Intro to Computational Sociolinguistics

Dr. Rachael Tatman
Data Scientist, Kaggle

@rctatman

The emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective.

(Nguyen 2017)

The emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective.

(Nguyen 2017)

All language contains sociolinguistic variation

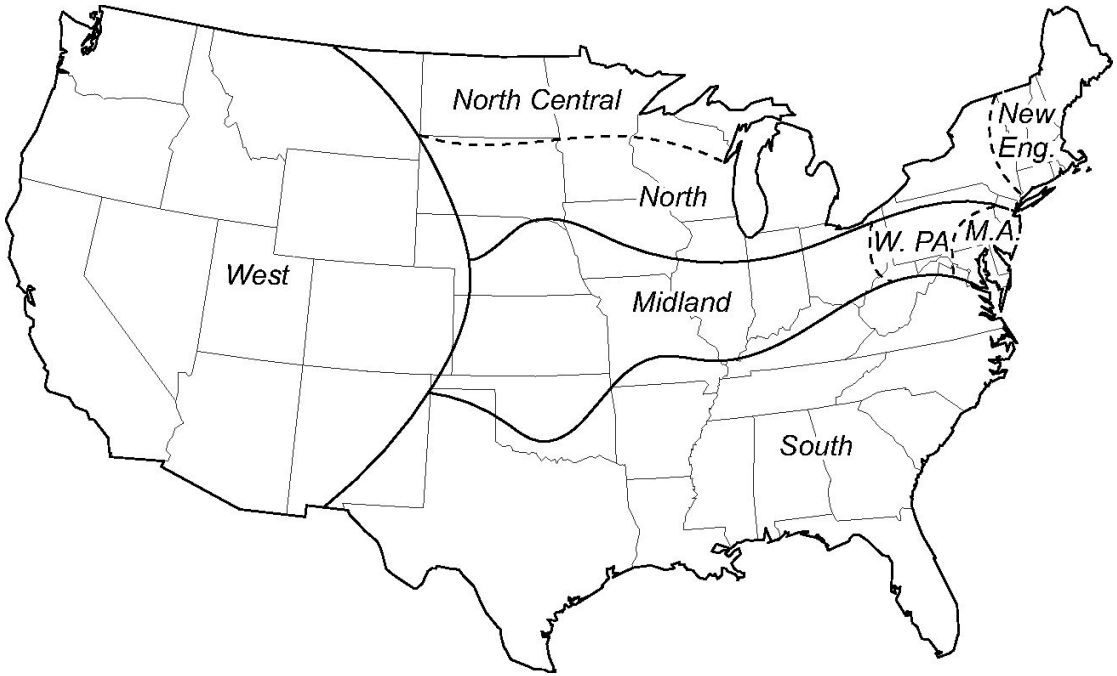
All language contains sociolinguistic variation

If you choose to ignore this it **will** eventually come back to bite you.

What matters?

- Regional Origin
- Age
- Socio-economic status/Social class
- Race/ethnicity
- Among many other factors!

Regional Origin



@rctatman

Age

In general, very young and very old talkers tend to use more regional variants or stigmatized forms.

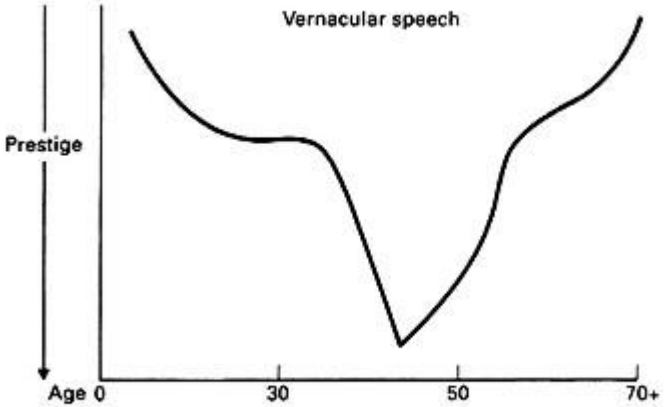
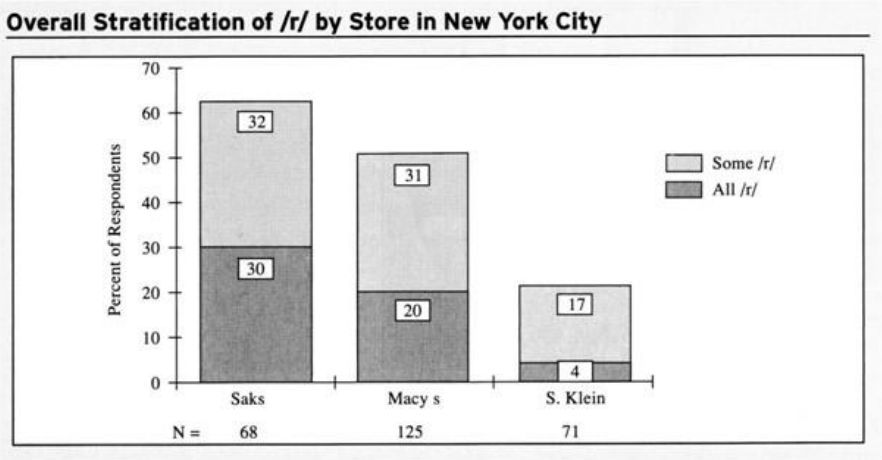


Figure 7.3 Relationship between use of vernacular forms and age. (Reproduced from Downes 1984: 191)

Socioeconomic Status

In general, speakers with higher socioeconomic status use fewer regional variants or stigmatized forms.



(Source: Finegan, 2004: 391)

Race/Ethnicity

- Many ethnolects in the US, including:
  - African American English
  - Chicano English
- Both have systematic structural differences from General American English

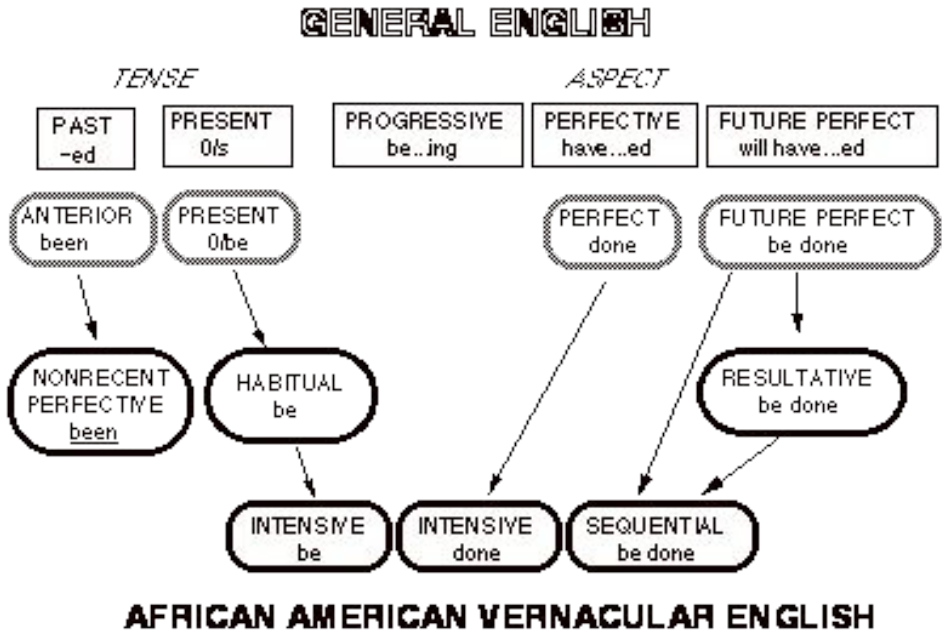


Figure 3. An overall view of the development of AAVE in the 20th century.

In S. Mufwene , J. Rickford, J. Baugh and G. Bailey (ed.), *The Structure of African-American English*, London: Routledge.1998. Pp. 110-153.

@rctatman

(Some) Other Factors

- Who you're talking to and how much you like them
  - Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. Annals of the International Communication Association, 10(1), 13-48.
- Who you think can hear you & what you want them to think about you
  - Bell, A. (1984). Language style as audience design. Language in society, 13(2), 145-204.
- Who you know & who they know (social networks)
  - Milroy, L., & Llamas, C. (2013). Social networks. The handbook of language variation and change, 407-427.

The emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective.

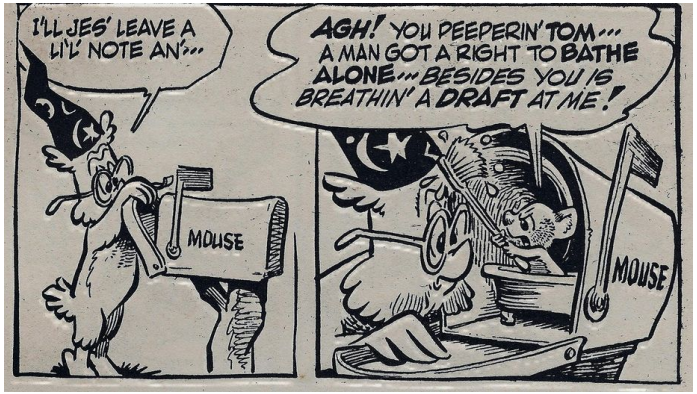
(Nguyen 2017)

Language + Computers: Existing Academic Disciplines

- Natural language processing automates the ability to recognize, understand and generate human language
  - Machine translation
  - Sentiment analysis, text summarization
  - Dialogue understanding and generation (e.g. chatbots)
- Other fields:
  - Automatic speech recognition (sometimes part of NLP, sometimes electrical engineering or signal processing)
  - Machine learning often uses language data
  - Computational linguistics is more focused on language as an object of study, less on applications

How did NLP handle language variation?

- Historically not a focus of the field
  - Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) started in 2014
- Variation removed through pre-processing
- Few tasks focused on language variation:
  - User profiling
  - Dialect adaptation (for speech recognition)



@rctatman

Sociolinguistics + NLP = Computational Sociolinguistics

Computational Sociolinguistics:

- Improves NLP systems
- Lets us ask new questions

Ignoring variation leads to preventable errors

- Lang ID performs reliably worse on African American English
  - Blodgett, S. L., & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. arXiv preprint arXiv:1707.00061.
- Part of speech taggers work better for older authors
  - Hovy, D., & Søgaard, A. (2015). Tagging Performance Correlates with Author Age. In ACL (2) (pp. 483-488).
- Part of speech taggers are less accurate on African American English
  - Jørgensen, A., Hovy, D., & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In Proceedings of the Workshop on Noisy User-generated Text (pp. 9-18).

Some approaches to correcting this

- Demographically-balanced training data
  - Blodgett, S. L., Wei, J., & O'Connor, B. (2017). A Dataset and Classifier for Recognizing Social Media English. In Proceedings of the 3rd Workshop on Noisy User-generated Text (pp. 56-61).
- Incorporate other sources of information (like geo-tagging for tweets)
  - Salehi, B., & Søgaard, A. (2017). Evaluating hypotheses in geolocation on a very large sample of Twitter. In Proceedings of the 3rd Workshop on Noisy User-generated Text (pp. 62-67).
- Generating models of variation
  - Nguyen, D., & Eisenstein, J. (2017). A Kernel Independence Test for Geographical Language Variation. *Computational Linguistics*.

New Questions & Applications

- Personalize machine translation
  - Rabinovich, E., Patel, R. N., Mirkin, S., Specia, L., & Wintner, S. (2017, April). Personalized Machine Translation: Preserving Original Author Traits. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 1074-1084).
- Study linguistic behavior of online communities
  - Shoemark, P., Sur, D., Shrimpton, L., Murray, I., & Goldwater, S. (2017). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (Vol. 1, pp. 1239-1248).
  - Tatman, R., Stewart, L., Paullada, A., & Spiro, E. (2017, August). Non-lexical features encode political affiliation on twitter. In Proceedings of the Second Workshop on NLP and Computational Social Science (pp. 63-67).

Computational Sociolinguistics:

- Improves NLP systems
- Lets us ask new questions

Some challenges

- Interdisciplinary work is held to the (sometimes conflicting) standards of different disciplines
  - Scale of data vs. careful manual analysis
  - Task-based vs. explanatory approaches
- Finding publication venues
  - Solution: Frontiers computational sociolinguistics research topic
  - Rolling submissions!
  - First paper: "Sources of microtemporal clustering in sociolinguistic sequences" by Meredith Tamminga (May 29, 2019)
    - Use of particular sociolinguistic variable (-ing vs -in') depends both on previous token and group of several previous tokens

The emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective.

(Nguyen 2017)

Thanks! Questions?
Slides: rctatman.com/talks