The different ways to write publishable research articles: An exploration using automated language processing tools

Weiyu Zhang & Yin Ling Cheung
Nanyang Technological University (Singapore)
NIE18ZHAN20@e.ntu.edu.sg & yinling.cheung@nie.edu.sg

Abstract

The exploration of linguistic profiles of research articles (RAs) has been underrepresented in the existing literature. This study utilizes two automated language processing tools and a cluster analysis approach to explore linguistic features and variation of published research articles (N=360) in two hard science disciplines (i.e., Biology and Medicine). Findings show five different profiles characterized by their use of distinct combinations of linguistic features. The identified profiles not only vary between the two disciplines, but also within each discipline. The distinct profiles within each discipline represent the potentially different ways for researchers to write publishable research articles. The study fills the research gap and contributes to a new understanding of linguistic features and variation of RAs.

Keywords: writing for publication, linguistic variation, writing profiles, cluster analysis, automated language processing.

Resumen

Las diferentes maneras de escribir artículos de investigación publicables: una exploración a partir de herramientas de procesamiento automático del lenguaje

La bibliografía especializada apenas se ha adentrado en el análisis de los diferentes perfiles lingüísticos de los artículos de investigación. El presente trabajo explora los rasgos lingüísticos y la variación que presentan 360 artículos de investigación publicados pertenecientes a dos disciplinas de ciencias puras (Biología y Medicina) a partir de dos herramientas de procesamiento automático del lenguaje y un enfoque de análisis cluster. Los resultados obtenidos revelan la

existencia de cinco perfiles que se caracterizan por su uso de diferentes combinaciones de ciertos rasgos lingüísticos. Los perfiles identificados no solo varían en función de la disciplina, sino que también se ha detectado variación dentro de una misma disciplina. Esa variación interna a la disciplina representa diferentes perfiles que parecen encarnan diversos modos de redacción a los que los investigadores podrían recurrir para producir artículos de investigación publicables. Con este estudio se pretende cubrir esta laguna de investigación y ofrecer una nueva mirada que ayude a profundizar en la comprensión de los rasgos lingüísticos y de la variación de los artículos de investigación.

Palabras clave: escritura académica, variación lingüística, perfiles de escritura, análisis *cluster*, procesamiento automático del lenguaje.

1. Introduction

Researchers around the world, especially those from English as an Additional Language (EAL) background, are under tremendous pressure to write publishable research articles (RAs) for refereed English-medium journals (Mur Dueñas, 2012; Flowerdew & Wang, 2016; Li, 2014; Moreno, Rey-Rocha, Burgess, López-Navarro & Sachdev, 2012). Many EAL writers face challenges due to their limited repertoire of linguistic resources and inadequate knowledge of the disciplinary writing conventions (Flowerdew, 2000). To inform the writing of these struggling researchers, there have been an increasing number of studies exploring the linguistic features of RAs and how they vary between different disciplines and sub-disciplines (Cortés, 2004, 2013; Hyland, 2005; Hyland & Tse, 2005; Afros & Schryer, 2009; Parkinson, 2013; Zhang & Cheung, 2017, 2018). Variation has been explored in terms of the frequencies of occurrence for individual features and interpreted in relation to the functional characteristics of different disciplines or sub-disciplines. The findings have improved our understanding of how to write publishable RAs and offered practical reference to the teaching and learning of English for Research Publication Purposes (ERPP). However, these studies may have oversimplified the picture of linguistic variation of RAs. First, meaningful variation may not lie in the occurrence of individual linguistic features, but rather in the combination of a collection of features. Second, most of the studies have not considered the possibility that there may be linguistic variation that is not functionally motivated. In fact, studies have already identified the use of different combinations of features, which they called "linguistic profiles", in functionally similar texts written by

students (Jarvis, Grant, Bikowski & Ferris, 2003; Crossley, Roscoe & McNamara, 2014).

The current study seeks to make a novel contribution to the field by exploring the linguistic profiles of RAs and how they vary. To do so, a corpus of 360 "Discussion" sections of RAs was compiled from two disciplines (i.e., Biology and Medicine) and four sub-disciplines (i.e., Genetics, Molecular Biology, Oncology as well as Immunology and Allergy) and two automated language processing tools were used to examine a large set of features beyond word-level. A cluster analysis approach was adopted to identify the linguistic profiles in the corpus. If the identified profiles vary according to the division of disciplines or sub-disciplines, the variation is likely to be functionally determined by disciplinary or sub-disciplinary characteristics. However, if different profiles are identified within one single discipline or sub-discipline, those profiles are less likely to be determined by functional factors but represent the different ways to write publishable RAs within that particular (sub)discipline.

The research questions to be addressed in this paper are:

1) What linguistic profiles of RAs are observable in the corpus?
2) To what extent do the linguistic profiles represent functional variation across disciplines and sub-disciplines?
3) To what extent do the linguistic profiles represent the different ways to write publishable RAs within the same disciplines and sub-disciplines?

2. Literature review

2.1. Linguistic profiles of RAs

The study builds on the assumption that successful writers may not depend on the use of individual linguistic features, but rather on how features are used in combination. The combination of features is called a "linguistic profile". A number of studies have explored this assumption and identified multiple profiles of writing in proficient students' texts (Jarvis et al., 2003; Crossley et al., 2014). Our research aims to further investigate this assumption and examine whether published RAs also demonstrate different profiles and how these profiles vary. Profiles can be identified via a cluster analysis approach, which can group RAs into clusters in a way that RAs in one cluster display within-group similarities with respect to the target linguistic

features, while revealing significant between-group differences regarding one or more features. The combination of features characterizing each cluster represents one linguistic profile. The exploration of linguistic profiles of RAs has been under-represented in the existing literature. The present study aims to fill this gap and contribute to a new understanding of linguistic features and variation of RAs.

2.2. Linguistic variation of RAs – A functional perspective

The linguistic variation of academic prose has been widely explored from a functional perspective. This perspective holds that the core characteristics of writing depend on its function, determined in turn by the communicative purpose and situational context of language use (Biber & Conard, 2009). The primary purpose of academic prose is to construct and disseminate knowledge, which distinguishes it from other types of writing. The unique characteristics of academic prose have been empirically investigated by many studies. That by Biber (1988), for one, draws attention to the nominal style of academic prose, which packs fairly complex information through a heavy reliance on nouns, noun phrases, and prepositions. This nominal style is distinct from other types of writing such as narrative, less informational and more involved, and with a higher use of verbs and personal pronouns. Academic prose is also found to be more structurally compressed and utilizes more phrasal (rather than clausal) modifiers embedded in noun phrases (Biber & Gray, 2010; Gray, 2015).

Academic prose is not a homogenous text type and undergoes functional variation among its subtypes. For example, book reviews and RAs have been found to deploy grammatical structures quite differently (Groom, 2005). Within the domain of RAs, variation is also determined by the situational context of writing for different disciplines and sub-disciplines, such as their epistemological beliefs and ontological assumptions, and their knowledge structures and research practices.

The functional variation of RAs has been explored at many different levels. At the level of lexis, research shows that the use of lexical bundles is closely linked to the functional moves of RAs, with a clear tendency to vary across disciplines (Cortés, 2004, 2013). At the level of grammatical structure, research findings not only point to variation according to disciplines but also among different sections of RAs with distinct communicative purposes (Hyland & Tse, 2005; Parkinson, 2013). In fact, studies that investigate single

lexical or grammatical features are relatively few. Rather, more research has been done on a collection of lexical and grammatical features used to achieve functional means. In recent years, the linguistic realization of interpersonal meanings has become an important research area. The use of intensifiers, personal pronouns, and self-citations to achieve the function of self-promotion in RAs has been found to vary between the disciplines of Linguistics and Literary Studies (Afros & Schryer, 2009). Other features such as hedges, boosters, and reader pronouns, which contribute to stance and reader engagement, have also been investigated across RAs of eight disciplines (Hyland, 2005). The results show that the frequencies of these features vary according to the traditional "hard" and "soft" division of disciplines. Research on writers' use of APPRAISAL (Martin & White, 2005) resources in RAs has uncovered meaningful variation not only between the disciplines of Computer Science and Applied Linguistics but also between the qualitative and quantitative research paradigms within Applied Linguistics (Zhang & Cheung, 2017, 2018). The paradigmatic variation can be interpreted in relation to the functional characteristics of the two research paradigms.

These studies have produced robust knowledge on the writing conventions of RAs for different disciplines and sub-disciplines. This knowledge is important for research writers, as they need to be aware of and conform to register and disciplinary conventions when writing for publication (Flowerdew, 2000). However, most of the studies quantified linguistic variation in terms of the frequencies of occurrence for individual features rather than the combination of features (or linguistic profiles). Moreover, informed by the functional perspective, the identified variation was often interpreted in relation to the distinct situational context of writing for different disciplines and sub-disciplines. Potential variation is usually not assumed within texts that share the same functional characteristics, for example, RAs within one discipline or sub-discipline. Hence, little research has been done to explore the potential linguistic variation within functionally equivalent texts in terms of style.

2.3. Different ways to write publishable RAs – A style perspective

The style perspective is distinct from the functional perspective. The analysis from the functional perspective seeks to compare functionally different writing with distinct communicative purposes or situational contexts. For example, the comparison between academic prose and narrative, and the

comparison between RAs of different disciplines or sub-disciplines. The linguistic variation identified is thus associated with the functional characteristics underlying the writings. The style perspective, however, seeks to identify linguistic variation within writings that share similar functional characteristics (Biber & Conard, 2009), for instance, within RAs of the same discipline or sub-discipline. Consequently, the identified variation is less likely to be related to the functions of writing but represents the optional alternatives that allow writers some freedom.

The stylistic variation of RAs has not been explicitly explored. In the research on student writing, findings have provided some evidence of potential linguistic variation in terms of style. For example, Jarvis et al. (2003) have identified multiple profiles of writing produced by L2 students under the same condition (timed) for the same purpose (exam) with the same results (highly-rated). Further analysis demonstrates that the variation in terms of the profiles cannot be fully explained by situational factors such as the writing topic and the students' L1 background. Therefore, the multiple profiles may represent the different ways for students to write highly-rated timed compositions for exams. Similarly, the study by Crossley et al. (2014) has identified four profiles of writing in a collection of highly-rated student essays and the different profiles are not related to the situational factors of the writings (prompt, grade level, and temporal condition). The finding gives further support to the existence of linguistic variation that is not functional but stylistic. To explain the variation, Jarvis et al. (2003) proposed the notions of "complementarity" and "compensation", which will be discussed later in more detail along with related findings of the current study.

The exploration of stylistic variation of RAs can reveal the alternative ways for researchers to write publishable RAs within a discipline or sub-discipline. This knowledge is particularly beneficial to those EAL researchers with a limited linguistic repertoire because it allows them some leniency to pragmatically choose the features that are more familiar to them while conforming to disciplinary or sub-disciplinary writing conventions.

2.4. Automated Language Processing Tools

The target linguistic features of the study were examined using two computational tools, namely, Coh-Metrix and Linguistic Inquiry and Word Count (LIWC). The majority of the features were explored through Coh-Metrix. Coh-Metrix provides multilevel analyses of discourse characteristics

(Graesser, McNamara & Kulikowich, 2011) and has been adopted in many studies to investigate linguistic features in a variety of discourse including RAs (McCarthy et al., 2007; Ye, 2013). The tool reports on a set of linguistic features, some of which have been widely studied, for example, pronouns, passives, negations, adjectives, adverbs, nominal forms, and prepositional phrases (Biber, 1988; Gray, 2015). It also allows the analysis of features beyond the word-level such as referential cohesion, syntactic patterns and complexity, and mental representation of causation and intentionality (Graesser & McNamara, 2011; McNamara et al., 2014). A complementary set of features was explored through LIWC (Pennebaker, Boyd, Jordan & Blackburn, 2015), which has been a popular tool to study student writing (e.g., Crossley et al., 2014; McNamara et al., 2015). In this study, the features investigated through LWIC include those that depict the writers' mental processes, time orientation, and relativity.

3. Methodology

3.1. Corpus construction procedures

The current study is based on a self-constructed corpus of 360 "Discussion" sections of published RAs from refereed English-medium journals. The discussion section serves the purpose to make sense of the findings and argue for the significance of the research. The discussion section was selected as the focus of our study because it is where "the current work is most vigorously 'sold'" (Hyland, 2009, p. 73) and was found to be among the most challenging sections to write for novice researchers (Uzuner, 2008). The discussion sections are from two different disciplines. Considering potential variation within disciplines (Zhang & Cheung, 2017, 2018), we selected two sub-disciplines from each discipline to represent their writing practices. The selection was made based on the journal ranking indicator provided by SCImago Journal Rank (SJR). The SJR indicator was calculated based on the Scopus database, which is the world's largest scientific database and the best representation of the structure of world science. Meanwhile, the SJR indicator has been found to strongly correlate with other journal metrics such as Journal Impact Factor (JIF) and Source Normalized Impact per Paper (SNIP), despite some differences in the ranking (Guerrero-Bote & Moya-Anegón, 2012).

To determine the discipline and sub-disciplines for study, the SJR metrics for all scientific journals during the period of 2012 to 2016 were downloaded.

All the journals were ranked in order according to the value of their SJR indicator. The top 15 scientific journals of all fields were identified by averaging the value of the SJR indicators from the year 2012 to 2016. By considering journal indicators from multiple years, the effect of drastic change in the ranking of some journals in a particular year was minimized to provide a more objective picture of the scientific value of the journals. The statistics reveal that most of the journals in the top 15 are from the two disciplines of Biology and Medicine. The statistics also reveal that the sub-disciplines of Genetics and Molecular Biology are most widely covered in Biology, whereas Oncology and Immunology & Allergy are most widely covered in Medicine.

For journal selection, we calculated the average value of the SJR indicator from 2012 to 2016 for all journals from the four sub-disciplines and ranked them in order. Six high-ranking journals from each sub-discipline were selected to form the journal pool where RAs were drawn to build the corpus. However, many top journals in the four sub-disciplines publish exclusively review articles and they were not included in the journal pool because the focus of the current study is on empirical RAs (see Appendix A for the list of selected journals from each subject area).

For article selection, computer-generated random numbers were used to select five articles from each year of the journals' publication in the period of 2015 to 2017. In total, 360 RAs were selected, which all have a separated and explicitly titled "Discussion" section. All the "Discussion" sections were then extracted and converted into plain text for analysis through Coh-Metrix and LIWC. See Table 1 for the descriptive statistics of the corpus.

	BIOLOGY				MEDICINE			
	Genetics (n=90)		Molecular (n=90)		Oncology (n=90)		Immunology & Allergy (n=90)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Paragraph count	6.81	3.20	5.91	2.50	7.26	2.41	7.09	2.68
Sentence count	41.16	17.05	34.82	14.49	41.32	14.06	42.34	14.09
Word count	1149.14	509.40	958.36	410.78	1070.32	344.17	1151.56	399.49
Paragraph length	6.33	1.85	6.17	2.00	5.85	1.63	6.44	2.63
Sentence length	28.06	3.76	27.69	4.11	26.34	3.80	27.47	4.33
Word length (syllable)	1.87	0.10	1.89	0.10	1.93	0.09	1.89	0.10
Word length (letter)	5.50	0.25	5.55	0.24	5.65	0.24	5.56	0.27

Table 1. Descriptive statistics of the corpus.

3.2. Statistical analyses

The statistical analyses were conducted based on the selected linguistic indices from Coh-Metrix and LWIC. We first conducted Pearson product-moment correlations between the initial 87 indices to make sure they are not assessing the same construct. Using a threshold of r > .900, 8 indices demonstrating multicollinearity were removed from the analysis, with 79 retained.

Then we used the z-scores of the 79 retained indices as independent variables to conduct an initial cluster analysis using hierarchical cluster analysis with squared Euclidean distance and Ward's method as the distance measure. The analysis can group the "Discussion" sections into clusters according to their shared linguistic features. The initial cluster analysis yielded an optimal five-cluster solution (See Figure 1 for the resulting dendrogram). We then used the selected five-cluster solution to conduct a follow-up cluster analysis, which resulted in the allocation of every "Discussion" to a specific cluster (see Table 2 and 3).

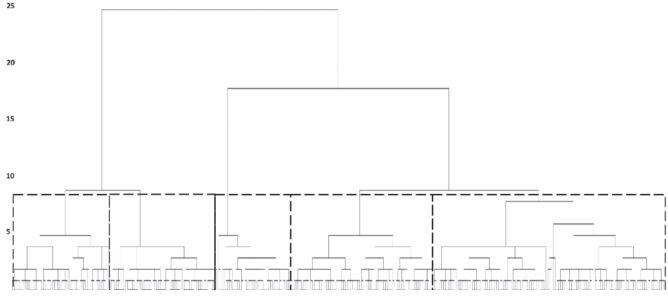

Figure 1. Dendrogram for cluster analysis.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Biology Genetics	15	28	22	24	1
Biology Molecular	27	34	22	7	0
Medicine Oncology	12	27	6	14	31
Medicine Immunology & Allergy	23	39	9	9	10
Total	**77**	**128**	**59**	**54**	**42**

Table 2. Make-up of the clusters.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Paragraph count	6.66	2.42	6.79	2.84	5.53	2.00	7.33	3.41	7.90	2.41
Sentence count	38.51	14.08	40.30	15.73	35.59	12.94	44.06	17.78	42.05	13.64
Word count	1135.52	414.67	1081.82	435.24	894.69	311.93	1171.67	531.71	1135.21	330.94
Paragraph length	6.14	2.46	6.22	1.98	6.70	2.16	6.27	1.74	5.43	1.53
Sentence length	29.86	4.12	27.09	3.78	25.45	2.85	26.55	4.12	27.58	3.97
Word length (syllable)	1.87	0.10	1.89	0.10	1.95	0.09	1.88	0.09	1.89	0.10
Word length (letter)	5.51	0.24	5.58	0.26	5.69	0.24	5.50	0.23	5.53	0.28

Table 3. Descriptive statistics of the cluster.

A multivariate analysis of variance (MANOVA) was conducted to test which linguistic indices demonstrate significant differences between the identified five clusters. MANOVA is based on the same conceptual framework as the univariate analysis of variance (ANOVA). ANOVA tests for statistical differences on one single dependent variable by an independent grouping variable. MANOVA extends the analysis by taking a combination of dependent variables into account. The use of MANOVA will compare whether the combination of variables differs by different groups. The indices were used as the dependent variables and the "Discussion" in each cluster as the independent variables. The results of the MANOVA demonstrate significant difference for 75 out of the 79 linguistic variables used in the cluster analysis (see Appendix B for the MANOVA results).

Following the MANOVA, the mean scores (z-scores) of the 75 linguistic indices that demonstrate significant differences were computed for each cluster to identify the clusters with the highest and lowest score for each index. This could provide a picture of the most distinctive linguistic features of a particular cluster. The mean scores are presented in Table 4.

Category	Index	Clusters 1	2	3	4	5
Lexical Features	Noun	+			-	
	Verb	-		+		
	Adjective	-		+		
	Adverb	-			+	
	Article			-	+	
	First person pronoun (plural)			+		-
	Third person	+				-
	Impersonal pronoun			+		-
	Content word frequency	-				+
	All words frequency			-		+
	Minimum frequency for content words	-				+
	Age of Acquisition			+		-
	Familiarity	-				+
	Concreteness	+			-	
	Meaningfulness		-			+
	Polysemy	-				+
	Hypernymy for nouns	-				+
	Hypernymy for verbs	+			-	
Lexical Diversity	Lexical Diversity - All Words			+		-
	Lexical Diversity - MTLD			+		-
	Lexical Diversity - VOCD			+		-
Syntactic Pattern Density	Noun phrase	+			-	
	Verb phrase	-		+		
	Adverbial phrase			+		-
	Preposition phrase			-		+
	Agentless passive voice	-			+	
	Negation			-		+
	Gerund			+		-
	Infinitive		-	+		
	Comparison	-				+
	Interrogative	-				+
	Number			-		+
	Quantifier	-				+
Syntactic Complexity	Left embeddedness			-		+
	Modifiers per noun-phrase	+				-
	MED (part of speech)	-			+	-
	MED (all words)	-		+		
	Syntax similarity (adjacent sentences)	-				+
	Syntax similarity (all sentences)	-				+

Category	Index	Clusters 1	2	3	4	5
Connectives	All connectives	-			+	
	Causal	-			+	
	Logical	-			+	
	Adversative and contrastive	-				+
	Temporal	-		+		
	Expanded temporal	-				+
	Additive		+			-
Referential Cohesion	Noun overlap (adjacent sentences)	+			-	
	Stem overlap (adjacent sentences)	+			-	
	Noun overlap (all sentences)	+			-	
	Content word overlap (adjacent sentences)	+		-		
	Content word overlap (all sentences)	+			-	
Latent Semantic Analysis	LSA overlap (adjacent sentences)	+			-	
	LSA overlap (adjacent paragraphs)	+			-	
	LSA given/new			-		+
Situational Model	Causal verb			+		-
	Causal verb and causal particles			+		-
	Intentional verb			+		-
	Intentional cohesion		-		+	
	LSA verb overlap	-				+
	WordNet verb overlap	-		-		+
Mental Processes	Positive emotion	-				+
	Negative emotion		-			+
	Insight	-			+	
	Causation			+		-
	Discrepancy	-		+		
	Tentativeness	-			+	
	Certainty	-			+	
	Differentiation	-				+
Time Orientation	Past focus			-		+
	Present focus	-			+	
	Future focus	-		+		
Relativity	Space			-		+
	Time		-			+

Table 4. Linguistic features with the highest and lowest mean score for all clusters.

Finally, a series of discriminant function analyses (DFA) were conducted to test the accuracy of the model. A discriminant analysis is a statistical procedure capable of predicting group membership. In the case of this study, the analysis examines whether the 75 indices can predict the cluster

membership. The DFA results demonstrate that the combinations of the 75 indices have successfully distinguished the five clusters ($\chi 2=781.78$, df=16, p<.001). Following typical procedures of discriminant analysis, the accuracy of the model is reported in terms of both "recall" and "precision". Recall is defined as the number of true positives (members of the cluster correctly identified) divided by the number of true positives plus the number of false negatives (members of the cluster incorrectly identified as non-members). Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives (non-members of the cluster incorrectly identified as members). The estimated accuracy of the model in predicting the membership of each cluster is presented in Table 5.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Recall	84.4%	64.8%	89.8%	68.5%	90.5%
Precision	74.7%	77.6%	76.8%	75.5%	79.2%
Estimated accuracy	**79%**	**71%**	**83%**	**72%**	**84%**

* *Estimated accuracy* $=2\times \frac{precison \times recall}{precision+recall}$

Table 5. Estimated accuracy of the model.

4. Results

4.1. Description of the clusters

The statistical analyses have grouped all the "Discussion" sections into five clusters, which are characterized by distinct combinations of linguistic features. The defining features of each cluster, as summarized in Table 4, are described in the following.

Cluster 1 is defined by high scores on features related to "noun", "word concreteness", "hypernymy for verbs", "third person pronoun", "noun phrase", "modifiers per noun-phrase", "referential cohesion", and "LSA". The lexical feature of this cluster is characterized by the more frequent use of nouns, concrete words, specific verbs (high hypernymy score), and third person pronouns. As nouns are the primary resource to convey referential meanings, the frequent use of these indicates high information density (Biber, 1988). The use of concrete words and specific verbs implies very precise lexical choice to present information content in an exact way. Third person pronouns are often used to refer to human participants in the study and are relatively rare in hard sciences such as Biology and Physics (Gray,

2015). The high frequency of "noun phrase" and "modifiers per noun phrase" are related to the nominal style of academic writing. The high score on "modifiers per noun-phrase" indicates that longer and more complex syntactic structures are used in this cluster because there are more words before the head noun. Such embedded noun phrases also allow more information to be packed into one sentence and contribute to the informationally dense and structurally compressed characteristics of academic prose (Biber & Gray, 2010). In terms of cohesion, Cluster 1 scores high on all indices of referential cohesion. These indices measure co-reference features, which refers to linguistic cues that link sentences, clauses, and propositions through the repetition of words or common lemma (McNamara & Kintsch, 1996). In addition to such explicit co-reference features, Cluster 1 is also high in conceptual overlap measured by LSA indices. These indices assess co-reference in terms of the similarity of implicit knowledge. For example, two words will be considered similar if they share similar surrounding words (Grasser & McNamara, 2011).

While Cluster 2 is the largest cluster, the linguistic characteristic of this cluster is the least distinctive. It only obtained a high score on one index of "additive connectives" (e.g., and, moreover, also), which serve to connect ideas and add information. This may indicate that RAs of this cluster make relatively balanced use of all features.

Cluster 3 is defined by high scores on "lexical diversity", "age of acquisition", "verb", "adjective", "first person pronoun (plural)", "impersonal pronoun", "temporal connectives", "situational model", "causation", "discrepancy", "MED (all words)" and "time orientation (future focus)". Overall, this cluster utilizes a more diverse range of vocabulary as it obtained high scores on all three indices of lexical diversity. The use of a more diverse vocabulary indicates the presentation of very specific meanings (Biber, 1988). It can also be linked to text difficulty and cohesion (Grasser & McNamara, 2011). Greater lexical diversity often adds to text difficulty because there are more unique words and new ideas integrated into the texts. This also means less repetition of words and thus less explicit cohesion features in the text. Moreover, the vocabulary used in this cluster tends to be more difficult, which is evident in the high score on "age of acquisition". This index specifies the age when the target word first appears in a child's vocabulary and a high score means the word is acquired at a later age. Unlike Cluster 1, which is information-oriented with the heavy reliance on nouns, Cluster 3 assumes an action-oriented style with the frequent use of verbs and

verb related terms such as verb phrases, gerunds, and infinitives. Consequently, there are more pronouns used in this cluster as the agent of the actions and more temporal connectives to sequence the actions. This cluster also obtained high scores on indices related to "situational model" ("causal verb", "causal verb and particles", and "intentional verb"). These indices are linked to the causality and intentionality of actions or processes, which help readers build a mental representation of the text (Grasser & McNamara, 2011). They contribute to readers' comprehension of the actions depicted in the RAs. This cluster employs more features related to "causation" (e.g., because, effect) and "discrepancy" (e.g., would, should). These features can also be linked to the action-oriented style because they can be used to explain the purpose and effect of actions. Last, the cluster scores high on one of the MED (Minimal Edit Distance) indices. These indices measure the consistency and uniformity of syntactic structures and a high score implies more complex structures.

Cluster 4 is defined by high scores on "adverb", "article", "all connectives", "causal and logical connectives", "agentless passive voice", "intentional cohesion", "insights", "tentativeness", "certainty", "time orientation (present focus)", and "MED (part of speech)". The lexical feature of this cluster is characterized by the frequent use of adverbs and articles. Adverbs give more information about the time and place mentioned in the texts. Articles give more information about the nouns they modify and can also contribute to the cohesion of writing (Jarvis et al., 2003). Other cohesive features include connectives, which provide cues of text organization by drawing links between ideas and clauses (McNamara et al., 2014). This cluster is high in the use of all connectives, especially the two types of "casual" (because, so) and "logical" (therefore, if). The frequent use of them may imply that the RAs in this cluster are more involved in the process of critical reasoning and deduction. This cluster scores high on agentless passives, which are more frequently used in hard rather than soft disciplines to de-emphasize the role of the researchers (Gray, 2015). This reflects the positivist-empirical epistemology of the hard disciplines, which suggests that the research would yield similar outcomes irrespective of the researchers conducting it (Hyland, 2005). While the role of the agents tends to be de-emphasized, more "intentional cohesion" features are used to help readers understand the agents' goal in performing actions. For features of mental processes, this cluster scores high on "insight" (think, know), "tentativeness" (maybe, perhaps), and "certainty" (always, never). All these features

contribute to the expression of writers' stance (Hyland, 2005; Zhang & Cheung, 2017). The use of "insight" can signal the subjectivity of the proposition that follows. The use of "tentativeness" implies the writer's reserved opinion towards the proposition, whereas the use of "certainty" indicates the writers' strong commitment. These two features serve similar functions as "hedges" and "boosters" and tend to co-occur in RAs (Hyland, 2005). Their use indicates that the researchers need to weigh up the commitment they invest in their arguments due to the existence of alternative interpretations. These features are relatively infrequent in hard science disciplines because hard knowledge is more objective and therefore provides fewer alternatives to consider (Hyland, 2005). This cluster is high in the use of present tense, which indicates general truths or describes the effects of phenomena (Jarvis et al., 2003; Gray, 2015). The use of present tense tends to be more frequent than that of past tense in a range of hard and soft disciplines, except for history (Gray, 2015). Moreover, similar to Cluster 3, this cluster scores high on one of the MED indices, which means more complex syntactic structures. Last, it is worth mentioning that Cluster 4 is defined by low scores on most indices related to co-reference ("referential cohesion" and "LSA"), which is in stark contrast to Cluster 1. The contrast will be further discussed later.

Cluster 5 is defined by high scores on indices of "lexical features", "syntactic structures", "adversative and contrastive connectives", "expanded temporal connectives", "LSA given/new", "verb overlap", "positive and negative emotion", "differentiation" "relativity", and "time orientation (past focus)". For lexical features, this cluster scores high on all three indices of word frequency ("content word frequency", "all word frequency", and "minimum frequency for content word"). High frequency words are those that appear in the English language more often than others and thus are easier words. The words in this cluster also obtain high scores on "meaningfulness", "familiarity", "polysemy", and "hypernym". Higher meaningfulness score indicates that the words have strong relations with other words. Familiarity assesses the extent to which a word is familiar to an adult. Polysemy score indicates the ambiguity of writing as the more potential interpretation of a word the higher its polysemy score. Meanwhile, it may also indicate easier words because high frequency words tend to have more meanings (McNamara et al., 2014). The words used in this cluster also tend to be more specific as suggested by high scores on the two indices of "hypernym". Taken together, all the lexical indices suggest that words used in this cluster

are comparatively easier to comprehend for readers than those in other clusters. In terms of syntactic features, this cluster employs a wider range of syntactic patterns including prepositional phrases, negations, comparison, interrogatives, numbers, and quantifiers. The syntactic structures of this cluster are more complex given the high score on "left embeddedness", which measures the number of words used before the main verb of the main clause in sentences. Further, the high scores on the two "syntactic simplicity" indices imply that the syntactic structures in the cluster are more consistent in style and form (McCarthy et al., 2009). The cohesive devices frequently deployed in this cluster include adversatives and contrastive connectives, expanded temporal connectives, and implicit co-reference features measured by "LSA given/new" and "verb overlap" indices (LSA verb overlap and WordNet verb overlap). This cluster is high in the expressions of positive and negative emotions as well as "differentiation" (but, else). This cluster also makes more reference to space (in, above) and time (end, until). The preferred tense is "past focus", which serves to situate one's research within established findings in the field, describe methodological procedures, or report findings that are not yet accepted as truth (Jarvis et al., 2003; Gray, 2015).

4.2. Variation according to disciplines and sub-disciplinary areas

As presented in Table 2, the majority of biological RAs fall into Cluster 1 to 4, whereas most of the medical RAs are in Cluster 1, 2, 4 and 5. This indicates variation of the linguistic profiles between the two disciplines. As Cluster 3 contains mostly biological RAs (75%), it is more representative of the discipline of Biology. Similarly, as Cluster 5 contains all (except one) medical RAs, it is more representative of the discipline of Medicine. There is no clear pattern of variation between the two sub-disciplines in each discipline.

4.3. Variation within disciplines and sub-disciplines

As presented in Table 2, the five linguistic profiles also vary within each discipline. To be more specific, there are four different linguistic profiles (Cluster 1 to 4) for biological RAs of both sub-disciplines; and four (Cluster 1, 2, 4 and 5) for medical RAs of both sub-disciplines. The variation of the linguistic profiles within each discipline is not related to sub-disciplinary division.

5. Discussion

5.1. Significance of the linguistic profiles

The study demonstrates the use of a cluster analysis approach to investigate the linguistic features and variation of RAs. This approach identifies five different linguistic profiles in our corpus. The profiles are characterized by their use of distinct combinations of linguistic features. So far, the research on linguistic profiles has been mostly restricted to student writing (Jarvis et al., 2003; Crossley et al., 2014). The current study provides initial evidence for the existence of multiple profiles of writing in RAs. Moreover, the study also uncovers that the profiles not only vary between disciplines but also within a single discipline and sub-discipline. The identification of the profiles reveals a more complex picture of linguistic variation of RAs, which will contribute to our more refined knowledge of the field.

5.2. Disciplinary variation in writing

The current study has identified linguistic variation between Biology and Medicine RAs. From the functional perspective, the variation is likely to be motivated by the different situational contexts of writing for the two disciplines. However, we were not able to conduct an in-depth qualitative analysis to explain the variation. While the automated language processing tools allow us to explore a relatively large set of features beyond word-level in a sizable corpus, it does not have an interface for researchers to scrutinize the analyzed instances. Therefore, we could only interpret our results according to established findings.

Cluster 3 contains mostly biological RAs. This cluster is high in the use of first person pronouns (plural), which usually refers to the writers themselves. The use of such interpersonal features has been found to be infrequent in academic writing, especially in RAs of hard science disciplines such as Biology and Physics (Hyland, 2005; Gray, 2015). The positivist-empirical epistemology underlying hard disciplines requires the researchers' role to be minimized in order to highlight the objectivity of the study (Hyland, 2005). However, the current finding suggests that the degree of "impersonality" varies between hard disciplines and some writers in Biology tend to make their own presence more visible to the readers. This points to their confidence in taking responsibility for the research and to their efforts to emphasize their own contribution to the

field. Such a practice is acceptable, considering that all the RAs have reached publication in high-ranking journals. This cluster is also high in the use of verbs and verb-related terms, activity verbs being the most popular type in RAs across a variety of disciplines (Gray, 2015). They are often used to describe data, concepts, methodology, or findings, and therefore may be more numerous in the discussion section. This cluster also uses the future tense more, probably to make claims concerning what needs to be done in future research.

Cluster 5 is more representative of the discipline of Medicine. Compared to Cluster 3, RAs in this cluster deploy more features related to time and space. The feature of space can be used in the form of internal references, as in "It was demonstrated above", characteristic of highly informational texts (Biber, 1988). References to space can also be used to state the research context, that is, the place where the research is conducted, or where the sample is drawn from. In contrast to biological RAs in Cluster 3, the medical RAs in this cluster score low on "first person pronoun (plural)", which reflects the researchers' intention to minimize their own presence. In this case, researchers in Medicine are more closely aligned to the common practice in the hard disciplines than their counterparts in Biology. RAs in this cluster also obtained high scores on the features of positive and negative emotions. One possible reason may be that part of our medical samples are from the sub-discipline of Oncology and deal with research topics related to cancer, which would likely provoke emotional responses.

5.3. Different ways to write publishable RAs

Our results show that the five linguistic profiles vary within each discipline, regardless of sub-disciplinary division. Therefore, variation is not likely to be motivated by the functional characteristics of the sub-disciplines, but may rather be attributable to mere stylistic preferences. The multiple profiles within each discipline may represent different ways for researchers in that discipline to write publishable RAs.

To interpret the findings from the style perspective, we adopt the two-dimensional framework of "complementarity" and "compensation" proposed by Jarvis et al. (2003). Complementarity refers to the idea that there may be a variety of linguistic resources that serve similar goals in RAs, and the high frequencies of some features may result in low frequencies of others. For example, RAs in Cluster 1 are defined by a high

frequency of co-reference features but by a low frequency of connectives, whereas RAs in Cluster 4 exhibit the opposite pattern. The use of co-reference features contributes to text cohesion by connecting propositions, clauses, and sentences and the lack of such features may lead to cohesion gaps in texts (McNamara et al., 2014). To bridge these gaps, writers can choose from a variety of cohesive devices such as "connectives", rather than relying solely on co-reference features. These two features may thus show a complementary distribution and their use in the same text may not be high.

Compensation means that writing deficiencies associated with certain features can be counterbalanced by the use of other features. In Cluster 3, for example, a vocabulary more diverse than in other clusters suggests that more ideas have been incorporated to the text. However, unlike Clusters 1 and 4, here the RAs do not show a high use of co-reference features nor of connectives to link ideas. This could potentially be a deficiency in writing but is compensated by a high use of features that contribute to the situation model, which assists readers in building a mental representation of the text for a better understanding. Also, RAs in Cluster 5 employ longer and more complex syntactic structures, which add to the readers' processing burden, however, this weakness may be compensated because the structures they use are more uniform and consistent.

Moreover, there is also variation that cannot be reasonably explained by this framework. For example, RAs in Cluster 3 show a high use of first person pronouns (plural) to make the researchers' presence more visible, whereas RAs in Cluster 4 tend to use more agentless passives to de-emphasize the role of the researchers. This difference is neither 'complementary' nor 'compensatory'. It may represent a style preference of the researchers, which reflects how they position themselves in relation to their research, their readers, and their discipline. Both styles are accepted for publication in high-ranking refereed journals.

6. Conclusion

The study reveals a complex picture of linguistic variation of RAs, which can offer useful implications for future ERPP research and pedagogy. First, the use of a cluster analysis approach identifies five different linguistic profiles in our corpus. The profiles are characterized by their use of

distinct combinations of linguistic features. This finding suggests that meaningful variation of RAs is more likely to exist in terms of different combinations of linguistic features or "linguistic profiles" rather than individual features. Previous research (e.g., Afros & Schryer, 2009; Hyland, 2005; Parkinson, 2013) that quantifies linguistic variation in terms of the occurrence of individual features may fall short in providing a comprehensive understanding of RAs in different disciplines or sub-disciplines. The cluster analysis approach utilized by this study has shown to be a productive approach to identify the different linguistic profiles of RAs. Future research may also adopt this approach to investigate RAs from other disciplines and sub-discipline to better understand the linguistic profile and variation of RAs. Second, the profiles are found to vary between the disciplines of Biology and Medicine. This finding corresponds with previous observations that the linguistic features of RAs vary according to disciplines (e.g., Cortés, 2004, 2013; Hyland & Tse, 2005; Zhang & Cheung, 2017, 2018). Such variation may represent the disciplinary writing conventions and is likely to be motivated by the situational contexts of writing for different disciplines, such as their epistemological belief, ontological assumption, knowledge structure, and research practice. To write publishable RAs, researchers need to conform to these disciplinary writing conventions. ERPP instructors need to raise the awareness of novice writers of such conventions and equip them with the necessary linguistic resources. Third, the profiles also vary within each discipline and there is no evidence that the variation is functionally related to the sub-disciplines. Therefore, the study concludes that linguistic variation of RAs can be stylistic and research writers are allowed to write publishable RAs in different ways while still conforming to the writing conventions of their discipline. ERPP instructors may introduce the strategies of 'complementarity' and 'compensation' to novice research writers, especially to those from EAL background. Through the use of these strategies, EAL writers with a limited repertoire of linguistic devices may strategically deploy the resources that are more familiar to them to complement or compensate for other resources that are unfamiliar to them to better achieve their writing goals.

Acknowledgements

The first author would like to thank the National Institute of Education, Nanyang Technological University Scholarship (HD-RSS Scholarship) for supporting the research work presented in this paper.

Article history:
Received 03 May 2019
Received in revised form 31 December 2019
Accepted 27 January 2020

References

Afros, E. & C.F. Schryer (2009). "Promotional (meta)discourse in research articles in language and literary studies". *English for Specific Purposes* 28(1): 58-68.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: CUP.

Biber, D. & S. Conrad (2009). *Register, Genre, and Style*. New York: CUP.

Biber, D. & B. Gray (2010). "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness". *Journal of English for Academic Purposes* 9(1): 2-20.

Cortés, V. (2004). "Lexical bundles in published and student disciplinary writing: Examples from history and biology". *English for Specific Purposes* 23(4): 397-423.

Cortés, V. (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.

Crossley, S.A., R. Roscoe & D.S. McNamara (2014). "What is successful writing? An investigation into the multiple ways writers can write successful essays". *Written Communication* 31(2): 184-214.

Flowerdew, J. (2000). "Discourse community, legitimate peripheral participation and the nonnative-English-speaking scholar". *TESOL Quarterly* 34(1): 127-150.

Flowerdew, J. & S.H. Wang (2016). "Author's editor revisions to manuscripts published in international journals". *Journal of Second Language Writing* 32: 39-52.

Graesser, A.C. & D.S. McNamara (2011). "Computational analyses of multilevel discourse comprehension". *Topics in Cognitive Science* 3(2): 371-398.

Graesser, A.C., D.S. McNamara & J.M. Kulikowich (2011). "Coh-Metrix: Providing multilevel analyses of text characteristics". *Educational Researcher* 40(5): 223-234.

Gray, B. (2015). *Linguistic Variation in Research Articles: When Discipline Tells Only Part of the Story*. Philadelphia: John Benjamins.

Groom, N. (2005). "Pattern and meaning across genres and disciplines: An exploratory study". *Journal of English for Academic Purposes* 4(3): 257-277.

Guerrero-Bote, V. P. & F. Moya-Anegón (2012). "A further step forward in measuring journals' scientific prestige: The SJR2 indicator". *Journal of Informetrics* 6(4): 674-688.

Hyland, K. (1999). "Disciplinary discourses: Writer stance in research articles" in C. Candlin & K. Hyland (eds.), *Writing: Texts, Processes and Practices*, 99-121. London: Longman.

Hyland, K. (2005). "Stance and engagement: A model of interaction in academic discourse". *Discourse Studies* 7(2), 173-192

Hyland, K. (2009). *Academic Discourse: English in a Global Context*. London/New York: Continuum.

Hyland, K. & P. Tse (2005). "Hooking the reader: A corpus study of evaluative *that* in abstracts". *English for Specific Purposes* 24(2): 123-139.

Jarvis, S., L. Grant, L., D. Bikowski & D. Ferris (2003). "Exploring multiple profiles of highly rated learner composition". *Journal of Second Language Writing* 12(4): 377-403.

Li, Y. (2014). "Chinese medical doctors negotiating the pressure of the publication requirement". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 28: 107-128.

Lin, L., & S. Evans (2012). "Structural patterns in

empirical research articles: A cross-disciplinary study". *English for Specific Purposes* 31(3): 150-160.

McCarthy, P.M., B.M. Lehenbauer, C. Hall, N.D. Duran, Y. Fujiwara & D.S. McNamara (2007). "A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British Scientists". *Foreign Languages for Specific Purposes* 6: 46-77.

McNamara, D.S., S.A. Crossley, R.D. Roscoe, L.K. Allen & J. Dai (2015). "A hierarchical classification approach to automated essay scoring". *Assessing Writing* 23: 35-59.

McNamara, D.S., A.C. Graesser, P.M. McCarthy & Z. Cai (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: CUP.

McNamara, D.S. & W. Kintsch (1996). "Learning from texts: Effects of prior knowledge and text coherence". *Discourse Processes* 22(3): 247-288.

Moreno, A.I., J. Rey-Rocha, S. Burgess, I. López-Navarro & I. Sachdev (2012). "Spanish researchers' perceived difficulty writing research articles for English-medium journals: The impact of proficiency in English versus publication experience". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 24: 157-184.

Mur Dueñas, P. (2012). "Getting research published internationally in English: An ethnographic account of a team of Finance Spanish scholars' struggles". *Ibérica, Journal of the European Association of Languages for Specific Purposes* 24: 139-156.

Ozturk, I. (2007). "The textual organisation of research article introductions in applied linguistics: Variability within a single discipline". *English for Specific Purposes* 26(1): 25-38.

Parkinson, J. (2013). "Representing own and other voices in social science research articles". *International Journal of Corpus Linguistics* 18(2): 199-228.

Pennebaker, J.W., R.L. Boyd, K. Jordan & K. Blackburn (2015). *The Development and Psychometric Properties of LWIC2015*. Austin, TX: University of Texas at Austin.

Uzuner, S. (2008). "Multilingual scholars' participation in core/global academic communities: A literature review". *Journal of English for Academic Purposes* 7: 250-263.

Ye, D. (2013). "A Coh-metrix analysis of language varieties between the journal articles of Chinese and American scientists. *International Journal of English Linguistics* 3(4): 63.

Zhang, W. & Y.L. Cheung (2017). "Understanding ENGAGEMENT resources in constructing voice in research articles in the fields of computer networks and communications and second language writing". *The Asian ESP Journal* 13(3): 72-99.

Zhang, W. & Y.L. Cheung (2018). "The construction of authorial voice in writing research articles: A corpus-based study from an APPRAISAL theory perspective". *International Journal of English Studies* 18(2): 53-75.

Weiyu Zhang is a PhD candidate at the English Language and Literature Academic Group at the National Institute of Education, Nanyang Technological University. She has published in *International Journal of English Studies* and *The Asian ESP Journal*.

Yin Ling Cheung is Associate Professor (English Language and Literature) and Associate Dean (Outreach and Engagement) at the National Institute of Education, Nanyang Technological University. She has published in journals such as *System* and *RELC Journal*.

Appendix A. Selected Journals for the current study

Subject Areas	Journals		SJR indicator (5-year average)
Biology: Genetics	1.	Nature Genetics	23.45
	2.	Genome Research	14.04
	3.	Cell Stem Cell	13.69
	4.	Genes and Development	12.10
	5.	Genome Biology	8.76
	6.	Systematic Biology	8.59
Biology: Molecular Biology	1.	Nature Methods	14.87
	2.	Molecular Cell	13.54
	3.	Nature Structural and Molecular Biology	10.97
	4.	Cell Metabolism	10.72
	5.	Cell Host and Microbe	8.09
	6.	EMBO Journal	7.85
Medicine: Oncology	1.	Cancer Research	15.06
	2.	Cancer Cell	13.41
	3.	The Lancet Oncology	8.87
	4.	Journal of National Cancer Institute	6.42
	5.	Journal of Clinical Oncology	5.56
	6.	Clinical Cancer Research	4.92
Medicine: Immunology & Allergy	1.	Immunity	15.76
	2.	Journal of Experimental Medicine	11.36
	3.	Journal of Allergy and Clinical Immunology	5.09
	4.	Mucosal Immunology	4.25
	5.	Arthritis and Rheumatology	3.88
	6.	Journal of Infectious Disease	3.70

Appendix B. MANOVA results

Index	F	p	η^2_p	Index	F	p	η^2_p
Noun Overlap (adjacent sentences)	46.882	0.000	0.346	Number	6.977	0.000	0.073
Stem overlap (adjacent sentences)	34.260	0.000	0.279	Quantifier	30.883	0.000	0.258
Noun overlap (all sentences)	48.031	0.000	0.351	First person pronoun (plural)	5.443	0.000	0.058
Content word overlap (adjacent sentences)	55.056	0.000	0.383	Impersonal pronoun	6.443	0.000	0.068
Content word overlap (all sentences)	47.472	0.000	0.348	Content word frequency	54.978	0.000	0.383
LSA overlap (adjacent sentences)	23.400	0.000	0.209	All words frequency	17.306	0.000	0.163
LSA given/new	21.331	0.000	0.194	Age of Acquisition	6.202	0.000	0.065
Lexical Diversity - All Words	18.285	0.000	0.171	Familiarity	36.628	0.000	0.292
Lexical Diversity - MTLD	45.777	0.000	0.340	Concreteness M	10.763	0.000	0.108
Lexical Diversity - VOCD	25.127	0.000	0.221	Meaningfulness	11.212	0.000	0.112
All connectives	9.481	0.000	0.097	Polysemy	26.620	0.000	0.231
Logical connectives	9.441	0.000	0.096	Hypernymy for nouns	27.905	0.000	0.239
Adversative and contrastive connectives	7.005	0.000	0.073	Hypernymy for verbs	16.574	0.000	0.157
Expanded temporal connectives	26.434	0.000	0.229	Hypernymy for nouns and verbs	9.329	0.000	0.095
Additive connectives	8.580	0.000	0.088	Discrepancy	11.674	0.000	0.116
Causal verb	5.701	0.000	0.060	Tentative	20.209	0.000	0.185
Intentional verb	12.444	0.000	0.123	Certainty	10.515	0.000	0.106
Intentional cohesion	10.066	0.000	0.102	Differentiation	13.916	0.000	0.136
LSA verb overlap	36.759	0.000	0.293	Past focus	44.910	0.000	0.336
WordNet verb overlap	13.909	0.000	0.135	Present focus	27.309	0.000	0.235
Left embeddedness M	15.343	0.000	0.147	Future focus	15.369	0.000	0.148
Modifiers per noun-phrase	5.338	0.000	0.057	Time	21.474	0.000	0.195
MED (part of speech)	10.567	0.000	0.106	Positive emotion	16.323	0.000	0.155
MED (all words)	44.921	0.000	0.336	Negative emotion	7.174	0.000	0.075
Syntax similarity (all sentences)	5.943	0.000	0.063	Causal verb and causal particles	4.501	0.001	0.048
Noun phrase	17.369	0.000	0.164	Insight	4.532	0.001	0.049
Verb phrase	37.238	0.000	0.296	Syntax similarity (adjacent sentences)	4.436	0.002	0.048
Adverbial phrase	10.777	0.000	0.108	Temporal connectives	4.184	0.003	0.045
Preposition phrase	21.900	0.000	0.198	Causal connectives	3.293	0.011	0.036
Agentless passive voice	11.719	0.000	0.117	Causation	3.325	0.011	0.036
Negation	20.558	0.000	0.188	Space	3.128	0.015	0.034
Gerund	17.389	0.000	0.164	LSA overlap (adjacent paragraphs)	2.939	0.021	0.032
Infinitive	13.598	0.000	0.133	Minimum frequency for content words	2.870	0.023	0.031
Noun	26.270	0.000	0.228	Third person	2.782	0.027	0.030
Verb	16.825	0.000	0.159	Adjective	2.714	0.030	0.030
Adverb	13.886	0.000	0.135	Temporal cohesion	2.360	0.053*	0.026
Article	6.038	0.000	0.064	Causal cohesion	2.298	0.059*	0.025
Conjunction	6.358	0.000	0.067	Perceptual	2.294	0.059*	0.025
Comparison	39.140	0.000	0.306	First person pronoun (single)	0.302	0.877*	0.003
Interrogative	5.436	0.000	0.058				

(*) marks non-significant difference.