Glottometrics 49, 2020, 32-51

Quantitative Analysis of Academic Writing as to Informality and Vocabulary Features

Ziqi Liu[1], *Haitao Liu*[2]

Abstract. What matters for a learner of English for academic purposes is to possess the ability to present results and achievements in international top journals. The ability is related to degrees of informality, vocabulary richness, and lexical complexity in academic writing. This study takes Chinese master degree candidates and advanced writers as research objects and concentrates on two research questions: (1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings? (2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices? The results are based on studying two datasets of the research objects. However, our results show that there is a complex picture for each informality indicator. Finally yet importantly, advanced writers show a higher level of vocabulary richness and complexity.

Keywords: informality features, vocabulary richness and complexity, Chinese master degree candidates, advanced writers, academic writing, EFL.

1. Introduction

Avoiding informality is necessary and essential for learners of English for academic purposes. Academic writing is characterized as an impersonal and objective reporting on independent and external reality (Lee, Bychkovska, & Maxwell, 2019; Hyland, 2001a). Thus, to avoid informality in academic writing is a key factor. Furthermore, academic writing is not just about the results, it is also relevant to the representation of writers (Hyland, 2002). Then, an important factor is how to employ vocabulary, and what words should be selected to convey the study content. Thus, exploring the gap in informality features, lexical richness, and complexity between “novices” and “experts” is indispensable.

As for the method, a quantitative approach should have a firm place and wide application in the study of academic writing. It can be employed to process a great amount of material not only with a lot of diverse features, but also in a short time. Furthermore, more details of datasets are acquired by exploiting the quantitative method. What is more, it is feasible to state the frequency of each informality feature and the figures of vocabulary richness and complexity indexes. Based on those data, further and more accurate analysis about the comparison is conducted.

[1] Department of Linguistics, Zhejiang University, Hangzhou, China.

[2] Institute of Quantitative Linguistics, Beijing Language and Culture University, Beijing, China; Department of Linguistics, Zhejiang University, Hangzhou, China. Correspondence to: Haitao Liu. Email address: htliu@163.com, ORCID-No.: https://orcid.org/0000-0003-1724-4418.

In recent decades, some studies have employed different indexes to compare informality features among diverse texts. The study by Petch-Tyson (1998) is carried out on writings of EFL students from different backgrounds, either language or cultural ones, including French, Dutch, Swedish, and Finnish. Aijmer (2002) concentrates on the situation of modality in Swedish learners' written interlanguage. It is difficult for second language students to express a suitable degree of doubt and certainty. Thus, Hyland and Milton (1997) study qualification and certainty in the writings of L1 and L2 students. Cobb (2003) explores the Québec learner corpus. As for sentence-initial *and* and sentence-initial *but*, the study by Bell (2007) is based on selections from 11 academic journals containing the domains of science, the humanities, and social science. Wang (2016) studies grammatical colloquial features through theses of EFL learners. In order to investigate the trend of informality, Hyland and Jiang (2017) examine 10 informal features (first-person pronouns, unattended anaphoric pronouns, split infinitives, sentence-initial conjunctions or conjunctive adverbs, sentence-final preposition, listing expressions, second-person pronouns/determiners, contractions, direct questions, and exclamations) across four disciplines, which are applied linguistics, sociology, electrical engineering, and biology. In the discipline of applied linguistics, Alipour and Nooreddinmoosa (2018) also investigate informality features. Besides, Lee et al.'s contribution (2019) is to compare informality features in the writing of L1 and L2 undergraduate students.

According to Fang and Liu (2015), the study of lexical richness was founded by Chotlos and Yule (Chotlos, 1944; Yule, 1944). It is complex either in linguistic or in mathematical aspects (Wimmer & Altmann, 1999). Lexical richness measurement belongs to one of the most traditional domains in quantitative linguistics (Kubát & Milička, 2013). The reason for employing it lies in the fact different groups of people use vocabulary with specific features. As for lexical diversity, Wen's research (2006) represents that the mean value of vocabulary richness of written English is higher than that of spoken English. In written language, writers have more time to avoid repeating some words. Thus, higher repeat rate and lower lexical richness are more likely to occur in colloquial English. In the study of Li and Liu (2019), they propose that written English and spoken English are two main types of style, and compared with written English, there exist informality features in spoken English according to Hickey (2014). Thus, vocabulary diversity is associated with informality features.

Vocabulary richness can be employed to investigate stylistic features (Smith & Kelly, 2002), to analyze different translation works (Fang & Liu, 2015), to explore genre analysis (Kubát & Milička, 2013) and authorship attribution (Jamak, Savatić, & Can, 2012; Hoover, 2003). As for lexical complexity – another indicator of writing style –, if there are more complex words, it means that the text is more sophisticated (Dai & Liu, 2019).

However, few studies compare the gap between Chinese master degree candidates and advanced writers based on both informality features (first/second-person pronouns, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs), and on vocabulary richness / complexity. In order to fill in the gap and to help Chinese English learners publish research articles in international top journals, this paper will employ the combination of the two aspects to study the following research questions:

(1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings?

(2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices?

The arrangement of this paper goes as follows. In the second section, the information about two self-built datasets and methodology is introduced. The third section is the results and discussion of the study, which relates to presentation and analysis of informality features and vocabulary richness / complexity of the datasets. In the final section, a conclusion is presented.

2. Methodology

2.1 Description of Material

In order to explore the gap in informality and vocabulary richness / complexity, two datasets, Chinese Master Thesis (CMT) and International Research Article (IRA), are established. They contain abstracts of Chinese master theses and of international research articles. An abstract is essential for academic writing. It summarizes the major aspects of the paper, which are introduction, methodology, results, and discussion. Besides, the abstract includes the research background and research questions, contains experimental design and methods used, and includes key results and their interpretations. The research of abstracts is also divergent. Abstracts are important materials in many studies concerning, for instance, publication (De Bruin, Treccani, & Sala, 2014; Scherer, Dickersin, & Langenberg, 1994; Snedeker, Totton, & Sargeant, 2010) or academic literacy practices (Starke and Bailer, 2019). Given the wide application of abstracts of research articles, they can also be employed to study the degree of informality and vocabulary features.

IRA contains abstracts of articles from 2014 to 2018 of three top journals in the domain of linguistics, which are *Journal of Memory and Language* (5-Year Impact Factor = 5.763), *Applied Linguistics* (5-Year Impact Factor = 4.516), and *Journal of Second Language Writing* (5-Year Impact Factor = 4.177). There are 75 abstracts in total (5 per journal each year). Besides, the amount of tokens for IRA is 13,555 and that of types is 2,570. To balance with IRA, CMT comprises 45 abstracts of Chinese master theses from the domain of Foreign Linguistics and Applied Linguistics in the same 5 years from three universities, which are ZJ (Zhejiang University), DL (Dalian Maritime University), and HN (Henan Normal University). ZJ belonged to Project 985[3] and Project 211[4]. DL was one member of Project 211. HN did not belong to either projects.

There are 15 abstracts for ZJ, 20 for DL, and 10 for HN. The tokens of CMT are 17,666 and the types are 2,507. Both the number of tokens and types have been acquired by software, QUITA (Kubát, Matlach, & Čech, 2014). The total of tokens of the two datasets is 31,221. In the study of Kalantari and Gholami (2017), 18,751 running words in the corpus are employed to investigate the lexical complexity development. Thus,

[3] Project 985 in China aims to construct world-class universities.

[4] Project 211 in China aims to strengthen about 100 institutions of higher education and key disciplines.

the amount of tokens of datasets in this study seems appropriate. Besides, the details of texts in CMT and IRA are listed in the appendix.

Table 1

Descriptions of CMT and IRA

Text	Types	Tokens
CMT	2,507	17,666
IRA	2,570	13,555

2.2 Data Analysis

2.2.1 Informality Features

In order to explore informality features of the two datasets, this research adopts an approach which is based on the revised version of other studies, which include Hyland and Jiang (2017), Lee et al. (2019), Aijmer (2002), and Petch-Tyson (1998). In the study of Hyland and Jiang (2017), first-person pronouns, second-person pronouns, unattended reference, and sentence-initial conjunctions / conjunctive adverbs are important indexes to indicate informality.

Next, academic writing should be semantically clear. If needless words are omitted, it will benefit achieving that goal. Employing unattended reference appropriately will make the expressions more concentrated, economical, and concise. Thus, unattended reference is not adopted as a feature of informality in this paper.

Last but not the least, academic writings prefer concrete and specific expressions. However, listing items are usual in the process of writing with vague and abstract meanings. Furthermore, it is also easily neglected. Thus, listing expression is taken into account as a feature of informality in the study.

All informality indexes employed in this paper to evaluate different degrees of informality in CMT and IRA are shown in Table 2.

Table 2

Description of informality features indexes

Category	Details
First-person pronouns	I, me, my, mine, we, us, our, ours
Second-personal pronouns	you, your
Sentence-initial conjunctions / Conjunctive adverbs	and, but, or, so, yet, again, also, besides, however, indeed, still, thus
Listing expressions	and so forth, and so on, etc.
Modal verbs	can, may, might, will, must, would, could, shall, should, ought to, have (got) to

To investigate whether the difference in each informality feature is significant or not, log-likelihood (*LL*) value is counted by the calculator (http://ucrel.lancs.ac.uk/llwizard.html) (Lee et al., 2019). At 5% level, $LL \geq 3.84$ means $p < 0.05$; at 1% level, $LL \geq 6.63$ is significant for $p < 0.01$; at 0.1% level, $LL \geq 10.83$ is equal to $p < 0.001$; at 0.01% level, $LL \geq 15.13$ represents $p < 0.0001$. As stated in Lee et al. (2019), *ELL* measure (Johnston, Berry, & Mielke, 2006), which represents effect size of log-likelihood measure, is also contained in the calculator. Besides, through the software AntConc (Anthony, 2011), the frequency of each informality feature in Table 2 is obtained.

2.2.2 Vocabulary Features

To study the gap of vocabulary features, which are richness and complexity, in CMT and IRA, different indexes (TTR, h-point, R_1, Repeat Rate, Entropy, and Average Tokens Length) are employed in this research.

TTR (V/N; the ratio of different words to all words) is an indicator of testing vocabulary richness (Yoon, 2017). Next, repeat rate (RR) and entropy are also indicators of vocabulary diversity, which are both based on the probability of occurrences of words in the text. In detail, the smaller the repeat rate, the greater the vocabulary richness; on the contrary, the greater the entropy, the greater the richness (Popescu, Čech, & Altmann, 2011).

The h-point is also calculated on the basis of word frequency (Dai & Liu, 2019). It is the point where the rank is equal to its frequency. Then,

$$r = f(r)$$

is applied to this situation. If there is no exact place like this, two neighbouring points will be adopted, which have $f(i)$ and $f(j)$. Under these circumstances,

$$f(i) > r_i \text{ and } f(j) < r_j,$$

and generally $r_i + 1 = r_j$ (Popescu et al., 2011). Here comes the formula of h-point:

$$(1) \qquad h = \frac{f(i) \times r_j - f(j) \times r_i}{r_j - r_i + f(i) - f(j)}.$$

The h-point is a critical point for the rank-frequency distribution of words in a text. Autosemantic words tend to appear after the h-point. In contrast, synsemantic words appear before the h-point. Hence, the h-point is an indicator for vocabulary richness.

$F(h)$ is the cumulative probability of words with the order from 1 to the h-point. With h and $F(h)$, another vocabulary richness index, R_1, has been proposed. R_1 is defined as follows:

$$(2) \qquad R_1 = 1 - \left(F(h) - \frac{h^2}{2N}\right).$$

TTR, h-point, R_1, repeat rate, and entropy are all the indicators to explore vocabulary richness. As for lexical complexity, word length is a common index (Dai & Liu, 2019). The larger the word length is, the more complex the text is. Thus, average tokens length is employed to investigate lexical complexity in this research. It is the mean of all the tokens lengths in the whole text. Those indicators of vocabulary richness and lexical complexity are measured by QUITA (Kubát et al., 2014).

3. Results and Discussion

3.1 Gap in Informality Features between CMT and IRA

Table 3 lists the overall frequencies and descriptions of five informality features. As shown, it is rather complicated to interpret these results. First-person pronouns are more likely used in IRA. However, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs occur more frequently in CMT.

For CMT and IRA, there is a significant difference in four indicators, except the index of second-person pronouns. Second-person pronouns (*you*, *your*) represent an obvious way to refer to readers as general or individual referents (Hyland, 2005). They are also the most visible acknowledgements of the reader's presence (Hyland, 2001b). Besides, there is a high percentage of occurrences of second-person pronouns in the texts of L2 learners (Petch-Tyson, 1998). Furthermore, in the study of Lee et al. (2019), second-person pronouns are also numerous in COLTE, the corpus of L2 learners. However, Hyland's research (2005) proposes that these reader pronouns (*you* and *your*) occur rarely in the student corpus. As for Table 3, the frequency of second-person pronouns in both databases is 0. It indicates that there may be a low frequency of *you* and *your* in abstracts, and even in the whole academic writing. For CMT, the writers are perhaps willing to present themselves in a relative junior status compared with the teachers, supervisors, and readers (Hyland, 2005). Thus, they try to avoid using second-person pronouns. As for the advanced writers or experts, informal features are also not suitable in their studies. Thus, the writers in CMT and IRA are likely not to use the second-person pronouns.

Table 3

Overall frequency and description of informality features

	CMT	IRA	*LL*	*ELL*	Significant
First-person pronouns	9	83	89.82	0.00078	√
Second-person pronouns	0	0	0	0	×
Sentence-initial conjunctions / conjunctive adverbs	70	26	10.96	0.00009	√
Listing expressions	4	0	4.56	-5.79	√
Modal verbs	117	55	9.45	0.00007	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

3.1.1 First-person Pronouns

As to Table 3, the frequency of first-person pronouns in CMT is 9 and that in IRA is 83. Besides, the log-likelihood value yields 89.82 ($p < 0.0001$), which means there is a significant difference; the effect size for log-likelihood is 0.00078.

What is more, the first-person pronoun is the only indicator that occurs more frequently in IRA than in CMT. In details, there is no occurrence of four pronouns (*me*, *my*, *mine*, *ours*) in both datasets. As a consequence, the difference between them is not significant. Compared to zero frequency of *us* in CMT, it occurs only once in IRA. As for *I*, *we*, and *our*, there is an obviously significant difference. Those data are presented in Table 4.

Table 4

Overall frequency and description of first-person pronouns

	CMT	IRA	*LL*	*ELL*	Significant
I	0	5	8.34	0.00034	√
me	0	0	0	0	×
my	0	0	0	0	×
mine	0	0	0	0	×
we	6	62	69.7	0.00066	√
us	0	1	1.67	-0.00006	×
our	3	15	12.23	0.00019	√
ours	0	0	0	0	×
total	9	83	89.82	0.00078	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

The results of previous studies for distribution of first-person pronouns are not unequivocal. Petch-Tyson's study (1998) reveals that non-native speakers of English adopt first-person pronouns more frequently than native speakers. Nevertheless, some studies show the opposite situation. According to Hyland (2002), experts or professional writers use more first-person pronouns than students. Besides, L1 writers are more likely to intervene with self-mentions (Lee & Deakin, 2016). Lee et al. (2019) also support the claim that L1 writers employ first-person pronouns/determiners more than ESL students. The study in this paper is in line with the second point – that advanced writers use them more in IRA. In the study by Leedham and Fernandez-Parra (2017), they find out that there is more occurrence of *we* for L1 Chinese and L1 Greek students than for L1 English students, and less frequency of *I* for L1 Chinese and Greek students than for L1 English students. However, in Table 4, *I*, *we*, and even *our* are used more frequently in IRA than in CMT.

First-person pronouns are considered to be a typical informality marker (Hyland & Jiang, 2017). Arguments in academic writings should be proposed in the most convincing way. Besides, acceptability, certainty, and plausibility of research require different and complex features, which include strong evidence, originality, and innovation of study, and an authoritative professional personality (Hyland & Jiang, 2017). Thus, an independent identity and the writer's voice need to be established. In the study of Hyland (2001a), employing first-person pronouns is a way to build and project a personal standing and authority. In addition, it is also a function to distinguish the writers from others. Thus, intervening with first-person pronouns appropriately benefits Chinese students in the constantly changing and competitive circumstances. Word choice also reveals the writers' social and psychological factors (Hyland, 2002); because of cultural background, some writers are more likely to avoid using first-person pronouns or to show modesty. Given the results of first-person pronouns, especially for *we*, it is essential to study further whether to make them the indicators of informality, or not.

3.1.2 Modal Verbs

As shown in Table 5, there is a significant difference in the modal verbs as a whole. The log-likelihood value is 9.45 ($p < 0.01$), with the effect size of 0.00007. Among them, significant difference only exists in four verbs, which are *may*, *will*, *could*, and *should*. Unlike *will*, *could*, and *should*, *may* is used more frequently in IRA rather than in CMT.

Table 5

Overall frequency and description of modal verbs

	CMT	IRA	*LL*	*ELL*	Significant
can	48	23	3.61	0.00003	×
may	14	22	4.54	0.00005	√
might	4	1	1.22	0.00005	×
will	21	1	17.45	0.00025	√
must	5	1	1.96	0.00007	×
would	3	4	0.53	0.00002	×
could	9	1	5.42	0.00012	√
shall	0	0	0	0	×
should	13	0	14.81	4.46	√
have (got) to	0	2	3.34	-0.00076	×
ought to	0	0	0	0	×
total	117	55	9.45	0.00007	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

Will is a predictive modal and is employed to predict future events with some certainty (Grant & Ginther, 2000; Aijmer, 2002). Besides, *may* is in the category of possibility modals with a lower certainty (Grant & Ginther, 2000; Aijmer, 2002; Hinkel, 2009). As for *could*, it is also a possibility modal with the meaning of probability (Grant & Ginther, 2000; Aijmer, 2002). According to Kennedy (1998), *may* occurs less in spoken corpus than in written corpus (879 versus 1,323); however, *could* (2,000 versus 1,744) and *will* (4,286 versus 2,804) are more highly employed in the spoken corpus. Hence, the lower occurrence of *may* in CMT represents higher informality. Besides, more uses of *will* and *could* also illustrate the lower formality of CMT. What is more, *should*, belonging to the obligation and necessity group, represents some actions with the meaning of desire and suggestion (Grant & Ginther, 2000; Aijmer, 2002). Biber et al. (2002) suggest that although obligation modals are usually suppressed, they are also exploited to express the meaning of personal obligation. Besides, some writings of non-native speakers seem brusque, dogmatic, too direct, and too tentative (Hyland & Milton, 1997). Thus, *should* is used less by advanced writers in IRA.

3.1.3 Sentence-initial Conjunctions / Conjunctive Adverbs

Alipour and Nooreddinmoosa (2018) illustrate that sentence-initial conjunctions are the most frequently used among those informality feature indexes in both native and non-native articles. As of Table 6, the significant difference lies in the total of sentence-initial conjunctions and conjunctive adverbs (*LL* = 10.96, *p* < 0.001, *ELL* = 0.00009) in CMT and IRA. More connectors in CMT may be a result of instructions. The writers are encouraged to use them to convey logic of academic writing and to show the connection between the preceding content and the following one. From this perspective, it means that the degree of informality for CMT is higher than for IRA. Besides, there exists significant difference in three main indicators, which are sentence-initial *and*, *so* and *besides*.

Table 6

Overall frequency and description of sentence-initial conjunctions / conjunctive adverbs

	CMT	IRA	*LL*	*ELL*	Significant
And	22	1	18.50	0.00026	√
But	1	0	1.14	-9.21	×
Or	0	0	0	0	×
So	5	0	5.69	-4.65	√
Yet	0	1	1.67	-0.00006	×
Again	0	0	0	0	×
Also	1	0	1.14	-9.21	×
Besides	7	0	7.97	-2.38	√
However	29	20	0.14	0	×
Indeed	0	0	0	0	×
Still	0	0	0	0	×
Thus	5	4	0	0	×
total	70	26	10.96	0.00009	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = *p* < 0.05; × = *p* > 0.05.

As shown in Table 6, sentence-initial *and*, which is second to *however*, is still highly employed in CMT. Bell (2007) also proposes three main roles of sentence-initial *and*, which are (i) to indicate the last item within the whole list; (ii) to develop arguments further; (iii) to represent shifts in authorial perspectives. In the study of Bell

(2007), sentence-initial *and* and sentence-initial *but* are most frequently used with additive and contrastive meanings respectively. In Table 6, sentence-initial *and* is still ranking first in its semantic group. However, sentence-initial *however* becomes the first rather than *but* for writers in both CMT and IRA. Furthermore, sentence-initial *however* is the most frequently adopted in both datasets, CMT and IRA. According to Hyland and Jiang (2017), the increases of sentence-initial *however* as well as declines of sentence-initial *but* and sentence-initial *and* also occur in the domain of applied linguistics and sociology. Some studies also prove that sentence-initial *however* is the most frequent used item of sentence-initial conjunctions and conjunctive adverbs (Lee et al., 2019; Alipour & Nooreddinmoosa, 2018).

3.1.4. Listing Expressions

Listing expression, a common index, is another type of informality features. As shown in Table 7, there is only one significant difference in the group. The log-likelihood value for the whole is 4.56 ($p < 0.05$), and the effect size is -5.79. However, no significant difference exists for individual listing expressions.

Table 7

Overall frequency and description of listing expressions

	CMT	IRA	*LL*	*ELL*	Significant
and so on	1	0	1.14	-9.21	×
and so forth	0	0	0	0	×
etc	3	0	3.42	-6.93	×
total	4	0	4.56	-5.79	√

Note: *LL* = log-likelihood value; *ELL* = effect size for log likelihood; √ = $p < 0.05$; × = $p > 0.05$.

Listing expressions in both datasets occur at a much lower frequency than the other four informality features. Compared with four occurrences of listing expressions in CMT, there is no hit in IRA. It may be due to the fact that advanced writers are aware of vagueness of listing expressions (Lee et al., 2019).

3.2 Gap in Vocabulary Richness and Complexity between CMT and IRA

As shown in Table 8, except the h-point and *RR*, values of vocabulary richness indicators – which are TTR, entropy, and R_1 – are higher in IRA than in CMT. The values of h-point and *RR* are higher in CMT than in IRA. All the data represent that there is more diversified and colourful vocabulary in advanced writers' material (IRA).

TTR is the type-token ratio. Compared with more tokens and fewer types in CMT, there are fewer tokens and more types in IRA. As to entropy and R_1, direct indicators of lexical richness, advanced writers have more word choices. The higher value of *RR* in CMT represents its lower lexical diversity. Besides, IRA (lower h-point) are more likely to possess more autosemantic words, which tend to come after h-point, and higher vocabulary richness.

Table 8

Description of vocabulary richness indexes

	CMT	IRA
TTR	0.141911	0.189598
h-Point	46	40
Entropy	8.902671	9.257164
R_1	0.635515	0.683807
RR	0.013138	0.008741

Note: CMT – types are 2,507; tokens are 17,666. IRA – types are 2,570; tokens are 13,555.

Average tokens length is an approach to test lexical sophistication approximately. It is seen in Table 9 that advanced writers in IRA are more likely to employ words with more complexity. Thus, a gap exists in lexical richness and complexity between master degree candidates and advanced writers.

Table 9

Description of vocabulary complexity indexes

	CMT	IRA
Average Tokens Length	5.455734	5.710144

Note: CMT: Types are 2,507; tokens are 17,666. IRA: Types are 2,570; tokens are 13,555.

4. Conclusions

This study employs two self-built datasets (CMT and IRA) to explore two research questions.

(1) To what extent do Chinese master degree candidates and advanced writers differ in the use of informal features in their writings?

(2) In what ways do Chinese master degree candidates and advanced writers differ in vocabulary choices?

In order to respond to the first research question, five informality features indexes (first-person pronouns, second-person pronouns, sentence-initial conjunctions / conjunctive adverbs, listing expressions, and modal verbs) – the choice based on the previous studies (Hyland & Jiang, 2017; Lee et al., 2019; Aijmer, 2002; Petch-Tyson, 1998) – are employed. The research is carried out by AntConc (Anthony, 2011), log-likelihood value and effect size calculator, and QUITA (Kubát et al., 2014). Details for each informal indicator provide a complex picture. On the one hand, advanced writers overuse first-person pronouns to express their identities and stances. On the other hand, Chinese master degree candidates frequently employ more modal verbs, sentence-initial conjunctions / conjunctive adverbs, and listing expressions. However, there is no occurrence of second-person pronouns in either group.

To answer the second research question, six indexes (type-token ratio, h-point, R_1, repeat rate, entropy, average tokens length) are selected to capture lexical diversity and vocabulary sophistication; the values are counted by the QUITA (Kubát et al., 2014) software. As for lexical richness, advanced writers possess higher type-token ratio, R_1, and entropy as well as lower h-point and repeat rate. It means that Chinese master degree candidates show a lower vocabulary diversity. Besides, experts in IRA also have a higher average tokens length. According to this result, Chinese master degree students are more likely to employ shorter words with less lexical sophistication than advanced writers.

Acknowledgements

This work is partly supported by the National Social Science Foundation of China (Grant No.17AYY021) and the Special Project for Humanities and Social Science, Zhejiang University (*Quantitative Analysis of Sentence Hierarchical Structure in English*).

References

Anthony, L. (2011). AntConc (Version 3.2. 4w). Tokyo: Waseda University. Available from http://www.laurenceanthony.net/software.

Aijmer, K. (2002). Modality in advanced Swedish learners' written inter-language. In: S. Granger, J. Hung & S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins Publishing Company, 57–76.

Alipour, M., & Nooreddinmoosa, M. (2018). Informality in Applied Linguistics Research Articles: Comparing Native and Non-Native Writings. *Eurasian Journal of Applied Linguistics* 4(2), 349–373.

Bell, D. (2007). Sentence-Initial *And* and *But* in Academic Writing. *Pragmatics* 17(2), 183–201.

Biber, D., Conrad, S., & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Chotlos, J. W. (1944). IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs* 56(2), 75–111.

Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Québec Replications of three European Studies. *The Canadian Modern Language Review* 59(3), 393–424.

Dai, Z., & Liu, H. (2019). Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages over 50 Years (1967–2018). *Glottometrics* 45, 63–88.

De Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive Advantage in Bilingualism: An Example of Publication Bias? *Psychological science* 26(1), 99–107.

Fang, Y., & Liu, H. (2015). Comparison of vocabulary richness in two translated Hongloumeng. *Glottometrics* 31, 54–75.

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing* 9(2), 123–145.

Hickey, R. (2014). *A Dictionary of Varieties of English*. Hoboken: Wiley-Blackwell.

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics* 41(4), 667–683.

Hoover, D. L. (2003). Another Perspective on Vocabulary Richness. *Computers and the Humanities* 37, 151–178.

Hyland, K. (2001a). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20(3), 207–226.

Hyland, K. (2001b). Bringing in the Reader Addressee Features in Academic Articles. *Written Communication* 18(4), 549–574.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics* 34(8), 1091–1112.

Hyland, K. (2005). Representing readers in writing: Student and expert practices. *Linguistics and Education* 16(4), 363–377.

Hyland, K., & Jiang, F. (2017). Is academic writing becoming more informal? *English for Specific Purposes* 45, 40–51.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6(2), 183–205.

Jamak, A., Savatić, A., & Can, M. (2012). Principal Component Analysis for Authorship Attribution. *Business Systems Research* 3(2), 49–56.

Johnston, J. E., Berry, K. J., & Mielke Jr, P. W. (2006). Measures of Effect Size for Chi-Squared and Likelihood-Ratio Goodness-of-Fit Tests. *Perceptual and Motor Skills* 103(2), 412–414.

Kalantari, R., & Gholami, J. (2017). Lexical Complexity Development from Dynamic Systems Theory Perspective: Lexical Density, Diversity, and Sophistication. *International Journal of Instruction* 10(4), 1–18.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA. Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Kubát, M., & Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20(4), 339–349.

Lee, J. J., Bychkovska, T., Maxwell, J. D. (2019). Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System* 80, 143–153.

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing* 33, 21–34.

Leedham, M., & Fernández-Parra, M. (2017). Recounting and reflecting: The use of first person pronouns in Chinese, Greek and British students' assignments in engineering. *Journal of English for Academic Purposes* 26, 66–77.

Li, T., & Liu, X. (2019). Ji yu yu liao ku gao zhong sheng ying yu shu mian yu kou yu hua te zheng yan jiu [A corpus-based study on the colloquial features in English writing produced by senior high students]. *Basic Foreign Language Education* 21(1), 3–11.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In: S. Granger (ed.), *Learner English on Computer*. London: Longman, 107–118.

Popescu, I.-I., Čech, R., & Altmann, G. (2011). *The Lambda-structure of Texts*. Lüdenscheid: RAM-Verlag.

Scherer, R. W., Dickersin, K., & Langenberg, P. (1994). Full publication of results initially presented in abstracts: A meta-analysis. *Jama* 272(2), 158–162.

Smith, J. A., & Kelly, C. (2002). Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities* 36, 411–430.

Snedeker, K. G., Totton, S. C., & Sargeant, J. M. (2010). Analysis of trends in the full publication of papers from conference abstracts involving pre-harvest or abattoir-level interventions against foodborne pathogens. *Preventive Veterinary Medicine* 95(1–2), 1–9.

Starke, M. D. D. J., & Bailer, C. (2019). Práticas de letramentos acadêmicos de alunos do Pibid interdisciplinar linguagens-Furb. *Revista EntreLínguas* 5(1), 195–209.

Wang, N. (2016). Investigating Grammatical Colloquial Features in EFL Learners' Theses by Chinese English Learners. *International Journal of English Linguistics* 6(6), 138–146.

Wen, Q. (2006). Ying yu zhuan ye xue sheng shi yong kou yu – bi yu ci hui de cha yi [Vocabulary variation across speech and writing produced by English majors]. *Foreign Languages and Their Teaching* 7, 9–13.

Wimmer, G., & Altmann, G. (1999). Review Article: On Vocabulary Richness. *Journal of Quantitative Linguistics* 6(1), 1–9.

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66, 130–141.

Yule, G. U. (1944). *A Statistical Study of Literary Vocabulary*. Cambridge: University Press.

Appendix: Texts Information

Texts in IRA

Text	Title
1	Unconventional Word Segmentation in Emerging Bilingual Students' Writing: A Longitudinal Analysis
2	Critical Analysis of CLIL: Taking Stock and Looking Forward
3	Discipline and Level Specificity in University Students' Written Vocabulary
4	An Investigation into Metaphor Use at Different Levels of Second Language Writing
5	Dynamics of Complexity and Accuracy: A Longitudinal Case Study of Advanced Untutored Development
6	Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy
7	Involvement in University Classroom Discourse: Register Variation and Interactivity
8	The Theoretical Research Article as a Reflection of Disciplinary Practices: The Case of Pure Mathematics
9	Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields
10	Marking Importance in Lectures: Interactive and Textual Orientation
11	Teacher Trainers' Beliefs About Feedback on Teaching Practice: Negotiating the Tensions Between Authoritativeness and Dialogic Space
12	An Activity-Theoretic Study of Agency and Identity in the Study Abroad Experiences of a Lesbian Nontraditional Learner of Korean
13	The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance
14	Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels
15	Implicit and Explicit Cognitive Processes in Incidental Vocabulary Acquisition
16	Individual Differences in Early Language Learning: A Study of English Learners of French
17	Exploring the Role of Phraseological Knowledge in Foreign Language Reading
18	Comprehension and Knowledge Components That Predict L2 Reading: A Latent-Trait Approach
19	A Longitudinal Study on the Impact of CLIL on Affective Factors
20	The Impact of Out-of-School Factors on Motivation to Learn English: Self-discrepancies, Beliefs, and Experiences of Self-authenticity
21	Fitting in or Standing out? A Conflict of Belonging and Identity in Intercultural Polite Talk at Work

22 Multilingual Children between Real and Imaginary Worlds: Language Play as Resignifying Practice

23 Evidence for the Adoption of the Flemma as an Appropriate Word Counting Unit

24 Building a Trustworthy Corporate Identity: A Corpus-Based Analysis of Stance in Annual and Corporate Social Responsibility Reports

25 Towards Ecological Validity in Research into Input-based Practice: Form Spotting Can Be as Beneficial as Form-meaning Practice

26 Effects of acoustic-phonetic detail on cross-language speech production

27 What can we learn from learning models about sensitivity to letter-order in visual word recognition?

28 *Nomina sunt consequentia rerum* – Sound–shape correspondences with every-day objects figures

29 Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese

30 The “sense boost” to dative priming: Evidence for sense-specific verb-structure links

31 Early semantic activation in a semantic categorization task with masked primes: Cascaded or not?

32 Examining the effects of active versus inactive bilingualism on executive control in a carefully matched non-immigrant sample

33 Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing

34 The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language

35 The effect of plausibility on eye movements in reading: Testing E-Z Reader’s null predictions

36 The persistence of syntactic priming revisited

37 Structural priming in artificial languages and the regularisation of unpredictable variation

38 Primed codeswitching in spontaneous bilingual dialogue

39 Unifying structural priming effects on syntactic choices and timing of sentence generation

40 The independence of syntactic processing in Mandarin: Evidence from structural priming

41 Phrase frequency effects in free recall: Evidence for redintegration

42 Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers

43 Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions

44 Serial position, output order, and list length effects for words presented on smartphones over very long intervals

45 The first- and second-language age of acquisition effect in first- and second-language book reading

46 Testing enhances memory for context

47 Voluntary language switching: When and why do bilinguals switch between their languages?

48 Listener sensitivity to probabilistic conditioning of sociolinguistic variables: The case of (ING)

49 How does foveal processing difficulty affect parafoveal processing during reading?

50 Semantic diversity, frequency and the development of lexical quality in children's word reading

51 Quantifying the development of phraseological competence in L2 English writing: An automated approach

52 Conceptualizing and measuring short-term changes in L2 writing complexity

53 Exploring multiple profiles of L2 writing using multi-dimensional analysis

54 Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners

55 L2 student–U.S. professor interactions through disciplinary writing assignments: An activity theory perspective

56 The effects of cognitive task complexity on writing complexity

57 What our students tell us: Perceptions of three multilingual students on their academic writing in first year

58 Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables

59 "We're drifting into strange territory here": What think-aloud protocols reveal about convenience editing

60 Exploring changes in FL writers' meaning-making choices in summary writing: A systemic functional approach

61 The relationship between lexical sophistication and independent and source-based writing

62 Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings

63 Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays

64 The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality

65 Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings

66 Motivation and feedback: How implicit theories of intelligence predict L2 writers' motivation and feedback orientation

67 Source text use by undergraduate post-novice L2 writers in disciplinary assignments: Progress and ongoing challenges

68 Using mind maps to reveal and develop genre knowledge in a graduate writing course

69	Emergent arguments: A functional approach to analyzing student challenges with the argument genre
70	Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis
71	Conceptualizations of language errors, standards, norms and nativeness in English for research publication purposes: An analysis of journal submission guidelines
72	An analysis of grammatical patterns in generation 1.5, L1 and L2 students' writings: A replication study
73	Balancing stability and flexibility in genre-based writing instruction: A case study of a novice L2 writing teacher
74	Articulating struggle: ESL students' perceived obstacles to success in a community college writing class
75	Synchronous and asynchronous teacher electronic feedback and learner uptake in ESL composition

Texts in CMT

Text	Title
1	A Corpus-based Study of English Verb Patterns in Marine Engineering English
2	A Corpus-based Study on Adjective Complementation
3	A Study of the Effect of Instruction under SCT on the Reading Achievement
4	Stylistic Analysis on Language Characteristics of Maritime Oral English
5	A Corpus-based Study on OVER in Maritime News English from the Cognitive Perspective
6	A Corpus-based Panchronic Study on Semi-auxiliaries
7	A Corpus-based Lexical Study in *the International Aeronautical and Maritime Search and Rescue Manual*
8	A Corpus-based Analysis of Stylistics of Headlines of Maritime News
9	The Study on Polysemous Word PUT from Cognitive Perspective
10	A Study on Three Metafunctions in *Marine News* Texts
11	A Corpus-based Study on Collocations of Key Words in Nautical English
12	Case Studies on the Effects of Text Summarization on Argumentation Writing Qualities of EFL Learners at Different Proficiency Levels
13	An Investigation into the Relationship between Junior High School Students' Foreign Language Anxiety, Emotional Intelligence and English Achievement
14	Semantic Features of Evaluative *It*-Clauses in the Research Articles by Chinese Writers
15	A Corpus-based Study on the Use of Shell Nouns in Marine Accident Investigation Report
16	A Study on Chinese College Students' Use of *Of*-Clusters and *Of*-Errors

17	A Corpus-based Comparative Study on *Turn* Verbs in MEC and BNC
18	A Case Study on the Impacts of Comprehensive Corrective Feedback on EFL Learners’ Written Accuracy
19	The Effect of Tasks on Collaborative Dialogue – An Empirical Study of English Majors at Dalian Maritime University
20	A Corpus-based Study on Nautical Lexis in Nautical Fiction
21	A New Analysis of *THERE* Sentences – from the Perspective of Copy Movement
22	A Contrastive Analysis of Hedges in English News from Chinese and American Newspapers
23	A Study of English Middle Construction Acquisition by Chinese EFL Learners
24	A Study on Characteristics of Contrastive Discourse Markers by Chinese English Majors
25	An Empirical Study on the Acquisition on English Verb Raising by Chinese College Students
26	An Empirical Study on the Acquisition of Wh-questions by Chinese EFL Learners
27	A Study on Chinese EFL Learners’ Article Acquisition from the Perspective of Syntax-pragmatics Interface
28	A Contrastive Study of Thematic Progression Patterns in English and Chinese Fairy Tales
29	A Corpus-based Study on the Use of Stance Adverbs in Chinese Learners’ Academic Writing
30	A Study on the Developmental Features of Lexical Bundles in Chinese English Learners' Academic Writing
31	Vocabulary Input in EFL Middle School Textbooks in China – A Corpus-based Study of Frequency, Complexity and Distribution
32	A Corpus-based Approach to the Interaction of English Verb Patterns with *it* and Registers
33	A Corpus-based Analysis of Recent Changes in American English Perfect Construction
34	Chinese Scholars' Perceptions of Writing for International Publication: An Applied Linguistics Perspective
35	Contextual Effects on the Processing Mechanism of Chinese 3VO Metaphor: An ERP Study
36	An Empirical Study on the Developmental Differences of Chinese EFL Learners' Implicit and Explicit Grammatical Knowledge
37	A Dependency Treebank-based Research on the Syntactic Complexity in Chinese EFL Learners’ Writings: A Developmental Perspective
38	Promotional Functions of Modal Verbs in the Introduction Sections of International Journal Articles: A Corpus-Based Analysis
39	Semantic and Affective Processing of Emotional Words in L1 and L2 in Unbalanced Bilinguals

40 Dependency Distance Differences across Interpreting Types

41 The Validity and Utility of the Component Unit in Chinese Sentence Chunking

42 A Study of Two Types of Writing Tasks in EFL Testing: Writing-only Tasks and Continuation Tasks

43 On Passive Voice Attrition among Non-English Majors from the Perspective of Dynamic Systems Theory

44 A Multidimensional Approach to Temporal Distance in Language

45 A Study on Demotivation of High School EFL Learners – A Case Study in a High School in Zhejiang Province