Scientometrics
https://doi.org/10.1007/s11192-022-04550-z

Linguistic complexity in scientific writing: A large-scale diachronic study from 1821 to 1920

Gui Wang[1] · Hui Wang[1] · Xinyi Sun[1] · Nan Wang[2] · Li Wang[1]

Received: 20 January 2022 / Accepted: 4 October 2022

Abstract
This study intends to describe the diachronic changes of linguistic complexity (i.e., overall, morphological, and syntactic complexity) in scientific writing based on Kolmogorov complexity, an information-theoretic approach. We have chosen the entire data (i.e., all the 24 text types including articles, letters, news, etc.) and two individual registers (i.e., the full texts and abstracts of articles) of *Philosophical Transactions of the Royal Society of London*, the world's oldest scientific writing journal. The Mann–Kendall trend tests were used to capture diachronic changes in linguistic complexity at three complexity levels, and the Pearson correlation coefficients were calculated to investigate the relationships between the three complexity metrics. Results showed that the overall and morphological complexity of both the entire data and full texts increased from 1821 to 1920, indicating a massive lexical expansion during this 100-year period, as evidenced by more and more word form variants in scientific writing. In contrast, the syntactic complexity of the entire data and full texts declined, suggesting a gradual shift towards grammatical simplification in the evolution of scientific writing, particularly in word order rules and syntactic patterns. A trade-off effect has also been found between syntactic and morphological complexity in the entire data. In addition, concerning abstracts, the overall and morphological complexity decreased while the syntactic complexity increased. Drawing from these results, researchers can better understand the changing linguistic complexity styles in scientific writing, thus making adjustments in their writing accordingly to garner greater attention in academia.

Keywords Linguistic complexity · Scientific writing · Kolmogorov · Register variations

✉ Li Wang
wanglily22@shnu.edu.cn

[1] Foreign Languages College, Shanghai Normal University, 100 Guilin Road, Xuhui, Shanghai, People's Republic of China

[2] School of International Chinese Studies, Beijing Foreign Studies University, Beijing, People's Republic of China

Introduction

Scientific writing, or the written language used to report original research, has been an important part of science's symbolic expression (Atkinson, 1998). Thus, looking at how scientific writing develops over time will provide insight into how science is evolving. In addition, investigating the diachronic features of scientific writing may not only reveal the general evolutionary pattern of the scientific world but also reflect the broader social-cultural changes.

In recent years, evidence has accumulated on the changes in certain linguistic features of scientific writing over time (Biber & Gray, 2016; Biber et al., 2014; Bizzoni et al., 2020; Degaetano-Ortlieb & Teich, 2018, 2019; Degaetano-Ortlieb et al., 2018; Sun et al., 2021). Degaetano-Ortlieb et al. (2018), for instance, have examined language features (e.g., model verbs and nominal compounding) involved in diachronic change in *Royal Society Corpus* from 1665 to 1869 by using relative entropy and average surprisal. They argue that over time scientific English has become increasingly dense (i.e., linguistic constructions allowing dense packing of information are progressively used). Similarly, with information-theoretic metrics, Bizzoni et al. (2020) trace the evolution of scientific English from 1675 to 1915. Results show that while the grammatical usage consolidates over time, the lexical use dynamically oscillates due to the new scientific discoveries and register diversification. In addition to relative entropy, Sun et al. (2021) have adopted another computational method (i.e., word-embedding concreteness and imageability) to examine the developmental patterns in scientific writing and demonstrate that the evolution of scientific writing has been characterized by increasing specialization and professionalization.

However, these studies only address certain linguistic features of scientific writing, such as lexicon information and part of speech, at the lexical or grammatical level. Little attention has been paid to the diachronic changes of linguistic complexity in scientific writing.

To this end, this study intends to explore the diachronic changes of linguistic complexity regarding scientific writing based on Kolmogorov complexity, an unsupervised and information-theoretic approach. This metric utilizes the text compression technique to approximate the information content of texts (Ehret, 2021). The Kolmogorov complexity of a text is equal to the length of its shortest possible description. That is, a text compressed more efficiently is considered to be less complex, and vice versa. It is worth noting that this metric is sensitive to structural surface redundancy and regularity rather than form-meaning relationships measured in traditional linguistic complexity research. As a result, this metric may uncover not only diachronic features of lexis and grammar but also other latent features in scientific writing.

Linguistic complexity

Previous research on linguistic complexity has mainly centered around the question of whether all languages are equally complex or not. Long-standing claims insist that all languages share an equal degree of linguistic complexity (Akmajian et al., 2017; Fortson, 2010; Wells, 1954). However, McWhorter (2001) proposes that some languages are simpler than others. More and more researchers begin to acknowledge the complexity distinction among languages and provide diverse evidence from sociocultural (Kusters, 2008), historical, or geographical perspectives (Nichols, 2013).

Some other studies have examined geographical varieties of the same language from a sociolinguistic-typological perspective. For example, Juola (2008) investigates the complexity of Bible translations in different languages; Sadeniemi et al. (2008) measure the complexity of translations of the European Constitution.

Meanwhile, researchers attempt to develop new measures or improve existing measures of language complexity (Bentz & Berdicevskis, 2016; Bentz et al., 2016; Ehret, 2014). These measures could be grouped into relative complexity and absolute complexity (Miestamo, 2004). Relative complexity metrics indicate that language complexity is relevant to a language user, and often is reflected by second language acquisition difficulty. These relative metrics are evaluated primarily by the following indicators: grammar elements of a language that second language learners are usually having difficulty learning (Kusters, 2003), transparency (Steger & Schneider, 2012), or processing efficiency (Hawkins, 2009). Absolute complexity metrics are associated with system-innate properties and are usually measured from three perspectives: (1) by the number of contrasts in a system (Nichols, 2016); (2) by the number of rules in grammar (McWhorter, 2001); (3) by the length that is required to describe a linguistic system (Ehret & Szmrecsanyi, 2016). It is worth noting that Bulté and Housen (2012) have provided a more elaborate classification to reveal the multifaceted nature of linguistic complexity compared with the traditional taxonomic framework that merely covers absolute and relative complexity.

As for studies on linguistic complexity of scientific writing, they are principally conducted from a synchronic perspective, and have seldomly been diachronically scrutinized. For example, Lu et al., (2019b) have investigated the association between the linguistic complexity of scientific writing and the author's cultural background. They find that articles produced by English ethnic authors consist of longer sentences with more clauses, longer nouns, and fewer nouns compared with that of non-English ethnic authors. Lu et al., (2019a) probe into the relationship between scientific impact and the linguistic complexity of scientific writing. Similarly, Chen et al. (2020) have examined the roles of linguistic complexity indicators (e.g., title length and average sentence length) in article views and downloads. These two studies suggest that linguistic complexity plays little role in either scientific impact or article views and downloads.

One of the studies that most resemble our sphere of interest is that of Juzek et al. (2020). They examine the evolutionary patterns of syntactic complexity in scientific English through an application of universal dependencies, a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different languages. This research observes a decrease in dependency length in the *Royal Society Corpus* (RSC) and proposes that scientific English develops specific syntactic choice preferences to increase efficiency in academic communication.

These studies greatly contribute to our understanding of the diachronic features of language complexity in scientific English. However, many of these studies' complexity measurements focus on one dimension, usually the syntactic level, of the theoretical ideas that they are designed to operationalize. For instance, syntactic complexity is observed by some structural and feature-specific measures, such as universal dependencies, nominal compounding, or modal verbs. Relative clauses and prepositional phrases as noun modifiers are regarded as markers of syntactic complexity as well.

In short, few studies have provided a holistic and global metric for examining language complexity from a diachronic perspective. Against this backdrop, this article adopts the Kolmogorov complexity, an innovative information-theoretic metric of complexity, to assess text complexity in scientific English writing.

Kolmogorov complexity

The concept of Kolmogorov complexity is derived from the information theory addressing the definition and measurement of information (Der, 1997). According to Shannon (1948), who proposed the first quantitative metric of information Shannon entropy, information could be defined based on its uncertainty, that is, the entropy involved in selecting a message from a set of possibilities. The content of the information is therefore unpredictable or unexpected. Drawing on the perspective of entropy, Kolmogorov complexity measures the information content of a string of symbols instead of a series of optional messages.

To be more specific, the Kolmogorov complexity can be measured by the length of the shortest description required to retell the original text (Juola, 2008; Li et al., 2004). Although this complexity metric cannot be computed directly due to some mathematical reasons (Kolmogorov, 1968), we can approximately calculate it with an entropy estimation approach. Such an approach can be realized by file compression programs like gzip, whose algorithms are based on the structural redundancies and regularities of the running texts (strings).

Strings A, B, C, and D are shown below to illustrate how the algorithm of Kolmogorov complexity works. Although both Strings A and B contain the same number of characters, i.e., ten characters, String A can be compressed as 5 × df, containing 4 characters, whereas String B cannot be compressed as it lacks any recurring pattern. As per Kolmogorov complexity, then, String A appears to be less complex than String B. Similarly, compared with String D, String C can be described more efficiently since the pattern *there are great* appears twice.

A. dfdfdfdfdf (ten characters) – 5 × df (four characters)
B. hegvshd3p9 (ten characters) – hegvshd3p9 (ten characters)
C. There are great holes and there are great caverns in an icy mountain.
   (56 characters; adapted from String D to facilitate understanding).
D. There are great holes and caverns which are made when the ice bursts.
   (56 characters; extracted from *Royal Society Corpus 6.0 Open*).

Therefore, in this study, we employed the gzip to measure linguistic complexity at the overall, morphological, and syntactic level in terms of the information content embodied in the texts. Texts which can be compressed more efficiently signify lower linguistic complexity, while texts which cannot be compressed so efficiently suggest a higher linguistic complexity.

In linguistic terms, Kolmogorov complexity is not fully in line with traditional views of linguistic complexity, which are based on structure and specific grammatical features. For example, some grammatical patterns such as dependent clauses and relative clauses indicate a more complex writing style (Biber & Gray, 2016). Kolmogorov complexity, on the contrary, is not feature-based, but global and holistic, as it takes the entire structural complexity of sample texts into consideration. In other words, this approach is agnostic about deep linguistic form-function pairings, but assesses the structural surface redundancy, which refers to the recurrence and repetition of orthographic character sequences (structures) in a text.

Kolmogorov complexity was firstly applied by the mathematician Juola (1998) and subsequently employed by several linguists. Initially, linguistic studies that rely on the Kolmogorov complexity use parallel corpora as their primary data source, which comprise the

original sample texts alongside their translations (Ehret & Szmrecsanyi, 2016; Juola, 2008; Sadeniemi et al., 2008). Such studies aim to examine cross-linguistic complexity variations in linguistic typology works.

Subsequently, it is important to note that Ehret and Szmrecsanyi (2016) investigate the applicability of Kolmogorov complexity in non-parallel newspaper texts to assess its applicability in naturalistic corpora. Furthermore, Ehret and Szmrecsanyi (2019) apply this compression technique to naturalistic second language acquisition data. Specifically, she investigates the relationship between the complexity level of English as a second language writings and the amount of instruction received by these learner writers. In addition, drawing on the British National Corpus (BNC), Ehret (2021) also utilizes Kolmogorov complexity to assess complexity variations, for example, across written and spoken registers of British English. These studies confirm that this complexity metric is applicable to both parallel and non-parallel naturalistic corpora. However, so far, this metric has never been used to examine the diachronic changes in scientific writing.

Therefore, this study intends to explore the diachronic changes in language complexity regarding scientific writing based on Kolmogorov complexity. Specifically, the scientific writing data are extracted from *the Philosophical Transactions of the Royal Society of London (PTRS)*, the world's oldest scientific writing journal. To be more specific, we attempt to examine the trend of language complexity in the dataset from 1821 to 1920 using Kolmogorov complexity at three complexity levels (i.e., overall complexity, morphological complexity, and syntactic complexity). Thus, we may gain insight into how scientific writing changes over time in terms of linguistic complexity. The research questions are as follows:

1. What changes did scientific writing undergo from 1821 to 1920 in terms of overall complexity, morphological complexity, and syntactic complexity, respectively?
2. Are there any correlations between these three metrics of complexity?
3. Do different registers (i.e., full texts and abstracts of articles) of scientific writing experience different trends in terms of the three complexity metrics?

Methodology

Corpus data

To trace the diachronic changes in scientific writing, we used the *Royal Society Corpus (RSC) 6.0 Open* as our corpus, which was built on the *Philosophical Transactions of the Royal Society (PTRS).* This corpus comprises 17,520 transcribed scientific articles published from 1665 to 1920 and amounts to approximately 78.6 million tokens.

The reason for choosing this corpus is threefold.

Firstly, *Philosophical Transactions of the Royal Society* is the world's first science journal with the longest history and can be taken as representative of scientific writing from the seventeenth century to the early nineteenth century. Over the past three hundred years, the journal has gone through the first and second industrial revolutions, the economic crisis, and several rounds of division and merger due to competition from other journals and its gradual specialization goals. Although having undergone these changes, the PTRS has published articles continuously. With such a long and continuous history, the corpus

provides us with a window into the language style of scientific writing (in this case, the linguistic complexity) in earlier times and how it has changed over time.

Secondly, this corpus comprises various types of texts such as research articles, reports, book reviews, and letters, thus providing a valuable resource for observing how different types of scientific writing have evolved.

Thirdly, all the full texts along with their relevant meta-data such as year, author, text type, and some statistical data (e.g., the number of tokens and sentences per text) can be freely downloaded from the Royal Society Corpus homepage (https://fedora.clarin-d.uni-saarland.de/rsc_v6/index.html). Based on the metadata, subcorpora could be further extracted to accomplish our research goals. In our study, two registers are specifically extracted, the full texts and the abstracts.

It is also important to note that publications between 1665 and 1820 are missing for historical reasons (e.g., the change of editors and the first industrial revolution). Considering that such a discontinuity of the data utilized will damage the accuracy of results in statistical tests, we decided to choose 11,485 texts published between 1821 and 1920 for our study. Table 1 shows the statistical overview of the final corpus by decade.

The calculation of Kolmogorov complexity

We used gzip (GNU zip, version 1.3.12., http://gnuwin32.sourceforge.net/packages/gzip.htm) to assess the Kolmogorov complexity of each text on the overall, morphological and syntactic plane. We followed Ehret (2017) and used Kolmogorov complexity as our metric of text complexity. Additionally, the scripts for calculating Kolmogorov complexity are accessible on GitHub: https://github.com/katehret/measuring-language-complexity.

(1) *Overall complexity* The Overall complexity refers to the global structural complexity of an original text (Ehret & Taboada, 2021). To calculate the overall complexity, we first obtained two measurements for each text: the file size (in bytes) before compression and the file size (in bytes) after compression. Then, we performed linear regression analysis by taking uncompressed file size as independent variable and compressed file size as dependent variable, thus eliminating the correlation between them. This step yielded the adjusted overall complexity score (i.e., regression residual), which indicates the overall complexity level of the given sample text: higher scores suggest

Table 1 Number of texts per decade in the final corpus

Years	Number of texts
1821–1830	609
1831–1840	508
1841–1850	632
1851–1860	966
1861–1870	1152
1871–1880	1388
1881–1890	1596
1891–1900	1575
1901–1910	1674
1911–1920	1385
Total	11,485

higher overall linguistic complexity; lower scores in analogy are indicative of lower complexity.

(2) *Morphological complexity* As noted by Juola (2008), linguistic complexity at the morphological and syntactic level can be indirectly measured by distorting text files before compression. Therefore, we randomly deleted 10% (a customary percentage utilized in previous literature; see Ehret & Taboada, 2021; Juola, 1998; Sadeniemi et al., 2008) of the characters in each text prior to applying compression. Subsequently, the distorted texts are compressed to determine how well or badly the compression technique deals with the distortion. The algorithm of morphological complexity is presented in Formula (1).

$$\text{Morphological complexity score} = -\frac{\text{m}}{\text{c}} \tag{1}$$

In Formula (1), *m* represents the compressed file size after morphological distortion, and *c* is the original compressed file size. Given that morphologically complex texts tend to have a larger number of different word forms, they will be less affected by distortion compared with morphologically simple texts, in which distortion may create new and random word forms. Hence, comparatively bad compression ratios after morphological distortion indicate low morphological complexity and vice versa.

(3) *Syntactic complexity* To calculate syntactic complexity, we randomly deleted 10% of all word tokens in each text. Similar to the measurement of morphological complexity, we then compressed the distorted texts and obtained the syntactic complexity scores of given texts according to Formula (2).

$$\text{Syntactic complexity score} = \frac{\text{s}}{\text{c}} \tag{2}$$

In Formula (2), *s* is the compressed file size after syntactic distortion, and *c* is the file size before distortion. It is necessary to note that syntactic complexity in the present study is measured by word order rigidity (Bakker, 1998): rigid word order signifies syntactically complex texts, whereas free word order is indicative of syntactically simple texts. Syntactic distortion, then, disrupts word order regularities, resulting in random noise. Syntactically complex texts are greatly affected, and their compression efficiency is compromised; syntactically simple texts, in contrast, are less affected due to a lack of syntactic interdependencies that could be compromised. As a result, comparatively bad compression ratios after syntactic distortion indicate high syntactic complexity.

Data processing

Figure 1 illustrates the data processing procedures, which comprise five steps: (1) data collection, (2) data cleaning, (3) Kolmogorov complexity calculation, (4) trend analysis and visualization, (5) correlation analysis. Homemade R scripts were coded to process the procedures.

Data collection

We first gathered all the texts in each year into a single text file since the Kolmogorov complexity requires relatively large texts (i.e., texts with at least 1000 words) (Ehret &

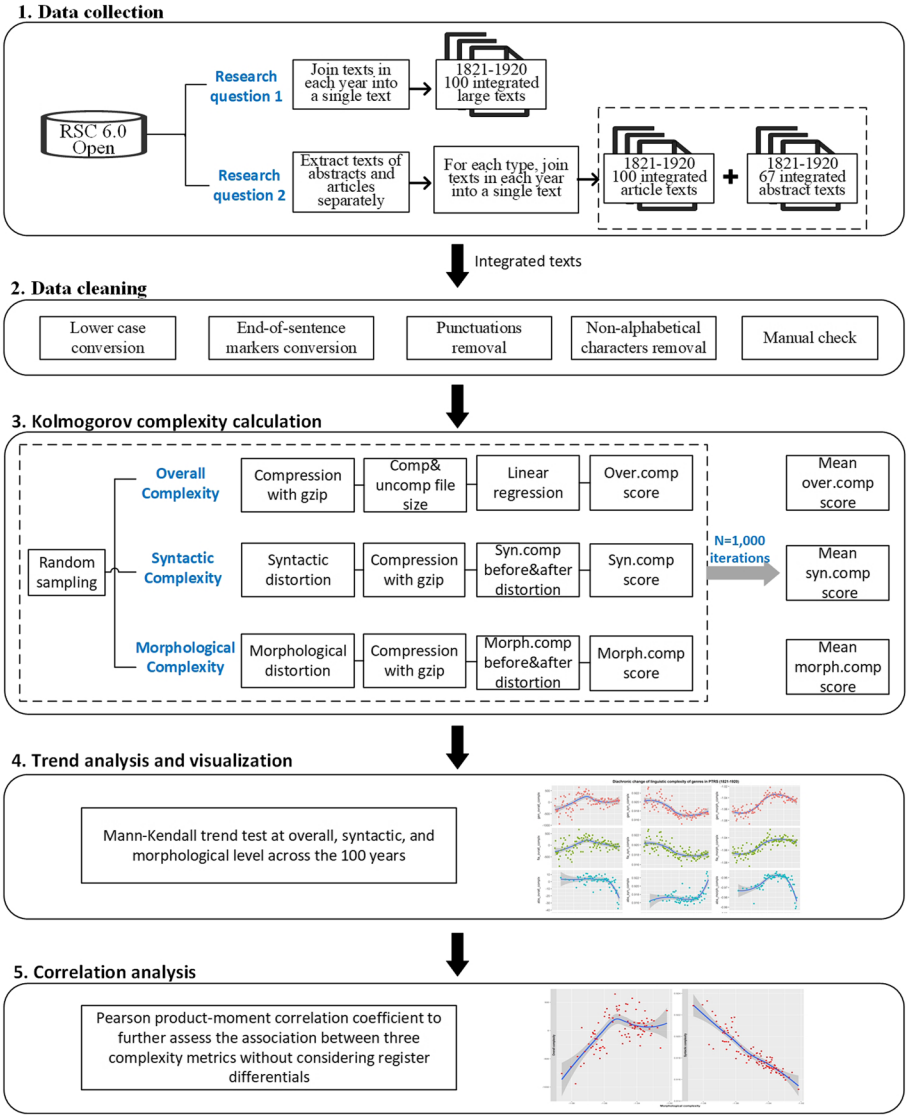

Fig. 1 Research procedures for this study

Szmrecsanyi, 2016). This step produced 100 large texts (1821–1920), and the subsequent steps were realized based on these 100 integrated texts.

We have also extracted the texts of two registers from the corpus, respectively: full texts and abstracts. We first combined texts of each micro-register in the same year into a large text, and we then obtained 100 sub-corpora for articles and 67 sub-corpora for abstracts. The reason why we have only extracted 67 sub-corpora for our present study is that there are missing abstracts in some years in this corpus.

It should be noted that there are 22 other different text types covered in the corpus, such as news, addendum, errata, letter, and advertisement. However, these types were not included in our study because on the one hand, all the 22 types account for a small proportion (9.39%) in total in terms of the number of tokens over the 100 years; on the other hand, there is a great paucity of the corresponding text data of these 22 types in many years due to some historical reasons.

Data cleaning

Data cleaning was performed by lowercasing all the running texts and removing non-alphabetical characters (e.g., numbers, UTF-8 characters, and corpus markups) and punctuations (e.g., dashes, commas, and hyphens). We did this because punctuations and non-alphabetical characters would compromise the compressibility of texts and thus increase their complexity. Notably, we retained the full stops and also replaced other end-of-sentence markings (e.g., question marks, exclamation marks, and semicolons) with full stops. This is because full stops serving as the end markers of sentences, are used to determine the linguistic units of random sampling in Step 3. In addition, we have also manually checked all the possible mistakes, especially the measurement units (e.g., *ml*, *mol*, and *cm*) due to the deletion of numbers and punctuations.

Kolmogorov complexity calculation

In order to generate a statistically robust result, we repeated the distortion and compression process for each text for 1000 times. In each iteration, we employed random sampling, that is, randomly selected 10% of the sentences per text. More precisely, the size of the random samples depends on the sentence number of the smallest text in the corpus. For example, if the smallest text in our corpus contains 500 sentences, we will then sample 50 (i.e., 500×0.1) sentences in each text per round. We did this because random sampling keeps sample size constant, thus ensuring the comparability of linguistic metrics among texts of different sizes.

To measure the overall Kolmogorov complexity, we calculated the mean file sizes before and after compression across all iterations. Subsequently, the linear regression was performed, and the adjusted overall complexity scores for each text were counted. With respect to the morphological and syntactic complexity, we firstly calculated their scores for each text file in each iteration, and the average morphological and syntactic complexity scores for each text were then computed across all iterations, respectively.

Then, we performed the same compression and iteration process of two registers as what we did for the entire data. The average adjusted overall complexity and the arithmetic mean of morphological and syntactic complexity were also calculated.

Trend analysis

Mann–Kendall trend test (Kendall, 1955; Mann, 1945), a nonparametric test, was conducted to capture the diachronic changes of linguistic complexity at three levels in English scientific writing.

To further account for the trend of morphological complexity, we have utilized MATTR version 2 (2007), a computer program, to measure the lexical diversity of all the texts. MATTR (the moving-average type-token ratio) is assumed to be a valid measure of the

lexical diversity of the entire text and is not affected by text length nor by any statistical assumptions (Covington & McFall, 2010), whereas the simple TTR (type-token ratio) is limited to the size of texts utilized (Cvrček & Chlumská, 2015).

In this study, MATTR was computed by choosing 500 words as a window length, which means that we calculated the TTR for words 1–500, 2–501, 3–502, and so on until reaching the last word of the text. The average TTR value was subsequently counted, serving as an indicator of the lexical diversity of the whole text. This step has been applied to the integrated text of each year from 1821 to 1920, which results in a total of 100 values of MATTR. Then, we performed the Mann–Kendall trend test to detect the developmental patterns of lexical diversity.

Correlation analysis

The Pearson product-moment correlation coefficients were calculated to further assess the association between overall, morphological and syntactic complexity without taking register differentials into consideration.

Results

In this section, we present our major findings. To begin with, we describe the diachronic changes in linguistic complexity (i.e., overall, morphological, and syntactic complexity) of scientific writing in the entire data. Second, we display the results of correlation analysis between three metrics of complexity in the entire data. Finally, we compare the complexity variations in terms of two registers (i.e., full texts and abstracts) of scientific writing in the corpus.

Diachronic changes in linguistic complexity

The summary statistics of the Mann–Kendall trend test in terms of the three complexity measures are presented in Table 2. Note that this trend test states that if *p* value is smaller than the significance level ($\alpha = 0.05$), the null hypothesis (H_0) will be rejected. Rejecting H_0 means that there is a significant trend in the time series data. On the other hand, if *p* value is greater than the significance level (0.05), H_0 will be accepted. Accepting H_0 indicates that no significant trend has been detected. Table 2 reveals that in all the three measures, H_0 is rejected, indicating that a significant trend in the time series data of linguistic complexity at all the three levels has been detected.

Table 2 Mann–Kendall trend test of the three complexity measures

Complexity measures	Mann–Kendall statistics	Kendall's Tau	Variance (S)	*p* value (two-tailed test)	Test interpretation
Overall complexity	754	0.152	112,750	0.025	Reject H_0 (STD)
Syntactic complexity	− 2294	− 0.463	112,750	0.000	Reject H_0 (STD)
Morphological complexity	2422	0.489	112,750	0.000	Reject H_0 (STD)

"STD" refers to "significant trend detected"

In addition, the Mann–Kendall correlation coefficient (Kendall's Tau) reveals the relationship between the time series data and complexity measures. The higher the Kendall's Tau value, the stronger the association between the time data and complexity measures will be, and vice versa. For overall (Tau = 0.152) and morphological complexity (Tau = 0.489), the Kendall's Tau values are both positive, indicating an upward trend over time, whereas the negative Kendall's Tau (Tau = − 0.463) of syntactic complexity suggests that syntax becomes less complex gradually.

The diachronic changes of three complexity metrics of the entire data across the 100 years (1821–1920) are plotted in Fig. 2. The results show that overall and morphological complexity exhibit an upward tendency from 1821 to 1920, whereas syntactic complexity undergoes a downward trend.

It is worth noting that, despite the fact that the overall complexity of scientific English has increased over the 100 years, the overall level fluctuates dramatically before 1880 but becomes stable subsequently, with only a few small fluctuations.

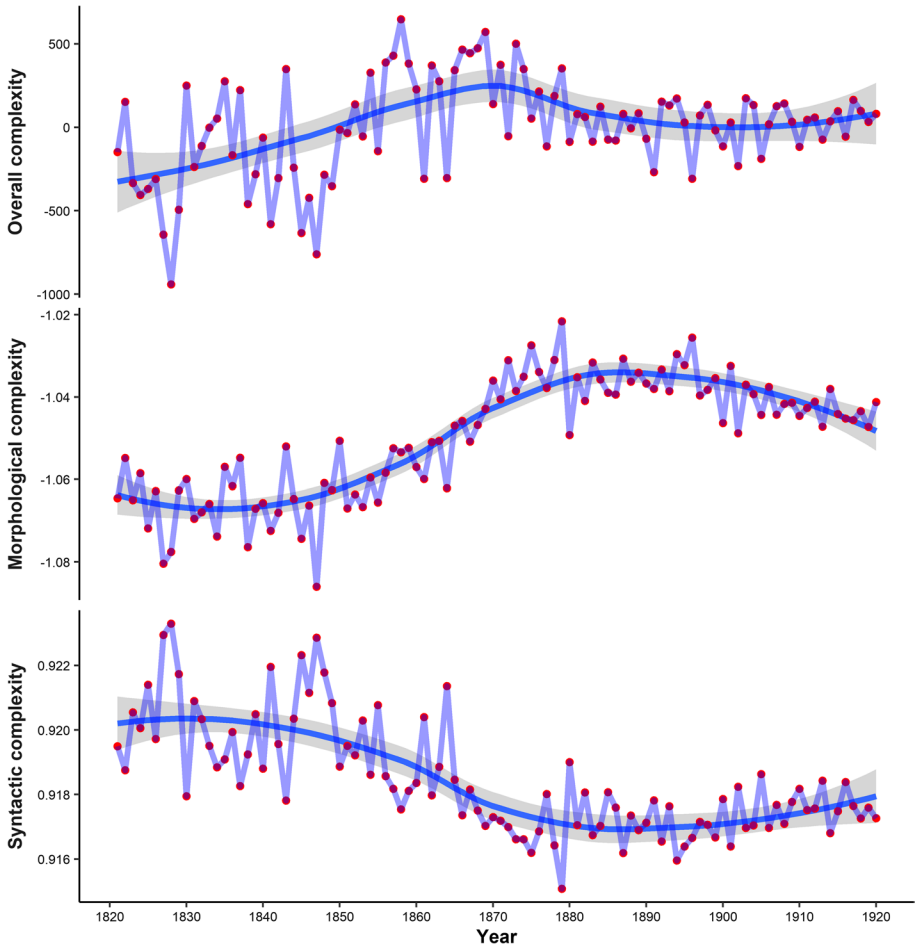

Fig. 2 Diachronic changes of overall complexity, morphological complexity, and syntactic complexity

A possible explanation for this phenomenon might lie in the occurrences of some significant historical events, which might have influenced the language use in scientific writing. To be more specific, the journal has experienced changes in editors, reviewing policies as well as the first industrial revolution before the year of 1880. Therefore, these historical events as well as the social environment might have exerted a profound impact on all facets of this scientific journal, particularly its language.

In addition, we have measured the MATTR of all the texts to further explain the increasing trend of morphological complexity. Figure 3 shows the diachronic changes of MATTR in the corpus across the 100 years. It is obvious that MATTR also shows an upward trend (Kendall's Tau = 0.133, p value = 0.05), indicating that lexical richness has increased from 1821 to 1920.

Correlation between the three complexity metrics

To further explore the interrelation between these three complexity metrics, we carried out the Pearson product-moment correlation tests. The descriptive statistics of intercorrelations in terms of the three complexity metrics are presented in Table 3. The correlations between overall complexity and morphological complexity as well as syntactic complexity and morphological complexity are plotted in Fig. 4.

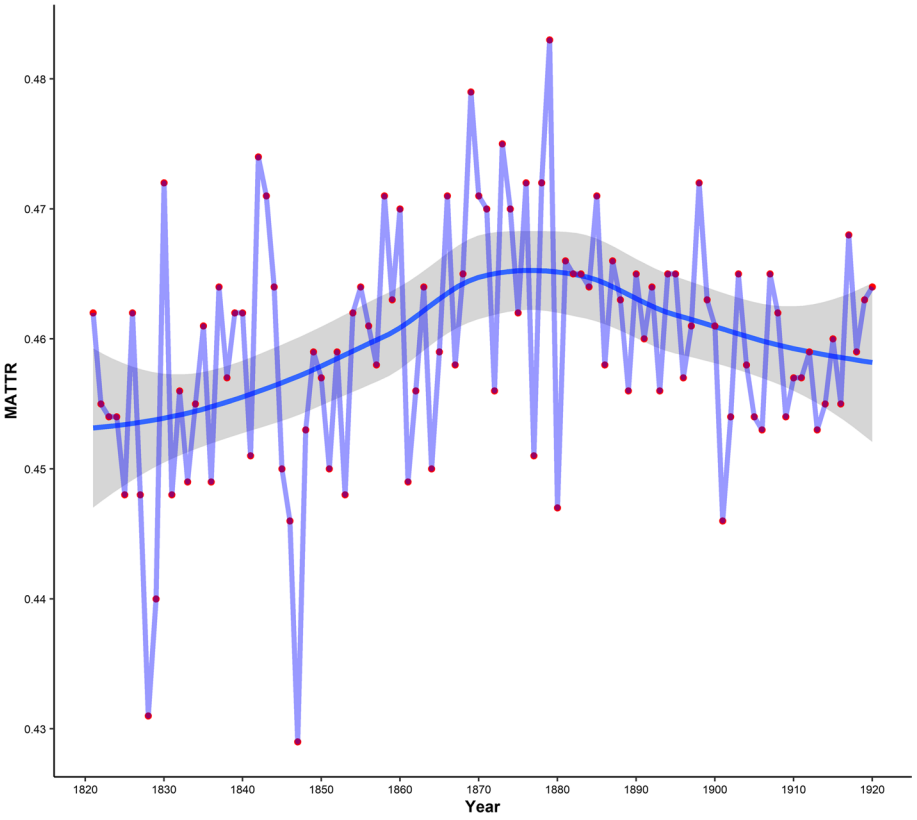

Fig. 3 Diachronic changes of MATTR in the corpus (1821–1920)

Table 3 Correlation tests of the three complexity measures

Complexity metrics	Syntactic	Morphological	Overall
Syntactic	–		
Morphological	− 0.90***	–	
Overall	− 0.72***	0.52***	–

*$p<0.05$; **$p<0.01$; ***$p<0.001$

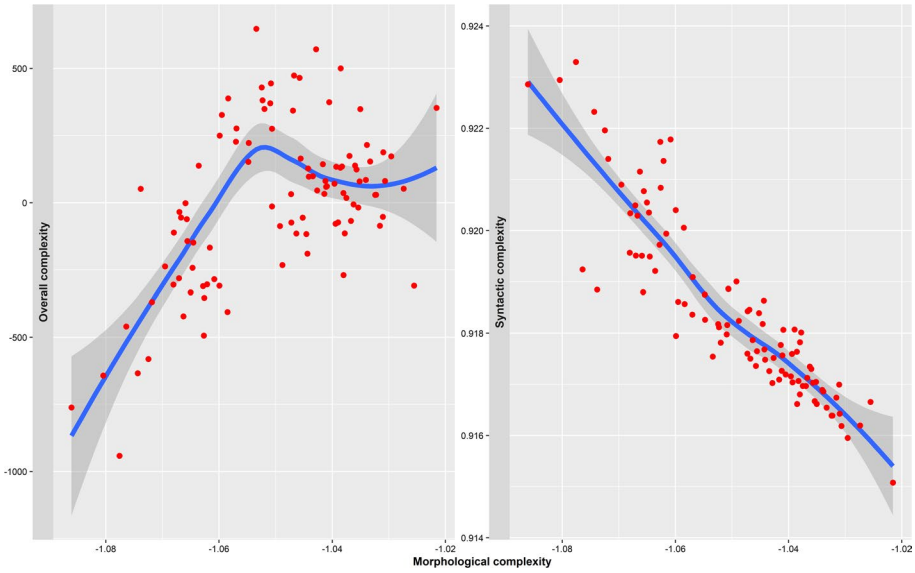

Fig. 4 Correlation between overall/syntactic complexity and morphological complexity

The results show that the three complexity metrics were significantly correlated. Specifically, significant negative correlations were found with large effect sizes ($r=-0.90$, $p<0.001$) between syntactic complexity and morphological complexity, while morphological complexity was found positively correlated with overall complexity, with medium effect sizes ($r=0.52$, $p<0.001$). In addition, a negative correlation was found between syntactic and overall complexity with a relatively large effect size ($r=-0.72$, $p<0.001$).

Complexity differences between full texts and abstracts

The analysis in the preceding section has sketched the whole picture, ignoring the specific registers of scientific writing. Thus, in this section, we attempt to reveal the historical development of two registers (i.e., full texts and abstracts) in the corpus, and the distinctions between their changes.

The diachronic changes of three complexity metrics, including trends of the entire data, full texts and abstracts are plotted in Fig. 5. Different colors of scatter points indicate different registers. Specifically, the first three plots in the first row show the changes of metrics of scientific writing of the entire data; three plots in the second row show the changes of metrics concerning full texts; the third row shows the plots of abstracts.

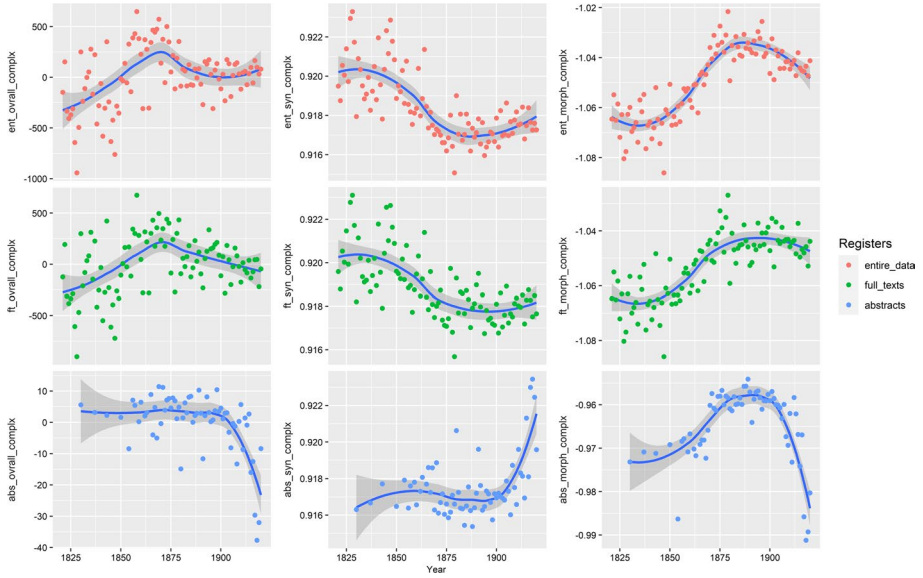

Fig. 5 Diachronic changes of complexity metrics regarding registers of scientific writing in the corpus (1821–1920)

It is evident that the trend of the three complexity measures in full texts appears to be the opposite of that in abstracts, yet it is closely aligned with that in the entire data. That is, overall and morphological complexity for full texts is increasing, whereas syntactic complexity is decreasing across the 100 years.

It is also important to note that all the three metrics of full texts fluctuate in a wild fashion before around 1870. After a closer observation of the metadata of *Royal Society Corpus (RSC) 6.0 Open*, we argue that such a violent fluctuation may result from the obscure constituents of full texts before 1870. Before 1870, the "full texts of article" genre includes not only what we define today as "full article" but also various other types of texts (e.g., letter, lecture, and astronomical observation) (Menzel et al., 2021). Therefore, the obscure nature in the corpus making of the so-called articles may have a significant impact on those complexity metrics.

However, it is interesting to find that the metrics of abstracts have experienced an almost opposite trend compared with those of articles and the entire data. To be specific, a steep decline has been observed in overall complexity and morphological complexity of abstracts, while syntactic complexity has experienced an upward trend.

Discussion

Based on the first and longest-running scientific journal in the world, the present study examined the diachronic changes of linguistic complexity in scientific writing. To our knowledge, this is probably the first study that utilizes an information-theoretic metric (here Kolmogorov complexity) to explore linguistic complexity in scientific writing from a diachronic perspective. Some possible explanations and implications of our findings will be discussed in detail below.

Standardization and professionalization in scientific writing

Results showed that the three complexity metrics of full texts are consistent with those of the entire data. Such a consistency is understandable, as full texts take the largest proportion of the whole texts of the corpus.

Specifically, the syntactic complexity experienced a downward trend in both the entire data and full texts from 1821 to 1920. Syntactic complexity, as explained in our method section, is relevant to word order rigidity and word order rules. To further illustrate, rigid word order is regarded as complex, whereas more varied word order is treated as simple. Therefore, the decreased trend of syntactic complexity in our study may indicate that the modern scientific writing is marked by more varied word order patterns than the scientific writing in the early stage.

This result coincides with the widely recognized notion that modern scientific writings are characterized by simple grammar (Gross et al., 2002; Mack, 2015). In other words, scientific writing is experiencing the evolution of grammatical simplification, particularly in word order rules and syntactic patterns. This result is in accordance with some works (Degaetano-Ortlieb et al., 2018; Juzek et al., 2020; Sun et al., 2021), in which grammatical simplification is measured from other perspectives such as universal dependencies and average surprisal for POS trigram.

Results also showed that the overall and morphological complexity of the entire data and full texts in the journal increased from 1821 to 1920, indicating that scientific writing contains more and more word form variations or has experienced a lexical expansion from 1821 to 1920. Furthermore, the upward trend exhibited by MATTR also confirmed that lexical diversity has increased over time. The present findings match our intuitive expectations about complexity in scientific writing, which is moving towards an increasingly standardized and professional genre, marked by increasing domain-specific terminologies, decreasing word order rules, and syntactic patterns (Casadevall & Fang, 2014; Houghton, 1975; Ure, 1982). These findings might be attributed to the massive use of technical terms in academic publications, as scientific writing needs efficient means of presenting and communicating its findings (Sun et al., 2021), which aids in the specialization of individual specific disciplines. Such a change also conforms to the idea that research domains in modern science are becoming more and more refined.

Trade-off between morphology and syntax

Our results show a strong negative correlation between morphological complexity and syntactic complexity, which may reveal a complementary relationship in the development of scientific communication. Specifically, while the morphology is becoming more and more complex, syntax, as a compensation, might be necessary to be simplified for the accelerated communication in academia.

Koplenig et al. (2017) also notice a statistical trade-off effect between the amount of information conveyed by the ordering of words and by internal word structure. In other words, if less information is carried within the word, more information has to be expressed through word order rules in order to communicate successfully. Sun et al. (2021) propose that as scientific writing gets professional, simplification in grammar may promote easy understanding and serve as compensation for the large-scale use of specialized terms. Thus, we argue that the decreasing syntactic complexity may serve as a counterbalance

to the increasing morphological complexity (Degaetano-Ortlieb & Teich, 2019). Such a trade-off has been observed not only in scientific communication, but also in natural language, as Yan and Liu (2021) propose the existence of the negative correlation between morphological richness and word order rigidity within Slavic languages.

Trend of linguistic complexity metrics in abstracts

Our results show that the linguistic complexity concerning abstracts undergoes almost an opposite trend compared with that of full texts and the entire data in our corpus. Specifically, morphological complexity shows an increasing trend before around 1900, whereas there has been a steep downward tendency after that. The rising tendency at the earlier stage may result from the terminology expansion due to the rapid specialization of individual academic disciplines, hence making it morphologically more complex over time. A possible explanation for the declining trend of morphological complexity after around 1900 is that more authors might promote more reader-friendly language use in the hope of attracting greater attention from both editors and non-academic groups.

Similar to morphological complexity, the overall complexity of abstracts has also seen a decreasing trend after around 1900. As argued by Ehret (2021), informal texts are marked by a lower overall complexity. Thus, the downward trend of overall complexity may indicate the gradual decreasing levels of formality in abstracts, which is congruent with Hundt and Mair (1999), who have pointed out the stepwise shift of academic writing toward gradual informality from the 1960s to the 1990s. Our finding is also partially consistent with Hyland and Jiang (2017). Their results show an increased use of informal elements such as first person pronouns and sentences beginning with conjunctions in the science and engineering disciplines or hard science. It is worth noting that Hyland and Jiang (2017) conclude their findings by examining the full texts of articles, whereas we pay attention primarily to the abstracts. The increasingly important role of science communication may account for the sudden decreased overall complexity as well. In other words, it is not reasonable even for a learned editor to have good acquaintance with much fine-grained subfields, let alone for the general public. Therefore, to boost scientific impact of their research, authors may adopt more accessible and reader-friendly (i.e., lower overall complexity) language in their abstracts, which are the most read part of an article (Pitkin, 1999) and may determine whether the general public continues reading the full text.

In contrast, the syntactic complexity of abstracts experienced an upward trend, which might result from the increasing standardization of abstract format. By standardization here, we mean the construction of abstracts (e.g., introduction, methodology, results, and implication) are becoming conventionalized and standardized. With the gradual standardization of abstracts, authors are accustomed to utilizing regular expressions and syntactic patterns in a structured abstract. This may result in the wide application of some grammatical patterns and word order rules, and thus lead to increased syntactic complexity to some degree.

Conclusions

Our study has employed a holistic and information-theoretic Kolmogorov complexity to explore the diachronic changes of linguistic complexity in scientific writing. Based on publications in *Philosophical Transactions of the Royal Society* from 1821 to 1920, we have

investigated the historical development of linguistic complexity at overall, morphological, and syntactic levels, and examined the correlation between these three complexity metrics. Furthermore, we explored the complexity differences of two registers (i.e., full texts and abstracts) in the corpus as well.

Results show that the entire data in scientific writing is moving towards formality and professionality, which is embodied in all three metrics. The increased morphological complexity indicates the increasing application of terminology due to the increasing refinement of disciplines. The decreased trend of syntactic complexity may result from the simplification of word order rules and syntactic patterns, which can speed scientific communication. The increased overall complexity reveals a trend toward formality in scientific writing, as formal registers are overall and morphologically more complex than less formal registers, but less complex regarding syntax (Ehret, 2021).

Besides, our exploration of the complexity differences concerning different registers of scientific writing also reveals some interesting results. Specifically, we found that the trend of full texts is closely consistent with the trend of the entire data, which is marked by standardization and professionality. However, the diachronic features of three metrics in abstracts are almost distinct from those of full texts and the entire data. The increasing syntactic complexity could be attributed to the gradual conventionalization of the construction of abstracts; the decreasing overall and morphological complexity may result from the authors' increasing awareness to make their articles more accessible to readers and thus more influential in academia.

The stylistic shifts reported in our study also have some significant implications for researchers. Specifically, researchers can refine and polish their language use accordingly to increase the chances of getting their articles published. In addition, the register differences between abstracts and full texts can also inspire researchers to use reader-friendly and accessible language with the hope of promoting their work to the general public.

The strength of our research lies in our research methodology. Specifically, Kolmogorov complexity is more holistic and comprehensive than the traditional measures such as type-token ratio and the number of subordinations, which only depict several given linguistic features or only cover certain aspects of linguistic level (e.g., grammatical and lexical level). This complexity measure does not capture the recurrence of arbitrarily selected features but reveals the linguistic complexity of texts as a whole.

The major limitation of this study is that Kolmogorov complexity does not work well with short texts, which count less than 1000 words. Thus, in this study, we are unable to further scrutinize all registers of scientific writing with small text sizes except full texts and abstracts. In addition, the Kolmogorov complexity scores are inherently relative. In other words, they are only meaningful when compared with other groups of scores or in the context of ranking. In addition, this study used the data from only one journal that focuses on the publications of hard science such as those in chemistry and astronomy.

In the future, our findings could be enhanced by examining other journals, notably those that centered on art and humanities, or by examining the diachronic changes of Kolmogorov complexity over a longer period of time, thus establishing the validity and reliability of this methodology.

Author contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by SXY, and data processing was carried out by all authors. The first draft of the manuscript was written by WG, SXY, WN, and WH. All authors especially WL commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by the National Social Science Foundation of China (No. 17BYY115).

Data availability All data supporting the conclusions of this article are included within the article (and its additional files).

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Ethical approval This study did not involve humans and/or animals; there is no need for institutional ethics review board approval.

References

Akmajian, A., Farmer, A. K., Bickmore, L., Demers, R. A., & Harnish, R. M. (2017). *Linguistics: An introduction to language and communication*. The MIT Press.

Atkinson, D. (1998). *Scientific discourse in sociohistorical context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Routledge.

Bakker, D. (1998). Flexibility and consistency in word order patterns in the languages of Europe. In A. Siewierska (Ed.), *Constituent order in the languages of Europe* (pp. 383–420). De Gruyter Mouton. https://doi.org/10.1515/9783110812206.383

Bentz, C., & Berdicevskis, A. (2016, December 1). *Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence*. ACLWeb; The COLING 2016 Organizing Committee. http://www.aclweb.org/anthology/W16-4125

Bentz, C., Ruzsics, T., Koplenig, A., & Samardžić, T. (2016, December 1). *A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora*. ACLWeb; The COLING 2016 Organizing Committee. http://www.aclweb.org/anthology/W16-4117

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English Linguistic change in writing*. Cambridge University Press.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence, 3*, 73. https://doi.org/10.3389/frai.2020.00073

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 23–46). John Benjamins.

Casadevall, A., & Fang, F. C. (2014). Specialized science. *Infection and Immunity, 82*(4), 1355–1360.

Chen, B., Deng, D., Zhong, Z., & Zhang, C. (2020). Exploring linguistic characteristics of highly browsed and downloaded academic articles. *Scientometrics, 122*(3), 1769–1790. https://doi.org/10.1007/s11192-020-03361-4

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Cvrček, V., & Chlumská, L. (2015). Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics, 39*(3), 309–325. https://doi.org/10.1007/s11185-015-9151-8

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., & Teich, E. (2018). An information-theoretic approach to modeling diachronic change in scientific English. In *From data to evidence in English language research* (pp. 258–281). Brill.

Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 22–33.

Degaetano-Ortlieb, S., & Teich, E. (2019). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0088

Der, V. (1997). *Information theory*. Cambridge University Press.

Ehret, K. (2014). Kolmogorov complexity of morphs and constructions in English. *Linguistic Issues in Language Technology*. https://doi.org/10.33011/lilt.v11i.1363

Ehret, K. (2017). *An information-theoretic approach to language complexity: Variation in naturalistic corpora. Doctoral dissertation*. Freiburg im Breis: University of Freiburg.

Ehret, K. (2021). An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory, 17*(2), 383–410. https://doi.org/10.1515/cllt-2018-0033

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, isolation, and variation* (pp. 71–94). de Gruyter.

Ehret, K., & Szmrecsanyi, B. (2019). Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research, 35*(1), 23–45. https://doi.org/10.1177/0267658316669559

Ehret, K., & Taboada, M. (2021). The interplay of complexity and subjectivity in opinionated discourse. *Discourse Studies, 23*(2), 141–165. https://doi.org/10.1177/1461445620966923

Fortson, B. W. (2010). *Indo-European language and culture: An introduction*. Wiley-Blackwell.

Gross, A. G., Harmon, J. E., & Reidy, M. (2002). *Communicating science: The scientific article from the 17th century to the present*. Oxford University Press.

Hawkins, J. A. (2009). *An efficiency theory of complexity and related phenomena*. Oxford University Press.

Houghton, B. (1975). *Scientific periodicals: Their historical development, characteristics and control*. Bingley.

Hundt, M., & Mair, C. (1999). Agile" and "uptight" genres. *International Journal of Corpus Linguistics, 4*(2), 221–242. https://doi.org/10.1075/ijcl.4.2.02hun

Hyland, K., & Jiang, F. (2017). Is academic writing becoming more informal? *English for Specific Purposes, 45*, 40–51. https://doi.org/10.1016/j.esp.2016.09.001

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics, 5*(3), 206–213. https://doi.org/10.1080/09296179808590128

Juola, P. (2008). Assessing linguistic complexity. In *Language Complexity: Typology, contact, change* (pp. 89–108). John Benjamins Publishing. https://doi.org/10.1075/slcs.94.07juo

Juzek, T. S., Krielke, M.-P., & Teich, E. (2020). Exploring diachronic syntactic shifts with dependency length: the case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 109–119.

Kendall, M. G. (1955). *Rank correlation methods second edition, revised and enlarged*. Charles Griffin & Co.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information*. *International Journal of Computer Mathematics, 2*(1–4), 157–168. https://doi.org/10.1080/00207166808803030

Koplenig, A., Meyer, P., Wolfer, S., & Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure: Large-scale evidence for the principle of least effort. *PLoS ONE, 12*(3), e0173614. https://doi.org/10.1371/journal.pone.0173614

Kusters, W. (2003). *Linguistic complexity: the influence of social change on verbal inflection*. Lot.

Kusters, W. (2008). Complexity in linguistic theory, language learning and language change. In *Language complexity: Typology, contact, change* (pp. 3–22). John Benjamins. https://www.jbe-platform.com/content/books/9789027291356-slcs.94.03kus

Li, M., Chen, X., Li, X., Ma, B., & Vitanyi, P. M. B. (2004). The similarity metric. *IEEE Transactions on Information Theory, 50*(12), 3250–3264. https://doi.org/10.1109/tit.2004.838101

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimoto, C. R., Paul, L., & Zhang, C. (2019a). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829. https://doi.org/10.1016/j.joi.2019.07.004

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019b). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology, 70*(5), 462–475. https://doi.org/10.1002/asi.24126

Mack, C. (2015). 350 years of scientific journals. *Journal of Micro/nanolithography, MEMS, and MOEMS, 14*(1), 010101. https://doi.org/10.1117/1.jmm.14.1.010101

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica, 13*(3), 245. https://doi.org/10.2307/1907187

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typol., 5*, 2–3. https://doi.org/10.1515/lity.2001.001

Menzel, K., Knappen, J., & Teich, E. (2021). Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics, 9*(1), 1–18. https://doi.org/10.32714/ricl.09.01.02

Miestamo, M. (2004). On the feasibility of complexity metrics. *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics, Tallinn*, 11–26.

Nichols, J. (2013). The vertical archipelago: Adding the third dimension to linguistic geography. In *Space in Language and Linguistics* (pp. 38–60). De Gruyter.

Nichols, J. (2016). Complex edges, transparent frontiers: Grammatical complexity and language spreads. In *Complexity, isolation, and variation* (pp. 117–138). de Gruyter.

Pitkin, R. M. (1999). Accuracy of data in abstracts of published research articles. *The Journal of the American Medical Association, 281*(12), 1110. https://doi.org/10.1001/jama.281.12.1110

Sadeniemi, M., Kettunen, K., Lindh-Knuutila, T., & Honkela, T. (2008). Complexity of European Union Languages: A comparative approach*. *Journal of Quantitative Linguistics, 15*(2), 185–211. https://doi.org/10.1080/09296170801961843

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Steger, M., & Schneider, E. W. (2012). Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 156–191). De Gruyter. https://doi.org/10.1515/9783110229226.156

Sun, K., Liu, H., & Xiong, W. (2021). The evolutionary pattern of language in scientific writings: A case study of philosophical transactions of royal society (1665–1869). *Scientometrics, 126*(2), 1695–1724. https://doi.org/10.1007/s11192-020-03816-8

Ure, J. (1982). Introduction: Approaches to the study of register range. *International Journal of the Sociology of Language, 1982*, 35. https://doi.org/10.1515/ijsl.1982.35.5

Wells, R. (1954). Archiving and language typology. *International Journal of American Linguistics, 20*(2), 101–107. https://doi.org/10.1086/464260

Yan, J., & Liu, H. (2021). Morphology and word order in Slavic languages: Insights from annotated corpora. *Voprosy Jazykoznanija, 4*, 131. https://doi.org/10.31857/0373-658x.2021.4.131-159