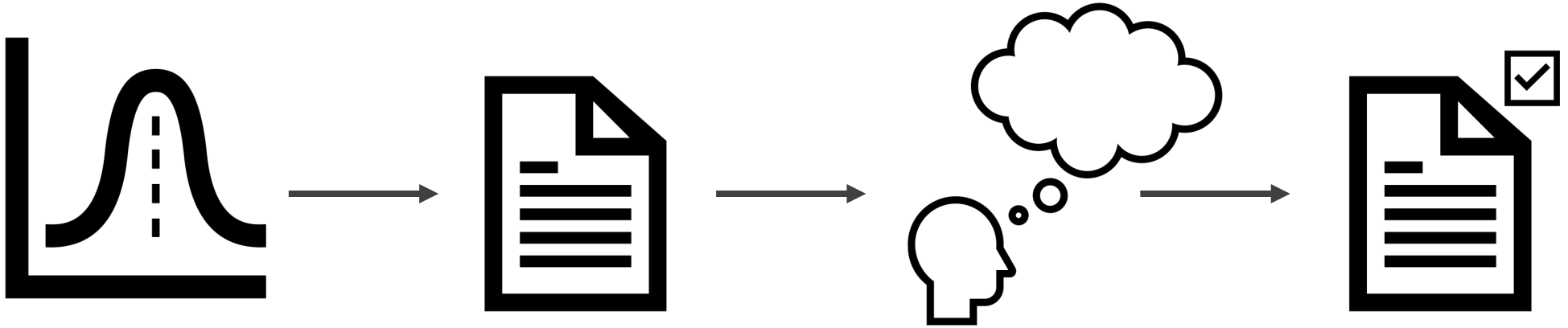
Carnegie Mellon University

Finding a statistical voice: Exploring rhetorical structure of student writing from introductory to advanced course levels

Meg Ellingwood
Department of Statistics & Data Science

(Thinking about) statistical writing is important



Assessing Writing 60 (2024) 100830

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Visualizing formative feedback in statistics writing: An exploratory study of student motivation using DocuScope Write & Audit

Michael Laudenbach *, David West Brown, Zhiyu Guo, Suguru Ishizaki, Alex Reinhart, Gordon Weinberg

Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15219, USA

Student-Written Reports from Two Statistics Courses

Total of 60 reports

- 30 each from an Introductory and an Advanced course

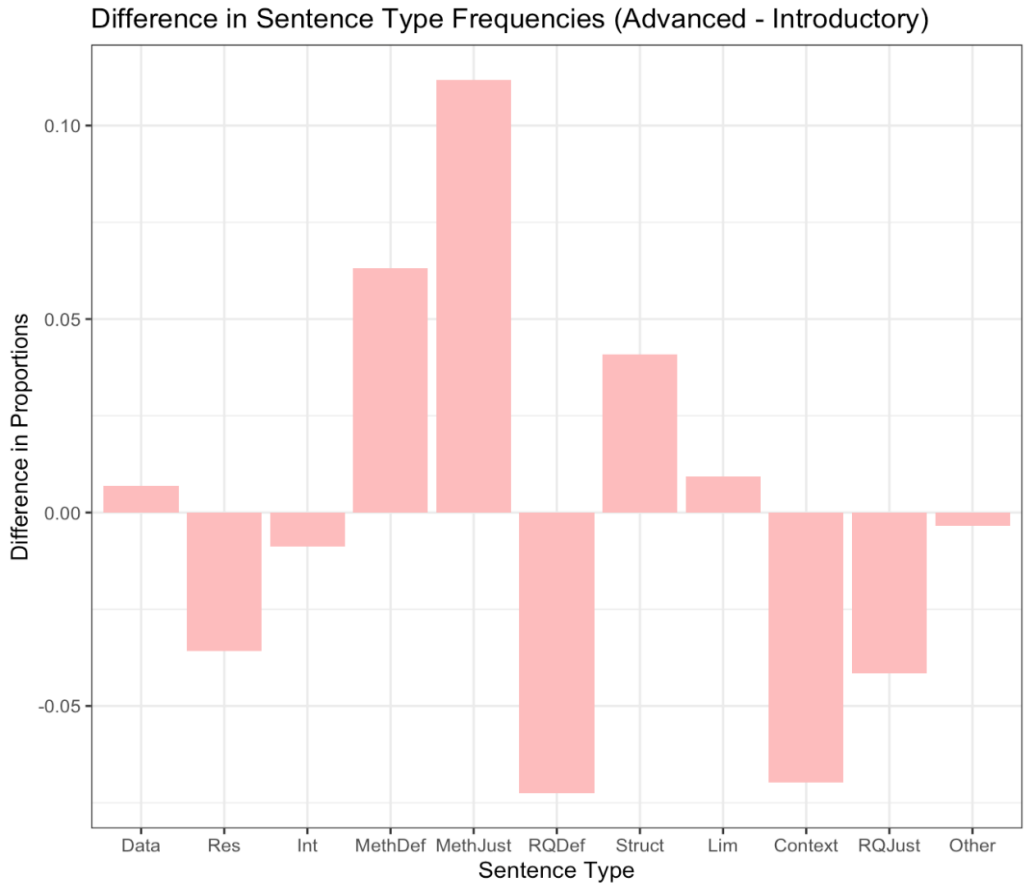
Tokenized by sentence

- 2007 sentences from Introductory course
- 2810 sentences from Advanced course

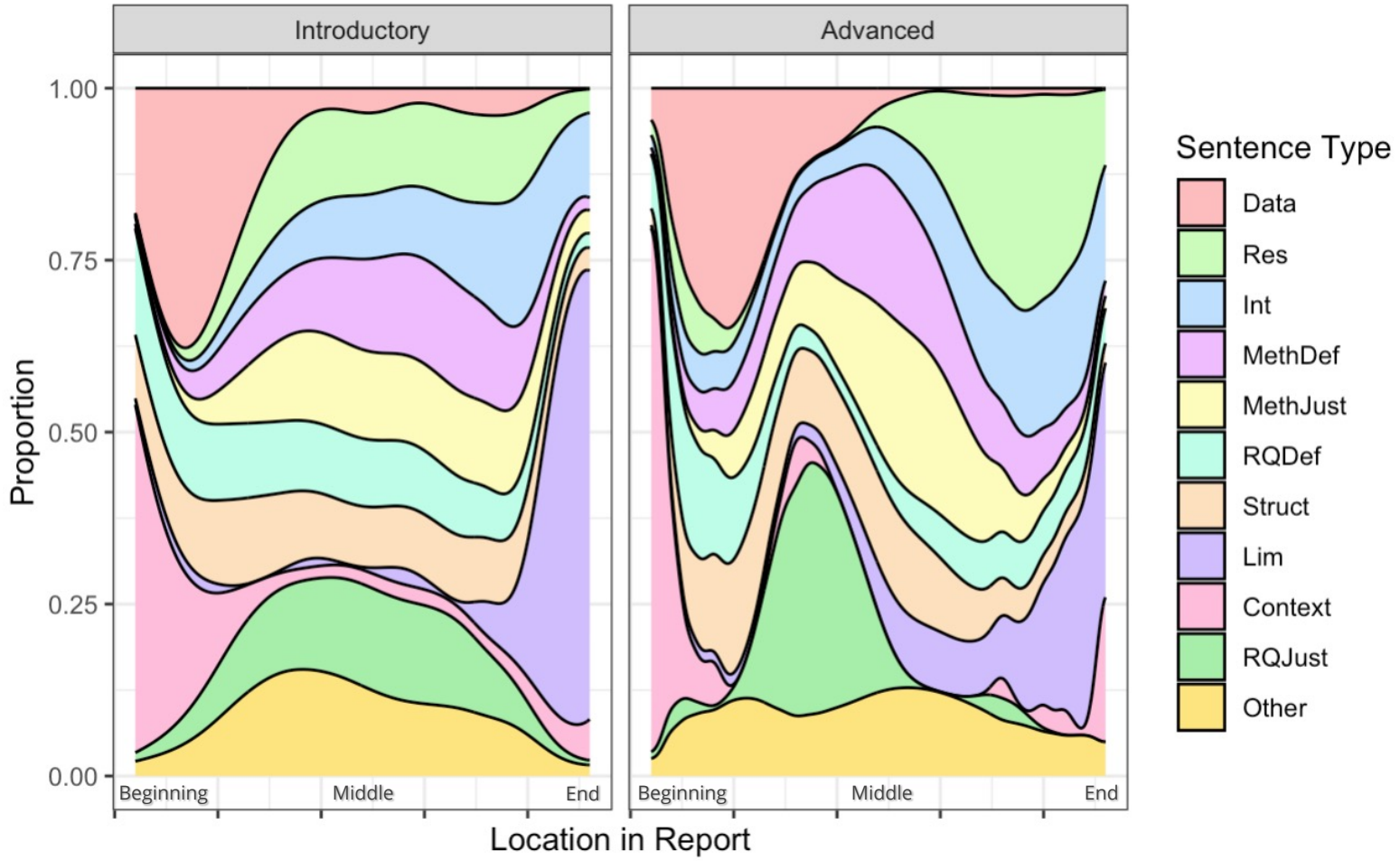
Rhetorical function: Sentences serve different purposes

Context	"The social media platform YouTube has grown significantly since its creation in 2005."	Results	"As the p-value is less than 0.05 we reject our null hypothesis and conclude there is a significant effect."
Data	"Other variables covered in the research questions include month initial spread index ISI and outside relative humidity RH."	Interpretation	"However we did not find enough evidence to determine that position affects the amount of time on the field that a player is given."
RQ Definition	"Our next research question is to look into whether mean total passes of players is associated with the position of the player."	Methods Def.	"We found that the outcome death can be expressed as rate data as the number of counts per day and is appropriate for Poisson regression removing the need for the predictor variable time."
RQ Justification	"While all of the days have different minimums and maximums their median and quartiles are all about the same suggesting that there may be no relationship between the two variables."	Methods Just.	"Furthermore several of our predictor variables like Metabolicbymasslog and Bodymassglog also appear much more normal as a result of the transformations."
Limitations	"A larger dataset with more observations and more features will help us improve our analysis."	Other	"We can form our formal hypothesis from this idea as well as choose a test."
Structural	"The following figure tabulates the varying levels of drug use."		Note: We are actively developing these labels and working on a classifier to tag sentences at scale

Students change writing structure over their statistics education



Sentence type distribution over the length of the report differs



Transitions that violate expectations lead to high surprisal

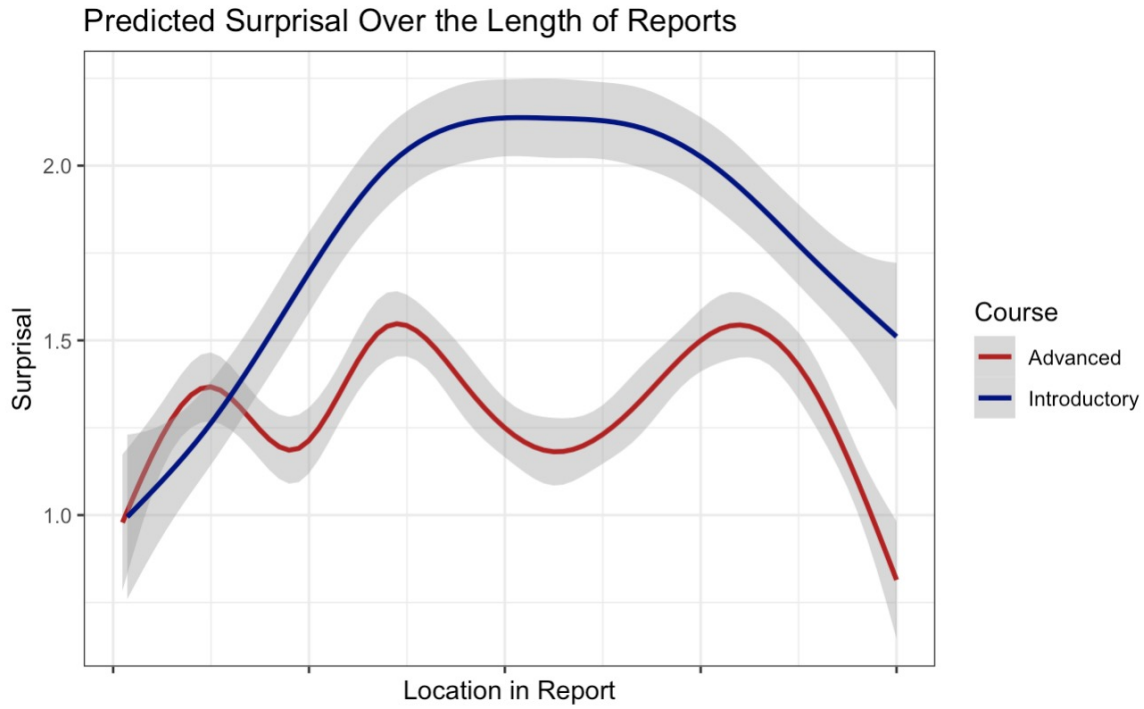
Built model to predict sentence tag based on previous tags, sBERT embeddings

$$\text{Surprisal} = -\log P(\text{tag} \mid \text{context})$$

Less common transition → more surprising

Introductory students: more surprising in general, but especially in main body of report

Advanced students: less surprising overall, especially at end of report



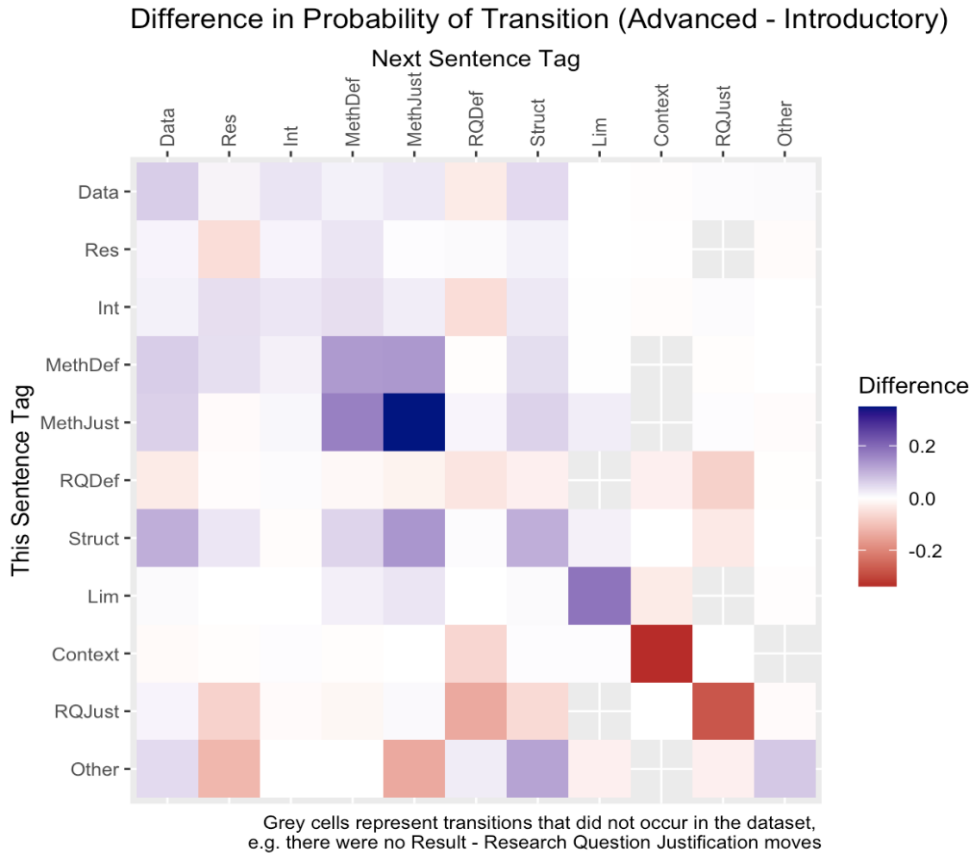
Student writing corpus has potential for future work

Modeling: predict written reports and compare to student versions

Feedback: help students develop understanding of reader expectations

Distributions: multiple “right ways” to construct a statistical report

Classification: tag sentences at scale



References and Related Works

- Michael Laudenbach, David West Brown, Zhiyu Guo, Suguru Ishizaki, Alex Reinhart, Gordon Weinberg. Visualizing formative feedback in statistics writing: An exploratory study of student motivation using DocuScope Write & Audit. *Assessing Writing*, Volume 60, 2024, 100830, ISSN 1075-2935, https://doi.org/10.1016/j.asw.2024.100830. (https://www.sciencedirect.com/science/article/pii/S1075293524000230)
- Clara Ye. Analyzing sentence-level rhetoric of writing in statistics. 2023

Advised by Alex Reinhart & Ronald Yurko

Carnegie Mellon University