **implications** of language variation on NLP models
- [Bonus] **Practical exercises** for students familiar with Python

Course structure

Day 1: Introduction
Overview of perspectives and methods

Day 2: Demographic factors (I)
Effects of geographic origin and socioeconomic status

Day 3: Demographic factors (II)
Effects of age and gender

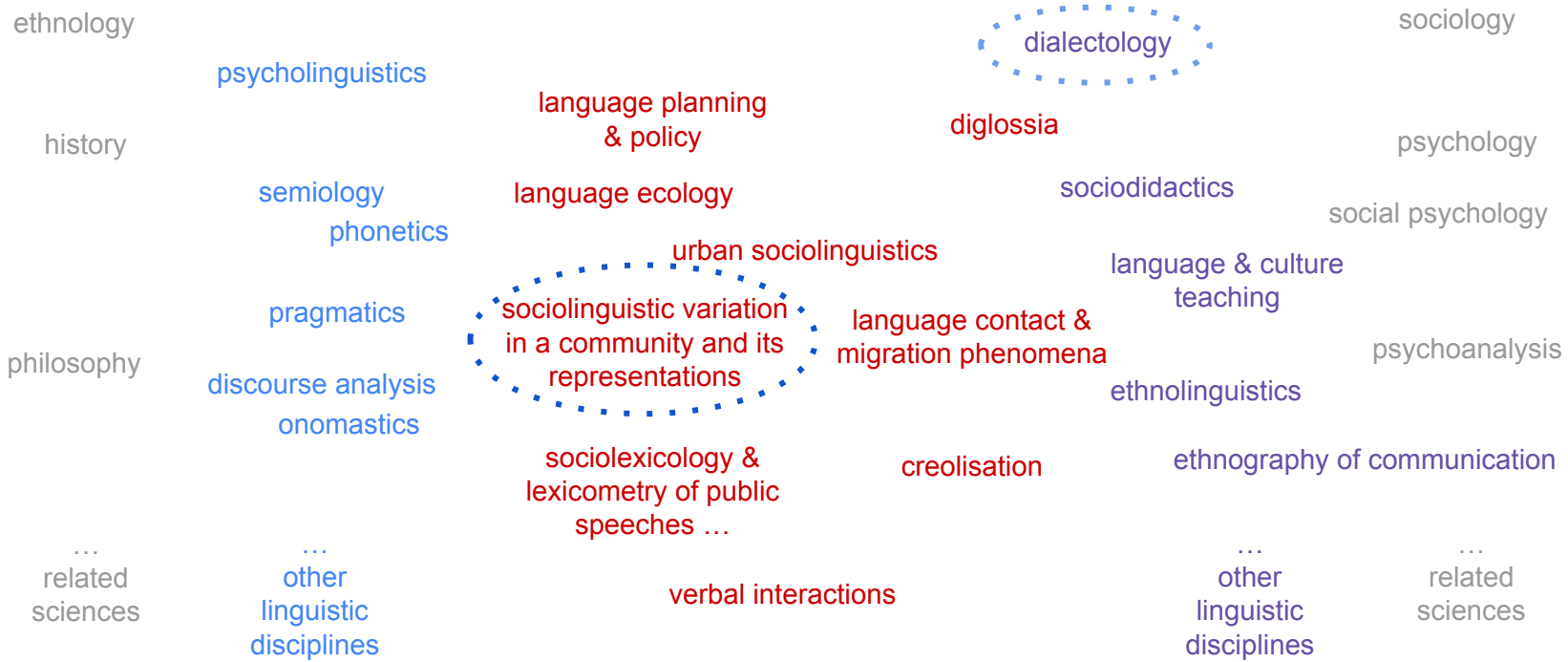
Day 4: Language variation in interaction
Adapting to interlocutors and conveying social meaning

Day 5: NLP applications and challenges
Ethical issues, applications, harms & biases

What is sociolinguistics?

Macrosociolinguistic perspective

has to do with the societal level,
the speech community in its entirety



group, network, family, or individual level

Microsociolinguistic perspective

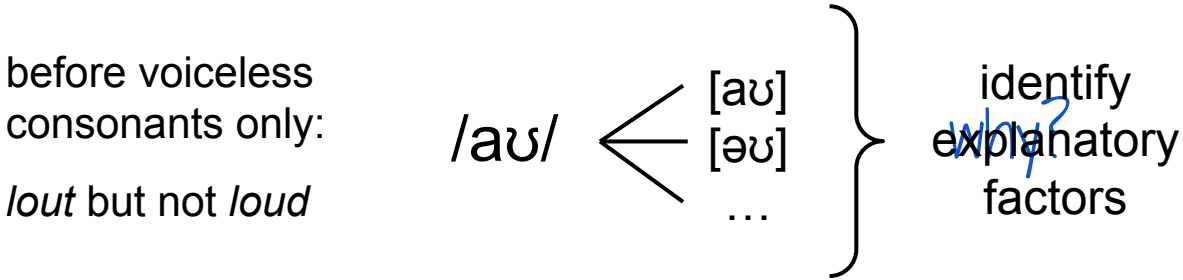
Adapted from Boyer (2019: 19)

Defining sociolinguistics

- In broad terms, sociolinguistics is “that part of linguistics which is concerned with language as a social and cultural phenomenon” (Trudgill, 2000: 21)
- In contrast to general linguistics, sociolinguistics is “the study of language structure and evolution within the social context of the speech community” (Labov, 1972: 184)
- Variationist sociolinguistics is “the study of the interplay between variation, social meaning and the evolution and development of the linguistic system itself” (Tagliamonte, 2006: 5)
- Computational sociolinguistics “integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective” (Nguyen et al., 2016: 540)

Main analytical steps

- Identify a linguistic variable: “two alternative ways of saying the same thing” (Labov, 2004: 7)
- Circumscribe the variable context (cf. Tagliamonte, 2006: 86–94)
- Apply the principle of accountability: “we will report values for every case where the variable element occurs in the relevant environments as we have defined them” (Labov, 1972: 72)



Types of variation

Levels of linguistic structure

- Phonetic / phonological
- Morphosyntactic
- Lexical
- Semantic
- Discourse-pragmatic
- …

Speaker perspective

- Interspeaker
- Intraspeaker
  - ↳ style

Internal explanatory factors

- Linguistic context
- Lexical frequency
- …

External explanatory factors

- Geographic origin
- Age & gender
- Socioeconomic status
- Ethnic origin
- Knowledge of languages
- …
- Time → language change

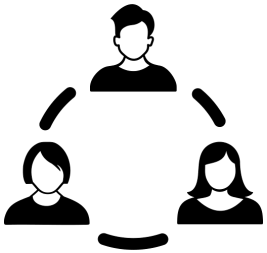
Bonus classification

- Diachronic → time
- Diaphasic → situation
- Diamesic → medium
- Diastratic → social factors

Method: Core principles

Focus on the **speech community:**

“defined [...] by participation in a set of shared norms; these norms may be observed in overt types of evaluative behavior, and by the uniformity of abstract patterns of variation which are invariant in respect to particular levels of usage” (Labov, 1972: 120–121)

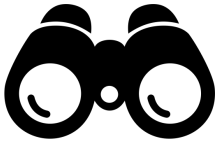


Careful **sampling** of speakers

Targets **the vernacular:**

“the style in which the minimum attention is given to the monitoring of speech” (Labov, 1972: 208)

The observer’s paradox:



“the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation” (Labov, 1972: 209)

Method: Approaches to data collection

- The sociolinguistic interview (Labov, 1992; Tagliamonte, 2006)
- Participant observation (Eckert, 2000)
- Rapid & anonymous survey (Labov, 1972)
- Surreptitious recording (cf. Milroy & Gordon, 2003: 83–83)
- Dialect survey (Dollinger, 2015; Preston, 2002)
- Matched guise technique (Lambert et al., 1960)

Why NLP and sociolinguistics?

Data sources in sociolinguistics

Traditional sociolinguistics

Small samples of data

Time consuming

The observer's paradox

Computational sociolinguistics

Where to get examples of language use on a big scale?

NLP to process and analyze the data

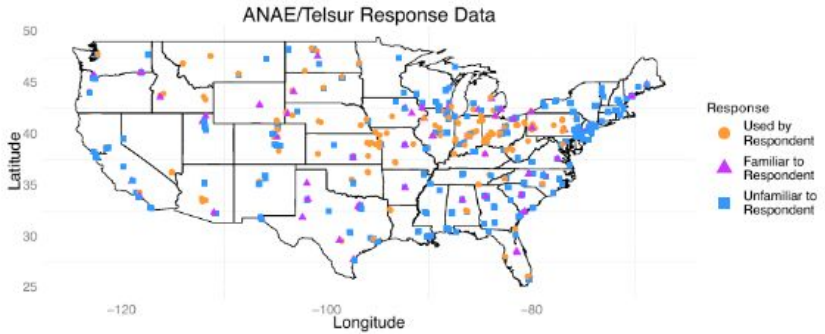
Source: slides by Dong Nguyen, 2017

Benefits of using large (social media) data

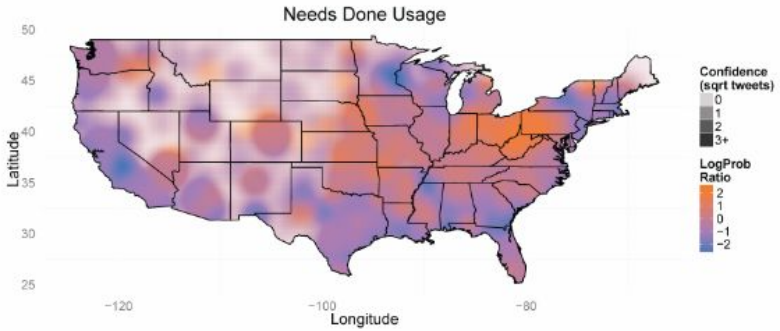
Testing and refining theories

Regional variation: ***needs* + past participle** as in *The car needs washed*

The *Atlas of North American English*



“Needs done” on Twitter



Source: Doyle, (2014)

→ Day 2: geographic origin

Analysis of rare and complex patterns

‘since I have no school I became very **lazy** :D’

German-Turkish code-switching

Source: Çetinoğlu (2016)

Improving and personalizing tools

Their music is pretty sick as well

I’m walking on sunshine <3 #and don’t you feel good

seitdem ich keine schule hab cok **faul** oldum ben ya:D

- Sentiment analysis?
- Translation?
- Question answering?
- (…)

Source: slides from slides by Dong Nguyen, 2017, Çetinoğlu (2016)

→ Day 5: NLP applications

Challenges of using social media data

Skewed version of the society

Twitter in 2024

- younger people (38% aged 18-29)
- most users in the US
- 66% men

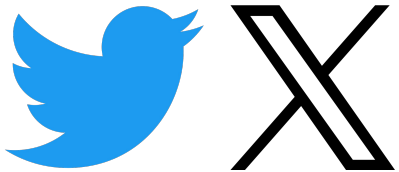
Reddit in 2024

- younger people (44% aged 18-29)
- 50% based in the US
- 66% men

Source: stats from Statista, Images from slides by Dong Nguyen, 2017

Changing platform design and restrictions

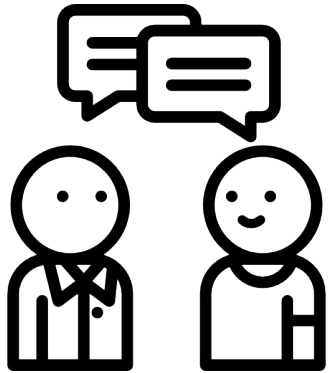
- Twitter API used to give access to 10% sample of public posts
- After 2019 – no geotagging
- After February 2023 – restricted access to Twitter API



- Big changes to the API announced in April 2023
- Smaller changes constantly happening:
  - e.g., now post “scores” take only “up” votes into account

Predominant modality of data

Traditional sociolinguistics (spoken)



Computational sociolinguistics (written)

Source: slides from Ian Stewart, 2020, conversation image

What can written language tell us about speakers?

Fill out the survey.

Tweet 1: I’m walking on sunshine <3 #and don’t you feel good

Tweet 2: lalaloveya <3

Tweet 3: @USER loveyou ;D

Source: Nguyen et al., (2014)

What can written language tell us about speakers?

Nguyen et al., 2014: “Why Gender and Age Prediction from Tweets is Hard”

- 77% of voters guessed female… **16-year old biological male**
- Linguistic markers usually associated with female style:
  - a heart represented as <3
  - emotions, emoticons
  - non-standard spelling
- Challenging task
  - no prosody, no voice
  - very short text
  - different markers: emoticons, spelling

Source: Nguyen et al., (2014)

Social media as a data source

Benefits:

- large scale
  - rare patterns
  - complex patterns
  - variety of patterns
- available in many languages
- use cases in NLP

Challenges:

- skewed demographics
- changing platform design and restrictions
- mostly written text

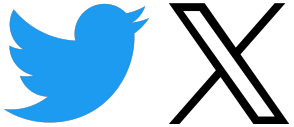
→ **Day 5: ethical considerations and biases**

How do we get demographic annotations for authors on social media ?

How to get demographic annotations for authors?

Source: slides by Dong Nguyen, 2017

Where to find **self-reported** annotations?



Twitter / X
(Nguyen et al. 2013)

Reddit
(Voigt et al., 2018)

Youtube comments
(Filippova et al., 2012)

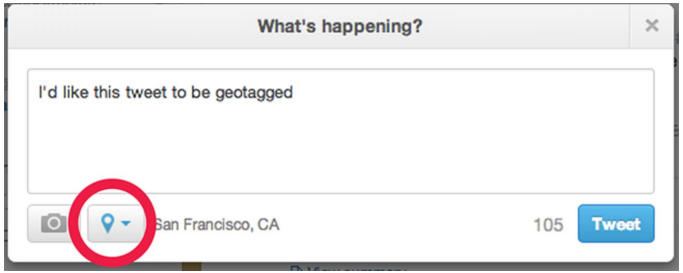
blogs
(Schler et al., 2006)

Online forums
(Voigt et al., 2018)

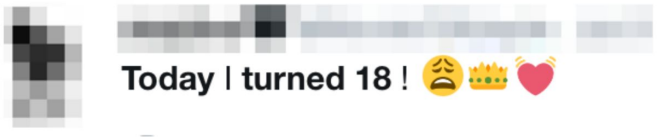
How do people self-report their information?

Profile information

Geotagging



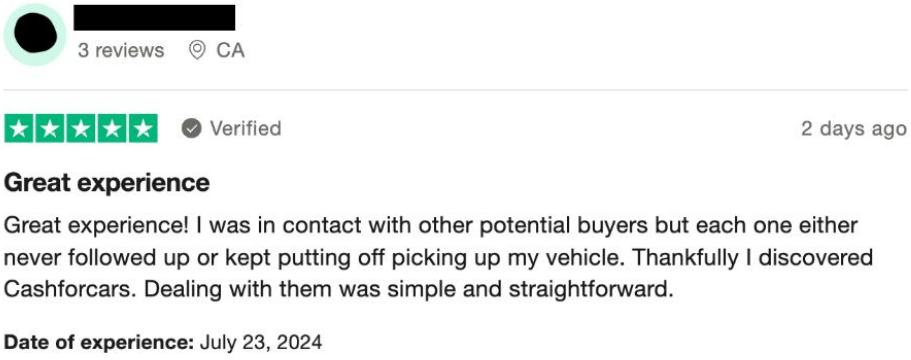
Explicit mentions



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

How frequent is demographic information in user profiles?

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



	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)

How frequent is demographic information in user profiles?

Twitter as a source of geotagged data (until 2019)

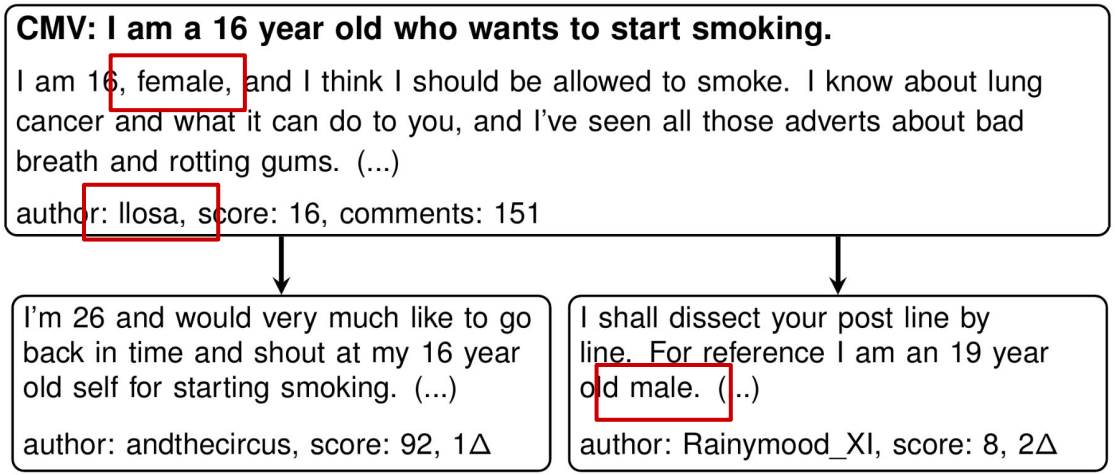
- Geotags – coordinates specified by a GPS system
- User-specified coordinates
- User specification of a home location

1-2% tweets with geotags (Duggan and Smith, 2013)

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

How frequent are explicit mentions?

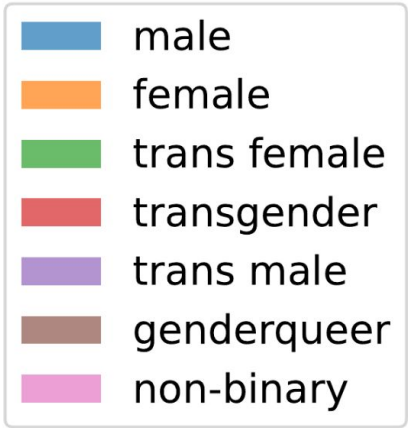
Falenska et. al, (2024): “Self-reported Demographics and Discourse Dynamics in a Persuasive Online Forum”



Source: Falenska et al., (2024)

How frequent are explicit mentions?

	Posts	Replies	Discussions	Authors
total	396	3,235	1,812	2,456
male	299	1,953	1,357	1,640
female	89	961	693	674
other	8	321	175	158



Source: Falenska et al., (2024)

Self-provided demographic annotations

Advantage:

- high accuracy

Disadvantages:

- biases
- availability
- truthfulness
- who “self-reports”?

How to get demographic information for authors?

Derived based on…

Wikipedia, for public figures
(Voigt et al., 2018)

Google

Google search, TED speakers
(Mirkin et al., 2015)

WikiData, youtubers
(Knupleš et al., 2024)

Approximated (?) based on…

First names
(Bamman et al., 2014)

Gender: morphology
(Falenska et al., 2018)

Pictures, LinkedIn, …
(Nguyen et al. 2013)

Derived demographic annotations

Advantage:

- high coverage

Disadvantages:

- quality
- manual filtering
- ethical considerations

How to analyze linguistic variation in textual data?

[In general]

Select data source and build corpora

1 I'm walking on sunshine <3 #and don't you feel good

2 lalaloveya <3

3 @USER loveyou ;D

4 Hiiiiii cutiesss!

5 Having fun with E[illegible] 20th class reunion!!

(...)

1 I went flat on my face. And the cabinet as well. Funny right? #Catshuis

2 Jeez. I almost can't follow it all.

3 Wow! That was a great concert! o_O

4 Well… what did we expect?

Group-generalized language use

Source: Nguyen et al., (2014)

Decide on the way to represent texts

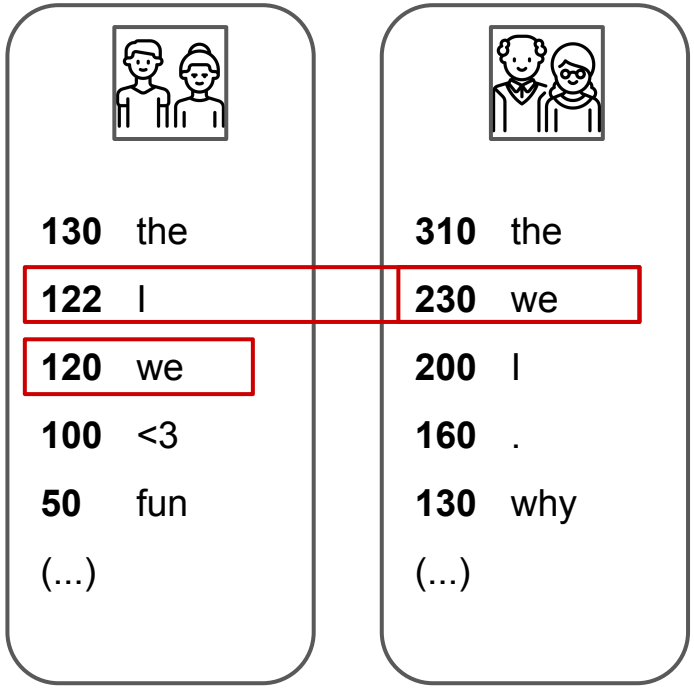
130	the
122	I
120	we
100	<3
50	fun
(...)	

310	the
230	we
200	I
160	.
130	why
(...)	

Other ways of representing text:

- n-grams of words:
  I love, we did, fun <3, …
- Characters
  a, b, c, d, …
  the, doi, unn, fun, …
- POS tags:
  NOUN, DET, VERB, …
- Syntactic relations:
  subj, obj, coord, …
- Vectors
  word2vec, ELMO, BERT, …

Analyze data

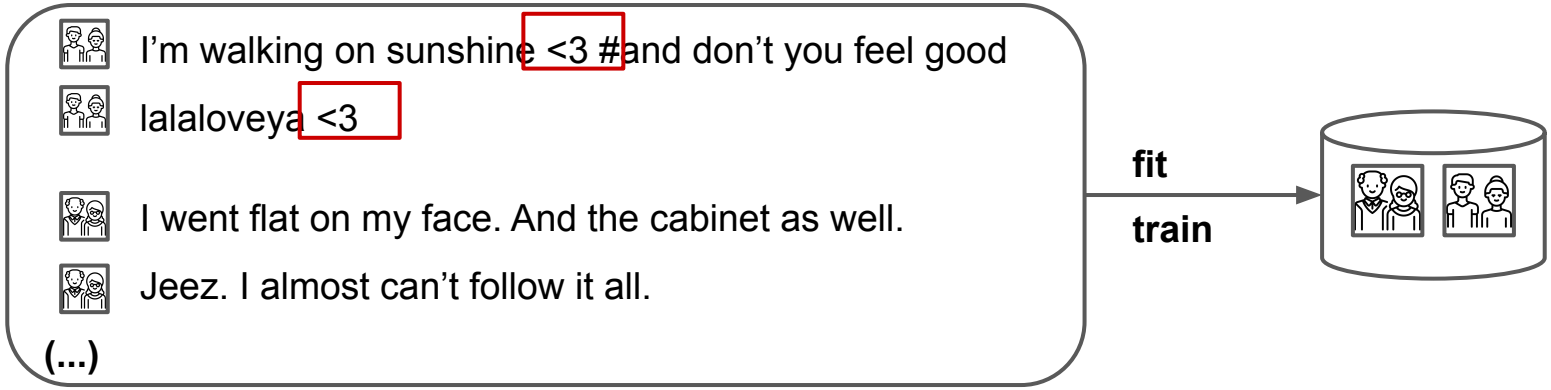


Goal: what differs between the corpora?

Examples of methods:

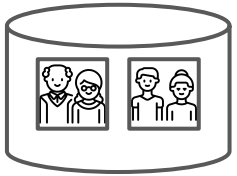
- Just frequency?
- Generative approaches P(“*we*” | *old*)
- Discriminative models: logistic regression

Terminology in discriminative models



Statisticians: fit models **Dependent variable**: age category **Independent variables**: features derived from the datasets (words, ngrams, …)	**ML:** train models **Task**: predict age category **Input**: features derived from the datasets (words, ngrams, …)	**Both:** **Models** that represent the data **Analyze** to find features with the strongest coefficients

Next three days: NLP for linguistic variation



- **Examples** of NLP methods for analyzing and discovering linguistic variation in texts
- Focus on **use cases**
- **No explanation** of the basic ML and statistic concepts
  - but you will survive without them :)

Last day: linguistic variation for NLP

Practical exercise

Day 1: Google Colab and Pandas dataframe

- In general, two types of notebooks:
  - Google Colab (all data and models will load automatically)
  - Local notebooks (search for the additional code to download the data and models in the repository)

- Today's goal: get familiar with data analysis tools
  - Open **Data Analysis** notebook
  - Follow all the steps and fill in the missing pieces of code (marked with TODO)
  - Get familiar with the trustpilot reviews

Trustpilot

Takehomes

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

Tomorrow…

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)

Example NLP methods for analysing linguistic variations
Logistic regression, SAGE, POS tagging, parsing, bleaching

Practical exercise
Exploring regional variation in a large-scale corpus

Universität Stuttgart

IRIS

Thank you for your attention!

Image credit: DALLE (2024)