RiCL Research in Corpus Linguistics

How is information content distributed in RA introductions across disciplines? An entropy-based approach

Wei Xiao – Jin Liu – Li Li
Chongqing University / China

Abstract – Recent years have witnessed a growing interest in research article (RA thereafter) introductions. Most previous studies focused on the macro structures, rhetorical functions and linguistic realizations of RA introductions, but few intended to investigate the information content distribution from the perspective of information theory. The current study conducted an entropy-based study on the distributional patterns of information content in RA introductions and their variations across disciplines (humanities, natural sciences, and social sciences). Three indices, that is, one-, two-, and three-gram entropies, were used to analyze 120 RA introductions (40 introductions from each disciplinary area). The results reveal that, first, in RA introductions, the information content is unevenly distributed, with the information content of Move 1 being the highest, followed in sequence by Move 3 and Move 2; second, the three entropy indices may reflect different linguistic features of RA introductions; and, third, disciplinary variations of information content were found. In Move 1, the RA introductions of natural sciences are more informative than those of the other two disciplines, and in Move 3 the RA introductions of social sciences are more informative as well. This study has implications for genre-based instruction in the pedagogy of academic writing, as well as the broadening of the applications of quantitative corpus linguistic methods into less touched fields.

Keywords – RA introductions; move analysis; CARS model; entropy; information theory; disciplinary differences

1. INTRODUCTION[1]

Research articles (RAs) are regarded as a central genre of knowledge production and dissemination, as well as a key medium for the legitimating of claims and disciplines (Berkenkotter and Huckin 1995: 3; Hyland 2000: 175). As a "crafted rhetorical artifact"

[1] This work was presented at the *12th International Conference on Corpus Linguistics* (CILC 2021). It was supported by the Social Science Foundation of Chongqing under Grant No. 2019QNYY51, the Fund of the Interdisciplinary Supervisor Team for Graduate Programs of Chongqing Municipal Education Commission under Grant No. YDSTD1923, the Graduate Research Innovation Program of Chongqing under Grant No. CYS20045 and the Fundamental Research Funds for the Central Universities under Grant No. 2021CDJSKZX07. We would like to extend our sincere gratitude and appreciation to the anonymous reviewers for their comments and suggestions.

and a “manifestation of rhetorical maneuver” (Swales 1990: 155), introductions play a critical role in RAs. The introduction section not only provides the interpretive structure that illustrates how readers may decode a study (Grant and Pollock 2011) but also functions to state the significance of the research field by concisely situating the actual research (Swales 1990: 142). Such salience and identifiability have made introductions the focus of a great number of researchers (e.g. Samraj 2002; Hirano 2009; Loi and Evans 2010; del Saz Rubio 2011).

One line of research in this regard has concentrated on the macro structures of RA introductions. Macroscopically, RA introductions can be regarded as composed of several moves, which work together as functional units in a given text (Connor *et al.* 1995) and can be illustrated by their specific communicative purposes and linguistic devices. Previous researchers have investigated the occurrences, sequences and patterns of moves of RA introductions (Samraj 2002; Fakhri 2004; Kanoksilapatham 2005; Hirano 2009; Loi 2010; Sheldon 2011; Lim 2012; Ahamad and Yusof 2012; Muangsamai 2018; Ye 2019) and have indicated that their macro-organization mainly follows Swales’ (1990, 2004) Create-A-Research-Space (CARS) model, i.e. (1) Establishing a Territory, (2) Establishing a Niche and (3) Occupying the Niche, with some conventional moves appearing more frequently than some others that are elective (Cortes 2013). These studies, from a macro perspective, have unraveled the combination and sequences of rhetorical strategies used in RA introductions.

Another branch of research has dealt with microscopic concerns. Some delved into metadiscourse features, such as interactive and reflexive features (Kashiha and Marandi 2019; Li and Xu 2020), the usage of metadiscourse markers in the rhetorical moves of RA introductions (del Saz Rubio 2011; Khedri and Kritsis 2018) and rhetorical and metadiscoursal variations (Kim and Lim 2013; Validi *et al.* 2016). Del Saz Rubio (2011) analyzed the metadiscoursal features in RA introductions, and found that evidentials, transition markers, and code glosses were the most omnipresent interactive categories. Kim and Lim (2013) examined the use of metadiscourse in RA introductions of educational psychology and found that far more interactive than interactional metadiscourse markers were used. Some others concentrated on lexico-grammatical features, such as lexical bundles (Cortes 2013; Esfandiari and Barbary 2017; Mizumoto *et al.* 2017), phraseological units (Liu and Lu 2020), and linguistic realizations utilized to fulfill a specific rhetorical function in the moves of RA introductions (see Lim 2012; Ädel

2014; Joseph *et al.* 2014; Tankó 2017; Lu *et al.* 2020). For instance, Lim (2012) examined how sophisticated writers in the field of management employ linguistic choices to establish research niches. The findings showed that gap indications in management RA introductions were realized by the employment of such expressions as negative verb phrases and attributive quantifiers demonstrating inadequacy. From the perspective of appraisal and evaluation, Wang and Yang (2015) investigated how the promotion was realized in RA introductions in applied linguistics, and what appeals and linguistic devices could be used to show the significance of the study. Their results show that claiming centrality occupied a prominent role in research promotion and worthiness indication. These explorations are of great significance in that they have deepened our understanding of how these rhetorical and linguistic features of RA introductions function in knowledge construction and communication.

Researchers have also been interested in the variations of RA introductions across disciplines (Holmes 1997; Samraj 2002; Ozturk 2007; Lin and Evans 2012; Kanoksilapatham 2015). Samraj (2002), for example, studied the disciplinary variations in move structures by comparing RA introductions of conservation biology and wildlife behavior. She found that the conservation biology introductions were more likely to use such steps as centrality claims and were more concerned with real-world matter instead of the epistemic world of research than those wildlife behavior introductions. Martín and León Pérez (2014) investigated how researchers in the fields of health sciences and social sciences promote their research in Move 3 (Occupying the Niche) and the corresponding linguistic realizations of each step as well. The results demonstrated that health sciences texts showed a higher degree of rhetorical promotion than those of social sciences. These cross-disciplinary investigations have revealed significant differences in move patterns of RA introductions. They have also promoted the awareness of disciplinary differences in the academic community and shed light on academic writing teaching and training.

Recently there have been a few attempts to analyze texts from the perspective of information theory, a field concerned with the concept and measurement of information (van der Lubbe 1997: 1). Entropy, an index widely used in information theory which allows to measure the information content of a message, has been introduced into language studies. This index is different from some widely used indices in corpus linguistics (e.g. TTR) in that it not only considers the variety of words but also their evenness of distribution (Zhu and Lei 2018). The relevant entropy-based studies have

covered such topics as linguistic complexity (Juola 2008; Lu 2012; Ehret and Szmrecsanyi 2014, 2019), cultural complexity (Juola 2013; Khany and Kafshgar 2016; Zhu and Lei 2018), and the information content of certain linguistic entities (Chen *et al.* 2016). RA introductions, regarded as essentially opening paragraphs for writers to prepare the ground for the research to come (del Saz Rubio 2011), are also worth investigating from this perspective. Considering that different moves in RA introductions are expected to achieve particular communicative functions (Shehzad 2010) and the salience of different moves appears dissimilar, their distribution of information content may differ. However, to the best of our knowledge, no entropy-based study has been conducted on RA introductions and their information distribution patterns, in particular, something that will be addressed in the present study. In addition, because of the shortage of relevant literature, there is not yet a clear understanding of the substantial significance of different entropies in academic texts, especially in RA introductions. For example, Juola (2013) managed to measure the complexity of the American culture by calculating the entropy of different grams in the *Google Books N-gram Corpus* and thought that one-gram entropy is the indicator of lexical complexity, two-gram entropy reveals relationships between two linguistic entities, and three-gram entropy reflects syntactic complexity. In another study, Zhu and Lei (2018) analyzed the speeches and debates from the British parliament and found that different N-grams may indicate distinct linguistic features, which confirmed Juola's (2013) claims. Despite their explorations, they have dealt with limited genres other than RA introductions. Their findings thus may not be conclusive. This study will then employ three indices to measure information content, hoping to uncover the correlates between different entropy indices and RA introduction features. Furthermore, as disciplinary variations have been widely documented in previous studies (e.g. Samraj 2002; Hyland 2000; Ozturk 2007; Kanoksilapatham 2015), there may be different information distribution patterns of RA introductions across disciplines, which awaits exploration. To summarize, the present study aims to answer the following three questions:

(1) How is information content distributed in RA introduction moves?
(2) Do different entropy indices reveal different linguistic features of RA introductions?
(3) Are there any variations in information distributional patterns across disciplines?

2. METHODS

2.1. Corpora

Research articles used in the present study were selected via the Web of Science (WoS) search engine. The collected articles were expected to meet the following criteria: 1) they were from high-ranking journals indicated by the 2019 SCI/SSCI/AHCI indices; 2) they were published in the latest years (2017–2020); 3) they were organized in a typical IMRD (Introduction-Methods-Results-Discussion) structure; 4) they were roughly similar in length (8,000 words); and 5) they were written in English. These standards ensured the representativity, reputation and accessibility of the paper collection (see Nwogu 1997). In the end, 120 research articles were randomly selected: 40 research articles in natural sciences,[2] social sciences,[3] and humanities.[4] *AntFileConverter* (Anthony 2017), a convenient and free software, was then applied to convert the articles into plain text format. The average length per text of natural sciences, social sciences, and humanities was 8,563, 9,163 and 8,873 words, respectively. One-way ANOVA[5] and post-hoc tests showed that there was no significant difference at the $p<0.05$ level in text length across the three disciplines [$F(2, 117) = 0.029$, $p=0.97$].

2.2. Information content and entropy

In information theory, the information content of a message can be measured by entropy, an index first introduced by Shannon (1948). Higher entropy indicates more information content. In language studies, the entropy of a given text can be calculated by the following formula, where P_i refers to the probability of the relative frequency of the ith word and N stands for the total number of word types in a given text.

$$H_t = -\sum_{i=1}^{N} P_i \log_2 P_i$$

[2] From *Chemical Engineering Journal*, *Computer Communications*, *Microporous and Mesoporous Materials*, *Future Generation Computer Systems* and *Atmospheric Environment*.
[3] From *Human Relations*, *International Journal of Research in Marketing*, *International Journal of Information Management*, *World Development* and *Telematics and Informatics*.
[4] From *Journal of Second Language Writing*, *Journal of Archaeological Science*, *Political Geography*, *Digital Applications in Archaeology and Cultural Heritage* and *Language Learning*.
[5] One-way ANOVA is a statistical test used to compare means for three or more groups. However, ANOVA can only tell us that there is a significant difference but cannot suggest between which groups the difference exists. The post-hoc test, therefore, should be conducted to make multiple comparisions between every two groups, so as to find out where exactly the difference lies.

As a word can also be considered as a one-gram (Zhu and Lei 2017), the entropy of a word can be regarded as one-gram entropy. Likewise, the formula above can be used to calculate two-gram or three-gram entropy if we replace the set of one-grams by two- or three-grams. Operationally, an n-gram can be identified by extracting every *n*-word in adjacency within a sentence, as illustrated in example (1).

(1) This is a sentence. And here is another sentence.

In the example above, all the one-grams are *this*, *is*, *a*, *sentence*, *and*, *here*, *is*, *another*, and *sentence.* All two-grams are *this is*, *is a*, *a sentence*, *and here*, *here is*, *is another*, and *another sentence*. And all three-grams are *this is a*, *is a sentence*, *and here is*, *here is another*, and *is another sentence*.

2.3. Move annotation

We adopted Swales' (2004) CARS model as the framework of annotation at the sentence level. Each sentence was analyzed and labeled as Establishing a Territory (Move 1), Establishing a Niche (Move 2), or Occupying the Niche (Move 3). This is illustrated in (2).

(2) The aim of this study is to establish the accuracy of different methods to obtain wave conditions in shallow water for nearshore studies, with a special focus on the wave direction. (de Swart *et al.* 2020: 2).

In example (2), *the aim of this study* refers to a prefacing and preparative expression, informing the readers to focus on the objectives of the study, which serves as a common way for authors to occupy the niche. Therefore, this sentence was labeled as Move 3.

To ensure the coding reliability, two researchers first annotated the sentences independently. The consistency coefficient was as high as 93 percent. The inconsistencies were then left for discussion until a final agreement was reached.

2.4. Data analysis

After the move annotation process, the *R* programming language (version 3.5.0, R Core Team 2018) was used for the extraction of one-, two-, and three-grams, the calculation of entropy values and further data analysis. First, we programmed to extract all the one-,

two-, and three-grams. By removing all the punctuations, we split the texts into words and obtained all the one-grams. For two-grams and three-grams, we split the texts into sentences with punctuation marks as boundaries and extracted all the two-grams and three-grams sentence by sentence. Second, we calculated the probabilities of every one-, two-, and three-gram by dividing the occurrence of a gram by the total grams. Third, we calculated the one-, two-, and three-gram entropies of each move according to the aforementioned formula.

After the calculation of one-, two-, and three-gram entropies of each move, we wrote an *R* script to conduct statistical analysis. First, we used the Shapiro-Wilk test[6] to check whether the original data were normally distributed. When the data were normally distributed, we applied the ANOVA test, in order to examine whether statistically significant difference exists across moves, grams, or disciplines; we then used the Tukey post-hoc analyses to check precisely where the significant difference appeared. When the data were not normally distributed, we then used the Kruskal-Wallis test to check whether there existed significant between-group difference (indicated by *p* value). The significance level was set at 0.05.

3. RESULTS

The group means of one-gram, two-gram and three-gram entropies across RA introductions moves are presented in Figure 1. The descriptive statistics and the corresponding post-hoc test results are displayed in Table 1. From Figure 1 (A), it can be observed that the information content of the three moves is unevenly distributed. The mean entropy of Move 1 is the highest, followed in sequence by Move 3 and Move 2. The entropy values of two-grams and three-grams, as represented in Figure 1 (B and C), suggest a similar distribution pattern in information content as one-gram entropy unravels. The Kruskal-Wallis test results in Table 1 demonstrate that significant differences appear across moves, and the post-hoc analysis further affirms significant differences across moves ($p<0.05$), independently of whether we are dealing with one-, two- or three-grams. These results indicate that Move 1 may carry more information and occupy a larger part in RA introductions than the other two moves, whereas Move 2 tends to be less

[6] The Shapiro-Wilk test is to test whether the data is normally distributed and then to indicate which hypothesis test method is appropriate. If $p>0.05$, it shows that the data is normally distributed and ANOVA should then be used. If $p<0.05$, it shows a non-normal distribution and Kruskal-Wallis test should be used.

informative in the introduction part.

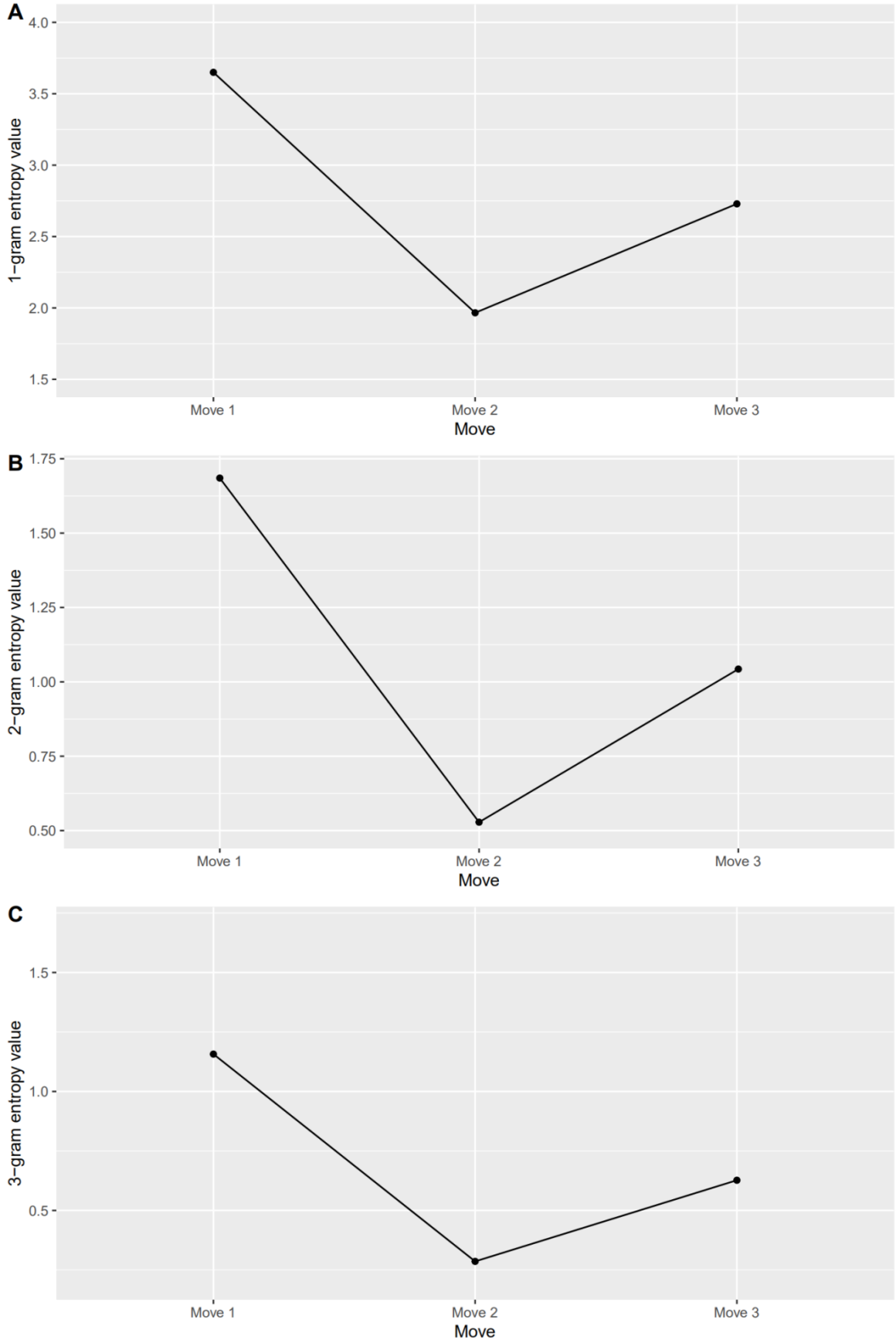

Figure 1: Mean entropy value of one-grams, two-grams, and three-grams across moves

	Mean (SD in parentheses)					Z		
	M1	M2	M3	X^2	*p*	M1-M2	M2-M3	M1-M3
one-grams	3.650(0.753)	1.966(0.605)	2.729(0.846)	174.27***	0.000	13.191***	-6.149***	7.042***
two-grams	1.685(0.761)	0.528(0.294)	1.043(0.575)	171.3***	0.000	13.081***	-6.904***	6.178***
three-grams	1.157(0.671)	0.286(0.191)	0.627(0.406)	176.7***	0.000	13.290***	-6.889***	6.401***

Table 1: Kruskal-Wallis test results of entropy values across moves[7]

The fact that different indices yield similar information distribution patterns can also be seen in Figure 2 and is indicated by the *r* values of correlation analyses in Table 2, where there are correlations among the values of one-, two-, and three-gram entropies (*r*s>0.8, *p*s<0.05), confirming the consistency of the three indices. Despite their similarities, the indices show some differences in terms of values, as indicated by the *p* values of pairwise t-tests in Table 3 (*ps*<0.05). These results suggest that entropies of different grams, though highly correlated, may reveal different linguistic features of RA introductions.

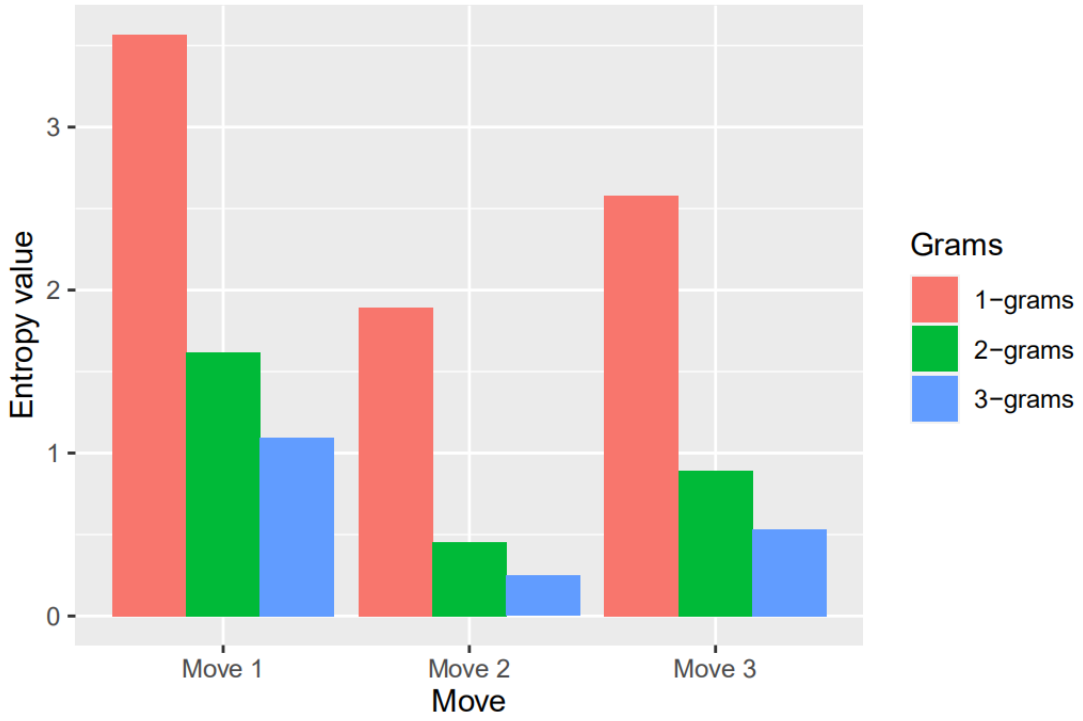

Figure 2: Comparisons among one-, two-, and three-gram entropies

[7] In this table, three asterisks (***) mean *p*<0.001.

Move	one-grams - two-grams	two-grams - three-grams	one-grams - three-grams
M 1	0.915***	0.975***	0.852***
M 2	0.897***	0.962***	0.841***
M 3	0.944***	0.964***	0.874***

Table 2: Pearson correlation coefficients between one-, two-, and three-gram entropy values in each move

Move	one-grams - two-grams	two-grams - three-grams	one-grams - three-grams
M 1	1.964***	0.529***	2.493***
M 2	1.437***	0.242***	1.679***
M 3	1.686***	0.416***	2.102***

Table 3: Pairwise t-tests between one-, two-, and three-gram entropy values in each move

Figure 3 represents the information distribution patterns of RA introductions across disciplinary areas (natural sciences, social sciences, and humanities). It can be observed that, in general, RA introductions across the three areas share a similar distribution pattern in information content. The mean of one-gram entropy of Move 1, given any discipline, is much higher than that of Move 3, followed by an even lower entropy of Move 2. The general informative distribution patterns of two-grams and three-grams, to a large extent, resemble that of one-grams. These results indicate that RA introductions from different domains share similarities in information distribution patterns.

With regard to cross-disciplinary variations, the information content of Move 1 of natural sciences is the highest, as indicated by the higher entropies of Move 1. The entropies of Move 1 of humanities and social sciences seem to be much lower. However, the *p* values of the ANOVA test in Table 4 suggest that there is no interdisciplinary variation in this move ($ps>0.05$). Despite higher entropy values of social sciences in Move 2, no significant cross-disciplinary variations were found in this move, neither ($ps>0.05$). In fact, significant differences across disciplines only occur in Move 3, as indicated by the *p* values of the ANOVA tests (see Table 4). Post-hoc tests further indicate significant differences among social sciences and the other two disciplines, where the mean entropy values of social sciences are considerably higher than that of natural sciences and humanities. Interestingly, there is no significant difference between natural sciences and humanities ($ps>0.05$).

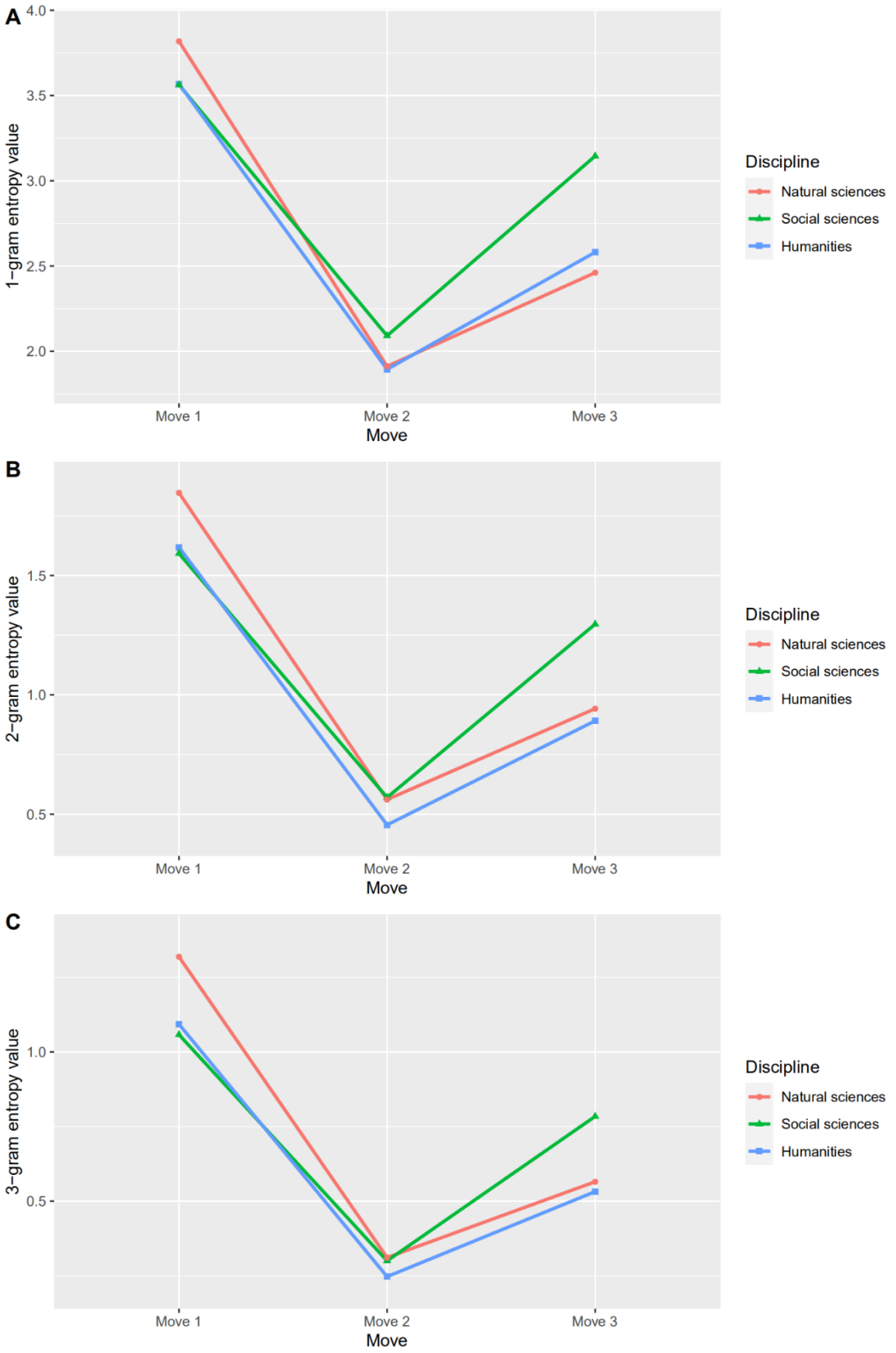

Figure 3: Entropy values of one-grams, two-grams, and three-grams across disciplines

		Mean (SD in parenthesis)					Mean difference		
		NS	SS	HM	*F*	*p*	NS-SS	NS-HM	SS-HM
one-grams	M1	3.818 (0.664)	3.563 (0.690)	3.568 (0.878)	1.507	0.226	0.255	0.249	-0.005
	M2	1.912(0.595)	2.091 (0.638)	1.894 (0.577)	1.294	0.278	-0.178	0.018	0.197
	M3	2.461 (0.655)	3.145 (0.880)	2.581 (0.841)	8.359***	0.000	-0.683***	-0.120	0.564***
two-grams	M1	1.846 (0.810)	1.592 (0.726)	1.618 (0.740)	1.354	0.262	0.254	0.228	-0.027
	M2	0.561 (0.293)	0.570 (0.319)	0.455 (0.261)	1.926	0.150	-0.009	0.106	0.115
	M3	0.942 (0.444)	1.296 (0.663)	0.891 (0.523)	6.407**	0.002	-0.354*	0.050	0.405**
three-grams	M1	1.319 (0.751)	1.058 (0.610)	1.093 (0.630)	1.808	0.167	0.261	0.226	-0.036
	M2	0.311 (0.190)	0.300 (0.202)	0.247 (0.178)	1.283	0.281	0.011	0.064	0.052
	M3	0.565 (0.302)	0.784 (0.500)	0.532 (0.352)	4.858**	0.009	-0.219*	0.033	0.252*

Table 4: Cross-disciplinary variations of entropy values[8]

4. DISCUSSION

The current study aimed to explore the information features of RA introductions, focusing on the information content distribution across moves, the linguistics features reflected by different entropy indices, and the cross-disciplinary variations. The moves of RA introductions were identified according to the CARS model, and their one-, two-, and three-gram entropies were calculated and compared. Our findings show that the information content is unevenly distributed across moves, where Move 1 takes up the largest proportion, followed in sequence by Move 3 and Move 2. Besides, the information measured by the three indices represents a similar pattern, although both similarities and variations were found across disciplines.

4.1. The informative distributional pattern of RA introduction moves

The information content in RA introductions was found to be unevenly distributed. The unevenness could be ascribed to the distinctive rhetorical functions of moves and their variations in salience in an RA introduction.

Move 1 (Establishing a Territory) was found to be the most informative one.

[8] In this table, one asterisk (*) means $p<0.05$, two asterisks (**) imply $p<0.01$, and three asterisks (***) are used to indicate $p<0.001$; NS stands for natural sciences, SS for social sciences, and HM for humanities.

According to Swales' (1990) CARS model, RA introductions often begin with a move that aims to assure readers that the general topic being discussed is worth investigating and that the field of the study is well established through an exhaustive review of the previous studies in the area. As Lim (2012) suggested, Move 1 constitutes a fundamental rhetorical move in RAs in various domains, and researchers are required to acquire sufficient background knowledge to meet the expectations of the academic community. Saricaoglu *et al*. (2021) also highlighted that reference to previous literature is a defining feature of typically all research writings. The literature review or background information thus plays a critical role and occupies a large amount of content in RA introductions. Therefore, it is rather understandable that Move 1 takes up the highest information proportion in RA introductions.

Move 2 (Establishing a Niche), serving as the hinge that connects Move 1 with Move 3 (Swales and Feak 2012: 348), seems to contain much less information, which can also be accounted for by the rhetorical function this specific move bears on. In Move 2, writers are supposed to utilize evaluative resources, mostly negative evaluations or even criticisms, to assess previous literature to create a research niche (Lim 2012). According to Ahmad (1997), Move 2 is used to identify the validation of the piece of research through the description of perceived limitations of the research field or summarizing the work of others. This primary function of Move 2 is often realized through short statements (Xie 2017). As Lindeberg (2004: 139) suggests, Move 2, functioning as a 'mini-critique', shows a preference for conciseness and often entails no more than one sentence, thus occupying a relatively short length in the introduction section. Swales and Feak (2004) also recommend that gap indications in Move 2 should be simple and fairly short due to their easy-to-follow and straightforward manner. Besides, although the way of revealing these limitations is variable, either counter-claiming, indicating a gap in previous studies, question-raising, or continuing a tradition, the phrases and expressions are relatively rigorous and tend to have more syntactic regularities. Shehzad (2008), for instance, pointed out that the gap statements were often realized by contrastive markers (e.g. *nevertheless*, *however*, etc.), quantifiers (e.g. *relatively few*), and negative statements (e.g. *none*, *no study*, etc.). This precise and rigorous nature may lead to the comparatively less informative proportion of Move 2.

Another possibility of the lower content of information in Move 2 could be related to socio-cultural factors. Due to the significance of face-saving in academic communities,

researchers may find it difficult and inappropriate to identify previous studies and point out the possible shortcomings and limitations that previous works might have (Taylor and Chen 1991; Kanoksilapatham 2005; Hamp-Lyons and Heasley 2006: 45). Taylor and Chen (1991), for example, found that the authors of Chinese papers were less likely to evaluate previous studies and provide less extensive discussions of other scholars' works owing to the unacceptability of argumentations and confrontations in the Chinese socio-cultural context. The reasons mentioned above may also contribute to the less information contribution of Move 2.

As for Move 3 (Occupying the Niche), the primary function is to turn the niche established in Move 2 into the research space that identifies the present research (Swales 1990). Essential components of this move include descriptions of the research purposes and the current work to be carried out, announcements of principal findings, and sometimes, presentations of the organization of the present paper. As Martín and León Pérez (2014) state, it is in Move 3 that scholars essentially present their research by outlining the research purpose. Move 3 plays a vital role in convincing peers of the relevance and validation of the study to be conducted, in which various promotional strategies are used to highlight the value of their work. Writers concentrate on the use of various rhetorical strategies to promote their 'selling point' (Hyland 2000; Shehzad 2010) through anticipating the principal findings or highlighting the significance, newness, and contributions that their work makes to the field. Shehzad (2010) also suggests that the introductions in computer sciences are result-oriented, where scholars would elaborate explanations in the description of results and highlight the writers' contributions in various ways. Scholars in this field conduct elaborate explanations in the description of results and in various ways highlight the writers' contributions to the field.

Since different moves play different roles in the organization of an RA introduction and the focus of differing moves is not identical, it is no wonder that variations in information distribution across moves exist.

4.2. The informative variations across different grams

The information distribution patterns revealed by different indices are of great similarity. This finding is in accordance with that of Zhu and Lei (2018), in which similar patterns were also found among the one-, two-, and three-gram entropies, thus corroborating the

usefulness of entropy as a measurement of information content. Despite the similarity, our results are different from those of Zhu and Lei (2018) in that we found larger one-gram entropy values than two- and three-gram ones, while they found the opposite. The possible reasons are twofold. First, the data employed in our study are academic articles while Zhu and Lei (2018) collected materials of spoken texts (speeches of the British Parliament). Thus, the divergence may be explained across genres. Second, this difference could be ascribed to the authentic and scientific nature of RAs, which are featured with prescriptive organization patterns (Ren and Li 2011; Taş 2008). There are a variety of recommended sentence patterns or conventional phrases for writers' reference (Cross and Oppenheim 2006). In order to be accepted by the academic communities, scholars have to conform to these academic conventions and increasingly show a preference for a restricted repertoire of identified moves (Holmes 1997). Hence, fixed patterns seem to lead to a lower syntactic complexity. As Juola (2013) suggested, one-gram entropy reflects lexical complexity, whereas two-gram entropy reveals relationships between two entities, and three-gram entropy is more concerned with syntactic complexity. Thus, it is explainable that the entropy value of two- and three-grams are considerably lower than that of one-grams.

4.3. The information distribution patterns of RA introduction moves across disciplines

With regard to disciplinary variations, significant differences can be found in Move 3, where the information content of social sciences was considerably higher than those of the other two disciplines. This finding could be accounted for by the discursive nature of social sciences. As Kuteeva and Airey (2014) state, unlike natural sciences (also regarded as hard sciences) that stress the quantitative and experimental nature of materials, social sciences (typically considered as soft sciences) emphasize the qualitative and interpretive disposition. In social sciences, knowledge is normally regarded as a process of making interpretations of a certain issue or some social phenomenon. The way of interpretation is very important in the process of persuasion of readers. Writers of social sciences may have diverse personal writing styles, which can be reflected through the employment of new words or innovative expressions, thus contributing to the informativeness of the text. Our findings are consistent with some previous studies. Lu *et al.* (2020), for example, investigated the rhetorical functions of syntactically complex sentences in RA introductions of social sciences and found that writers in social sciences tend to elaborate

on announcing the piece of research by the use of a variety of structures at the phrasal and clausal levels, giving rise to greater syntactic complexity. The occurrence of more new word types in Move 3 could be an indication of the discursive nature of social sciences, which is captured by the higher entropy values.

It should also be noted that the information content of Move 1 in natural sciences is higher than that of social sciences and humanities, although no significant difference appears. The emphasis on Move 1 can be explained by the accumulative and iterative nature of natural sciences. In natural sciences, the processes of knowledge construction rely heavily on solid foundations accumulated by experimental support or empirical data (Kuteeva and Airey 2014). Compared with those in social sciences and humanities, works in natural sciences are typically featured with shared paradigms, in which references to previous literature are of great importance. Our findings are in accordance with some previous studies. Samraj (2008), for instance, explored the introductions from three disciplines (linguistics, philosophy, and biology) and found that biology students tend to establish stronger links to previous works. Another possibility could be related to the richness of content words in natural sciences. Due to the typical feature of ready-made paradigms in natural sciences, researchers have to resort to an essential number of prefabricated concepts and technical terms, hoping to establish a common background with peers in the same field, and to be accepted by the academic community (Xiao and Sun 2020). In addition, as science and technology are rapidly developing, a flood of new terms and concepts keeps emerging in this field, thus contributing to the complexity in establishing a territory. These reasons may give rise to a higher information proportion in the Move 1 of natural sciences.

5. Conclusion

The present study investigated the information contents of RA introductions and the variations across disciplines. An entropy-based approach was employed and three quantitative indices, that is, one-, two-, and three-gram entropies, were adopted. It was found that the information content is unevenly distributed in introductions, where Move 1 tends to be more informative than the other two moves. It was also found that the one-gram entropy values were higher than the other two indices. Furthermore, disciplinary variations were also attested. In Move 1, the RA introductions of natural sciences are more informative than those of the other two disciplines, and in Move 3 the RA

introductions of social sciences are more informative. These differences may be explained by the rhetorical and linguistic features of individual moves, the different aspects possibly reflected by different indices and the very nature of distinctive disciplines.

This study is a preliminary attempt to explore RA introductions from the perspective of information theory. It demonstrates the promising prospects of using quantitative linguistics methods in the study of RA introductions and many other genres, where traditional qualitative methods and basic statistics still prevail. The findings suggest how information content is supposed to be distributed and arranged in RA introductions, which is difficult to find with mere qualitative methods. This study also sheds light on the pedagogy of academic writing, in that it can help students to be aware of the structure of introductions and the prominence of each move, as well as the potential conventions and paradigms of their own disciplines. With this awareness in mind, students would be better equipped in the organization of introductions, the allocation of information and the promotion of research, which will eventually lead to successful academic writing and reputation winning.

There are also several limitations in this study. First, due to the laborious manual coding processes, we only analyzed 120 introductions. Future studies may enlarge the corpus size to ensure the stableness of results. Second, different disciplines and even sub-disciplines may have their own features, which threats the homogeneity of constitution and challenges the representativeness and generalizability of sampling and corpus building. Future studies may use data from a wider range of disciplines to cross validate our results. Third, although we managed to reveal the significant differences of entropies of different grams in RA introductions, more fine-grained measurements, that is, the four-, five-, six-gram entropies were not explored, which awaits further research. Fourth, we only investigated the introduction part of RAs. Future studies may extend the focus to other parts, such as abstracts, to unravel more features of information distribution in research articles.

REFERENCES

Ädel, Annelie. 2014. Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes* 16: 68–80.

Ahamad, Mohamed I. and Amira M. Yusof. 2012. A genre analysis of Islamic academic research article introductions. *Procedia - Social and Behavioral Sciences* 66: 157–168.

Ahmad, Ummul. 1997. Research article introductions in Malay: Rhetoric in an emerging research community. In Anna Duszak ed. *Culture and Styles in Academic Discourse*. Berlín: Mouton de Gruyter, 273–303.

Anthony, Laurence. 2017. *AntFileConverter* (version 1.2.1). Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/ (01 May, 2020.)

Berkenkotter, Carol and Thomas N. Huckin. 1995. *Genre knowledge in disciplinary communication: Cognition/Culture/Power*. New Jersey: Lawrence Erlbaum Associates.

Chen, Ruina, Haitao Liu and Gabriel Altmann. 2016. Entropy in different text types. *Digital Scholarship in the Humanities* 32/3: 528–542.

Connor, Ulla, Kenneth Davis and Teun de Rycker. 1995. Correctness and clarity in applying for overseas jobs: A cross-cultural analysis of US and Flemish applications. *Text & Talk* 15/4: 457–475.

Cortes, Viviana. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes* 12/1: 33–43.

Cross, Cate and Charles Oppenheim. 2006. A genre analysis of scientific abstracts. *Journal of Documentation* 62: 428–446.

De Swart, Rinse, Francesca Ribas, Daniel Calvete, Aart Kroon, and Alejandro Orfila. 2020. Optimal estimations of directional wave conditions for nearshore field studies. *Continental Shelf Research* 196: 104071.

Del Saz Rubio, M. Milagros. 2011. A pragmatic approach to the macro-structure and metadiscoursal features of research article introductions in the field of Agricultural Sciences. *English for Specific Purposes* 30/4: 258–271.

Ehret, Katharina and Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. In Raffaela Baechler and Guido Seiler eds. *Complexity, Isolation, and Variation*. Berlin: Boston De Gruyter, 71–94.

Ehret, Katharina and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in SLA production data. *Second Language Research* 35/1: 23–45.

Esfandiari, Rajab and Fatima Barbary. 2017. A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes* 29: 21–42.

Fakhri, Ahmed. 2004. Rhetorical properties of Arabic research article introductions. *Journal of Pragmatics* 36/6: 1119–1138.

Grant, Adam M. and Timothy G. Pollock. 2011. Publishing in AMJ—part 3: Setting the hook. *Academy of Management Journal* 54/5: 873–879.

Hamp-Lyons Liz and Ben Heasley. 2006. *Study Writing: A Course in Writing Skills for Academic Purposes*. Cambridge: Cambridge University Press.

Hirano, Eliana. 2009. Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English. *English for Specific Purposes* 28: 240–250.

Holmes, Richard. 1997. Genre analysis, and the social sciences: An investigation of the structure of research article discussions sections in three disciplines. *English for Specific Purposes* 16: 321–337.

Hyland, Ken. 2000. *Disciplinary Discourse: Social Interactions in Academic Writing*. London: Longman.

Joseph, Renu, Jason Miin-Hwa Lim and Nor Arifah Mohd. 2014. Communicative moves in forestry research introductions: Implications for the design of learning materials. *Procedia - Social and Behavioral Sciences* 134: 53–69.

Juola, Patrick. 2008. *Assessing Linguistic Complexity.* Amsterdam: John Benjamins.

Juola, Patrick. 2013. Using the Google N-Gram corpus to measure cultural complexity. *Literary and Linguistic Computing* 28/4: 668–675.

Kanoksilapatham, Budsaba. 2005. Rhetorical structure of biochemistry research articles. *English for Specific Purposes* 24/3: 269–292.

Kanoksilapatham, Budsaba. 2015. Distinguishing textual features characterizing structural variation in research articles across three engineering sub-discipline corpora. *English for Specific Purposes* 37: 74–86.

Kashiha, Hadi and Susan S. Marandi. 2019. Rhetoric-specific features of interactive metadiscourse in introduction moves: A case of discipline awareness. *Southern African Linguistics and Applied Language Studies* 37/1: 1–14.

Khany, Reza, and Neda Babanezhad Kafshgar. 2016. Analyzing texts through their linguistic properties: A cross-disciplinary study. *Journal of Quantitative Linguistics* 23/1: 278–294.

Khedri, Mohsen and Konstantinos Kritsis. 2018. Metadiscourse in applied linguistics and chemistry research article introductions. *Journal of Research in Applied Linguistics* 9/2: 47–73.

Kim, Loi Check and Jason Miin-Hwa. 2013. Metadiscourse in English and Chinese research article introductions. *Discourse Studies* 15/2: 129–146.

Kuteeva, Maria and John Airey. 2014. Disciplinary differences in the use of English in higher education: Reflections on recent policy developments. *Higher Education* 67/5: 533–549.

Li, Zhijun and Jinfen Xu. 2020. Reflexive metadiscourse in Chinese and English sociology research article introductions and discussions. *Journal of Pragmatics* 159: 47–59.

Lim, Jason Miin-Hwa. 2012. How do writers establish research niches? A genre-based investigation into management researchers' rhetorical steps and linguistic mechanisms. *Journal of English for Academic Purposes* 11/3: 229–245.

Lin, Ling and Stephen Evans. 2012. Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes* 31/3: 150–160.

Lindeberg, Ann C. 2004. *Promotion and Politeness: Conflicting Scholarly Rhetoric in Three Disciplines*. Pargas, Finland: Åbo Akademi University Press.

Loi, Chek Kim and Moyra Sweetnam Evans. 2010. Cultural differences in the organization of research article introductions from the field of educational psychology: English and Chinese. *Journal of Pragmatics* 42/10: 2814–2825.

Lu, Xiaofei. 2012. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45/1: 36–62.

Lu, Xiaofei, J. Elliott Casal and Yingying Liu. 2020. The rhetorical functions of syntactically complex sentences in social science research article introductions. *Journal of English for Academic Purposes* 44: Article 100832.

Martín, Pedro and Isabel K. León Pérez. 2014. Convincing peers of the value of one's research: A genre analysis of rhetorical promotion in academic texts. *English for Specific Purposes* 34/1: 1–13.

Mizumoto, Atsushi, Hamatani Sawako and Imao Yasuhiro. 2017. Applying the bundle-move connection approach to the development of an online writing support tool for research articles. *Language Learning* 67/4: 885–921.

Muangsamai, Pornsiri. 2018. Analysis of moves, rhetorical patterns and linguistic features in New Scientist articles. *Kasetsart Journal of Social Sciences* 39/2: 236–243.

Nwogu, Kevin Ngozi. 1997. The medical research paper: Structure and functions. *English for Specific Purposes* 16/2: 119–138.

Ozturk, Ismet. 2007. The textual organization of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes* 26: 25–38.

R Core Team. 2018. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (01 May, 2020.)

Ren, Hongwei, and Yuying Li. 2011. A comparison study on the rhetorical moves of abstracts in published research articles and master's foreign-language theses. *English Language Teaching* 4/1: 162–166.

Samraj, Betty. 2002. Introductions in research articles: Variations across disciplines. *English for Specific Purposes* 21/1: 1–17.

Samraj, Betty. 2008. A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes* 7/1: 55–67.

Saricaoglu, Aysel, Zeynep Bilki, and Lia Plakans. 2021. Syntactic complexity in learner-generated research paper introductions: Rhetorical functions and level of move/step realization. *Journal of English for Academic Purposes* 53: Article 101037.

Shannon, Claude E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27/3: 379–423.

Shehzad, Wasima. 2008. Move two: Establishing a niche. *Iberica* 15/1: 25–50.

Shehzad, Wasima. 2010. Announcement of the principal findings and value addition in computer science research papers. *Iberica* 19/1: 97–118.

Sheldon, Elena. 2011. Rhetorical differences in RA introductions written by English L1 and L2 and Castilian Spanish L1 writers. *Journal of English for Academic Purposes* 10/4: 238–251.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Swales, John M. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Swales, John M. and Christine B. Feak. 2004. *Academic Writing for Graduate Students: Essential Tasks and Skills*. Ann Arbor: University of Michigan Press.

Swales, John M. and Christine B. Feak. 2012. *Academic Writing for Graduate Students*. Ann Arbor: University of Michigan Press.

Tankó, Gyula. 2017. Literary research article abstracts: An analysis of rhetorical moves and their linguistic realizations. *Journal of English for Academic Purposes* 27: 42–55.

Taş, Elvan Eda. I. 2008. *A Corpus-based Analysis of Genre-specific Discourse of Research: The PhD Thesis and the Research Article in ELT*. Ankara, Turkey: Middle East Technical University dissertation.

Taylor, Gordon and Chen Tingguang. 1991. Linguistic, cultural, and subcultural issues in contrastive discourse analysis: Anglo-American and Chinese scientific texts. *Applied Linguistics* 12/3: 319–336.

Validi, Mahmood, Alireza Jalilifar, Zohreh G. Shooshtari and Abdolmajid Hayati. 2016. Medical research article introductions in Persian and English contexts: Rhetorical and metadiscoursal differences. *Journal of Research in Applied Linguistics* 7/2: 73–98.

Van der Lubbe, Jan C. A. 1997. *Information Theory*. Cambridge: Cambridge University Press.

Wang, Weihong and Chengsong Yang. 2015. Claiming centrality as promotion in applied linguistics research article introductions. *Journal of English for Academic Purposes* 20: 162–175.

Xiao, Wei, and Shuyi Sun. 2020. Dynamic lexical features of PhD theses across disciplines: A text mining approach. *Journal of Quantitative Linguistics* 27/2: 114–133.

Xie, Jianping. 2017. Evaluation in moves: An integrated analysis of Chinese MA thesis literature reviews. *English Language Teaching* 10/3: 1–20.

Ye, Yunping. 2019. Macrostructures and rhetorical moves in energy engineering research articles written by Chinese expert writers. *Journal of English for Academic Purposes* 38: 48–61.

Zhu, Haoran and Lei Lei. 2017. British cultural complexity: An entropy-based approach. *Journal of Quantitative Linguistics* 25/2: 190–205.

Zhu, Haoran and Lei Lei. 2018. Is modern English becoming less inflectionally diversified? Evidence from entropy-based algorithm. *Lingua* 216: 10–27.

Corresponding author
Wei Xiao
Chongqing University
School of Foreign Languages and Cultures
No.55 Daxuecheng South Rd.
401331. Chongqing
China
e-mail: xiaoweiyyx@126.com

received: August 2021
accepted: November 2021