Statistical Trends in Senior Independent Study Theses

Patrick May

December 5, 2023

1 Introduction

As students and faculty at the College of Wooster (and as a current senior myself), everyone comprehends the weight that exists behind the concept of our "Independent Studies" – involved, individual, senior capstone projects that result in a (typically) 50 to 200 page long thesis paper. In the last year or so, I discovered many independent studies (ISes) are indexed and *downloadable* online in their full PDF form [7]. Hence, the data for an analysis and model building about Independent Studies is possible, but first the multiple gigabytes of IS rawtexts must be flattened into a manageable dataset. There are many reasons why this data transmogrification and analysis is worthwhile. As a senior myself, takeaways about overall writing structure may be useful in improving the writing quality of my independent study. For novelty's sake, as all Wooster students share (coerced) interest in IS, discovering broad trends about it is *interesting*. Finally, the data collection and transformation pipeline is an involved exercise in software, but once complete, the *dataset* persists and can be used by aspiring statisticians without the need for understanding of lexical analysis techniques, natural language processing, webscraping, etc.

Given a moderate corpus of academic writing, a few possibilities exist for discerning rank, order, and relationship between writings. A common research method for academic writing is that of determining *impact* of a piece of research. Impact analysis typically comes through traversing the citation network of papers, where each subsequent paper that cites an original has a directed edge connection [9]. For independent studies, this type of analysis is unfortunately limited by the publication platform and amorphous goals of IS. Independent studies are not terse conference papers that are easy to cite – they are long undergraduate theses that have a much larger emphasis on researcher learning. This is seen in how ISes are "published" online, through Wooster's OpenWorks. Unlike academic journal pages, OpenWorks does not automatically link prior resources that each IS references. Hence social impact of an IS would be an incredibly difficult metric to calculate without manual citation extraction, which sits well outside the scope of this project.

Another textual analysis option is that of *lexical* analysis. For Independent studies, this means stepping back from the academic research label and considering these publications as large collections of natural language. *Lexical* analysis comes from converting a block of text into tokens. Each token is either a single word or a part of a word, and has various information associated with it, such as its part of speech, a complexity value, etc. Then, using the tokenized text, one computes various metrics about the underlying IS [5]. Further analysis comes from looking at how tokens fit together to construct sentences [5, 8]. For example, processing what category a token is based on the context of the surrounding sentence (Part-Of-Speech/POS) [5]. Given a body of text, metrics

concerning complexity, variability, density, diversity, and rarity can be computed using various lexical and semantic techniques. The specific methods behind the computation are outside the scope of this work.

Lexical analysis is a rapidly evolving field, with new analysis methods appearing frequently. There is substantially more depth analytical lexical statistics in terms of how they are calculated and various considerations that are included into their result. Those specifics are well outside the bounds of a undergraduate semester project, so background information is minimal such that it simplifies computation significantly to provide a more accessible, mostly accurate survey of these various lexical analysis categories (this work is concerned with **what** these statistics mean, and not **how** one gets there). Text **complexity** arises from the idea of how "difficult" a word is. This is typically done by comparing a word or token to a pre-computed dictionary that has associated each token with a difficulty value. Alternative methods exist to account for difficulty for speakers of alternative languages or difficulty of a word based on the surrounding sentence [5]. **Rarity** of text comes from comparing the words in the text under analysis with a broad sample of the overall language [8]. Rarity has different forms, typically either based on if a word is rarely used in the entire base language, or rarely used in a specialized subset/"genre". For this work, rarity is concerned with comparison to the entire English language. **Variability** is a lexical statistic interested in the size of the vocabulary used in the text. A text with more repeated words has less variability than a text with diverse word selection. Finally, **density** quantifies content word usage. Essentially, lexical density is a calculation for how many fluff words are used compared to content words.

With these more advanced lexical statistics, this opens the IS dataset to a large amount of exploration. The first goal of this work is to explore this new, novel dataset. How do Independent Studies compare over time? Do different majors have different writing styles? A small subset of the data was listed as "exemplar" ISes (highest honors) on OpenWorks. The second goal of this work is specifically concerned with the split between exemplar and non-exemplar ISes. Can a meaningful statistical model be built to predict exemplary status of an Independent Study? This study explores a novel independent studies theses dataset, by looking at categories of comparison (*primary major* and *publication date*) along with a large amount of lexical observations computed for each thesis.

2 Methodology

2.1 Data Collection

Wooster has an *Openworks* website that serves as an online repository for many pieces of writing. One subset of this are independent studies, both *exemplar* and *normal* ISes. Every IS that has an entry on the website has an associated metadata page, with information associated with the paper, such as title, abstract, author, advisors, departments, publication year, etc. 11,600 independent studies are listed, however only $\sim$ **5,600** ISes were downloadable in full text, due to embargos on

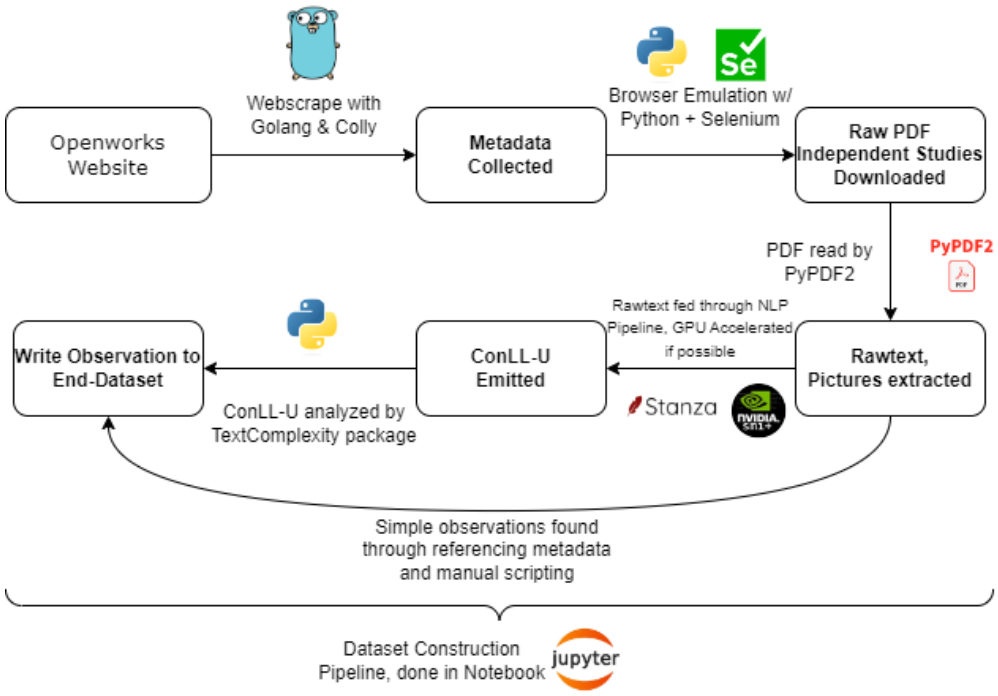

Figure 1: Dataset Collection and Transmogrification Pipeline

every IS earlier than 2006.

As seen in figure 1, a few tools were utilized in condensing hundred-page research theses into lexical analysis statistics. An existing PDF reader library was used to parse text from the IS PDFs. From there, a large natural language processing pipeline was used to convert the extracted text into a known lexical analysis format, `.conllu`. Finally, the processed `CoNLLu`'s were fed through a python library called `textcomplexity` that outputted many observations. These observations were then aliased into more usuable dataframe names and written to an end file. For further information about the data collection process, refer to the collection postmortem [3].

2.2 Data Cleaning

Of the total possible **5,544** independent studies fully collected, a few presented issues that meant they were dropped from the final data used for exploration and model building. **221** values were removed because they IS text was unable to be processed by the lexical analysis tool. **1** IS was dropped due to improper titling and parsing issues, which was caused by the author listing themselves as a septuple major student. **34** ISes were dropped due to being impossible outliers that arose due to text extraction errors. For example, ISes that were analyzed to be less than 100 words long, or have many tens of thousands of single-character words. After removing obvious outliers, there are **5,289** individual observations, **330** of which are *exemplar* ISes, with the remaining being *normal*. Refer to the appendix for exact processes used in data filtering. An overview of key variables can be seen in Table 1. For a breakdown of more/all observations, refer to the appendix.

Variable	Description	Type	Mean (μ)	St. Dev (σ)
publication date	year IS was published	Numerical*	-	-
isExemplar	(0) - not exemplar (1) - exemplar	Categorical	-	-
department1	first department listed on IS	Categorical	-	-
department2	second department listed on IS	Categorical	-	-
lexDensity	fluff to content words	Numerical	0.419	0.071
lexRarity	uncommon-ness of word choice	Numerical	0.325	0.089
lexVariability	measure of token reuse	Numerical	0.017	0.019
lexDispersion	type of words throughout text	Numerical	0.642	0.027
lexEvenness	how tokens of diff types appear	Numerical	0.871	0.019

Table 1: Abbreviated Variables of Interest in Theses Dataset [4]
*Treated as a categorical variable in some instances

2.3 Analytical Methods

For simple exploration of IS data, barplots for categorical variables were utilized. One observation is *publication date*, which can be treated as a temporal variable over which trends in numerical variables can be observed. Boxplots were also utilized for comparing categorical dependent variables, such as the *department/major* of each student to the lexical values observed in their work. Logistic regression was utilized to model outcomes of *exemplar* status within Independent Studies.

When attempting to model a multiple outcome categorical variable, **multinomial logistic** regression was attempted. Multinomial logistic regression is used when the dependent variable has multiple (> 2) categories. Adjacent to building a model with multiple, discrete, non-ordered outcomes is wanting to test if a numerical variable (any of the lexical observations) has an effect on a categorical variable (such as *publication date, isexemplar, major,* etc. This was done using a **Kruskal-Wallis H test**, which tests if different samples are independent. An alternative would be to use a simple **One-Way Analysis of Variance (ANOVA)** test, however some lexical statistics did not follow a normal distribution, so a non-parametric test was utilized instead [1].

3 Results

3.1 Exploration

The Theses data can be divided in multiple ways, although the most prevalent are through *exemplar* status, *department* of the IS, or **publication date**. Of the 5,300 individual theses, the most prevalent primary department listings were **History** (514), **Psychology** (449), **Political Science** (443), **Sociology & Anthropology** (420), **English** (357), **Communication Studies** (354), **Biology** (219), and **Mathematics** (213). The least prevalent primary majors were **Statistical & Data Sciences** (13), **East Asian Studies** (9), **Russian Studies** (8), **Comparative Literature** (7), **Global Media & Digial Studies** (4), **Chemical Physics** (2), **Film Studies** (2), and **Chemistry** (1). Many independent studies at the school are written by students with double majors, however many are not. The most prevalent secondary majors in the dataset are

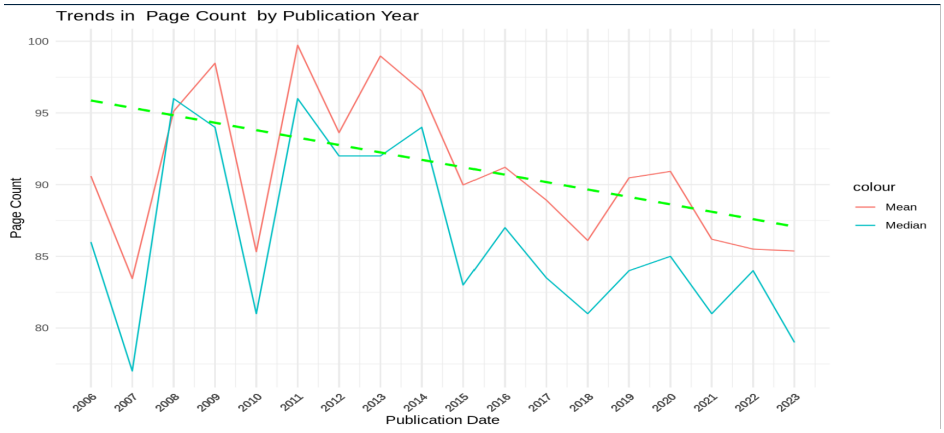

Figure 2: Page Counts of Independent Studies over Time

NA/Single Major Only (4565), **Mathematics** (93), and **History** (74). The least prevalent secondary majors are **Africana Studies**, **Archaeology**, **BCMC**, **Biology**, **Chemistry**, **Music**, and **Urban Studies**, all tied for 1 secondary major listing. Refer to the appendix for a full breakdown of all majors within the dataset.

For publication metrics, all years between (2006, 2023) are well represented in the dataset. ISes for graduation years between 2012 and 2022 all have between 300 - 450 studies present in the dataset, while 2023 has $\sim$ 270. The earlier years of accessible studies (2006 - 2011), each have between 100 - 200 theses collected from each year.

3.2 Trends

Theses with Communication Sciences and Disorders as the primary major have the largest median for page length, while Art & Art History majors have the lowest median for page length. Further dissection of page lengths by majors can be seen in Figure 3.

Instead looking at page length over publication years, there is not significant data to suggest a trend in either direction, although a simple line of best fit shows that IS page lengths have been decreasing, on average, over the past 17 years. This can be seen in Figure 2.

In playing with the various lexical statistics in the dataset, it was discovered that significant differences in medians exist for **Lexical Rarity**, which informally is a measure of esotericism within the text. Departments with higher rarities tended to be STEM majors, such as BCMB, Chemistry, Biology, (Chemical) Physics, Mathematics etc. Interestingly, Spanish, Music, and Communication Sciences & Disorders also had comparatively high lexical rarity. Less lexically rare majors were Sociology & Anthropology, Education, Global Media and Digital Studies, Urban Studies, and History [4]. Boxplots for lexical rarities of all primary major's independent studies can be seen in Figure 4. Aside from rarity, **Lexical Density**, a measure of concentration of content words, was found to be loosely increasing over publication year. Computer Science, SDS, Physics, and Mathematics had the *lowest* median *mtld* (measure of textual lexical density) values. Computer science ISes also had notably lower **variability**, specifically a higher median *Simpson's S*, which roughly encodes the

likelihood of randomly selecting two tokens from a text that are the same token [8]. **Evenness**, a measure of how tokens are distributed among different types, which is measured by *entropy*, showed no significant difference in medians delineated by primary major. **Dispersion**, a measure of how different types of tokens are distributed throughout a body of text, showed a slight decreasing trend from higher dispersion in humanities/social sciences (English, History, International Relations) to lower dispersion in harder sciences (Chemical Physics, Computer Science, Physics).

3.3 Exemplary Status Modeling

The best logistic regression for predicting if an Independent Study *isexemplar* (highest honors) or not is simple for such a large possibility space of observation combinations.

Variable	Coefficient	St. Err.	p-value	Exponentiated 95% Conf. Int.
Intercept	$-2.27 * 10^2$	$2.85 * 10$	$1.6 * 10^{-15}$	$(5.10 * 10^{-124}, 1.93 * 10^{-75})$
Lexical Density	-3.15	6.37	$7.43 * 10^{-7}$	$(1.27 * 10^{-2}, 1.56 * 10^{-1})$
Pace Count	$8.83 * 10^{-3}$	$1.25 * 10^{-3}$	$1.40 * 10^{-12}$	(1.006, 1.011)
Publication Date	$1.113 * 10^{-1}$	$1.414 * 10^{-2}$	$3.57 * 10^{-15}$	(1.09, 1.15)
Punctuation Per Sentence	4.266	1.33	0.00131	$(4.90, 9.14 * 10^2)$

Table 2: Regression Coefficients, Errors, P-Values, and Intervals for a Logistic Model Predicting Exemplar Status

The AIC (Akaike Information Criterion) for the model presented in Table 2 is 2338.1. This model predicts that the odds of an *exemplar* independent study increases as *page count, publication date,* and *punctuation per sentence* increases, and as *lexical density* decreases. The reference state of this model is non-exemplars. A 95% exponentiated confidence interval for *page count* is between

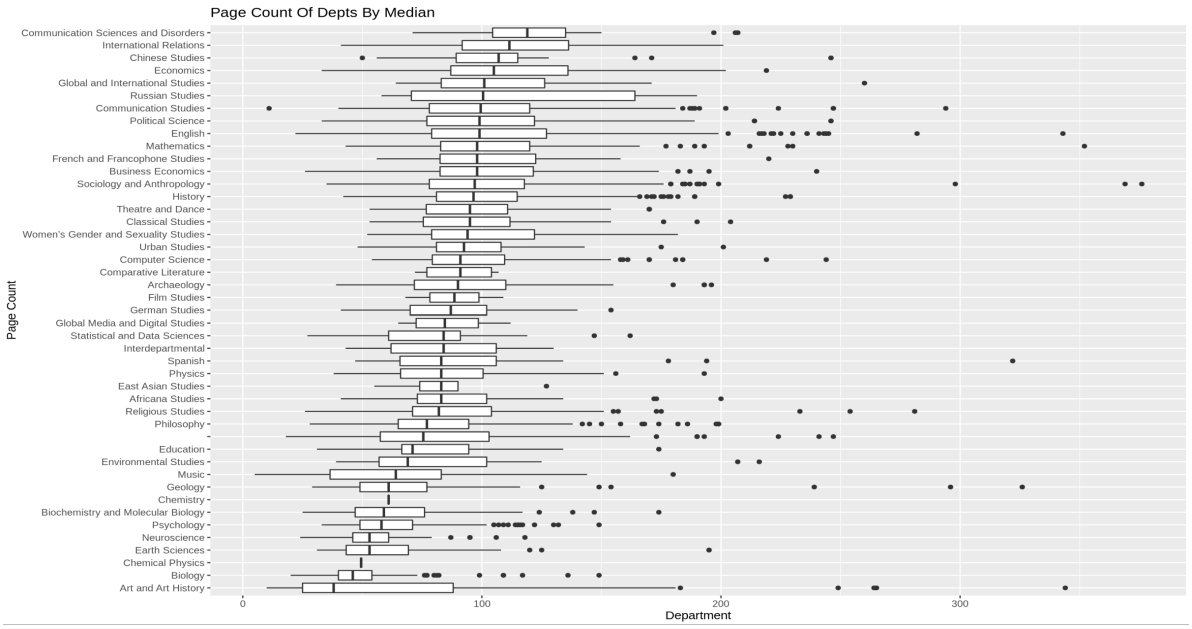

Figure 3: Boxplots of Page Length by Department

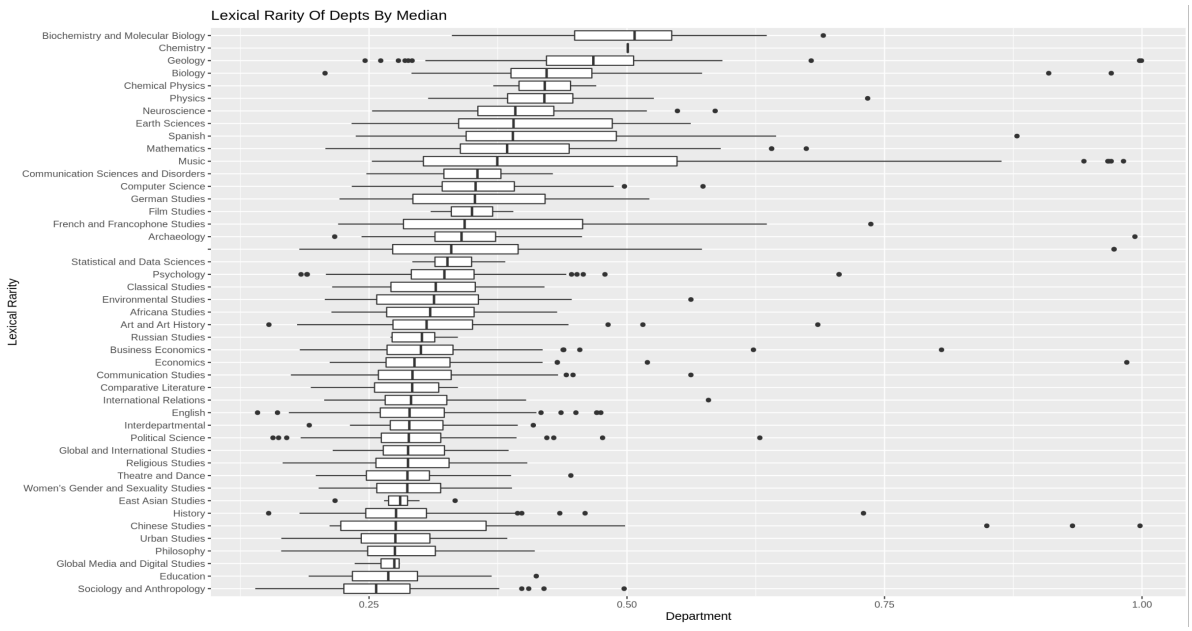

Figure 4: Boxplots of Lexical Rarity By Department

$(1.006, 1.011)$. This means that all else held equal, a thesis with one additional page has between (1.006, 1.011) times the odds a thesis without an additional page of being an exemplar IS. A 95% confidence interval for the true coefficient of *lexical density* is $(1.27 * 10^{-2}, 1.56 * 10^{-1})$. All else held equal, the odds of a paper with increased lexical density being *exemplar* are (0.0127, 0.156) times the odds of a paper without said increased lexical density.

The misclassification rate for falsely guessing non-exemplar for exemplar thesis was approximated 6.2 %. Falsely predicting exemplar for non-exemplar theses occured 0.0189% of the time. It is worth noting that this dataset is heavily biased towards non-exemplar theses ($\sim$ 95% of dataset), with the model accurately predicting an exemplar IS **0 % of the time**. These misclassification statistics were computed with a 50 % cutoff rate for model success. If this value is increased, the model never predicts any thesis will be an exemplar.

4 Discussion

Many trends of various strength exist in the independent study dataset based when looking at changes over time or by department. There **are** differences between different majors and the associated lexical rarity, density, and variability of theses. Less variance exists for evenness and dispersion. I have also demonstrated that regression models can be built off the dataset, however their accuracy leaves much to be desired. These takeaways should be generalizeable to all Independent Studies at the College of Wooster. However, changes in structure and requirements in theses gives pause to applying the takeaways of this work to all undergraduate/graduate thesis papers at other institutions. This work finds that differences exist in writing styles of different departments

and of different class years for ISes.

4.1 Threats To Validity

The dataset itself must be treated with a level of understanding in its possible fallacies. The extraction of text from PDFs automatically is imprecise, especially for ISes that are not fully standardized. A primary issue is that of **kerning**, which is the alignment of letters next to each other. Many computer fonts are not monospaced, which can lead to parsing words incorrectly. Since the entire text of these ISes were extracted autmatically, that means all information in tables, figures, and appendicies was included. For some ISes that included 300 pages of numerical data, this can skew the lexical analysis metrics in an undesireable fashion. The lexical analysis was done using **stanza**, but due to computational limits, papers were processed with nonstandardized parameters for the natural language processing (NLP) pipeline. This means that theses analyzed with the aid of a graphics card accelerator may be biased differently than analyzed by CPU without help from a graphics card. There is a reliance on accuracy and only mild understading of the meanings of the various included lexical statics, so lexical data that may be abnormal is unable to be discerned to the non-expert eyes of the author.

Statistically, the dataset does meet most conditions for inference. The sample size is large enough for inference. The dependent variable is binary, so linearity is automatic. Each independent study is reasonable independent from each other, they have no effect. The ISes in the dataset are not random, as it was every possible IS present on the OpenWorks website. However, the sample is large enough and not so skewed that I believe the randomness condition is met reasonably well enough for exploration and simple model building. The model presented in Table 2 does have *publication date* as a significant predictor. Publication date is a time-series value, however it was not treated as such in the analysis. Since the period is yearly, there is no apparent seasonal or cyclic piece, however it means that predicting exemplary status of ISes published 2024 and beyond is extrapolation. The dataset is limited to automated lexical observations about an IS and associated information about a student. For the model presented, confounding variables may exist that influence a paper's exemplar status, such as a student's prior academic record, their advisor(s), etc.

4.2 Future Work

This is a new dataset, and with it comes many areas for extension. An obvious area of improvement would be reducing biases within the existing data, through solving some of the reliance on existing, fallible software libraries. Alternative exploratory work could be in expanding the observations of the dataset, such as building a citation network or further breaking overall ISes into smaller segmented chunks. Of course, there exists plenty of space within the existing data to be further investigated. Fitting multinomial models to predict departments of majors is one possible avenue. Alternatively, if departments are imbued some heirarchy between social and hard sciences allowing the category to be ordinal, other model building approaches could be used.

References

[1] Daniel, W. W. *Applied Nonparametric Statistics*, 2nd ed. PWS-Kent, Boston, 1990.

[2] Fergadiotis, G., Wright, H. H., and Green, S. B. Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research 58*, 3 (2015), 840–852.

[3] May, P. Collection and Processing Steps of IS Data. https://patrick-may.github.io/projects/Independent-Study-Statistical-Model/, 2023.

[4] May, P. Wooster Independent Study Theses Lexical Dataset. Read the collection + construction article: https://patrick-may.github.io/projects/Independent-Study-Statistical-Model/, 2023. Not Public. Contact pmay24@wooster.edu if you are a Wooster Student to ask for access.

[5] North, K., Zampieri, M., and Shardlow, M. Lexical complexity prediction: An overview. *ACM Comput. Surv. 55*, 9 (jan 2023).

[6] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020).

[7] The College of Wooster. Open works at the college of wooster, 2023. Accessed: 14 October 2023.

[8] Tobias Sproisl. TextComplexity: A Python Library for Text Complexity Analysis, 2023. Accessed: 14 October 2023.

[9] Zhang, F., and Wu, S. Predicting future influence of papers, researchers, and venues in a dynamic academic network. *Journal of Informetrics 14*, 2 (2020), 101035.