Magic Vs Evolution*

Or more specifically language in Harry Potter vs Darwin's Natural Selection

Shreya Sakura Noskor

April 17, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: HarryDarwin

Training machine learning models with language. the purpose of this text is not to discover a new thing but to helpl people understand the way machine learning models use these informations to train them.

1 Introduction

Research Question: Does the topic we talk about influence our language, or does the laguage influence the theme of the topic.

2 Data

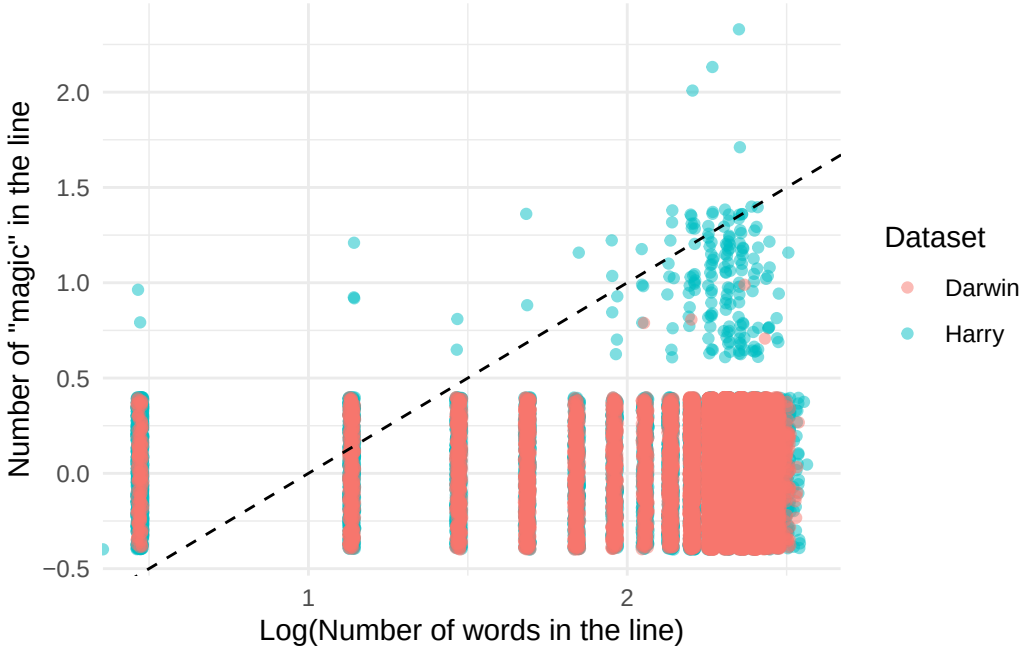
2.1 Source

The data utilized was from Project Gutenburg and PDF Drives and with the help of R (R Core Team 2023) we were able to create this paper. Also code for making the models were made referencing Telling Stories by Rohan Alexander(Wickham et al. 2019a). Other R packages were used to clean, process and model the data such as, Wickham et al. (2019b), Johnston and Robinson (2023), Goodrich et al. (2022), Ooms (2023), Wickham et al. (2023), Richardson et al. (2024), Wickham (2023), Arel-Bundock (2024).

2.2 Variables

2.2.1 Distribution of each word with each text

Table 1: **?(caption)**



3 Model

3.1 Natural and the 2 texts

Define y_i is the number of times "natural" appeared in each text and the explanatory variable is the number of words in the line. This means that we have 4 models in total with the y_i's being, number of times the word "natural" and "species" showed up in The book of Evolution or Harry Potter and the prisoner of azkaban. Also the number of times the word "magic", "wizard" and "miracle" showed up in The book of Evolution or Harry Potter and the prisoner of azkaban. We predict to see that there is a positive correlation in Darwin's text but not in Harry Potter due to the difference in topics.

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{1}$$
$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{Number of Words}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\tag{5}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

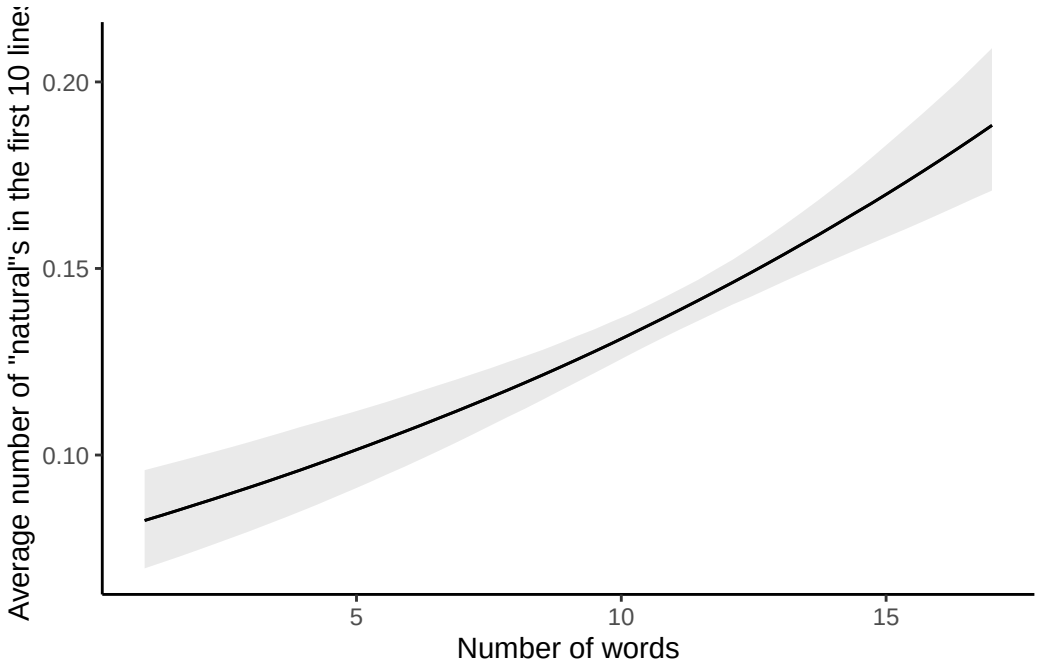


Figure 1: Predicted number of "nature"-realted words in each line based on number of words in Darwin's Book

positive correlation it seems

No correlation at all

3.1.1 Model justification

We predict to see that there is a positive correlation in Darwin's text but not in Harry Potter due to the difference in topics.

3.2 "Magic" and the 2 texts

Define y_i is the number of times "magic" or "miracle" appeared in the text and the explanatory variable is the number of words in the line.

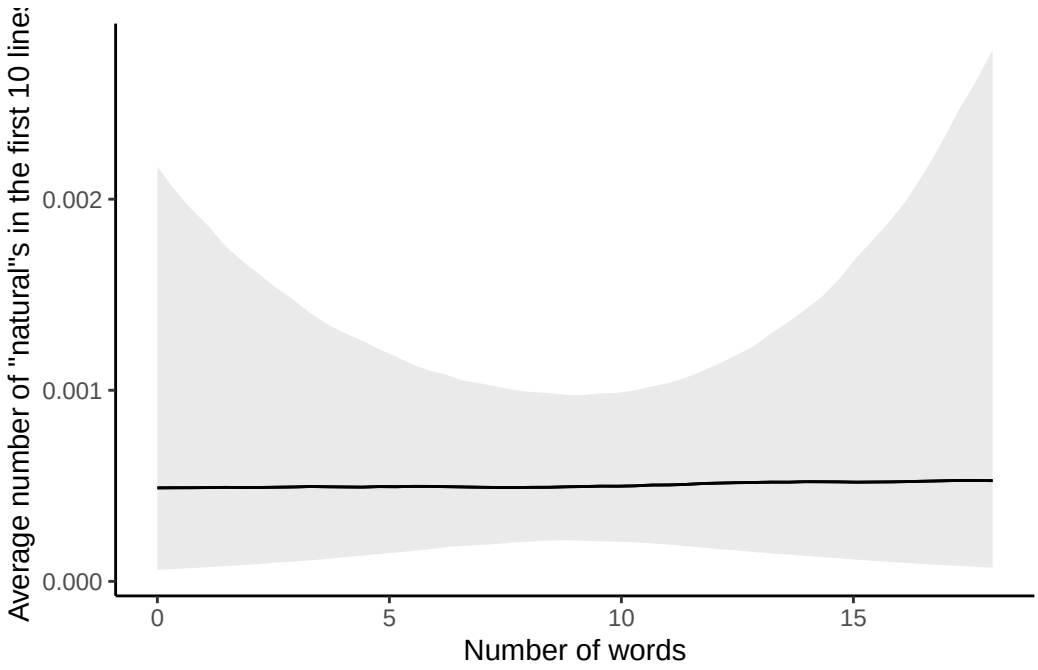


Figure 2: Predicted number of "nature"-realted words in each line based on number of words in Harry Potter

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i) \tag{6}$$
$$\log(\lambda_i) = \beta_0 + \beta_1 \times \text{Number of Words}_i \tag{7}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{8}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{9}$$
$$\tag{10}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

As we can see here there is almost no correlation at all.

3.2.1 Model justification

We predict to see that there is a positive correlation in Harry Potter but not in Darwin's text due to the difference in topics.

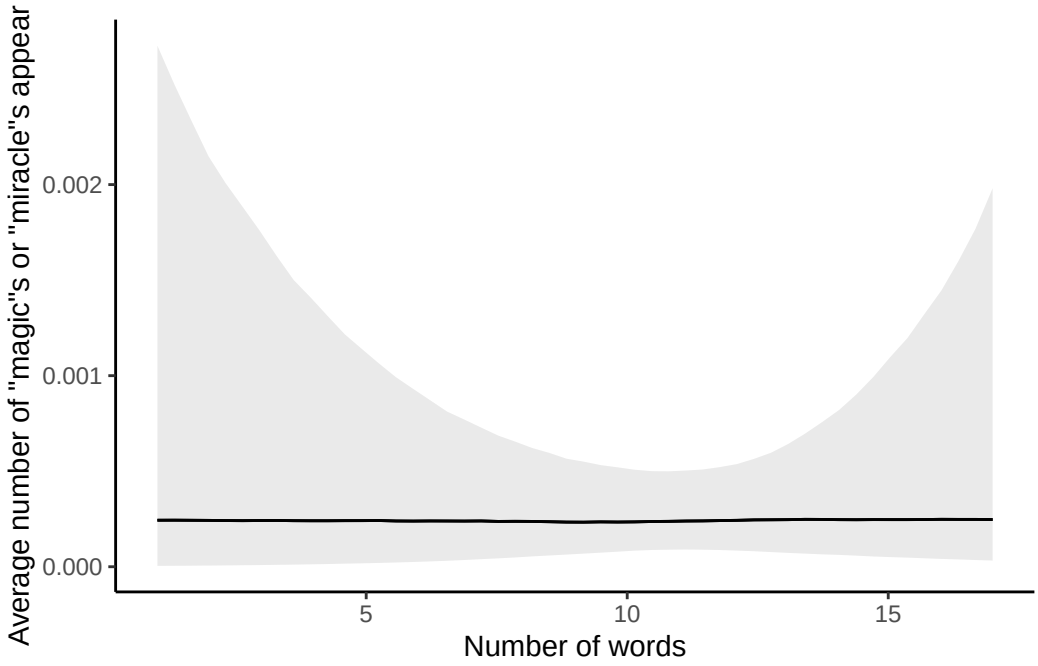


Figure 3: Predicted number of “magic”-realted words in each line based on number of words in Darwin’s Book

4 Results

Our results are summarized below.

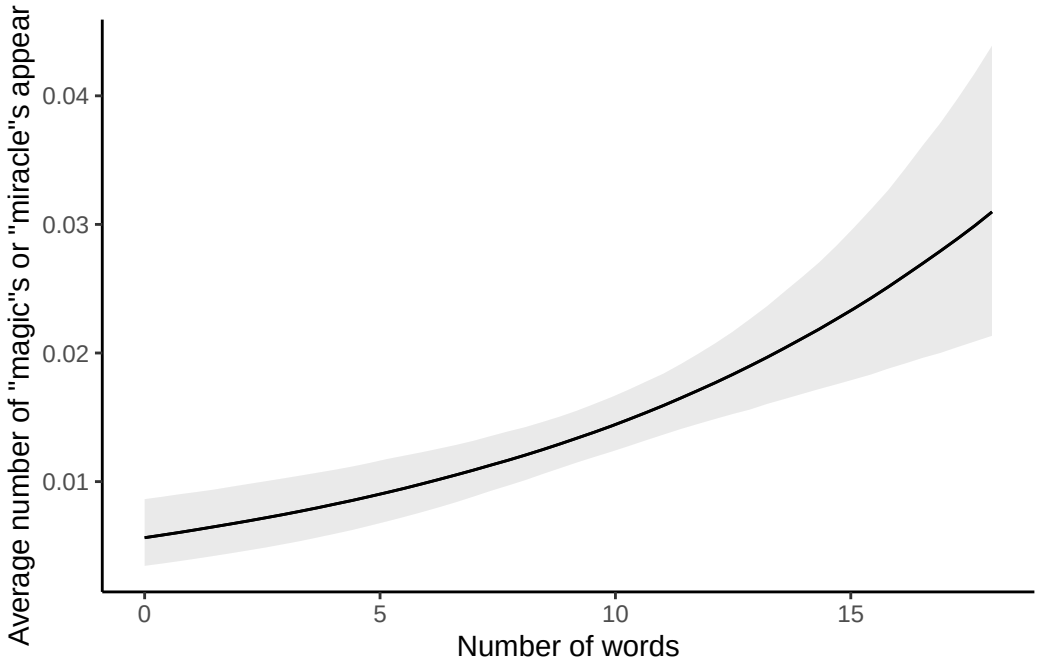


Figure 4: Predicted number of “magic”-realted words in each line based on number of words in Harry Potter

Table 2: Model Summary showcasing the correlation coeffcient for Darwin’s evolution book and the word nature

	darwin and nature
(Intercept)	−2.55
	(0.09)
word_count	0.05
	(0.01)
Num.Obs.	19 411
Log.Lik.	−8095.192
ELPD	−8096.9
ELPD s.e.	106.3
LOOIC	16 193.7
LOOIC s.e.	212.7
WAIC	16 193.7
RMSE	0.37

Table 3: Model Summary showcasing the correlation coeffcient for Darwin's evolution book and the word magic

	darwin and magic
(Intercept)	−8.31
	(1.78)
word_count	0.00
	(0.16)
Num.Obs.	19 411
Log.Lik.	−38.004
ELPD	−39.7
ELPD s.e.	17.2
LOOIC	79.4
LOOIC s.e.	34.4
WAIC	79.4
RMSE	0.01

Table 4: Model Summary showcasing the correlation coeffcient for Harry Potter and The Prisoner of Azkaban and the word nature

	harry and nature
(Intercept)	−7.62
	(0.93)
word_count	0.00
	(0.10)
Num.Obs.	12 767
Log.Lik.	−52.022
ELPD	−53.6
ELPD s.e.	19.0
LOOIC	107.2
LOOIC s.e.	38.0
WAIC	107.2
RMSE	0.02

Table 5: Model Summary showcasing the correlation coeffcient for Harry Potter and The Prisoner of Azkaban and the word magic

	harry and magic
(Intercept)	−5.18
	(0.22)
word_count	0.09
	(0.02)
Num.Obs.	12 767
Log.Lik.	−938.888
ELPD	−940.6
ELPD s.e.	57.8
LOOIC	1881.1
LOOIC s.e.	115.7
WAIC	1881.1
RMSE	0.12

5 Discussion

5.1 Why is the result the way it is

5.2 Importance

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

References

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Ooms, Jeroen. 2023. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.