Decoding Academic Language: Investigating How Scientific Language Across Academic Disciplines Differs From General Language

Capstone

Author: Nora Svensson Hahr (AUC)*
Supervisor: Giovanni Colavizza (UvA)†,
Major: Science

Date: *June 9, 2023* Reader: *Jelke Bloem*
Word Count: *7336* Tutor: *Anco Lankreijer*

*nora.svensson.hahr@student.auc.nl
†g.colavizza@uva.nl

Decoding Academic Language: Investigating How Scientific Language Across Academic Disciplines Differs from General Language

Nora Svensson Hahr
nora.svensson.hahr@student.auc.nl

Abstract

The United Nations and the European Parliament have both acknowledged the value of engaging the public in scientific research to tackle challenges pertaining to climate change, public health, and global economics. Yet, several studies have found academic English to be inaccessible to a general audience, thus obstructing public engagement and understanding. Simultaneously, new research suggests that there are substantial differences in language across academic disciplines. Thus, some academic disciplines might be more accessible to a general audience than others, and disciplines may be difficult to understand in different ways. To understand these differences and to be able to facilitate academic communication with lay audiences, this study aims at a macroscopic analysis of how language in various academic disciplines differ from general language. Five aspects of language are analyzed using four language models: analyzing overlap in lexis and grammatical structure using Kullback-Leibler Divergence, measuring the use of technical jargon using word sense disambiguation, comparing overall textual similarity using word embeddings, and measuring textual coherence through Latent Semantic Analysis. The results confirm that academic language is discipline-specific but also suggest a stylistic macro-field divide between the social and natural sciences.

Keywords – Metascience, Science Communication, Academic English, Computational Language Models, Linguistic Comparison.

1 Introduction

The purpose of academic communication is not only to share scientific developments with others in the field. Arguably, an equally important aspect of such communication is to make disciplinary knowledge accessible to laymen and academics from other areas of research. During the past decade, international organizations such as the United Nations (UN), the Organisation for Economic Cooperation and Development (OECD), and the European Parliament have acknowledged the value of engaging the public and people across various academic disciplines in research, when tackling big societal challenges such as climate change, public health, food distribution, and sustainable energy production [1]. Such broad collaborations necessitate a language that is accessible to a wide range of readers.

However, there is a large body of scholarly work indicating that academic English is difficult for a lay audience to comprehend and may act as a barrier for engagement due to its distinct style (e.g [2–5]). Indeed, serval studies show how academic English has developed into its own style that diverges from general English both in terms of lexis and grammar [6, 7]. Simultaneously, recent literature also shows great stylistic variations amongst academic disciplines [8–10]. This suggests that although a lay audience generally finds academic English difficult, academic English from different disciplines might vary in difficulty and may be perceived as difficult for different reasons. These differences between discipline-specific language and general language have yet to be explored in academic literature. Thus, the research question for this paper is: How does the language across academic disciplines differ from general language?

This paper proposes a data-driven approach, analyzing five aspects of language using four language models to provide a macroscopic view of

the stylistic differences between general English and discipline-specific academic English. The five aspects of language that will be measured are vocabulary and grammatical overlap, the use of technical jargon, textual coherence, and holistic similarity. These features will be measured using Kullback-Leibler Divergence (KLD), Latent Semantic Analysis (LSA), word sense disambiguation via EWISER, and word embeddings through Word2Vec. The research will be conducted using the OA CC-BY corpus [11] to represent academic articles and the Corpus of Contemporary American English (COCA) to represent general language.

It is hypothesized that the social sciences and arts will employ a language closer to that of general language, compared to the natural sciences. This is partially motivated by the fact that social science topics such as economy and international politics are frequently covered in general news and magazines, and we therefore expect a larger lexical overlap between the social sciences and general language. Furthermore, considering Comte's hierarchy of the sciences, which suggests a higher degree of consensus within the natural sciences compared to the social sciences and arts, it is anticipated that the use of language will reflect this disparity. Consequently, we expect that the natural sciences may employ more jargon-laden language and grammatical structures to enhance objectivity, while the social sciences and arts may adopt a more subjective and general language to elucidate and justify a particular interpretation of concepts.

The indented contribution of the paper is twofold: First, to provide a broad overview of the variations that exists in contemporary academic English across disciplines, in relation to a general baseline. Second, under the assumption that language that is more similar to general English will be easier for a lay audience to understand, to provide a macroscopic review of what disciplines are more or less comprehensible to a general audience, and in what regard.

The remainder of this paper is structured as follows. Section 2 provides the research context for the project. In particular, it surveys the relevant work related to the evolution of academic English, the differences in English across academic disciplines, as well as the state-of-the-art of computational linguistic models for measuring stylistic differences between corpora. Section 3 describes the datasets used and explains the selected methodology. Section 4 is dedicated to the research findings and a summary of the main results. Lastly, section 5 presents a brief discussion of the findings as well as an assessment of the benefits and shortcomings of the project.

2 Research Context

Several studies have shown how academic English has diverged, and continues to diverge, from general English [8]. With its beginnings in the Early Modern period, this separation between academic and general English primarily occurs though the diversification and specialization of lexis, in parallel with the conventionalization of grammar. As academic fields have divided and transformed into new academic fields, new theories and accompanying terminology have appeared. This has given rise to a diverse and highly specialized vocabulary across academic English. At the same time, the strive towards precise and non-ambiguous communication has resulted in a conventionalization of grammar. These concurrent trends have separated academic English from general English and marked it as its own distinct style of writing, often characterized by high information density, an authoritative tone, a high level of abstraction, and technical jargon [2]. However, when analyzing the trends and characteristics that separate academic English from general English, academic English is often considered a monolith without regard for stylistic variations across academic disciplines.

Nevertheless, much research suggests that significant style variations exist amongst academic disciplines. For instance, a recent study by Lucy et al. [12] found that the use of jargon differs significantly between academic disciplines. The study makes use of BERT-based word sense induction and normalized pointwise mutual information (NPMI metrics) to analyze the use of jargon across 12.0 million academic abstracts spanning 19 disciplines. The results show that the natural sciences generally use more words specific to the discipline (word types) whereas the social sciences, alongside mathematics and technology, tend to adopt general words to use in specialized contexts (word senses). Furthermore, several studies have used citation-based methods to measure the levels of disagreement in academic

fields [9, 10, 13]. In particular, Lamers et al. [9] combined citation tracking with a simple version of sentiment analysis to discern disagreement within academic fields at macro- and meso-levels. The findings suggest that although the social sciences and humanities generally contain more disagreement than the natural sciences do, at a meso-level the disagreement within fields vary substantially. Consequently, the researchers conclude that disagreement is discipline-specific, based in epistemic characteristics of the field and its local culture. Moreover, when investigating the similarities and evolution of scientific disciplines, Dias et al. [10] utilize a linguistic similarity measure alongside citation- and expert-based classification methods. Comparing the three measures, the authors infer that linguistic measurements are related to, but distinct from, both citation- and expert classification. Yet, both methods of linguistic and citation comparisons suggest that academic disciplines distinguish themselves from each other. Notably, the research findings suggest considerable linguistic differences amongst academic disciplines.

These findings are also aligned with Comte's hierarchy of the sciences, a hierarchical arrangement of scientific disciplines based on their level of consensus and complexity, which is supported by both bibliometrics [14, 15] and theory [16]. According to Comte, the natural sciences, particularly physics and astronomy, hold the highest position in the hierarchy due to their greater consensus and objective nature [17]. The social sciences and arts occupy lower positions as they involve more subjective interpretations and exhibit a lesser degree of consensus. Comte's hierarchy implies that the natural sciences rely on specialized language and grammatical structures to enhance objectivity, while the social sciences and arts may employ a more general language to convey subjective interpretations. Although these findings strongly suggest that academic English differs across disciplines, there is still a limited body of quantitative research regarding how discipline-specific academic English differs from general language.

In terms of methods, various metrics and models can be used in the task of analyzing stylistic differences in corpora. Analyzing the Kullback-Leibler Divergence (KLD) of ngram models has proven useful in a diverse range of stylistic comparisons of a variety of corpora. For instance, both Bochkarev, Solovyev, and Wichmann [18] and Kim et al. [19] used KLD to analyze variations in word frequency distributions within and across languages. Moreover, Degaetano-Ortlieb and Teich [6] use KLD on both word unigrams and part-of-speech trigrams to track diachronic changes and separations between scientific and general language. In a similar vein, Bizzoni et al. [8] use KLD on word ngrams to trace the development of science through its language use over the past 250 years. As an information-based metric it has been favored over frequency-based style comparisons (e.g. [20–22]) because it does not require pre-defined features for comparison. Furthermore, whereas frequency-based comparisons rely on features specific to the corpora of analysis, KLD can be applied to a wide range of corpora and is transferable across domains. Throughout this paper, the methodology proposed by Degaetano-Ortlieb and Teich [6] will be adopted, analyzing both lexical and grammatical overlap in the analyzed corpora.

Although KLD measures general overlap in lexis, it cannot capture the use of technical jargon. Since technical jargon is identified as a central aspect of the academic writing style [2], it will be analyzed separately. As Justeson and Katz [23] describe it, technical terms are used in academic English to denote specific concepts or phenomena. As such, although certain technical terms overlap with general language, when used as a technical term a word is only used in a specific context. For instance, although "air" can be used as a synonym to both "express", "ventilate", and "broadcast" it would not be used as such in a paper on aerodynamics. There have been many proposed methods for identifying jargon in text. Justeson and Katz provide an algorithm for identifying technical terms based on the patterns in which jargon normally occurs whereas others have used predetermined vocabulary lists. Nevertheless, traditional computational approaches, such as keyword matching or frequency analysis, may not capture the nuanced usage and context-specific interpretations of jargon accurately [12]. In this paper an alternative method is proposed, not to identify and count the exact number of jargon-specific words in a corpus but to capture the general specificity and technicality with which words are used in a text. The methodology presumes that an author that

writes in a technical tone would refrain from using technical jargon alongside several homonyms, as that would detract from the initial purpose of using technical terms. Consequently, a highly technical text would contain fewer homonyms than a less technical text. Under this assumption, word sense disambiguation (WSD) can be employed to discern the level of technicality in writing, where a text that contains ambiguous use of words, and several homonyms, could be considered less technical than a text that uses words in distinct contexts.

The analysis will also include a measure of textual coherence using Latent Semantic Analysis (LSA). A text can be considered coherent if a reader can create an internal representation of the contents provided in the text [24] and as described by Wilson and Corlett [25], the expected level of coherence for a text varies per audience. A more advanced reader is more engaged by a text that has a high information density which requires the reader to make connections on their own, whereas a novice reader prefers a less dense text with more explanation and a slower progression of ideas. This also aligns with Fang's description of scientific language, describing it as particularly informationally dense with a high number of content words per clause [2]. Textual coherence can be measured in several ways including traditional readability formulas [26] and topic modelling through Latent Dirichlet Allocation (LDA) [27]. However, Latent Semantic Analysis (LSA) has become one of the most popular methods for measuring textual coherence and has proven useful across a variety of corpora [28]. It builds on the assumption that when adjoining segments of text have a similar focus, that is an indication of overall textual coherence. Originally introduced by Foltz, Kintsch, and Landauer [24] in 1998, it has for instance been used by Tulsieram, Arocha, and Lee [29] to measure the coherence of Canadian HPV information, by Crossley et al. [30] to measure coherence in second language discourse, and was adopted by Wang and Sui [31] as a measure for coherence in Chinese EFL student's writing. In this project, the original methodology is adopted, using cosine similarity as a distance measure and sentences as the unit of analysis.

Word embeddings will also be employed to model overall differences between corpora and work as a complement to the other methods that capture specific linguistic differences. Word embeddings have become a popular method for measuring linguistic variation in corpora and is often used to capture summative changes between corpora rather than distinctly lexical or grammatical variations. For instance, Hamilton, Leskovec, and Jurafsky [32] use word embeddings to measure semantic shifts in the English language over time and Grieve, Nini, and Guo [33] use it to analyze emerging word-forms in tweets. Moreover, Bizzoni et al. [8] compare word embedding spaces to trace scientific discoveries and diversification of the scientific field. Word embeddings can be constructed using various methods with the two most frequently used architectures being Word2Vec [34] and transformer-based models such as BERT [35]. Word2Vec embeddings are based either on a Continuous Bag of Words model (CBOW) or a Skip-gram model, which both have a single hidden layer and learn word embeddings through simple feedforward neural networks. In contrast, transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), employ a multi-layered self-attention mechanism that allows contextual information capture and modeling relationships between words in a sentence. Although transformer-based models have shown state-of-the-art performance for a multitude of NLP-related tasks [36, 37], Word2Vec is favored in this project due to their ability to capture syntactic and semantic relationships between words and because of the increased interpretability of the embeddings [38].

3 Methodology

3.1 Datasets

There are two datasets used in the research project, one of academic articles and one of contemporary general English. For academic articles, a variety of corpora were considered, among which the S2ORC dataset [39] and the Elsevier OA CC-BY corpus [11] showed to be the most relevant for the project as they both contain recently published full text articles from a variety of academic disciplines. Nevertheless, the Elsevier OA CC-BY corpus was selected as it exhibits a more balanced distribution of full texts of recently published academic articles across academic disciplines. The dataset includes the body text (with mathematical equations removed), abstract, title, publication meta-

data (publication year, issue, volume, etc.), subject classification, and citations for each article. Only the full body text and subject classifications were used throughout the analysis. The subject classification codes were used to separate the texts into their respective academic disciplines. Although the corpus spans 27 classification codes, one code regards "general" texts. Due to the vague nature of this classification, the class was removed from the analysis. If a text contained multiple classification codes, the texts was included in the corpora for all the academic disciplines it was classified as. Thus, the anlysis spans 26 academic disciplines, all of which can be seen in Table 1 alongside their abbreviations and document counts.

For general English, the Corpus of Contemporary American English was employed [39]. The study makes use of the version released 2016, containing over one billion tokens gathered 1990-2015. The texts come from a wide range of genres including fiction, spoken language, movie subtitles, and newspapers which is the main reason it was favored above other contemporary corpora of general language such as the CC-News dataset [40]. The corpus also contains academic articles but considering the project aim of comparing academic language to other genres of text, these were not included in the general baseline. Due to computational limitations, only a subsection of the COCA was used in the analysis. Namely, all the texts from the year 2015. This subsection was chosen as it covers the most recent texts from the corpus and has the most overlap with the time period during which the Elsevier OA CC-BY articles were published.

Both datasets were tokenized, part-of-speech tagged and annotated for word sense using the spaCy *en_core_web_sm* model with the EWISER word sense disambiguator added to the preprocssing pipeline. Punctuation was removed but stopwords were included during preprocessing since they play an important role when analyzing the grammatical patterns of each text.

3.2 Methods

As previously mentioned, scholarly English is primarily distinguished from general language by its use of jargon and specialized vocabulary, particular grammatical structures, and its high information density [2]. The more scholarly the tone of the text, the more foreign it is assumed to be for a general reader. Thus, the methods for this analysis is designed around measuring these stylistic differences compared to a general language baseline.

3.2.1 Kullback-Leibler divergence

Kullback-Leibler Divergence (KLD), or relative entropy, is a statistical measurement of how much two probability distributions differ from each other [41]. It is an information-based measure that calculates how much extra information is required to predict an observation generated by distribution A, using distribution B. In this project, KLD will be used to measure the grammatical and lexical differences between two corpora. KLD is an asymmetric measure, meaning that the distance from distribution A to distribution B differs from the distance from distribution B to distribution A. This is particularly useful in this analysis: for instance, a physics researcher might understand general language better than a general audience would understand physics publications.

In this implementation, word unigrams and part-of-speech trigrams will be used as the units of analysis, following the methodology presented by Degaetano-Ortlieb and Teich [6]. Calculating the distance from distribution A to disribution B is done accordingly:

$$KL(A \parallel B) = \sum_i A(i) * \log \frac{A(i)}{B(i)}$$

.

To compensate for the differences in corpus size across academic disciplines, bootstrapping was employed with 1000 iterations. That is, each distance was calculated a thousand times and for each iteration the academic discipline corpora were sampled with replacement. These measurements were then averaged to reach a KLD estimate, including a 95% confidence interval. For the lexical analysis, stopwords were removed from each corpus. However, for the grammatical analysis they were included due to their important role in grammatical structures. For both measures, units of analysis that occurred less than 10 times in the corpus were removed.

3.2.2 Word Sense Disambiguation

Multiple methods have been suggested for tackling the task of word sense disambiguation (WSD). For the aims of this research, EWISER (Enhanced WSD Integrating Synset Embeddings

Discipline	Abbreviation	Document Count
Agricultural and Biological Sciences	AGRI	4840
Arts and Humanities	ARTS	982
Biochemistry, Genetics and Molecular Biology	BIOC	8356
Business, Management and Accounting	BUSI	937
Chemical Engineering	CENG	1878
Chemistry	CHEM	2490
Computer Science	COMP	2039
Decision Sciences	DECI	406
Earth and Planetary Sciences	EART	2393
Economics, Econometrics and Finance	ECON	976
Energy	ENER	2730
Engineering	ENGI	4778
Environmental Science	ENVI	6049
Immunology and Microbiology	IMMU	3211
Materials Science	MATE	3477
Mathematics	MATH	538
Medicine	MEDI	7273
Neuroscience	NEUR	3669
Nursing	NURS	308
Pharmacology, Toxicology and Pharmaceutics	PHAR	2405
Physics and Astronomy	PHYS	2404
Psychology	PSYC	1760
Social Sciences	SOCI	3540
Veterinary	VETE	991
Dentistry	DENT	40
Health Professions	HEAL	821

Table 1: Distribution of academic disciplines in the Elsevier OA CC-BY corpus.

and Relations) has been identified as the most suitable architecture for the task due to its state-of-the-art performance in token-tagging and its simple implementation using the python spaCy library.

EWISER is supervised neural architecture that treats WSD as a classification problem for which each token is matched with the most appropriate synset. It builds on the previous EWISE neural architecture that incorporates prior knowledge using synset embeddings created by training a gloss encoder through triplet loss on WordNet [42]. EWISER uses a similar approach but generalizes the process, showing that off-the-shelf pretrained embeddings can be used and presents a novel structured logits mechanism which allows for the utilization of concept relatedness, expressed as edges in a Lexical Knowledge Base (LKB), in the classification process [43]. In short, the architecture uses BERT Large (cased) and takes the sum of the outputs from the last 4 layers as input for each word to disambiguate. A 2-layer feedforward network with swish activation function then computes the logit scores. When computing the logit scores the model leverages the explicit information present in a weighted knowledge graph, where nodes represent synsets and edges represent concept relatedness, when computing the probability distribution vector for a target word. "Hidden" logits of related synsets are multiplied by the corresponding edge weights and are subsequently added to the "hidden" logits of the related synsets. These sums then represent the "final" logits. By embedding information from the WordNet LKB graph within the neural architecture, and by exploiting pretrained synset embeddings, the model can disambiguate homonyms even for words that are not included in the training dataset.

The implementation used in this project makes use of the model trained on the SemCor in union with tagged WordNet glosses and WordNet examples. The sense tagging of each token was included in the preprocessing pipeline. Subse-

quently, the number of senses associated with each token in a text was tallied, with stopwords and single-use words within a text excluded from the count. Next, the average number of senses per token was computed for each document, and these document-averages were further averaged to generate a corpus-level average. To address the variability in document distribution across different discipline-corpora, the process was bootstrapped with 1000 iterations, resulting in an estimated corpus-average along with a 95% confidence interval.

3.2.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a mathematical technique for representing words as numerical vectors where words that are used in similar contexts appear closer together in the vector space. In this project, the method is used as a measure of textual coherence by measuring the similarity between adjoining sentences in a text [24]. If a text is coherent, it is assumed that adjoining sentences follow from each other both in vocabulary and in topic and this would show itself in the LSA vectorization of these sentences [24]. To create the LSA matrix, a word by document co-occurrence matrix is first constructed using TF-IDF vectorization. The TF-IDF vectorization normalizes each vector to give more weight to informative low-frequency words and reduce the weight of uninformative high-frequency [44]. The LSA matrix is then produced by singular value decomposition (SVD) of the document-term co-occurrence matrix to reduce dimensionality and uncover latent semantic structures within the term-document matrix [45].

For each corpus, the LSA matrix was created with each sentence in each text treated as its own document. The matrix was reduced to 100 dimensions. The average distances between LSA vectorizations of adjoining sentences were then calculated per text using cosine similarity. The cosine similarity measures the distance between two angles in a vector space and is calculated as:

$$sim(w_1, w_2) = \cos(w_1, w_2) = \frac{w_1 * w_2}{|w_1| * |w_2|}$$

The text averages were then used to create a corpus average where a large average indicates that there is high degree of textual coherence across a corpus and a lower average indicates a low degree of textual coherence.

The procedure was once again bootstrapped to create an estimated average with a 95% confidence interval. However, due to time constraints the number of iterations was decreased to 300.

3.2.4 Word Embeddings

Word embeddings are representations of words as multi-dimensional vectors, where words that share a similar syntagmatic context are closer together in the embedding space. In this project, the word embedding space is used as holistic representations of the corpora, where differences in two embedding spaces indicate holistic differences between the associated corpora. The original Word2Vec approach [34] aims at maximizing the likelihood of a word given its context. Formally it does so by maximizing the function

$$L = \frac{1}{T} \sum_{t \in T} \sum_{-c \leq j \leq c} \log p(w_{t+j} | w_t)$$

where T is a given document and c is the number of context words taken into consideration in the algorithm. Several models have been suggested to improve the original model. For instance, Ling et al. [46] have shown that word order in context words is valuable when capturing words with grammatical functions and their structured skip-gram approach was later adopted by Bizzoni et al. [8] in their corpus comparisons. Such alterations could certainly have been adopted for this research project as well. However, due to computational and time constraints, the traditional gensim implementation of Word2Vec was used instead.

As word embeddings create models of a specific corpus, using them for comparative purposes requires a measurement for them to be compared by. This study adopts the cosine distance as measurement, outlined in the previous subsection. A smaller distance between same-word vectors in different embedding spaces indicate that the word is used in similar contexts in both corpora. The cosine distance between the general language embedding space and the discipline-specific embedding spaces was measured for 1000 randomly selected tokens present in all embedding spaces. The average cosine distance between tokens in the general language corpus and each discipline corpus was then calculated.

The process of creating the academic embedding spaces was bootstrapped to produce an estimated average distance between tokens, with a

95% confidence interval. However, due to the aforementioned constraints on time and computational power the number of iterations was capped at 100.

4 Results

The aim of this study is to analyze how English across academic disciplines differs compares to general English from the five perspectives of lexical overlap, grammatical overlap, the use of technical jargon, textual coherence, and overall syntagmatic similarity. The analysis builds on the assumption that academic disciplines that shows similarity with general language would be more accessible and understandable to a lay audience than disciplines where language use is significantly different.

4.1 Lexical Overlap

Kullback-Leibler divergence (KLD) was used to measure overlap in both lexis and grammar between each corpus of academic texts and general language. As KLD is an asymmetric measure, how corpus A diverges from corpus B differs from how corpus B diverges from corpus A. Observing the range of lexical divergence for academic corpora vs. general language and comparing it to the range of divergence for general language vs. academic corpora in Figure 1, it becomes clear that there is a larger variability in lexical overlap when comparing the lexis of academic texts to general language than vice versa. This suggests that the level of familiarity a lay person would have with the vocabulary used in an academic article would differ significantly depending on the discipline.

Furthermore, the results also show that when comparing how lexis in scholarly language diverges from that of general language there is a clear division amongst the macro-classifications of academic disciplines. The social sciences have the lowest divergence, followed by the arts & humanities and mathematics & engineering. The natural sciences distinguish themselves with comparatively high divergences and the with the top nine highest divergences all coming from disciplines within the natural sciences. The same trend can be noticed when comparing general language to academic language, however not as clearly. Not only are the academic disciplines not as clustered according to macro-classifications. The confidence intervals are also larger in relation to the overall range of values, making an ordering less obvious. Looking at the pointwise KLD for each token, to analyze what tokens contribute the most to the overall KLD, it becomes clear that it is (parts of) chemical compounds, particles, and mathematical terminology that mostly contribute to the high KLD's for the natural sciences.

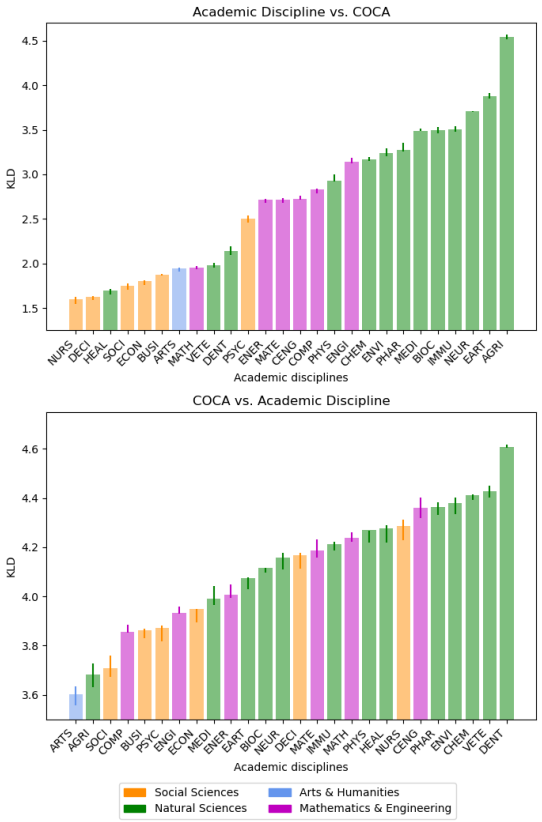

Figure 1: Average KLD per academic discipline based on token unigrams.

On a discipline-specific level, some further interesting observations can be made. BUSI as well as ECON both show consistently low divergence alongside SOCI. This implies not only that the language used in these subjects has much overlap with general language but also that general language contains much of the vocabulary used within these fields. On the contrary, PHAR together with ENVI and CHEM show consistently high divergences.

4.2 Technical Jargon

However, it is important to note that audience familiarity with a word does not necessarily imply familiarity with the concepts associated with that word within a specific academic discipline. As previously mentioned, jargon can both take the

form of word types and word senses [12]. Jargon as word types indicate a use of specialized terms not frequently used outside of the field and jargon as word senses refers to words that are frequently used in everyday language but are overloaded with context-specific meanings within an academic discipline. To understand how jargon and technical tone is used in each corpus, the number of senses associated with each word in each text is analyzed.

As hypothesized, the Corpus of Contemporary American English (COCA) exhibits a higher number of word senses per token than any of the discipline-specific corpora. It averages at 1.556 word senses per word which could be interpreted as more than half of all words in COCA being used with two distinct senses per text. Furthermore, the results show that there is a larger difference in word specificity between general language and academic language overall, than there is a distinction within the academic disciplines themselves. The difference between the COCA average and the highest average within the academic disciplines, ARTS, is 0.161 based on the reported average and 0.152 based on the upper confidence interval. And the difference between the highest and lowest average within the academic disciplines, ARTS and NURS respectively, is 0.135 based on the reported averages and 0.147 based on the lower confidence interval of the lowest average and the upper confidence interval of the highest average. These results suggest that general academic English distinguishes itself from everyday language more so than academic disciplines distinguishes themselves from each other with regards to word specificity and technical style.

Figure 2 shows the average number of word senses per word in each text per academic discipline, as well as how much each average differs from that of the COCA average. Upon examination, it becomes evident that the level of word specificity follows the same macro-field trends as the previous KLD results: the arts & humanities are closets to general language, followed by the social sciences, mathematics & engineering, and with the natural sciences being the most different from general language.

By combining the results regarding lexical overlap with the token sense results, we can gain insight into the potential nature of how technical jargon is being used on a discipline-specific level. MEDI, ENVI, CHEM, and BIOC all exhibit high

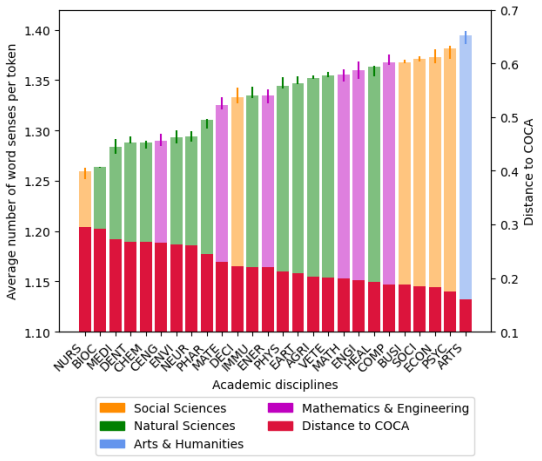

Figure 2: Average number of word senses per token per academic discipline alongside the difference between reported averages and the average number of word senses per token in COCA.

KLD values, indicating a significant difference in vocabulary usage compared to general language. These disciplines also show a low average number of word senses per token, suggesting that words are used in a more specialized technical style. Together these results imply that these disciplines use a jargon that differs from general language with regards to word type. That is, the jargon and terms used within these disciplines are specific to the respective disciplines. In contrast, when examining the NURS and DENT corpora the KLD values are high but the average number of word senses per token is low. This indicates that common language is utilized within the disciplines but used with a level of specificity higher than that in everyday language, suggesting that the jargon within these fields is dominated by word sense alterations of pre-existing common words. Lastly, the disciplines of BUSI, SOCI, and ECON demonstrate lower KLD values along with a higher average number of word senses per token. This suggests that many of the words used in these disciplines are frequent in general language and are likely to be used with a specificity more similar to that of general language than other academic disciplines, indicating a potentially lesser use of technical jargon overall.

4.3 Grammatical Overlap

To estimate grammatical differences between general language and academic English, KLD has been employed on part-of-speech trigrams. By in-

specting Figure 3, displaying a summary of the average KLD per academic discipline, previous macro-field patterns become apparent with regards to grammar as well. In particular, CHEM and chemistry-related fields such as PHAR and CENG display high divergence from general language.

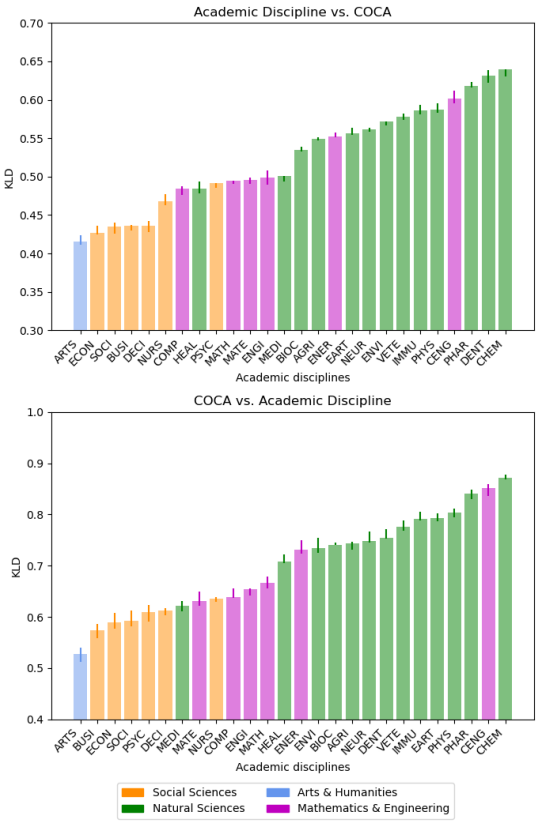

Figure 3: Average KLD per academic discipline based on part-of-speech trigrams

By inspecting the pointwise KLD related to these fields it becomes clear that proper nouns (NN), especially in combination with present and past verbs followed by a past participle (e.g. NN-VBD-VBN and NN-VBZ-VBN), have a large impact on the total KLD. Degaetano-Ortlieb and Teich [6] attribute such trigrams to the use of passive voice alongside determiner–noun–which trigrams (DT-NN(S)-WDT) and noun-in-which trigrams (NN(S)-IN-WDT). Such trigrams also show high pointwise KLD within these disciplines and other natural sciences such as PHYS and IMMU. Considering the lower pointwise KLD associated with such trigrams in the social sciences and arts & humanities, the results suggests that the use of passive voice is more widespread within the natural sciences than in the social sciences and humanities. This is further supported by the high number of personal pronouns (PRP) used in the arts & humanities and to a certain degree in the social sciences, indicating the use of an active voice.

4.4 Textual Coherence

To approximate textual coherence, LSA was employed and cosine similarity between the vector representation of adjoining sentences in a text was measured. In contrast to previous results, general language does not set itself apart from academic language. Instead, the distance between adjoining sentences in COCA is fairly centered within the cosine distances of the academic disciplines. Thus, accessibility to a general audience can be considered from two distinct viewpoints. On the one hand, an academic discipline can be viewed as more accessible if it exhibits a higher level of textual coherence, as a higher coherence indicates a text that would be easier to follow and potentially have a lower information density. On the other hand, an academic text can be considered more accessible if it is written with a similar level of textual coherence to that of general language, since that is the level of coherence a general audience would be most familiar with. Figure 4 shows the average cosine similarity between adjoining sentences per academic corpus as well as their absolute distance from the average cosine similarity of COCA.

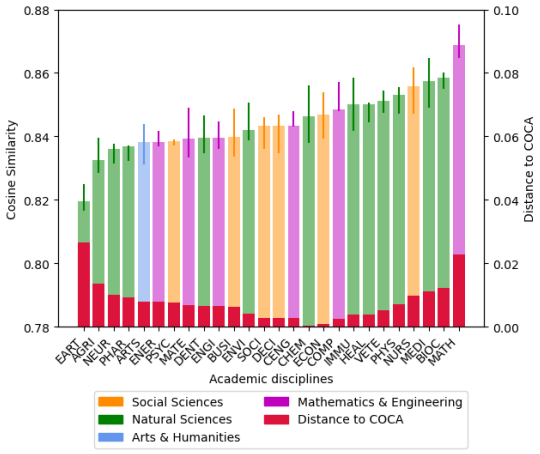

Figure 4: Average cosine similarity between adjoining sentences in LSA-matrix per academic discipline alongside the difference between reported averages and the average cosine similarity between adjoining sentences in COCA.

The results suggest that the MATH corpus is the most textually coherent, followed by MEDI

and NURS and the least textually coherent seemingly being ENGI, EART, and AGRI. Furthermore, observing the distance to the COCA average, CHEM, ECON, and COMP are the closest. This clearly breaks from trends in previous results, where the social sciences and humanities have shown the greatest similarity with general language. However, due to the size of the confidence intervals, an exact ordering is difficult to establish.

4.5 Holistic Differences

As a holistic complement to previous metrics, the overall syntagmatic differences between general and academic English are measured using Word2Vec. This was done by training an embedding space on COCA and measuring the cosine similarity between 1000 randomly selected words in that embedding space, with the same words in the embedding spaces trained on each academic corpus. The average cosine similarities between same-word vectors are presented in Figure 5. The results reinstate previous findings, once again suggesting that the social sciences, alongside the arts & humanities, display very high levels of cosine similarity with general language. In particular, for the disciplines ECON, BUSI, and SOCI, a previously noted high degree of lexical and grammatical overlap with general language in combination with a high cosine similarity for words used in all corpora suggest that the language use within these respective disciplines is similar to general language both in terms of what language is used and in what situations it is used. In contrast, but once again adhering to previous results, the natural siences and chemistry-related disciplines show the greatest average distance to general language. Considering that CHEM and chemistry-related fields show a low lexical and grammatical overlap with general language, the word embedding results are indicative of the following: in the language used in CHEM (and related disciplines) even common words are used in contexts and in grammatical structures not usually employed in everyday language.

4.6 Summary of Results

Figure 6 provides a summary of all results in terms of each academic disciplines ranking in each metric (not taking confidence intervals into account). As hypothesized and as concluded when considering each result separately, SOCI, ECON,

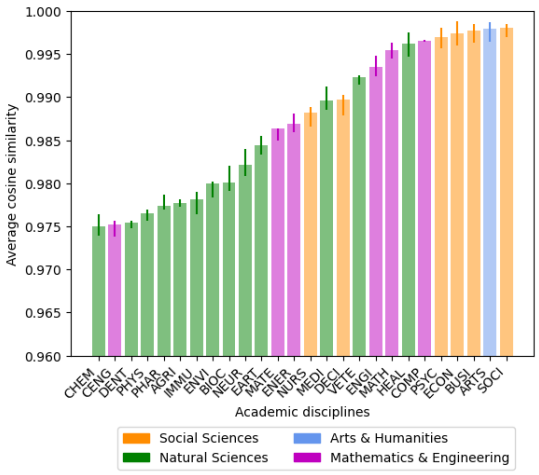

Figure 5: Average cosine distance to same-word vector in COCA, per academic discipline.

and BUSI show consistently high similarities with general language. The findings suggest that out of all academic disciplines, these three use a lexis most similar to that of everyday language, with a similar level of technical style and textual coherence. HEAL also displays a generally high ranking in all metrics. This is not surprising considering that many articles published within the field are aimed at practitioners within the health profession and not academics per se. On the contrary, many natual sciences place low with fields such as NEUR and ENVI showing consistently high differences to general language use. Interestingly, both MATH and COMP place in the mid- to upper-mid range in most rankings. This would suggest that the language used within these disciplines in many aspects is more linguistically accessible to a general audience than that of many natural sciences and engineering fields. However, this notion is not entirely sound considering that both MATH and COMP rely heavily on mathematical notations to mediate knowledge and that such information is not contained within the corpora. For instance, much of what would be considered mathematical jargon is expressed through notation that would not be present, which skews the measurement within this analysis. Thus, these results should be met with a particular degree of skepticism.

By analyzing how discipline-rankings relate to each other there is some indication that academic disciplines that are conceptually similar to one another use language that mimics this similarity. For instance, CHEM, CENG and IMMU follow a sim-

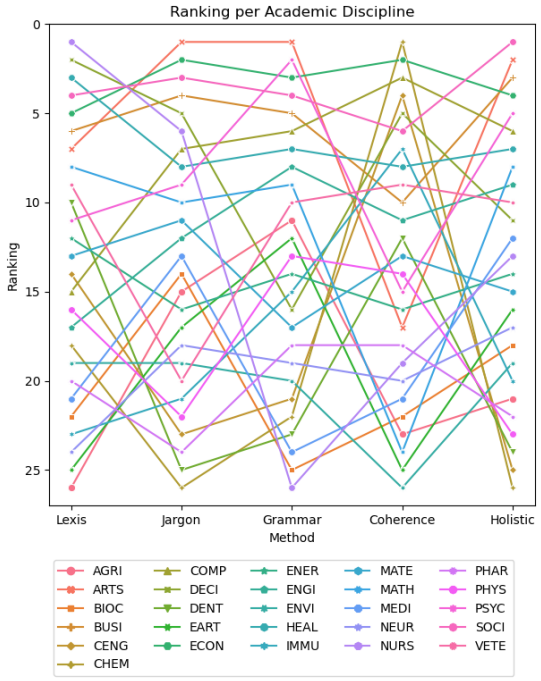

Figure 6: Summary of results by ranking each academic discipline by similarity to general language. Horizontal axis: *Lexis* denotes the ranking for KLD using unigrams (Discipline vs. COCA), *Jargon* denotes the ranking for average number of word senses per token, *Grammar* denotes the ranking for KLD using part-of-speech trigrams (Discipline vs. COCA), *Coherence* denotes the smallest distance to the COCA measure of textual coherence using LSA, and *Holistic* denotes the average cosine distance for measuring same-word vectors in COCA, per each academic discipline.

ilar trajectory for each ranking: a low mid-range ranking for lexical overlap, a decrease for jargon, a slight increase for grammatic overlap, a top ranking for textual coherence, and a bottom ranking in the holistic measure. BIOC and MEDI also seem to follow a ranking pattern similar to each other's. This is not surprising, considering that intellectual exchange occurs more frequently between closely related fields than between conceptually disparate disciplines [47].

When labeling the rankings according to macro-field classifications, as seen in Figure 7, the separation between macro-fields is once again visualized. Although not completely stratified, the language used within the natural sciences tend to be furthest away from, and the language used within the social sciences tend to be most similar to, general language in all aspects measured. The arts & humanities show a very high similarity with everyday language in all aspects except for textual coherence. However, considering the great uncertainty that coupled the textual coherence results, the remaining rankings still provide strong evidence to suggest that the language used within the ARTS is indeed very similar to that of general language. The mathematics & engineering disciplines are less unified in their language than other macro-fields. This disparity can most likely be attributed to the field consisting of a diverse group of disciplines that exhibit strong links to other disciplines outside of the field. For instance, ENER has a fairly similar trajectory in the rankings as PHYS and as previously mentioned CHEM and CENG show very similar ranks. However, considering previous visualizations the disciplines within the field tend to rank between the social and natural sciences.

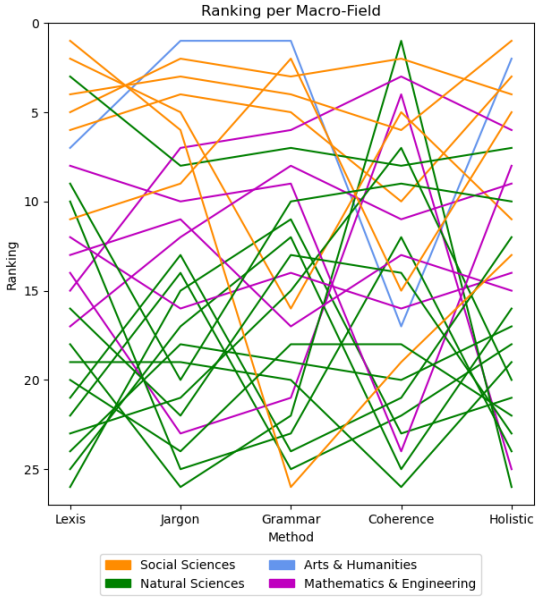

Figure 7: Summary of results by ranking each discipline by similarity to general language, according to its macro-field classification. Labels for horizontal axis are the same as for Figure 6.

5 Discussion

This study centers around understanding how the language from various academic disciplines differs from general language, based on five linguistic metrics. It was initially hypothesized that the social sciences and the arts & humanities would be more similar to general language than the natu-

ral sciences and mathematics & engineering. The results align well with this expectation, suggesting that the language within SOCI, BUSI, and ECON is the most similar to general language in all aspects measured. Furthermore, disciplines such as ENVI and PHAR, alongside most other natural sciences, show consistently high differences with general language. For mathematic & engineering, the results are less conclusive. However, considering individual disciplines within the field and comparing them to Ramage, Manning, and McFarland [47]'s mapping of intellectual exchange in academia, the disciplines that are most closely aligned with general language in this study show the greatest interaction with the social sciences and humanities in Ramage, Manning, and McFarland [47]'s investigation. For instance, COMP and ENGI which both exhibit generally high similarity with general language in relation to other disciplines within the mathematics & engineering field whilst also showing strong intellectual exchange with Communication & Information Science and Public Affairs & Public Policy respectively. On the other hand, CENG which exhibits a language very different from general language, has the strongest intellectual ties with Chemistry and few exchanges with the social sciences or Humanities. This indicates the existence of a linguistic spectrum where the proximity of an interdisciplinary discipline to general language is influenced by the balance between the social sciences/humanities and the natural sciences.

This project was conducted with a background of various voices in academia and civil society warning that science communication might act as a barrier of access for general audiences to understand, contribute to, and implement scientific research [2–5]. Under the assumption that language that is more similar to general English will be easier for a lay audience to understand, the results in this study have significant implications. First of all, the evidence suggests that all academic language is significantly different from general language both in terms of lexis, grammar, and jargon. This indicates that language is a barrier to public engagement and understanding in all disciplines. Secondly, the results suggest that many of the natural sciences are particularly inaccessible to a general audience, primarily due to technical terms and jargon, as well as grammatical structures that are uncommon in everyday English. This is alarming considering that many findings and technological developments that have implications for all of society emerge within the natural sciences. In particular, the linguistic differences between everyday English and the language used within ENVI are interesting in the light of the current climate crisis which historically has been underreported in news media outlets globally [48]. Although speculation, the difference between general language and ENVI-specific language might have contributed to the lack of media coverage of early findings and signs of the climate crisis. This does of course not suggest that environmental scientists should have, or ought to sacrifice linguistic precision and academic rigor for public appeal when publishing academic papers. As previously mentioned, academic English has in large part developed specifically to communicate scientific findings in an objective and efficient manner. Instead, these linguistic differences can work as a guide to understand the extra resources that are required to explain and mediate the importance of scientific findings from certain academic disciplines, in particular the natural sciences, to a lay audience.

While this study focuses on how language in academia differs from general English, it is important to consider that other factors may act as stronger barriers to academic participation than the language itself. For instance, the cost of accessing academic journals and the fundamental barrier of language proficiency can significantly hinder engagement with scholarly work. Furthermore, it is essential to recognize that the analysis primarily examines language use in academic articles. Although these articles are a crucial component of scholarly communication, they may not be the most suitable source for understanding how science is mediated and communicated to the general public. To gain a more comprehensive understanding, future research could explore how language in popular science writing and scientific journalism differs from other forms of general language, providing further insights into science communication with a broader audience. Different genres and contexts within academia, such as conference presentations, grant proposals, or academic discussions, may exhibit variations in language use that also warrant separate investigations.

6 Conclusion

In this paper, the language used in various academic disciplines has been explored as compared to a general language baseline. The exploration has been conducted using five linguistic metrics using four computational language models: Kullback-Leibler Divergence has been used to measure lexical and grammatical overlap using token unigrams and part-of-speech trigrams, technical style has been measured using word sense disambiguation through EWISER, Latent Semantic Analysis has been used to measure textual coherence, and overall syntagmatic differences have been measured using Word2Vec. The results suggest a general trend with the social sciences and arts being the most similar to general language and the natural sciences being more different. Under the assumption that academic disciplines that employ language more similar to general language is more accessible to a lay audience, the results thus present an accessibility divide between the natural and social sciences. However, the results also suggest that linguistic characteristics are discipline-specific and at least partially influenced by interdisciplinary intellectual exchange, thus reducing the impact of macro-field generalizations and highlighting the nuanced nature of language use within academia.

7 Data Availability Statement

All code to conduct the reasearch as well as all figures are available at: https://github.com/norahahr/Capstone2023. The Elsevier OA CC-BY Corpus is available for download at https://elsevier.digitalcommonsdata.com/datasets/zm33cdndxs/2 and the COCA is available for purchase at https://www.corpusdata.org/purchase.asp.

Works Cited

[1] Ruben Vicente-Saez and Clara Martinez-Fuentes. "Open Science now: A systematic literature review for an integrated definition". en. In: *Journal of Business Research* 88 (July 2018), pp. 428–436. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2017.12.043. URL: https://www.sciencedirect.com/science/article/pii/S0148296317305441 (visited on 03/15/2023).

[2] Zhihui Fang. "Scientific literacy: A systemic functional linguistics perspective". en. In: *Science Education* 89.2 (2005), pp. 335–347. ISSN: 1098-237X. DOI: 10.1002/sce.20050. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.20050 (visited on 03/16/2023).

[3] Catherine E. Snow. "Academic Language and the Challenge of Reading for Learning About Science". In: *Science* 328.5977 (Apr. 2010). Publisher: American Association for the Advancement of Science, pp. 450–452. DOI: 10.1126/science.1182597. URL: https://www.science.org/doi/abs/10.1126/science.1182597 (visited on 03/16/2023).

[4] Tal August et al. "Explain like I am a Scientist: The Linguistic Barriers of Entry to r/science". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376524. URL: https://doi.org/10.1145/3313831.3376524 (visited on 06/09/2023).

[5] Yang Liu, Alan Medlar, and Dorota Głowacka. "Lexical ambiguity detection in professional discourse". en. In: *Information Processing & Management* 59.5 (Sept. 2022), p. 103000. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2022.103000. URL: https://www.sciencedirect.com/science/article/pii/S0306457322001133 (visited on 06/09/2023).

[6] Stefania Degaetano-Ortlieb and Elke Teich. "Toward an optimal code for communication: The case of scientific English". en. In: *Corpus Linguistics and Linguistic Theory* 18.1 (Feb. 2022). Publisher: De Gruyter Mouton, pp. 175–207. ISSN: 1613-7035. DOI: 10.1515/cllt-2018-0088. URL: https://www.degruyter.com/document/doi/10.1515/cllt-2018-0088/html (visited on 09/22/2022).

[7] Douglas Biber and Edward Finegan. "Drift and the Evolution of English Style: A His-

tory of Three Genres". In: *Language* 65.3 (1989). Publisher: Linguistic Society of America, pp. 487–517. ISSN: 0097-8507. DOI: 10.2307/415220. URL: https://www.jstor.org/stable/415220 (visited on 10/13/2022).

[8] Yuri Bizzoni et al. "Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach". In: *Frontiers in Artificial Intelligence* 3 (Sept. 2020), p. 73. ISSN: 2624-8212. DOI: 10.3389/frai.2020.00073. URL: https://www.frontiersin.org/article/10.3389/frai.2020.00073/full (visited on 09/19/2022).

[9] Wout S Lamers et al. "Investigating disagreement in the scientific literature". In: *eLife* 10 (Dec. 2021). Ed. by Peter Rodgers. Publisher: eLife Sciences Publications, Ltd, e72737. ISSN: 2050-084X. DOI: 10.7554/eLife.72737. URL: https://doi.org/10.7554/eLife.72737 (visited on 09/22/2022).

[10] Laércio Dias et al. "Using text analysis to quantify the similarity and evolution of scientific disciplines". In: *Royal Society Open Science* 5.1 (Jan. 2018). Publisher: Royal Society, p. 171545. DOI: 10.1098/rsos.171545. URL: https://royalsocietypublishing.org/doi/full/10.1098/rsos.171545 (visited on 01/23/2023).

[11] Daniel Kershaw. *Elsevier OA CC-BY Corpus*. Aug. 2020. DOI: 10.17632/ZM33CDNDXS.2. URL: https://data.mendeley.com/datasets/zm33cdndxs/2 (visited on 09/20/2022).

[12] Li Lucy et al. *Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications*. arXiv:2212.09676 [cs]. May 2023. DOI: 10.48550/arXiv.2212.09676. URL: http://arxiv.org/abs/2212.09676 (visited on 05/30/2023).

[13] Jeroen Bruggeman, V. A. Traag, and Justus Uitermark. "Detecting Communities through Network Data". en. In: *American Sociological Review* 77.6 (Dec. 2012). Publisher: SAGE Publications Inc, pp. 1050–1063. ISSN: 0003-1224. DOI: 10.1177/0003122412463574. URL: https://doi.org/10.1177/0003122412463574 (visited on 01/27/2023).

[14] Daniele Fanelli and Wolfgang Glänzel. "Bibliometric Evidence for a Hierarchy of the Sciences". en. In: *PLOS ONE* 8.6 (June 2013). Publisher: Public Library of Science, e66938. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0066938. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0066938 (visited on 06/01/2023).

[15] Daniele Fanelli. ""Positive" Results Increase Down the Hierarchy of the Sciences". en. In: *PLOS ONE* 5.4 (Apr. 2010). Publisher: Public Library of Science, e10068. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0010068. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0010068 (visited on 06/01/2023).

[16] Stephen Cole. "The Hierarchy of the Sciences?" In: *American Journal of Sociology* 89.1 (July 1983). Publisher: The University of Chicago Press, pp. 111–139. ISSN: 0002-9602. DOI: 10.1086/227835. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/227835 (visited on 06/01/2023).

[17] Auguste Comte. *The Positive Philosophy of Auguste Comte*. en. Google-Books-ID: ktM3AQAAMAAJ. Blanchard, 1858.

[18] V. Bochkarev, V. Solovyev, and S. Wichmann. "Universals versus historical contingencies in lexical evolution". In: *Journal of The Royal Society Interface* 11.101 (Dec. 2014). Publisher: Royal Society, p. 20140841. DOI: 10.1098/rsif.2014.0841. URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2014.0841 (visited on 03/22/2023).

[19] Yoon Kim et al. *Temporal Analysis of Language through Neural Language Models*. arXiv:1405.3515 [cs]. May 2014. DOI: 10.48550/arXiv.1405.3515. URL: http://arxiv.org/abs/1405.3515 (visited on 03/22/2023).

[20] Douglas Biber and Bethany Gray. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press, 2016.

[21] Stefania Degaetano-Ortlieb et al. "SciTex – A Diachronic Corpus for Analyzing the Development of Scientific Registers." In: Nov. 2013.

[22] Aminul Islam and Diana Inkpen. "Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity". In: *TKDD* 2 (July 2008). DOI: 10.1145/1376815.1376819.

[23] John S. Justeson and Slava M. Katz. "Technical terminology: some linguistic properties and an algorithm for identification in text". en. In: *Natural Language Engineering* 1.1 (Mar. 1995). Publisher: Cambridge University Press, pp. 9–27. ISSN: 1469-8110, 1351-3249. DOI: 10.1017/S1351324900000048. URL: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/technical-terminology-some-linguistic-properties-and-an-algorithm-for-identification-in-text/D5F076938C4E3F24B11EDC2E831216AF (visited on 03/20/2023).

[24] Peter W. Foltz, Walter Kintsch, and Thomas K Landauer. "The measurement of textual coherence with latent semantic analysis". en. In: *Discourse Processes* 25.2-3 (Jan. 1998), pp. 285–307. ISSN: 0163-853X, 1532-6950. DOI: 10.1080/01638539809545029. URL: http://www.tandfonline.com/doi/abs/10.1080/01638539809545029 (visited on 03/21/2023).

[25] John R. Wilson and Nigel Corlett. *Evaluation of Human Work, 3rd Edition*. en. Google-Books-ID: dSmKYLp82b4C. CRC Press, Apr. 2005. ISBN: 978-1-4200-5594-8.

[26] Pontus Plavén-Sigray et al. "The readability of scientific texts is decreasing over time". In: *eLife* 6 (Sept. 2017). Ed. by Stuart King. Publisher: eLife Sciences Publications, Ltd, e27725. ISSN: 2050-084X. DOI: 10.7554/eLife.27725. URL: https://doi.org/10.7554/eLife.27725 (visited on 06/01/2023).

[27] Hemant Misra, Olivier Cappé, and François Yvon. "Using LDA to detect semantically incoherent documents". en. In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning - CoNLL '08*. Manchester, United Kingdom: Association for Computational Linguistics, 2008, p. 41. ISBN: 978-1-905593-48-4. DOI: 10.3115/1596324.1596332. URL: http://portal.acm.org/citation.cfm?doid=1596324.1596332 (visited on 11/07/2022).

[28] Mohamad Abdolahi and Morteza Zahedi. "An overview on text coherence methods". In: *2016 Eighth International Conference on Information and Knowledge Technology (IKT)*. Sept. 2016, pp. 1–5. DOI: 10.1109/IKT.2016.7777794.

[29] Kurt Lomas Tulsieram, Jose Frank Arocha, and Joon Lee. "Readability and Coherence of Department/Ministry of Health HPV Information". en. In: *Journal of Cancer Education* 33.1 (Feb. 2018), pp. 147–153. ISSN: 1543-0154. DOI: 10.1007/s13187-016-1082-6. URL: https://doi.org/10.1007/s13187-016-1082-6 (visited on 03/21/2023).

[30] S. Crossley et al. "LSA as a measure of coherence in second language natural discourse". In: 2008. URL: https://www.semanticscholar.org/paper/LSA-as-a-measure-of-coherence-in-second-language-Crossley-McCarthy/1ce17e161d38dd94e1acb75c3ab1%2005600cd158c9 (visited on 03/22/2023).

[31] Huili Wang and Danni Sui. "Measuring Coherence in Chinese EFL Majors' Writing through LSA (Latent Semantic Analysis)". In: (Mar. 2023).

[32] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. *Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change*. arXiv:1606.02821 [cs]. Sept. 2016. DOI: 10.48550/arXiv.1606.02821. URL: http://arxiv.org/abs/1606.02821 (visited on 03/20/2023).

[33] Jack Grieve, Andrea Nini, and Diansheng Guo. "Analyzing lexical emergence in Modern American English online1". en. In: *English Language & Linguistics* 21.1 (Mar. 2017). Publisher: Cambridge University Press, pp. 99–127. ISSN: 1360-6743, 1469-4379. DOI: 10.1017/S1360674316000113. URL: https://www.cambridge.org/core/journals/english-language-and-linguistics/article/analyzing-lexical-emergence-in-modern-american-english-online1/73E2D917856BE39ACD9EE3789E2BE597 (visited on 03/20/2023).

[34] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781 [cs]. Sept. 2013. URL: http://arxiv.org/abs/1301.3781 (visited on 09/19/2022).

[35] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. May 2019. DOI: 10.48550/arXiv.1810.04805. URL: http://arxiv.org/abs/1810.04805 (visited on 10/01/2022).

[36] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (visited on 06/07/2023).

[37] Diksha Khurana et al. "Natural language processing: state of the art, current trends and challenges". en. In: *Multimedia Tools and Applications* 82.3 (Jan. 2023), pp. 3713–3744. ISSN: 1573-7721. DOI: 10.1007/s11042-022-13428-4. URL: https://doi.org/10.1007/s11042-022-13428-4 (visited on 06/07/2023).

[38] Dhivya Chandrasekaran and Vijay Mago. "Evolution of Semantic Similarity—A Survey". In: *ACM Computing Surveys* 54.2 (Feb. 2021), 41:1–41:37. ISSN: 0360-0300. DOI: 10.1145/3440755. URL: http://doi.org/10.1145/3440755 (visited on 10/12/2022).

[39] Mark Davies. *The Corpus of Contemporary American English (COCA)*. 2008. URL: https://www.english-corpora.org/coca/.

[40] Sebastian Nagel. *Cc-news*. 2020.

[41] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951). Publisher: Institute of Mathematical Statistics, pp. 79–86. ISSN: 0003-4851. URL: https://www.jstor.org/stable/2236703 (visited on 10/13/2022).

[42] Michele Bevilacqua et al. "Recent Trends in Word Sense Disambiguation: A Survey". en. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4330–4338. ISBN: 978-0-9992411-9-6. DOI: 10.24963/ijcai.2021/593. URL: https://www.ijcai.org/proceedings/2021/593 (visited on 11/07/2022).

[43] Michele Bevilacqua and Roberto Navigli. "Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2854–2864. DOI: 10.18653/v1/2020.acl-main.255. URL: https://aclanthology.org/2020.acl-main.255 (visited on 11/07/2022).

[44] Juan Ramos. "Using TF-IDF to Determine Word Relevance in Document Queries". en. In: *Proceedings of the first instructional conference on machine learning* 242.1 (2003), pp. 29–48.

[45] Thomas K Landauer, Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis". en. In: *Discourse Processes* 25.2-3 (Jan. 1998), pp. 259–284. ISSN: 0163-853X, 1532-6950. DOI:

10 . 1080 / 01638539809545028. URL: http : / / www . tandfonline . com / doi / abs / 10 . 1080 / 01638539809545028 (visited on 03/22/2023).

[46] Wang Ling et al. “Two/Too Simple Adaptations of Word2Vec for Syntax Problems”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May 2015, pp. 1299–1304. DOI: 10 . 3115 / v1 / N15 - 1142. URL: https : / / aclanthology . org / N15-1142 (visited on 03/22/2023).

[47] Daniel Ramage, Christopher D. Manning, and Daniel A. McFarland. *Mapping Three Decades of Intellectual Change in Academia*. arXiv:2004.01291 [cs, stat]. June 2020. DOI: 10 . 48550 / arXiv . 2004 . 01291. URL: http : / / arxiv . org / abs / 2004 . 01291 (visited on 09/30/2022).

[48] Valerie Hase et al. “Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018) - ScienceDirect”. In: *Global Environmental Change* 70 (2021). URL: https : / / www . sciencedirect . com / science / article/pii/S0959378021001321 (visited on 06/09/2023).