Meaningful levels of analysis in (corpus) linguistics

Jesse Egbert

Northern Arizona University

Introduction

- Corpus linguistics = **computers** for **linguistics**
- Computational linguistics = **linguistics** for **computers**

Online language

“ […] **our predictions may be more prone to failure in the era of Big Data**. As there is an exponential increase in the amount of available information, there is likewise an exponential increase in the number of hypotheses to investigate. […] **there isn’t any more truth in the world than there was before the internet** or the printing press. **Most of the data is just noise**, as most of the universe is filled with empty space.”

Nate Silver

Language sample (corpus)

Texts

The text

- Definition: a written or spoken unit of discourse that is:
  - Naturally occurring
  - Recognizably self-contained
  - Functional
- The text is the ideal unit of observation for corpus linguistic research.
  1. Fundamental unit of discourse
  2. Important social construct
  3. Situational and linguistic integrity

Egbert, forthcoming; Biber & Conrad (2009)

Language sample (corpus)

Texts

Linguistic characteristics

Levels of analysis

- Levels of analysis *within* texts (i.e. leaves)
  - Discourse
  - Syntax
  - Lexico-grammar
  - Phraseology
  - Lexis
  - Morphology
  - Phonology

Language sample (corpus)

What meaningful levels of analysis exist between the text and the corpus?

Texts

Linguistic characteristics

Levels of analysis

- Levels of analysis *containing* texts (i.e. species)
  - Defined by *user*
    - Geographic region
    - Socioeconomic status
    - Gender
    - Age
    - Race
  - Defined by *use*
    - Register

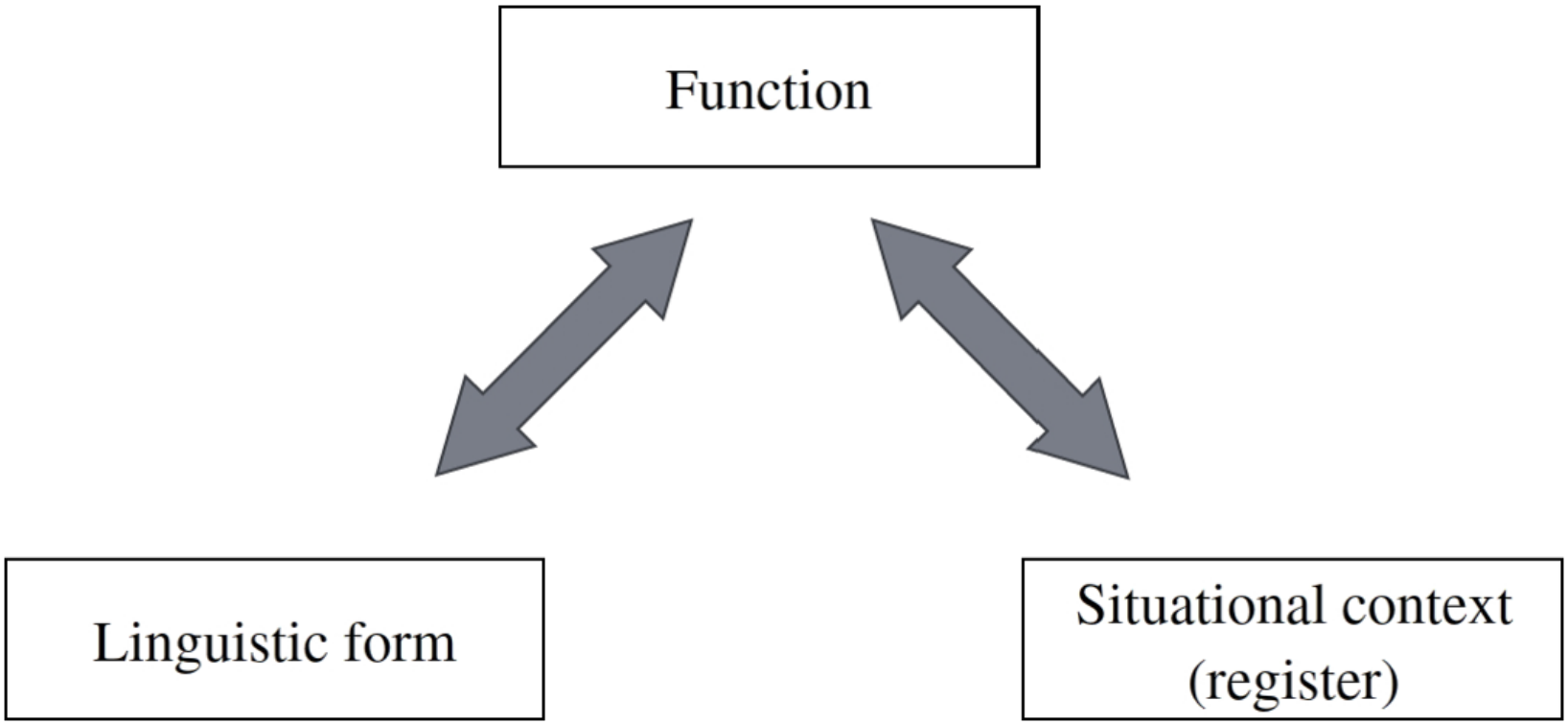
Register

- Definition: Varieties of language defined by their situation of use (Biber & Conrad, 2009)

- Functional link between situation and language (Egbert & Biber, 2017)
- Valid social construct (Egbert, Biber & Davies, 2015)
- Strong(est?) predictor of linguistic variation (Biber, 2012)

Register—functionally interpretable

- Functional link between situation and language



Adapted from Biber & Conrad (2009)

Register and probability

"Register variation can in fact be defined as systematic variation in probabilities"

Halliday (1991)

- Language varies across registers *at every linguistic level*
- Probabilities based on "general" language are inaccurate

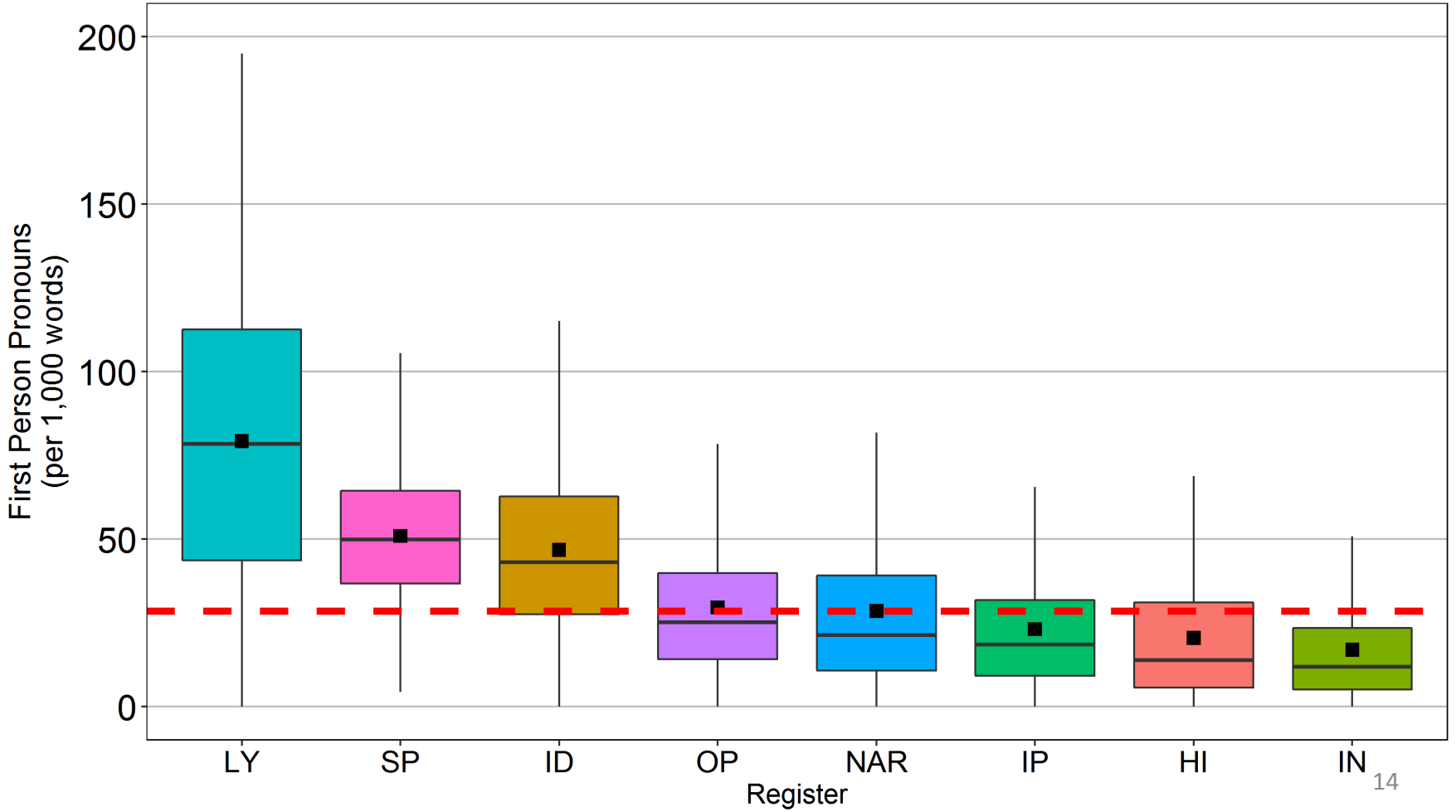
CORE: Corpus of Online Registers of English

- Large corpus of English web documents
  - ~50,000 documents
  - ~50 million words
- Random sample from the searchable web
- Situational characteristics coded by non-experts
  - 8 register categories
    - At least 3-way agreement: 69.2%
  - 33 sub-register categories
    - At least 3-way agreement: 51.4%

Biber & Egbert (in press)

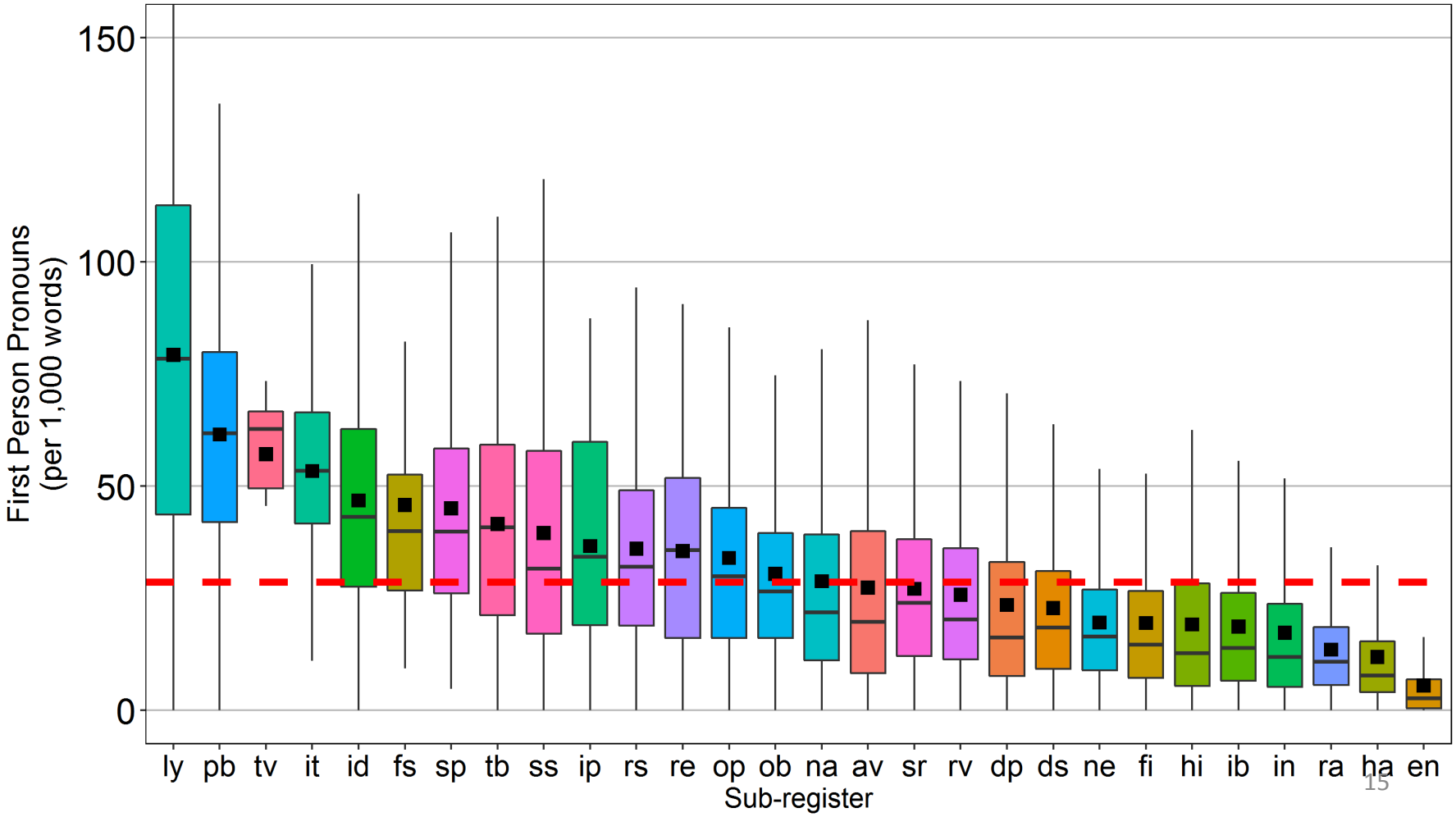
Register—strong predictor of variation

- 1st person pronouns across **registers**



Register—strong predictor of variation

- 1st person pronouns across **sub-registers**

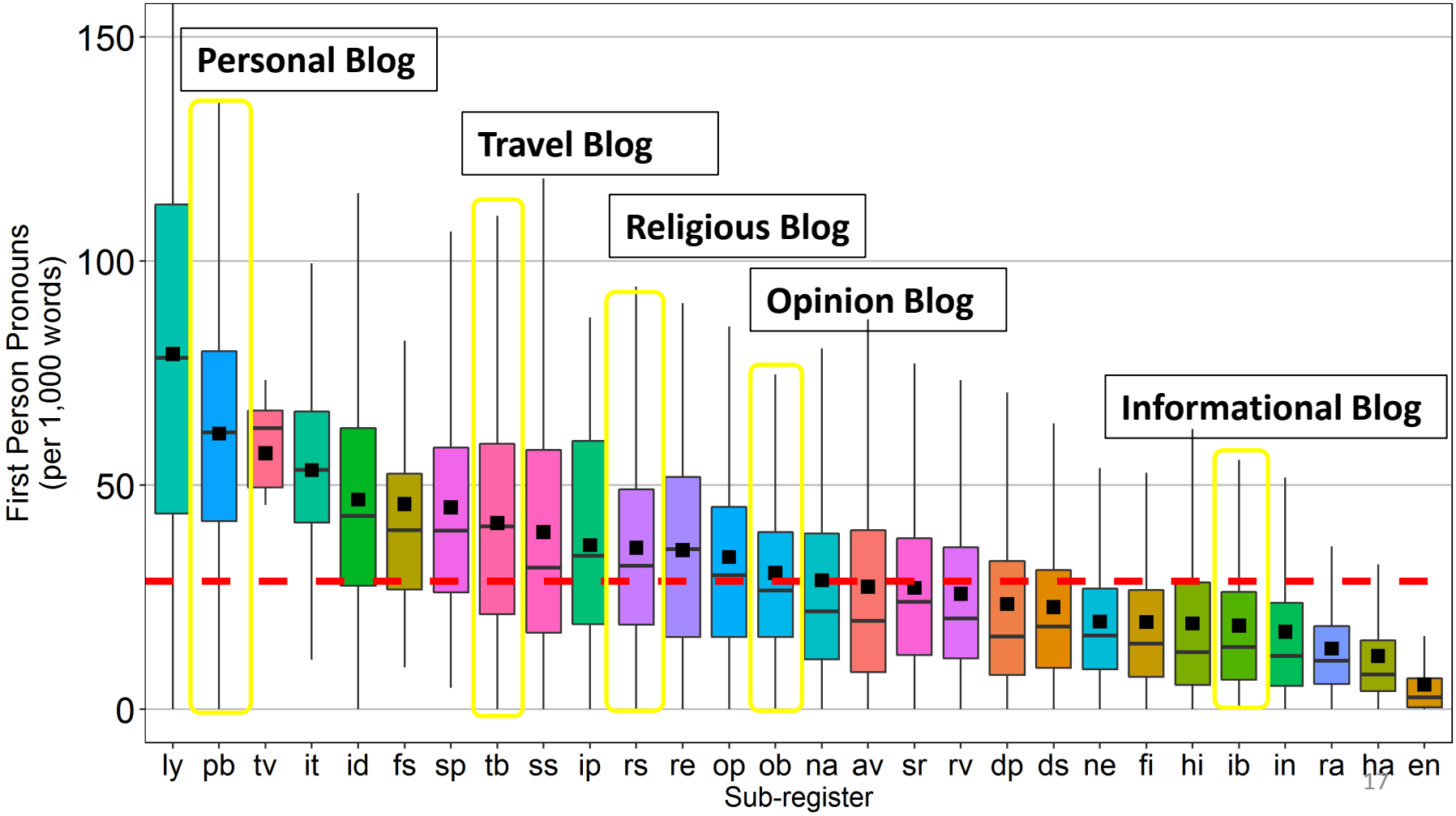


Blog registers

	Personal Blog	Travel Blog	Religious Blog	Opinion Blog	Info. Blog
Purpose	Narrative	Narrative/ Description	Opinion	Opinion	Description
Subject	Author's life	Travel	Religion	Author's stance	Topic to be explained
Audience	Friends/ Family/ Followers	Travelers	Religious adherents	Various	Students/ Non-experts

Register—strong predictor of variation

- 1st person pronouns across **sub-registers**



Register and probability

"Register variation can in fact be defined as systematic variation in probabilities"

Halliday (1991)

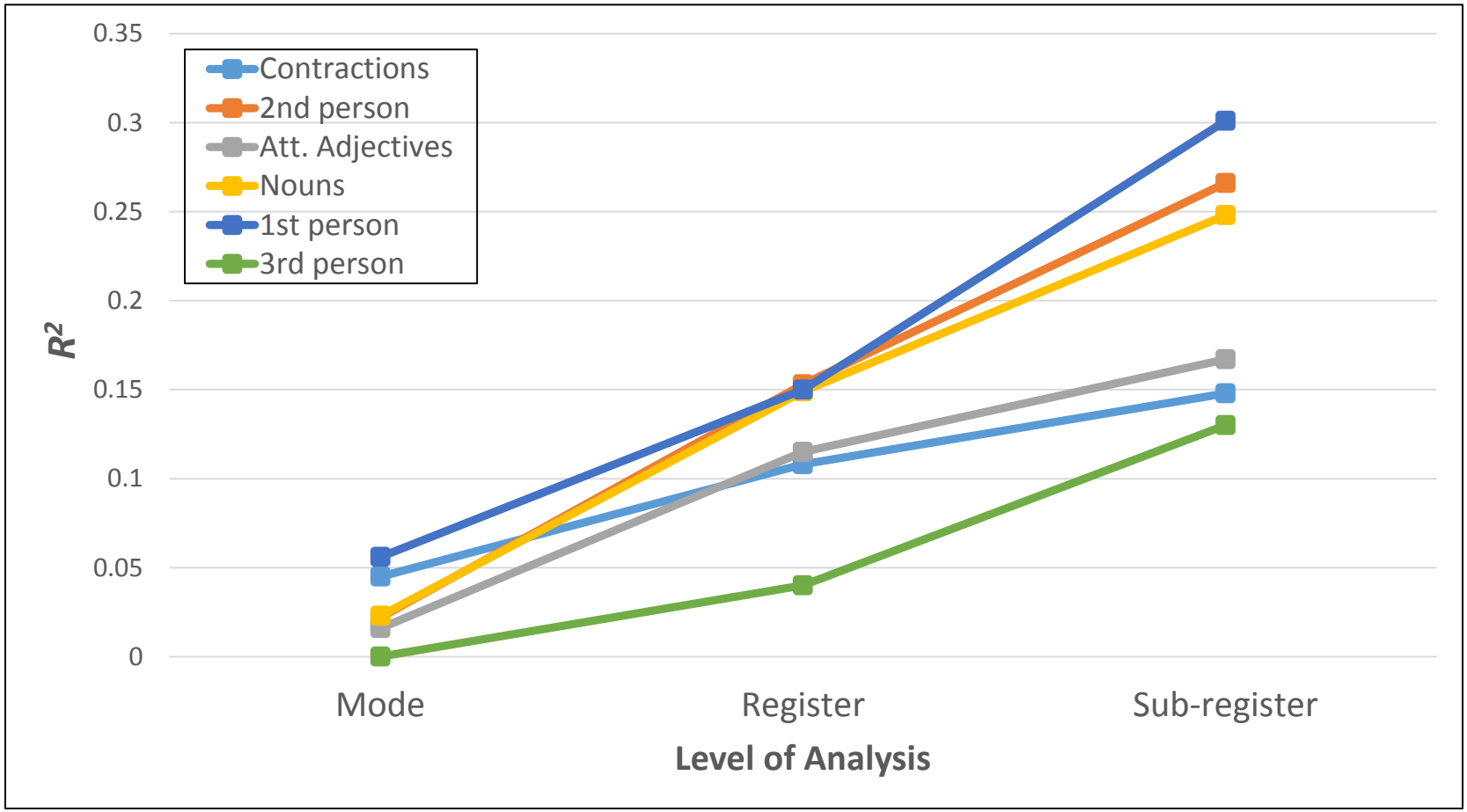
- Language varies across registers *at every linguistic level*
- Probabilities based on "general" language are inaccurate
- Baseline probabilities should be conditioned on register:

$$P(\text{FREQ}_{\text{TEXT}} \mid \text{FREQ}_{\text{REGISTER}})$$

Determining the ideal level of analysis

- Which level accounts for the most variance?
  - Mode (spoken v. written)
  - Register (8 levels)
  - Sub-register (33 levels)
- Six linguistic variables
  - Contractions
  - 1st person pronouns
  - 2nd person pronouns
  - 3rd person pronouns
  - Nouns
  - Attributive adjectives
- Coefficient of determination (R^2)

Determining the ideal level of analysis



Analyzing multiple levels of analysis

- Multi-Dimensional analysis (Biber, 1988)
- Cluster analysis (Biber & Egbert, in press)
- Factorial designs (Egbert, 2014)
- Hierarchical mixed effects models (Gries, 2015)
- Machine learning (Argamon, Koppel & Pennebaker, 2007)

Take away messages

- The text is the ideal unit of observation
- (Online) language is noisy; register can provide signal
- Accuracy improves when linguistic probabilities are conditioned on register
- Statistical methods can:
  - help identify the ideal level of analysis
  - simultaneously account for multiple levels of analysis
- Keep the “linguistics” in computational linguistics!

References

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday, 12(9).

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8: 9-37.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1): 7–34.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D. & Egbert, J. (in press). *Register variation online*. Cambridge: Cambridge University Press.

Egbert, J. (2015). Sub-register and discipline variation in published academic writing: Investigating statistical interaction in corpus data. *International Journal of Corpus Linguistics*. 20(1): 1-29.

Egbert, J. & Biber, D. (2017). Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis. *Corpus Linguistics and Linguistic Theory*.

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*. 66(9): 1817-1831

Egbert, J. & Schnur, E. (forthcoming). Missing the trees for the forest: The role of the text in corpus and discourse analysis. In Anna Marchi and Charlotte Taylor (Eds.*), Corpus Approaches to Discourse: A Critical Review*, New York: Routledge.

Gries, S. Th. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora*, 10(1), 95-125.

Halliday, M. (1991). *Corpus studies and probabilistic grammar*. English Corpus Linguistics: Studies in Honour of Jan Svartvik.

Thank you

Jesse Egbert

Jesse.Egbert@nau.edu

http://oak.ucc.nau.edu/jae89

This research was funded in part by the National Science Foundation (Grant No. 1147581)