





ARTICLE



<https://doi.org/10.1057/s41599-025-04742-x>

OPEN

The impact of research paradigms on encoding and decoding difficulties of applied linguistics articles

Jiayi Wu¹, Hongzhong Chen²  & Lang Chen³ 

The diachronic change in the information load and the textual readability of research articles has recently attracted much scholarly interest. However, existing research has either focused on disciplinary variations or treated discipline as a homogeneous category. Little attention has been given to the potential interaction between the research paradigm (quantitative vs. qualitative) and the diachronic change within a specific discipline. To address this research gap, this study investigates the cognitive encoding and decoding difficulties in 160 Applied Linguistics research articles cutting across two historical periods (1981–1985 vs. 2011–2015) and two research paradigms (quantitative vs. qualitative). These research articles were randomly selected from four prestigious journals in Applied Linguistics, and the cognitive encoding and decoding difficulties were operationalized as information entropy and mean dependency distance of a research article, respectively. Statistical analyses with a MANOVA and two follow-up univariate ANOVAs show that time and research paradigm combined can significantly explain a large proportion of the variance in the two cognitive difficulty indices. Specifically, qualitative research articles consistently exhibit higher cognitive encoding complexity than their quantitative counterparts in both periods while they both experienced significant increases in this metric. However, regarding the cognitive decoding difficulty, only quantitative research articles have experienced a significant rise. As a result, quantitative research articles have become higher in cognitive decoding difficulty than qualitative ones in the second historical period whereas no paradigmatic difference is found in the first period. These findings are discussed by considering the distinctive epistemological assumptions of the two research paradigms and against the background of ever-growing publication pressure. Hypotheses regarding the effect of the promotional language on both cognitive encoding and decoding difficulties are proposed to address the discrepancy between findings in this study and those in previous research. This study contributes to existing literature by revealing the nuanced patterns in academic writing within a specific discipline, thereby deepening our understanding of disciplinary writing.

¹School of Foreign Languages and Cultures, Nanjing Normal University, Nanjing, China. ²Institute for Language Sciences, Utrecht University, Utrecht, The Netherlands. ³Solearn International Pte. Ltd, Singapore, Singapore. ✉email: michaellangchen@gmail.com

Introduction

Academic writing, particularly in the form of research articles (RAs), plays a crucial role in disseminating knowledge and advancing scholarly discourse. Recently, there has been growing interest in understanding how academic writing has evolved, especially in response to the increasing pressure to publish in an era of “attention economy” in academia (Hyland, 2023). This changing landscape influenced how researchers construct and present arguments, potentially affecting both the information load and readability of RAs. Consequently, diachronic changes in the complexity of academic texts have become a subject of extensive research. Most studies have examined cognitive decoding difficulty—the challenges readers face during interpretation—through metrics such as lexical complexity (Heng et al., 2022; Kalantari and Gholami, 2017; Zareva, 2019), syntactic complexity (Kyle and Crossley, 2018; Lei et al., 2023; Yin et al., 2023), and grammatical complexity (Biber and Gray, 2016; Biber et al., 2024), with the exception of Xiao et al. (2023) and Zhao et al. (2023), which incorporate the cognitive encoding difficulty—the challenges writers encounter in text creation and information encoding—in their examination of the disciplinary variations and diachronic changes in RA abstracts. However, all these studies have either focused on broad disciplinary variations or treated disciplines as homogeneous categories, overlooking potential intra-disciplinary differences.

One significant and underexplored factor is the interaction between research paradigms (quantitative vs. qualitative) and diachronic changes within a specific discipline. Previous research has demonstrated that RAs within the same discipline but adopting different research paradigms exhibit distinct textual characteristics due to differing epistemological assumptions (Chen and Hu, 2020; Hu and Cao, 2015). Given that the underlying drive behind the disciplinary difference in academic writing is also largely distinctive epistemological assumptions between disciplines, it could be hypothesized that quantitative and qualitative RAs may also exhibit distinctive changing patterns in the encoding and decoding difficulties. To verify this hypothesis, the present study investigates diachronic changes in the cognitive encoding and decoding difficulties of Applied Linguistics RAs across two time periods (1981–1985 and 2011–2015) and two research paradigms (quantitative and qualitative). By focusing on a single discipline while differentiating between research paradigms, this study aims to reveal nuanced patterns in academic writing that may have been obscured in broader cross-disciplinary analyses.

Following Zhao et al. (2023), we operationalize the cognitive encoding difficulty as information entropy (IE) and the decoding difficulty as mean dependency distance (MDD). These measures quantify the information load and structural complexity of texts, respectively. Our analysis compares these metrics across quantitative and qualitative RAs, revealing how different research traditions within Applied Linguistics have evolved in their cognitive demands on both writers and readers. We examined the same two metrics as those investigated in Zhao et al. (2023) because their study inspires the present one for two reasons: first, Zhao et al. (2023) have so far been the only study that analyses the changing patterns of both encoding and decoding difficulties in academic texts and we consider it necessary to integrate difficulties encountered by writers with those by readers for a complete understanding of the changing patterns in academic writing. Second, Zhao et al. (2023) observe a trade-off between cognitive encoding and cognitive decoding difficulty in abstracts of social sciences, and we would like to argue that further research is needed to find out whether their finding applies to full-length RAs adopting different research paradigms within a specific social science, in our case, Applied Linguistics.

This study can potentially contribute to our current understanding of academic writing practices by adding the nuanced effect of research paradigms on the changing patterns of textual complexity. The findings have implications for academic writing instruction, editorial practices, and our broader conceptualization of how knowledge is constructed and communicated in different research traditions.

Literature review

Text complexity: Relative versus absolute. Text complexity can be classified into two categories, namely relative complexity and absolute complexity (Bulté and Housen, 2012, 2014; Housen et al., 2019; Li and Yang, 2023), even though little consensus has been reached on their precise definition (Biber et al., 2024). According to Bulté and Housen (2012), the relative complexity refers to linguistic features or systems that impose cognitive demands on users and learners, particularly in terms of the mental effort or resources required for processing and internalizing these linguistic elements. Bulté and Housen (2012) also term the relative complexity as cognitive complexity. However, we decide not to use the latter term for two reasons. First, any type of complexity imposes cognitive load on language users, and thus, logically speaking, all types of complexity are cognitive complexity. Second, the two complexity measures used in Zhao et al. (2023), which are IE and MDD, are also considered cognitive difficulty measures, but according to their definitions, they are absolute complexity measures (we will discuss absolute complexity below). To avoid terminology confusion, therefore, in the present study, we will use the terms of relative and absolute complexities for the opposition between two grand groups of complexity measures. Regarding the two complexity indices to be investigated in this study, we will refer to them as cognitive encoding and decoding difficulties, respectively, to align with the terms adopted in Zhao et al. (2023).

As the relative complexity is defined in relation to language users and focuses on language features or systems themselves instead of their frequency distribution in a text, it is often assessed experimentally with a focus on the acquisition order or reading comprehension of specific linguistic features. For example, psycholinguistic studies (e.g., Byrnes and Sinicrope, 2008; Diessel, 2004) have shown that complex structures like embedded clauses and passives are more challenging to process and are acquired later than simpler structures such as coordinate and active constructions. Similarly, Spooren and Sanders (2008) found that coherence relations of lower cognitive difficulty are acquired earlier than those of higher cognitive difficulty based on two empirical studies on child language development. Focusing on causal relations, Mak and Sanders (2013) observed a differential effect of causal expectations on the processing of referential and relational coherence with two eye-tracking experiments. Specifically, they found that causal expectations based on the first sentence led to faster processing of pronouns referring to the expected referent at the start of the second sentence, even without an explicit causal connective. Additionally, causal relatedness between sentences facilitated integration throughout the second sentence, but only when the content actually allowed a causal interpretation, suggesting readers do not proactively compute causal coherence before the relation can be inferred from the content.

The absolute complexity, on the other hand, is defined as the quantitative measure of discrete components within a language feature or system and the interconnections between these components (Bulté and Housen, 2012). As absolute complexity measures can be obtained quantitatively from the frequency

distribution of target linguistic features, that is, they are learner-independent, research in this domain often employs corpus-based methods to examine the relationships between complexity indices and writing quality. For example, based on a corpus of second-language students' writing essays, Lu (2011) found that the complex nominals per clause and mean length of clause are the best syntactic complexity measures among the 14 indices he tested for predicting writing proficiency. Similarly, after comparing three types of syntactic complexity indices in a corpus of second-language argumentative essays, Kyle and Crossley (2018) concluded that fine-grained phrasal indices (e.g., number of dependents per prepositional object and dependents per object of the preposition) perform better in predicting the academic writing quality than traditional clausal indices (e.g., mean length of clause and coordinate phrases per clause) and fine-grained clausal indices (e.g., number of subjects per clause and adjective complement). Recently, MDD has been considered as an even more effective indicator of writing proficiency compared to traditional syntactic complexity indices as Ouyang et al. (2022) have shown with a corpus-based study comparing the MDDs of texts produced by beginner, intermediate, and advanced EFL (English as a Foreign Language) learners.

Given that academic RAs are produced and consumed by highly proficient writers, research on text complexity of RAs typically focus on absolute complexity measures. In the next subsection, we will review some representative studies on complexity studies of RAs, identify research gaps based on the review, and list our research questions accordingly.

Textual complexity of RAs. Studies on the absolute complexity of RAs have primarily examined disciplinary variations and diachronic change. Disciplinary differences often stem from distinctive epistemological assumptions, particularly those between soft and hard disciplines, which shape linguistic choices. For example, Parviz et al. (2020) conducted a comparative analysis of phrasal complexity features (PCFs) in the results sections of research articles from Applied Linguistics and Physics, revealing several notable disciplinary differences. Specifically, they found that Applied Linguistics articles employed significantly more nominalizations and appositive noun phrases compared to Physics articles. In contrast, Physics articles exhibit more complex patterns of pre-modification, including quintipartite pre-modifiers and multitudinous hyphenated adjectives, which were absent in Applied Linguistics texts. In a similar vein but focusing on the discussion sections, Ziaieian and Golparvar (2022) compared RAs from applied linguistics, chemistry and economics in terms of 16 clausal and phrasal features and found that applied linguistics and economics showed higher clausal complexity, while chemistry demonstrated greater phrasal complexity. Adopting a more nuanced discipline classification and with a focus on hard sciences, a recent study by Zhou et al. (2023) found no significant variation in overall syntactic complexity between hard-pure (Chemistry, Mathematics, and Physics) and hard-applied (Chemical Engineering, Computer Science, and Mechanical Engineering) disciplines after investigating the relationship between syntactic complexity features and rhetorical functions in science research article introductions across these six disciplines. However, they pointed out that comparison to previous studies still suggests variation between science and social science disciplines. Recently, researchers began to investigate complexity in RAs from the writers' perspective, examining the cognitive encoding difficulty in disciplinary writing. For example, Xiao et al. (2023) employed an entropy-based approach to examine disciplinary differences in research article introductions. Their analyses reveal distinctive patterns across disciplines.

Specifically, natural sciences introductions were found to exhibit higher information content in Move 1 (Establishing a Territory), reflecting the field's emphasis on cumulative knowledge and strong links to previous work. In contrast, social sciences introductions boast significantly higher information content in Move 3 (Occupying the Niche), which the authors attribute to the more discursive and interpretive nature of social sciences. Humanities introductions generally fall between natural and social sciences in terms of information content across moves.

The other frequently researched aspect of complexity in RAs is how it has evolved over time. For example, Zhou et al. (2023) examined diachronic changes in three lexical complexity indices, namely lexical density, lexical sophistication, and lexical diversity, of articles published in *Nature Biology Letter* from 1929 to 2019 and found significant increases in all three dimensions across this 100-year period. The authors interpreted their findings as a trend of academic language towards greater compression and conventionalization. There have also been studies that investigated the diachronic change in text complexity of RAs by considering multiple disciplines. For instance, Biber and Gray (2010), through analysing a custom-built corpus of academic research articles spanning from 1965 to 2005, demonstrated that academic writing has shifted away from clausal elaboration toward more phrasal/nominal structures over time. This shift is evident in the increased use of nouns as pre-modifiers, prepositional phrases as noun modifiers, and appositive noun phrases, particularly during the 20th century. Considering both cognitive encoding and decoding difficulties, Zhao et al. (2023) investigated the diachronic change in text difficulty of research article abstracts across natural sciences, social sciences, and humanities using a novel cognitive information-theoretic approach. Their results show that over the past two decades, the cognitive encoding difficulty of abstracts as represented by IM has generally increased, while the cognitive decoding difficulty measured by MDD has decreased. Disciplinary variations were also observed, with humanities showing no significant change in encoding difficulty and social sciences showing no significant change in decoding difficulty. The authors attributed these findings to factors such as the informative and promotional traits of abstracts, the accumulative nature of academic knowledge, and the mechanism of dependency distance minimization in human languages.

While the studies reviewed above (and other similar studies) have offered valuable insights into broad disciplinary differences and overall temporal trends, their primary focus on comparisons between major disciplinary categories may result in overlooking important variations within disciplines, particularly those stemming from different research paradigms - specifically, quantitative versus qualitative approaches. Many academic fields, especially in the social sciences and some areas of the humanities, employ both quantitative and qualitative methodologies. Given previous research showing that RAs within the same discipline but adopting different research paradigms exhibit different textual characters (Chen and Hu, 2020; Hu and Cao, 2015), it can be hypothesized that these distinct approaches to research may well be associated with different patterns of linguistic complexity, information structure, and rhetorical organization in RAs. To address this research lacuna, thereby providing a more nuanced understanding of how diverse research approaches within a discipline shape the complexity and presentation of academic writing, this study investigates the cognitive encoding and decoding difficulties in 160 Applied Linguistics RAs cutting across two historical periods (1981–1985 vs. 2011–2015) and two research paradigms (quantitative vs. qualitative). These RAs were randomly selected from four prestigious journals in Applied Linguistics, and, following (Zhao et al., 2023), the cognitive

Table 1 Descriptive statistics on the corpus ¹ .					
	Quantitative		Qualitative		Total
	1981-1985	2011-2015	1981-1985	2011-2015	
No. of RAs	40	40	40	40	160
Total words	150,900	313,869	216,027	308,388	989,184
Mean word count	3773	7847	5401	7710	6406
Std. of word counts	1143	1559	1988	1356	2366
Total sentences	7002	13,342	8987	11,640	40,971
Mean sentence count	175.050	333.550	224.675	291.000	256.069
Std. of sentence counts	47.976	75.216	83.180	65.827	91.818

encoding and decoding difficulties were operationalized as IM and MDD of an RA, respectively.

The present study aims to answer the following four questions:

1. Is there a significant difference in the IM between RAs in Applied Linguistics published from 1981 to 1985 and those published from 2011 to 2015?
2. Do qualitative and quantitative RAs in Applied Linguistics show the same changing pattern in IM?
3. Is there a significant difference in the MDD between RAs in Applied Linguistics published from 1981 to 1985 and those published from 2011 to 2015?
4. Do qualitative and quantitative RAs in Applied Linguistics show the same changing pattern in MDD?

Methods

Corpora. This study utilized a corpus of 160 full-length empirical RAs in Applied Linguistics, compiled as part of a larger project examining metadiscourse use across disciplines and research paradigms (Chen and Hu, 2020; Hu and Chen, 2019). We decided to focus on one specific discipline, instead of incorporating more disciplines, for two reasons. Firstly, as mentioned in the literature review section, the intra-disciplinary difference, potentially resulting from paradigmatic effect, is largely an overlooked topic. Secondly, some recent studies have found that discipline may not be an optimal indicator for variations in academic writing. For example, using cluster analysis on appraisal markers, Zhang and Cheung (2023) found research articles from four disciplines, which are Chemistry, Geoscience, Education, and Management, can be regrouped by their use of appraisal features, calling for caution against “the foregrounding of disciplinary influence in writing research”. Similarly, a conference reported at AAAL2024 by Thompson and Gray (2024) also identified complex cross-discipline cluster among 14,024 research articles from 10 disciplines, challenging the assumption that the primary drive of variation among RAs is discipline.

The 160 RAs were randomly sampled from four prestigious journals in applied linguistics: Applied Linguistics, Modern Language Journal, TESOL Quarterly, and Language Learning. All the four selected journals take language learning as their primary scope, which further reduces the effect of confounding variables. The corpus was stratified by time period and research paradigm. We selected 80 articles from 1981–1985 and 80 from 2011–2015, representing a 30-year span. Due to the limited availability of qualitative or quantitative articles in certain years - such as only 24 qualitative studies in 1981 and 10 in 1985—we opted for five-year spans instead of annual data points. The initial period, 1981–1985, was chosen because the journal Applied Linguistics was established in 1980, and excluding such an influential journal in applied linguistics would omit significant data. Following recent studies in English for Academic Purposes (e.g., Chen & Hu, 2020; Hyland & Jiang, 2021, 2023; Wang & Hu, 2024), we selected a 30-year interval

between periods, deemed sufficient for observing changes. Thus, our time point selection was guided by both practical constraints and established practices.

To stratify the corpus by research paradigm, we first coded all empirical articles from the four journals within the two selected time periods as qualitative, quantitative, or mixed methods based on the characteristics of their data collection and analysis methods (Creswell, 2013; Dörnyei, 2007; Wen, 2001). There are also non-empirical articles, including some book reviews and conceptual articles. They were coded accordingly but were not included in the pools for random sampling. To ensure coding accuracy, another scholar in Applied Linguistics was invited to code a subset of the 1,149 articles with the third author. A Cohen’s kappa was calculated to assess the inter-rater agreement. Since the result ($k = 0.978$ (95% C, $p < 0.0005$)) indicated an almost perfect inter-rater agreement, the third author went on to code the rest of the articles. After the paradigm coding was finished, mixed-methods studies were excluded because there were only 21 mixed-methods RAs in the four journals between 1981 and 1985. Our initial data coding and sampling resulted in four pools of 40 articles each:

1. 1981–1985 Qualitative
2. 1981–1985 Quantitative
3. 2011–2015 Qualitative
4. 2011–2015 Quantitative

The corpus totals approximately one million words. Table 1 below provides descriptive statistics on the corpus.

Measurements. Following Zhao et al. (2023), we calculated two measures for each text: IE and MDD, which represent the cognitive encoding and decoding difficulties, respectively. According to Zhao et al. (2023), they capture different aspects of text difficulty from a cognitive information-theoretic perspective.

IE. IE, originating from information theory (Shannon, 1948), measures the average amount of information contained in each word of a text, based on the probabilities of words occurring, representing the unpredictability or randomness of a text. A line of studies has shown that higher-entropy texts are less predictable, making them more difficult to process and encode (Bentz et al., 2017; Sayood, 2018; Shannon, 1948). Following Zhao et al. (2023), we calculated IE using the Miller-Madow (MM) entropy algorithm (Hausser and Strimmer, 2014), which counterbalances the underestimation bias in small text sizes. The MM entropy is calculated as Formula 1:

$$H(MM) = H(T) + \frac{V - 1}{2N}$$

where $H(T)$ is the classic Shannon entropy, V represents word types, and N refers to word tokens. The Shannon entropy is

calculated as Formula 2:

$$H(T) = - \sum_{i=1}^V p(w_i) \cdot \log_2 p(w_i)$$

where $p(w_i)$ is the probability of the occurrence of the i^{th} word type in a text T .

MDD. MDD is an average of dependency distances within a text, where dependency distance (DD) refers to the linear distance between two words within a syntactic dependency relation (H. Liu, 2008; X. Liu et al., 2022). A higher MDD indicates higher syntactic complexity and heavier cognitive load, suggesting more effort required in cognitive decoding (Zhao et al., 2023).

The DD between two dependency-related words is calculated as Formula 3:

$$DD = |P_G - P_D|$$

where P_G refers to the position of the governor and P_D to the position of the dependent.

The MDD of a sentence is then calculated as Formula 4:

$$MDD = \frac{\sum_{i=1}^N |DD_i|}{N}$$

where DD_i is the dependency distance of the i^{th} dependency pair in the sentence, and N is the total number of dependency relations in the sentence. For example, in this short sentence, “She is in the classroom”, there are four dependency relations. Table 2 below shows the starting and ending points of each dependency relation and its distance. According to Formula 4, the MDD for this sentence is $(1 + 1 + 1 + 2)/4 = 1.25$. The MDD of a text is the result of dividing the sum of dependency distance by the total number of dependency relations.

Data processing and statistical analysis. Custom Python (version 3.11.5) scripts were developed to calculate IE and MDD for each article in the corpus¹. The built-in tokenizer and dependency parser in spaCy (version 3.7.5), which is a Python library for natural language processing, were used for tokenization and dependency parsing with the transformer-based language model for English, namely “en_core_web_trf” (version 3.7.3). The “en_core_web_trf” model in spaCy was selected because of its

high accuracy in dependent parsing. As indicated in its Hugging Face page², this model achieved a benchmark of 95.26% on the Unlabelled Attachment Score (UAS) and 93.91% on the Labelled Attachment Score (LAS), which is slightly higher than the 94.26% (for UAS) and 92.41% (for LAS) by the latest version of Stanford CoreNLP (Weiss et al., 2015), which was adopted in Zhao et al. (2023). All dependency types were taken into consideration except for that involving punctuation (“punct”), which was excluded in line with previous studies (e.g., Lei and Wen, 2019; H. Liu, 2008). To examine the relationship between historical period and research paradigms on IE and MDD, we conducted a two-way multivariate analysis of variance (MANOVA). Historical period (1981–1985 vs. 2011–2015) and research paradigm (qualitative vs. quantitative) were the independent variables, with IE and MDD as the dependent variables. We used Wilks’ Lambda, one of the most commonly used and well-established multivariate test statistics, as the test statistic to assess the effects of the independent variables on the combined dependent variables. For all the statistical analyses, the alpha values were set at 0.05. When post-hoc analyses were necessary, Bonferroni corrections were applied. Levene’s Test of Equality of Error Variances (based on mean) was conducted to test the assumption of homogeneity of variances. The results were not significant for both IE ($F(3, 156) = 2.493, p = 0.062$) and MDD ($F(3, 156) = 0.601, p = 0.615$), indicating that the assumption was met. Additionally, Box’s Test of Equality of Covariance Matrices was conducted to assess the assumption of homogeneity of covariance matrices. The result was not significant ($F(9, 27885.865) = 1.845, p = 0.055$), indicating that this assumption was also satisfied.

Results

The effects of different factors on IE and MDD. Multivariate tests based on Wilks’ Lambda return a statistically significant difference in the combined dependent variables across different levels of time ($F(2155) = 50.385, p < 0.001$, partial $\eta^2 = 0.394$), paradigm ($F(2155) = 9.065, p < 0.001$, partial $\eta^2 = 0.105$) and their interaction ($F(2155) = 9.414, p < 0.001$, partial $\eta^2 = 0.108$). Table 3 below presents detailed statistics. To further explore how each independent variable influences each dependent variable individually, tests of between-subjects effects were conducted for IE and MDD.

IE across time periods and research paradigms. The mean IE of the articles in our corpus is 8.740 ($SD = 0.297, N = 160$). As can be seen from Table 4 below, qualitative RAs consistently exhibit higher IE than quantitative RAs in both periods, with both research paradigms experiencing a rise in IE over time. Results from the tests of between-subjects effects indicate that both the diachronic change ($F(1) = 78.406, p < 0.001$, partial $\eta^2 = 0.334$) and the paradigmatic difference ($F(1) = 18.206, p < 0.001$, partial $\eta^2 = 0.105$) are statistically significant. No significant interaction is observed between the time period and the research paradigm ($F(1) = 0.229, p = 0.663$, partial $\eta^2 = 0.001$), indicating qualitative RAs maintain a consistently higher IE than quantitative RAs, with

Table 2 Illustration of dependency distance calculation example.

Word	Position	Governor	Relationship Description	Distance
She	1	is	subject	1
is	2	-	root node	-
in	3	is	prepositional phrase marker	1
the	4	classroom	determiner	1
classroom	5	in	object of preposition	2

Table 3 Results of multivariate tests^a.

Effect	Value	F	Hypothesis DF	Error DF	Sig.	Partial η^2
Intercept	0.001	118159.757 ^b	2	155	<0.001	0.999
time	0.606	50.385 ^b	2	155	<0.001	0.394
paradigm	0.895	9.065 ^b	2	155	<0.001	0.105
time * paradigm	0.892	9.414 ^b	2	155	<0.001	0.108

^aDesign: Intercept + time + paradigm + time * paradigm.

^bExact statistic.

both paradigms following a uniform rising trend. The very large effect size for the model as measured by partial eta squares (0.383) shows that the two independent variables can account for a large portion of the variance in IE, with time being a more significant predictor than research paradigm. Table 5 gives the statistical results of the tests of between-subjects effects on IE. The diachronic change in each research paradigm is plotted in Fig. 1³ as separate lines for clearer visualization.

MDD across time periods and research paradigms. The average MDD of the RAs in our corpus is 2.804 (SD = 0.236, $N = 160$). Table 6 below shows the descriptive statistics for MDD across research paradigms and time periods. The univariate analysis shows that the two independent variables can explain 24.2% of the variance in MDD. Unlike IE, which differs significantly between times and research paradigms, MDD shows no significant paradigmatic difference ($F(1) = 0.028, p = 0.868$, partial $\eta^2 < 0.001$). However, there is significant diachronic change ($F(1) = 30.705, p < 0.001$, partial $\eta^2 = 0.164$) and significant interaction between the time periods and the research paradigms ($F(1) = 18.938, p < 0.001$, partial $\eta^2 = 0.108$). A summary of the tests of between-subjects effects for MDD is provided in Table 7. Post-hoc pairwise comparisons with independent sample t-tests were conducted to assess the simple effect of each independent variable by keeping the other constant. Cohen's d was calculated for each independent sample t-test to obtain the effect sizes. The statistical results of pairwise comparisons are presented in Table 8 below. As can be seen from Table 8, quantitative RAs in applied linguistics experienced significant increase in MDD ($p < 0.0125$, the corrected p -value through Bonferroni methods) whereas the MDD of qualitative RAs did not change much between the two time periods ($p > 0.0125$, the corrected p -value through Bonferroni methods). As a result, during the second time period, quantitative RAs had significantly higher MDD than their qualitative counterparts, a complete

reversal of the situation in the first time period (see Fig. 2⁴ for an illustration).

Discussion

The present study examined the diachronic changes in two information-cognitive complexity indices, namely IE and MDD, of Applied Linguistics RAs across quantitative and qualitative paradigms. Our findings reveal a general trend towards increased difficulty over time, with notable differences between paradigms. Both qualitative and quantitative RAs showed significant increases in cognitive encoding difficulty, while only quantitative RAs experienced a significant rise in decoding difficulty. To the best of our knowledge, the present study is the first one to investigate the diachronic changes in encoding and decoding difficulties of RAs across research paradigms within a single discipline. Significant paradigmatic differences indicate that research paradigm needs to be taken into consideration for a better understanding of the cognitive complexity in RAs. Since Zhao et al. (2023) is the only study currently available that examines diachronic changes of both IE and MDD in academic texts, comparison between the present study and Zhao et al. (2023) will be incorporated into discussion.

The cognitive encoding difficulty of RAs in Applied Linguistics. It was found that IE of both qualitative and quantitative RAs in our corpus had increased significantly in the past three decades. This finding is in line with Zhao et al. (2023), which observes a rising trend in IE in the abstracts of social sciences. Since higher IE indicates greater information content, the increase in IE may be attributed to the accumulative nature of academic knowledge. Corresponding to the increasing volume of knowledge produced within the academia, RAs published recently need to package more concepts, more methodological details, and more diverse topics, resulting in higher information content than those published in the past (Biber and Gray, 2010; Zhou et al. (2023)). Another contributing factor to the increased IE may be Applied Linguistics being a relatively new field with significant recent growth (Hyland & Jiang, 2021). Combined with the typical social science character of being cross-disciplinary (Zhao et al. 2023), Applied Linguistics may have been incorporating diverse terminologies and concepts from other disciplines, leading to the rise in the information load of RAs in this field. Our findings suggest that these factors are consistent across both qualitative and quantitative studies. However, the accumulation of knowledge and the cross-disciplinary features may not be the only drive behind this upward trend in the cognitive encoding difficulty of RAs, a point we will return when discussing the findings on MDD.

Between the two research paradigms, a significant difference was observed: qualitative RAs exhibit higher IE than quantitative RAs in both time periods. This result shows that qualitative RAs

Table 4 Descriptive statistics on IE across time periods and research paradigms.				
Time	Paradigm	Mean	Std. Deviation	N
1981-1985	qualitative	8.664	0.297	40
	quantitative	8.487	0.237	40
	Total	8.576	0.281	80
2011-2015	qualitative	8.976	0.195	40
	quantitative	8.835	0.198	40
	Total	8.905	0.208	80
Total	qualitative	8.820	0.295	80
	quantitative	8.661	0.278	80
	Total	8.740	0.297	160

Table 5 Results of tests of between-subjects effects on IE.						
Source	Type III Sum of Squares	DF	Mean Square	F	Sig.	Partial η^2
Corrected Model	5.362 ^a	3	1.787	32.281	<0.001	0.383
Intercept	12223.190	1	12223.190	220755.463	<0.001	0.999
time	4.341	1	4.341	78.406	<0.001	0.334
paradigm	1.008	1	1.008	18.206	<0.001	0.105
time * paradigm	0.013	1	0.013	0.229	0.633	0.001
Error	8.638	156	0.055			
Total	12237.189	160				
Corrected Total	14.000	159				

^aR Squared = 0.383 (Adjusted R Squared = 0.371).

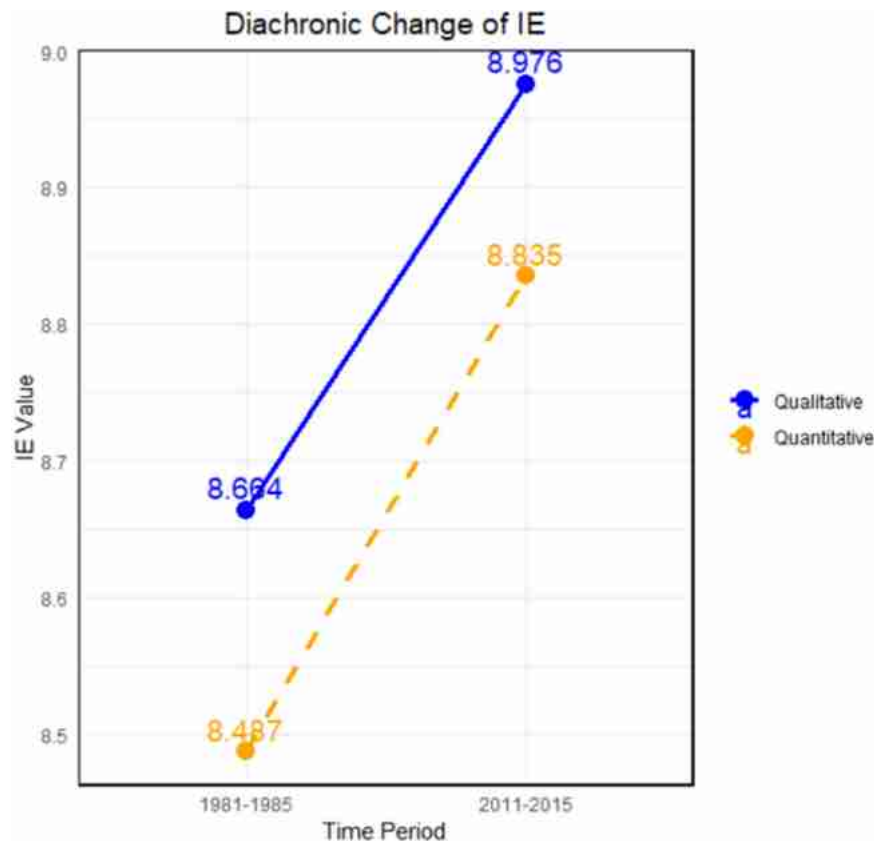


Fig. 1 Diachronic change of IE across research paradigms. This figure illustrates the diachronic change in information entropy (IE) values in Applied Linguistics research articles across two time periods, 1981-1985 and 2011-2015, and adopting two research paradigms, qualitative and quantitative approaches.

Table 6 Descriptive statistics on MDD.				
Time	Paradigm	Mean	Std. Deviation	N
1981-1985	qualitative	2.787	0.208	40
	quantitative	2.639	0.174	40
	Total	2.713	0.205	80
2011-2015	qualitative	2.827	0.241	40
	quantitative	2.964	0.203	40
	Total	2.895	0.232	80
Total	qualitative	2.807	0.224	80
	quantitative	2.802	0.249	80
	Total	2.804	0.236	160

require more effort in cognitive encoding (Sayood, 2018) than quantitative RAs. This difference may stem from the distinctive epistemological assumptions of these two research paradigms, which shape how research is conducted and reported, respectively. Rooted in positivist epistemology, the quantitative paradigm seeks to establish general laws of causality that are immutable and universal (Cao and Hu, 2014). Correspondingly, quantitative research is typically conducted in controlled conditions with standard procedures, resulting in research reports that are more predictable in structure and content. In addition, reports of quantitative findings usually follow standard formats, further reducing their IE.

In contrast, the qualitative paradigm is grounded in constructivist epistemology, focusing on context-specific understanding. Contrary to quantitative research, qualitative research tends to be conducted in natural settings with flexible procedures, leading to reports that include richer descriptions of research

contexts and nuanced interpretation of findings. These characteristics necessitate more complex and varied linguistic expressions, contributing to higher encoding complexity. Furthermore, the absence of uniform procedures reduces information predictability, further increasing the cognitive encoding difficulty of qualitative RAs.

The cognitive decoding difficulty of RAs. Regarding the cognitive decoding difficulty, a significant interaction between the research paradigm and time period was observed. Specifically, during the first time period, qualitative RAs had significantly higher MDD than quantitative RAs. However, over the 30-year span, quantitative RAs not only closed the gap but also significantly surpassed qualitative RAs in MDD.

This finding contrasts, to some degree, with Zhao et al. (2023), which found no significant diachronic change of MDD in RA abstracts of social sciences and a decreasing trend in natural sciences. Although the present study and Zhao et al. (2023) are not directly comparable because different variables were investigated, the different relationships between IE and MDD observed in these two studies are worth discussing and may pose challenge to some prevailing cognitions regarding the development of academic writing in recent years.

In Zhao et al. (2023), abstracts in natural sciences exhibited an opposite trend between IE and MDD: IE increased significantly while MDD decreased significantly. Against the backdrop of the ever-increasing promotional traits in academic writing due to ever-growing publishing pressure (Hyland, 2023), Zhao et al. interpreted this pattern as authors' effort to strike a balance between being informative and promotional. However, this contrast is not observed in our data: we found both IE and

Table 7 Results of tests of between-subjects effects on MDD.

Source	Type III Sum of Squares	DF	Mean Square	F	Sig.	Partial η^2
Corrected Model	2.146 ^a	3	0.715	16.557	<0.001	0.242
Intercept	1258.218	1	1258.218	29125.102	<0.001	0.995
time	1.326	1	1.326	30.705	<0.001	0.164
paradigm	0.001	1	0.001	0.028	0.868	0.000
time * paradigm	0.818	1	0.818	18.938	<0.001	0.108
Error	6.739	156	0.043			
Total	1267.103	160				
Corrected total	8.885	159				

^aR Squared = 0.242 (Adjusted R Squared = 0.227).

Table 8 Results of independent sample t-tests.

Variables		t	DF	Sig. ^a	Cohen's d	95% confidence interval	
						Lower	Upper
Qual ^b	1980s-2010s	−0.777	78	0.439	−0.174	−0.139	0.061
Quan ^c	1980s-2010s	−7.683	78	< 0.001	−1.718	−0.409	−0.241
1980s ^d	Qual-Quan	3.466	78	0.003	0.775	0.063	0.234
2010s ^e	Qual-Quan	−2.760	78	0.007	−0.617	−0.237	−0.038

^a2-tailed.
^bqualitative.
^cquantitative.
^d1981-1985.
^e2011-2015.

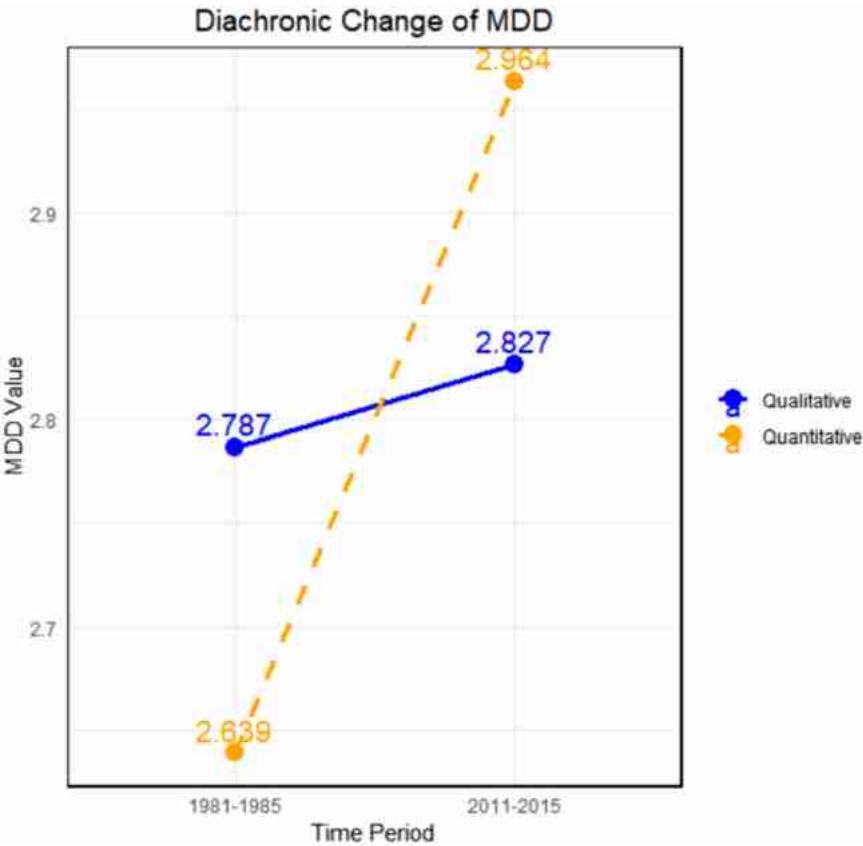


Fig. 2 Diachronic change of MDD across research paradigms. This figure illustrates the diachronic change in mean dependency distance (MDD) values in Applied Linguistics research articles across two time periods, 1981-1985 and 2011-2015, and adopting two research paradigms, qualitative and quantitative approaches.

Table 9 Illustration of dependency distance calculation example with more promotional language.				
Word	Position	Governor	Relation Description	Distance
She	1	is	subject	1
is	2	-	root node	-
in	3	is	prepositional phrase marker	1
the	4	classroom	determiner	2
fantastic	5	classroom	adjective modifier	1
classroom	6	in	object of preposition	3

MDD went up significantly for quantitative RAs while MDD in qualitative RAs did not undergo significant change.

While the increase in IE across both paradigms can be attributed to the accumulation of knowledge, the divergence in the changing pattern of MDD likely stems from notable changes in quantitative RAs, which were not mirrored in qualitative RAs. As diachronic changes in RAs subscribing to a specific research paradigm is currently lacking in literature, we believe that comparative analyses between RAs in natural sciences and those in social sciences in terms of the diachronic changes of some linguistic features may be of referential values. This comparison is not a far-fetched one, given the similarity between the epistemological distinctions between quantitative and qualitative research and those between natural sciences and social sciences (Chen and Hu, 2020). In this context, we would like to direct readers' attention to the rising trend in the use of promotional markers in natural sciences.

For example, some recent studies on the changing patterns of metadiscourse (Hyland and Jiang, 2016, 2018a, 2018b) reveal that natural sciences published in recent years employ much more promotional interactional metadiscourse, such as attitude markers and boosters, than those published a few decades ago, whereas social sciences exhibit relatively stable use of the same linguistic markers. Similarly, the use of academic hypes has increased in both natural and social sciences, with a much more dramatic rise observed in the natural sciences (Hyland and Jiang, 2021). In addition, the frequency of positive words in academic discourse has risen across both hard and soft disciplines, but such terms are more prevalent in hard disciplines (Xie and Mi, 2023). These findings are in line with the observation of greater incorporation of promotional language in hard sciences than softer ones (Hyland and Jiang, 2023; Martín and León Pérez, 2014). This trend is often attributed to fiercer competition and higher market value associated with hard sciences (Monteiro and Hirano, 2020).

Our speculation that the increasing use of promotional language may drive up MDD is tied to the algorithm of MDD. Take the same short sentence mentioned above ("She is in the classroom"). as an example. If we add an attitude marker *fantastic* before *classroom*, the sentence becomes "She is in the fantastic classroom". This modification alters the dependency relations in the sentence, as illustrated in Table 9:

Using Formula (4), the modified sentence yields an MDD of 1.6 ((1 + 1 + 2 + 1 + 3)/5), greater than 1.25 for the original sentence without the attitude marker. However, since promotional markers may assume diverse syntactic roles, their effect on MDD requires further validation with future empirical studies. Ideally, such studies should include computational simulations that control for text content to isolate the specific effects of promotional markers on MDD.

Abstract vs. full-length articles. An alternative explanation for the discrepancy between findings of the present study and those

in Zhao et al. (2023) may be the genre difference: we analysed full-length articles whereas Zhao et al. (2023) focused solely on abstracts. Previous research has shown that significant differences in terms of linguistic features exist between part genres of academic texts (Deng et al., 2024; Yin et al., 2021). In our case, this difference is clearly manifested in the fact that the mean IE value obtained in the present study is much higher than that observed in Zhao et al. (2023) whereas the MDD shows the opposite pattern.

Since Zhao et al. (2023) do not have data for 1981–1985, we limit our comparison to data from 2011 to 2015. Drawing on the epistemological similarities between natural sciences and quantitative studies, and between social sciences and qualitative studies, we compare quantitative RAs in our study to natural sciences in Zhao et al. (2023) and qualitative RAs to social sciences. In the present study, the mean IE for qualitative RAs published between 2011 and 2015 is 8.487 whereas in Zhao et al. (2023) the mean IE for abstracts in social sciences is only 6.794. For quantitative versus natural sciences, the ratio is 8.835 to 6.80.

By contrast, in our study, MDD for quantitative studies increased from 2.594 to 2.905 whereas in Zhao et al. (2023), the same index for abstracts in natural sciences decreased from 3.029 in 2000 to 2.881 in 2015. In other words, the IE of full-length articles is constantly higher than that of abstracts whereas MDD of full-length articles has just caught up with that of abstracts, which is only possible when the former increases and the latter decreases. While full-length articles are understandably more diverse in content, thereby showing higher IE, the complex patterns for MDD require further research.

In summary, this complicated landscape in cognitive complexity of academic texts should serve as a reminder that cautions must be taken when generalizing findings on one part-genre to another or to the entire genre.

Conclusion

The present study analysed the temporal evolution of two complexity measures rooted in information-cognitive theory, namely IE and MDD, within Applied Linguistics research articles (RAs) across quantitative and qualitative approaches. Our results reveal an overall pattern of increasing complexity over time, with marked variations between the two paradigms. RAs from both qualitative and quantitative traditions exhibited significant growth in encoding difficulty (IE), whereas a significant increase in decoding difficulty (MDD) was observed exclusively in quantitative RAs.

These findings advance our understanding of academic writing in several ways. Firstly, they highlight the importance of considering research paradigms when examining diachronic changes in academic writing, even within a single discipline. The significant paradigmatic differences observed indicate that research paradigm plays a crucial role in shaping the cognitive complexity of RAs. Secondly, our findings challenge some prevailing assumptions about the relationship between information load and readability in academic writing, suggesting a more nuanced relationship, particularly when examining full-length articles rather than abstracts. Thirdly, our study also underscores the necessity of considering part-genre differences in academic writing research. The notable differences in IE and MDD between our full-length article corpus and previous studies on abstracts suggest that caution must be taken when generalizing findings from one part genre to another or to the entire genre. Lastly, perhaps most importantly, our findings suggest the need to re-evaluate the role of promotional language in academic writing. While promotional language may enhance the visibility of scientific discoveries, its increasing use may also contribute to higher

cognitive decoding difficulty of RAs. Further empirical research is needed to find out the complex relationship between the use of promotional language and cognitive difficulty indices.

Due to resource constraints, we recourse to a two-time point design instead of a time series design that incorporates more time points, preferably with yearly data. While our choice of the time span aligns with common practices in EAP studies, the relatively sparse time points may result in overlooking more subtle, even hidden patterns. Future large-scale research incorporating finer-grained sampling is therefore encouraged to capture a more detailed picture of diachronic changes in academic writing.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 23 August 2024; Accepted: 11 March 2025;

Published online: 24 March 2025

Notes

- 1 The Python script for calculating IE and MDD and the corpus used in this study can be accessed through this GitHub link: https://github.com/truemichaelwolf/calculate_entropy_mmd_RAAL.
- 2 Hugging Face is a leading open-source platform in the field of artificial intelligence, specializing in natural language processing (NLP). It provides pre-trained models for various NLP tasks along with documentation of their benchmarks. The Hugging Face page for “en_core_web_trf” is as follows: https://huggingface.co/spacy/en_core_web_trf?utm_source=chatgpt.com.
- 3 The figure of diachronic change of IE across research paradigms will be provided in a file titled Fig. 1.
- 4 The figure of diachronic change of MDD across research paradigms will be provided in a file titled Fig. 2.

References

- Bentz C, Alikaniotis D, Cysouw M, Ferrer-i-Cancho R (2017) The entropy of words-learnability and expressivity across more than 1000 languages. *Entropy* 19(6):275
- Biber D, Gray B (2010) Challenging stereotypes about academic writing: complexity, elaboration, explicitness. *J Engl Acad Purp* 9(1):2–20
- Biber D, Gray B (2016) *Grammatical complexity in academic English: linguistic change in writing*. Cambridge University Press, Cambridge
- Biber D, Larsson T, Hancock GR (2024) The linguistic organization of grammatical text complexity: comparing the empirical adequacy of theory-based models. *Corpus Linguist Linguist Theory* 20(2):347–373
- Bulté B, Housen A (2012) Defining and operationalising L2 complexity. In Housen A, Kuiken F, Vedder I (eds) *Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA*. Benjamins, Amsterdam, p 21–46
- Bulté B, Housen A (2014) Conceptualizing and measuring short-term changes in L2 writing complexity. *J Second Lang Writ* 26:42–65
- Byrnes H, Sinicrope C (2008) Advancedness and the development of relativization in L2 German: a curriculum-based longitudinal study. In Ortega L, Byrnes H (eds) *The longitudinal study of advanced L2 capacities*. Routledge, New York, p 109–138
- Cao F, Hu G (2014) Interactive metadiscourse in research articles: a comparative study of paradigmatic and disciplinary influences. *J Pragmat* 66:15–31
- Chen L, Hu G (2020) Mediating knowledge through expressing surprises: a frame-based analysis of surprise markers in research articles across disciplines and research paradigms. *Discourse Process* 57(8):659–681
- Creswell JW (2013) *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage, London
- Deng L, Cheng Y, Gao X (2024) Promotional strategies in English and Chinese research article introduction and discussion/conclusion sections: a cross-cultural study. *J Engl Acad Purp* 68(3):101344
- Diessel H (2004) *The acquisition of complex sentences*. Cambridge University Press, Cambridge
- Dörnyei Z (2007) *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press, Oxford
- Hausser J, Strimmer K (2014) Entropy: estimation of entropy, mutual information and related quantities. R Package Version 1
- Heng R, Pu L, Liu X (2022) The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. *Front Psychol* 13:1082228
- Housen A, Clercq B, de, Kuiken F, Vedder I (2019) Multiple approaches to complexity in second language research. *Second Lang Res* 35(1):3–21
- Hu G, Cao F (2015) Disciplinary and paradigmatic influences on interactional metadiscourse in research articles. *Engl Specif Purp* 39:12–25
- Hu G, Chen L (2019) To our great surprise ...: a frame-based analysis of surprise markers in research articles. *J Pragmat* 143:156–168
- Hyland K (2023) Academic publishing and the attention economy. *J Engl Acad Purp* 64:101253
- Hyland K, Jiang F (2016) We must conclude that...: a diachronic study of academic engagement. *J Engl Acad Purp* 24:29–42
- Hyland K, Jiang F (2018a) ‘We Believe That ...’: changes in an Academic Stance Marker. *Aust J Linguist* 38(2):139–161
- Hyland K, Jiang F (2018b) In this paper we suggest: changing patterns of disciplinary metadiscourse. *Engl Specif Purp* 51:18–30
- Hyland K, Jiang F (2021) Our striking results demonstrate ...: persuasion and the growth of academic hype. *J Pragmat* 182:189–202
- Hyland K, Jiang F (2023) Hying the REF: promotional elements in impact submissions. *Higher Education*
- Kalantari R, Gholami J (2017) Lexical complexity development from dynamic systems theory perspective: lexical density, diversity, and sophistication. *Int J Instr* 10(4):1–18
- Kyle K, Crossley SA (2018) Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Mod Lang J* 102(2):333–349
- Lei L, Wen J (2019) Is dependency distance experiencing a process of minimization? A diachronic study based on the State of the Union addresses. *Lingua* 239(2):102763
- Lei L, Wen J, Yang X (2023) A large-scale longitudinal study of syntactic complexity development in EFL writing: a mixed-effects model approach. *J Second Lang Writ* 59:100962
- Li Y, Yang R (2023) Assessing the writing quality of English research articles based on absolute and relative measures of syntactic complexity. *Assess Writ* 55:100692
- Liu H (2008) Dependency distance as a metric of language comprehension difficulty. *J Cogn Sci* 9(2):159–191
- Liu X, Zhu H, Lei L (2022) Dependency distance minimization: a diachronic exploration of the effects of sentence length and dependency types. *Humanit Soc Sci Commun* 9(1):420
- Lu X (2011) A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers’ language development. *TESOL Q* 45(1):36–62
- Mak WM, Sanders TJM (2013) The role of causality in discourse processing: effects of expectation and coherence relations. *Lang Cogn Process* 28(9):1414–1437
- Martín P, León Pérez IK (2014) Convincing peers of the value of one’s research: a genre analysis of rhetorical promotion in academic texts. *Engl Specif Purp* 34:1–13
- Monteiro K, Hirano E (2020) A periphery inside a semi-periphery: The uneven participation of Brazilian scholars in the international community. *Engl Specif Purp* 58:15–29
- Ouyang J, Jiang J, Liu H (2022) Dependency distance measures in assessing L2 writing proficiency. *Assess Writ* 51:100603
- Parviz M, Jalilifar A, Don A (2020) Phrasal discourse style in cross-disciplinary writing: a comparison of phrasal complexity features in the results sections of research articles. *Círculo De Lingüística Aplicada Comunicación* 83:191–204
- Sayood K (2018) Information theory and cognition: a review. *Entropy* 20(9):709
- Shannon C (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
- Spooren W, Sanders TJM (2008) The acquisition order of coherence relations: on cognitive complexity in discourse. *J Pragmat* 40(12):2003–2026
- Thompson P, Gray B (2024) “Beyond the research article”: Identifying research article sub-registers through multi-dimensional and text-type analysis. Paper presented at the 2024 American Association for Applied Linguistics Conference, Houston, Texas, 16–19 March 2024
- Wang Q, Hu G (2024) Expressions of confusion in research articles: A diachronic cross-disciplinary investigation. *Scientometrics* 129(1):445–471
- Weiss D, Alberti C, Collins M, Petrov S (2015) Structured training for neural network transition-based parsing. In: Zong C, Strube M (ed) *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, 2015
- Wen QF (2001) *Applied Linguistics: Research Methods and Thesis Writing*. Foreign Language Teaching and Research Press, Beijing
- Xiao W, Li L, Liu J (2023) To move or not to move: an entropy-based approach to the informativeness of research article abstracts across disciplines. *J Quant Linguist* 30(1):1–26
- Xie S, Mi C (2023) Promotion and caution in research article abstracts: the use of positive, negative and hedge words across disciplines and rankings. *Learn Publ* 36(2):249–265

- Yin S, Gao Y, Lu X (2021) Syntactic complexity of research article part-genres: differences between emerging and expert international publication writers. *System* 97:102427
- Yin S, Gao Y, Lu X (2023) Diachronic changes in the syntactic complexity of emerging Chinese international publication writers' research article introductions: a rhetorical strategic perspective. *J Engl Acad Purp* 61:101205
- Zareva A (2019) Lexical complexity of academic presentations. *J Second Lang Stud* 2(1):71–92
- Zhang W, Cheung Y (2023) The different ways to write publishable research articles: Using cluster analysis to uncover patterns of APPRAISAL in discussions across disciplines. *J Engl Acad Purp* 63:101231
- Zhao X, Li L, Xiao W (2023) The diachronic change of research article abstract difficulty across disciplines: a cognitive information-theoretic approach. *Humanit Soc Sci Commun* 10(1):194
- Zhou W, Li Z, Lu X (2023) Syntactic complexity features of science research article introductions: rhetorical-functional and disciplinary variation perspectives. *J Engl Acad Purp* 61:101212
- Zhou X, Gao Y, Lu X (2023) Lexical complexity changes in 100 years' academic writing: evidence from *Nature Biology Letters*. *J Engl Acad Purp* 64:101262
- Ziaeeian E, Golparvar SE (2022) Fine-grained measures of syntactic complexity in the discussion section of research articles: the effect of discipline and language background. *J Engl Acad Purp* 57:101116

Acknowledgements

We would like to thank Dr Liu Yanhua for inter-rating the research paradigm. We would also want to express our gratitude to Mr Li Xiuhai, Mr Hao Xiaodong, Dr Li Longxing, and Ms Wang Wenpu for cleaning a large proportion of text files, which is a very time-consuming and arduous task. This work was supported by the Medici Project [<https://www.chenlang.info/medici-project>].

Author contributions

First author: Writing—original draft, Methodology, Formal analysis, Data curation, Conceptualization. Second author: Writing—review & editing, Conceptualization. Corresponding author: Writing—revision & editing, Software, Supervision, Conceptualization.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-025-04742-x>.

Correspondence and requests for materials should be addressed to Lang Chen.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025