5 mistakes you'll probably make with language data (and how to recover)

Dr. Rachael Tatman (@rctatman)
Senior Developer Advocate, Rasa HQ

1. Transcribed speech != text
2. Not expecting variation
3. Doing too much text cleaning
4. Not using meta-data
5. Anglo-Centrism

Transcribed speech != text

- People do not speak in full sentences (your parser will break!)
- Punctuation will depend on your ASR
- Speech includes disfluencies (uh, um, etc), they are natural and normal **and** contain information
- Token frequency is very different between speech & text

Tokens

Speech

- More pronouns
- More skewed frequency

Text

- More rare words

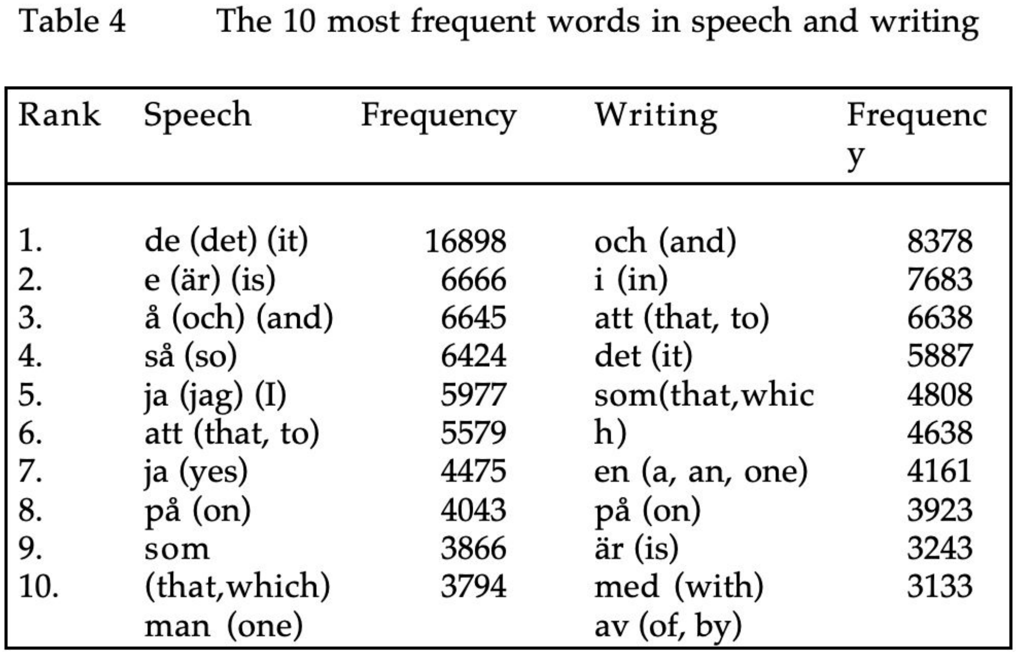


Table 4 The 10 most frequent words in speech and writing

Rank	Speech	Frequency	Writing	Frequency
1.	de (det) (it)	16898	och (and)	8378
2.	e (är) (is)	6666	i (in)	7683
3.	å (och) (and)	6645	att (that, to)	6638
4.	så (so)	6424	det (it)	5887
5.	ja (jag) (I)	5977	som(that,which)	4808
6.	att (that, to)	5579	en (a, an, one)	4638
7.	ja (yes)	4475	på (on)	4161
8.	på (on)	4043	är (is)	3923
9.	som (that,which)	3866	med (with)	3243
10.	man (one)	3794	av (of, by)	3133

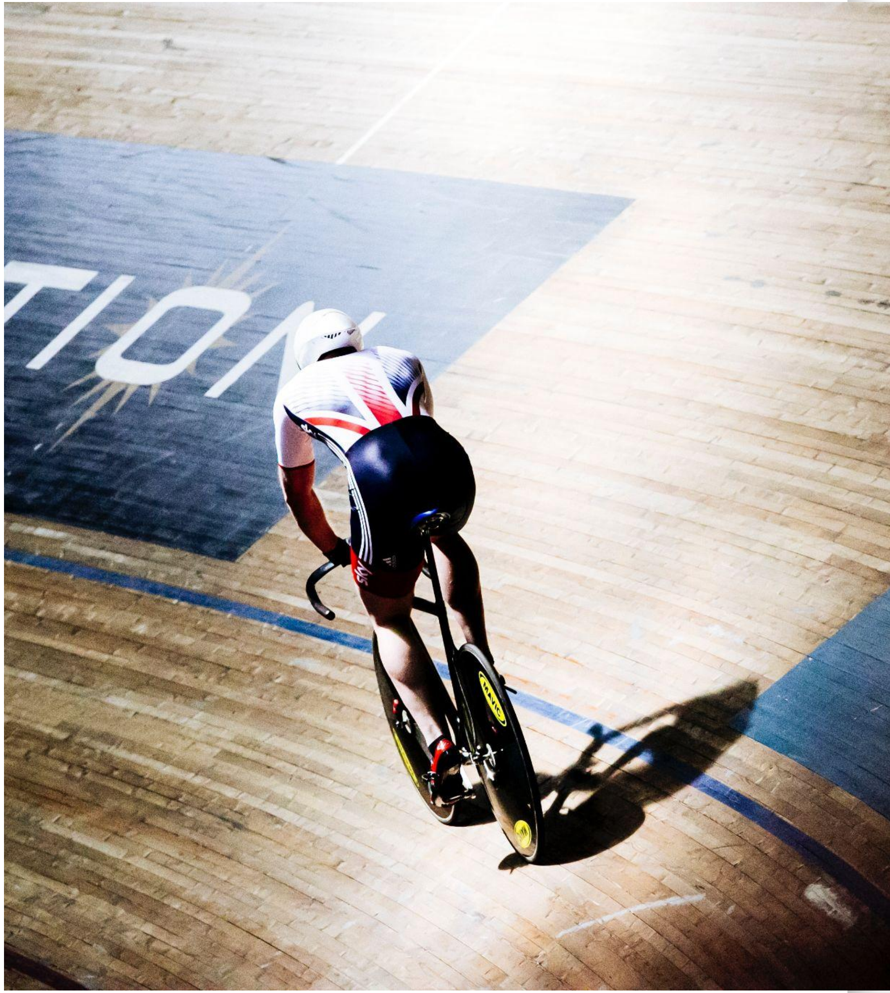
Allwood, J. (1998). Some frequency based differences between spoken and written Swedish. In *Proceedings of the 16th Scandinavian Conference of Linguistics, Turku University, Department of Linguistics* (pp. 18-29).

What should you do?

- Do not try to use a speech model for text without tuning (and vice versa)
- Budget extra time for multimodal work

Underestimating variation

- Social variation
  - who is saying/writing this?
- Genre/register
  - How are the communicating? Formal, informal, professional?
  - What are they talking/writing about? (Velodromes or finance?)
- Models trained on one variety don’t often transfer easily



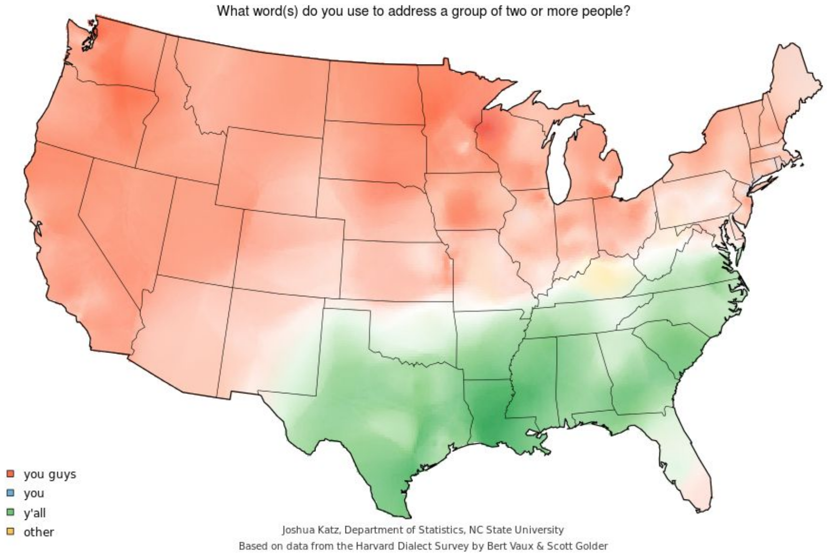
Sociolinguistic variation exists at every level of the grammar

a. Phonetics & sub-lexical features:
   i. cot/caught merger
   ii. :) vs. :-)

b. Lexical variation
   i. firefly vs. lightning bug

c. Syntactic variation
   i. Needs washed
   ii. We stay home anymore

d. Semantic variation
   i. lift, torch, boot



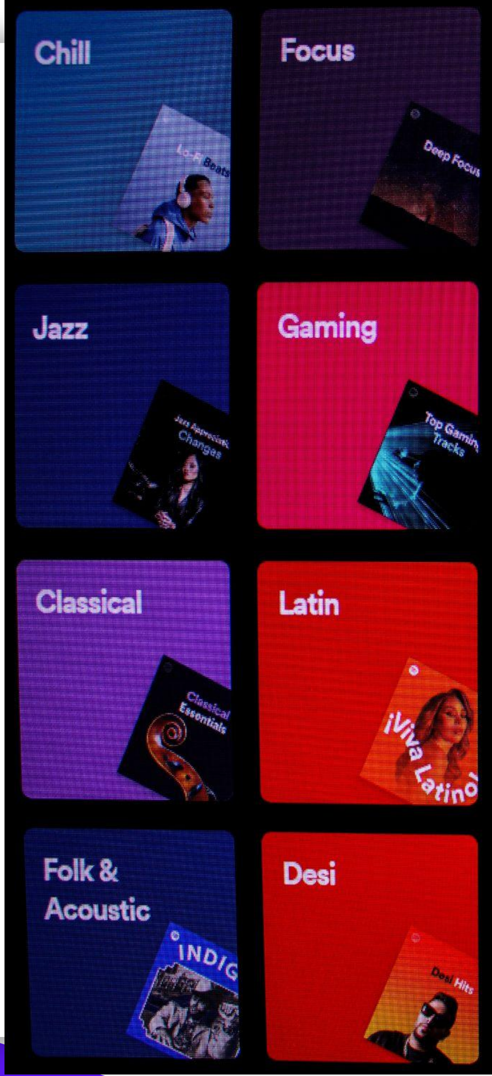
What word(s) do you use to address a group of two or more people?

Joshua Katz, Department of Statistics, NC State University
Based on data from the Harvard Dialect Survey by Bert Vaux & Scott Golder

Genre/register?

Some big ones you might run into:

- “Noisy user-generated text” (Twitter, Reddit, etc.)
- Phone calls/conversations
- Reviews (Yelp, Amazon, etc.)
- The Wall Street Journal (it’s a very popular corpus)
- **The one you’re trying to analyze/use**



What should you do?

- Know about language variation
- Use socially-stratified validation to make sure your model works across groups
- Find data as close to your target a possible

Overdoing the text cleaning
(baby out with the bathwater)

“Cheesecake! 😍” vs “Cheesecake! 🤮”

“i love cheesecake” vs “I LOOOOVE CHEESECAKE!”

Overdoing the text cleaning
(baby out with the bathwater)

- You can and should keep stop words if you’re not using a frequency-based technique (e.g. tf-idf)
- Pre-trained LM’s do *not* remove stop words, if you’re using them you’ll need to keep them in

Overdoing the text cleaning
(baby out with the bathwater)

- keep stop words using frequency-based technique (e.g. tf-idf)
- Pre-trained LM's remove stop words, using need keep

NLTK's English stopwords

What should you do?

- Only do the text cleaning you absolutely need to for the specific methods you’re using
- Work on a copy of the data so you can get to the original

Only looking at the text

- WHO is saying/writing this?
  - Has this customer just called ten times in a row? 😬
- WHEN did they say/write it?
  - Language change is constant!
  - The world & things we talk about change!
- WHERE/HOW did they say/write it?
  - On what platform?
- WHY did they say/write it?
  - Especially important for task-oriented systems

What should you do?

- Make sure you’re looking at that sweet sweet metadata

Anglo-centrism

Not all languages...

- Are written or spoken
- Having big training corpora are available
- Use white space
- User words as the smallest unit of meaning
- Have canonical spellings
- Are used one at a time (code-switching)

Orgs & Resources for Non-English!

- Masakhane (https://www.masakhane.io/)
  - A grassroots NLP community for Africa, by Africans
- NLP en ES (https://nlp-en-es.org/)
  - "NLP en ES 🤗" es la comunidad de hispanohablantes de la iniciativa "Languages at Hugging Face"
- Natural Language Processing group from Universidad de la República (UdelaR) Uruguay 🇺🇾 (https://www.fing.edu.uy/inco/grupos/pln)
  - Resources: https://github.com/pln-fing-udelar/pln-inco-resources
- Search for target language on https://aclanthology.org/ (free peer reviewed papers from all ACL conferences, which are the big NLP research venues!)

What should you do?

- Keep in mind that most people don’t use English
- Hire a language consultant for your target language(s)
- Budget additional time

1. Transcribed speech != text
2. Not expecting variation
3. Doing too much text cleaning
4. Not using meta-data
5. Anglo-Centrism

Some last tips

- It pays to be a little paranoid
- LOOK at your data at every stage
- Do as little as possible to your data
- Modular, flexible pipelines
- Budget for better (i.e. more relevant) data
- Testing, validation & evaluation!
- Don’t reinvent the wheel!

Thanks!

@rctatman