「研究ノート」

Construction of Medical Research Article Corpora with AntCorGen: Pedagogical Implications

Motoko ASANO

Abstract

This paper examines the usefulness of a novel corpus generating tool AntCorGen (Anthony, 2017a) for the compilation of medical research article corpora for use in pedagogical settings. A corpus comprising approximately 1,500 *PLOS ONE* medical research article abstracts was compiled successfully using AntCorGen. A larger corpus was also built using 400 research articles encompassing approximately 1.73 million words from four disciplines of medicine and was explored from the viewpoint of English for Specific Purposes (ESP). The findings suggest much promise for the use of corpora in the classroom.

1. Introduction

With the emergence of "English as the *lingua franca* of scientific communication in general[,] and of medicine in particular" (Salager-Meyer, 2014: p. 49), English is being used in approximately eight million health-related peer-reviewed articles published globally each year (Salager-Meyer, 2014). Medical students and researchers need to be able to efficiently use English to become members of their disciplinary discourse communities (Mauranen, 2017).

Linguistic support for apprentice scientists has been provided along with research on medical discourse (Maher, 1986). According to Salager-Meyer (2014), genres within ESP contexts that have been studied most extensively are research articles (RAs) and case reports. Following the framework of rhetorical movement or "moves" in RA introductions, which was termed "*Create a Research Space* (CARS) model" (Swales, 1990, p. 140), Nwogu (1997) analyzed 15 medical research articles from international journals such as *The Lancet* and suggested 11 moves. Salager-Meyer (1989) analyzed

linguistic features of 51 medical texts including RAs, case reports, and editorials. In all these studies, small, specialized corpora were built and used. This approach has been considered legitimate as Lee (2008, p. 94) maintains that "discourse analysts who work with specialized discourses can benefit from compiling their own corpora and applying some of the techniques of corpus-based linguists to support their analyses."

For teaching, corpora have been "extremely useful for ESP teachers in that they are able to show how language is used in the context of particular academic genres" (Paltridge, 2013, p. 351). The findings from textual analyses have been incorporated into various textbooks for RA writing (Feak & Swales, 2012; Nakatani & Bucsis (Ed.), 2016; Noguchi, Matsuura, & Haruta, 2015).

Recently, many science and engineering RAs come from China and Japan (Flowerdew & Wang, 2017). For science students writing in their second language, "a certain extent of language re-use from other texts is *acceptable*" (Flowerdew & Li, 2007, p. 442). Flowerdew and Li (2007) maintain that "re-use of formulaic structures at the syntactic level and formulaic chunks at the lexical level are basic learning strategies upheld by corpus-based pedagogy, while the formulaicity of various texts at the rhetorical level has been the foundation of genre-based pedagogy" (p. 460).

In classrooms, learners use corpora to discover linguistic patterns. Such "inductive approaches to the learning of grammar and vocabulary" (Jones & Dimant, 2013, p. 395) are known as data-driven learning (DDL) (Johns, 1991). The DDL approach has been used for students from disciplines such as educational technology (Lee & Swales, 2006) and biological sciences (Noguchi, 2004); learners download articles from their target journals and create mini corpora. Concordance texts are observed for "the simulation of inductive learning strategies" (Johns, 1991, p. 30). However, building corpora can be an arduous task even for those well versed in English for research publication purposes (Swales, 1990; Nwogu, 1997; Maswana, Kanamaru, & Tajino, 2015).

To offer help with corpus building, Anthony (2017b, p. 71) developed the novel freeware AntCorGen (Anthony, 2017a), which enables rapid, automatic generation of discipline-specific corpora from articles in the *PLOS ONE* research database, thus "making the tool ideally suited for use in Data-Driven Learning." This paper reports on the use of AntCorGen to build discipline-specific corpora of *PLOS ONE* medical RAs and examines possible applications for classroom use.

2. Examination—The CAPHYRA corpus

2.1 Corpus construction

A corpus of cardiovascular physiology RA abstracts (CAPHYRA) was prepared using AntCorGen (Version 1.1.0), which was the latest version at the time of the study. The field of cardiovascular physiology was chosen because physiology is one of the basic fields of medicine (Caze, 2011) and because cardiovascular RAs have been used for linguistic studies (Coates, Sturgeon, Bohannan, & Pasini, 2002). As AntCorGen indicated that the number of physiology RAs was 50,101, the RA category was limited to "cardiovascular physiology," and only the abstract portion of the RAs was collected using the following procedure:

1. A new folder was created on the desktop of a computer.
2. AntCorGen was launched by left clicking the AntCorGen icon and selecting run as an administrator, which was a safer way to avoid freezing of the software during the operation.
3. The corpus storage folder, which was prepared in procedure 1, was selected to store the files.
4. The research article category was chosen by checking the [cardiovascular physiology] box.
5. The target field of collection was selected by clicking the abstract box.
6. The [create corpus] button was pressed to download the article abstract texts.

The CAPHYRA corpus, comprised of 1,551 abstracts, was built within a few seconds. The file name of each abstract text, such as "_10_1371_journal_pone_0012983.txt," represents the URL of the original RAs. Therefore, DDL instructors and students should be able to easily locate the original RAs.

2.2 Methods for examining the CAPHYRA corpus

The CAPHYRA corpus was examined using the concordance software AntConc (Anthony, 2014) because the tool has been used in RA writing classrooms (Noguchi, 2004) and is recommended in a textbook (Feak & Swales, 2012). The types and tokens were counted. The most frequent words were identified, and the most frequent four-

word expressions, called "4-grams" (Nesi, 2013, p. 418), were extracted because "[m]ost researchers have chosen to examine four-word combinations" and because "[t]hey usually reveal more about the genre of the corpus than its topic" (p. 418). Concordance lines were obtained for some of the most frequent words. As Hunston (2013: p. 158) indicates that concordancing tools "only find and organize the data. Interpretation is a human activity," the concordance lines were observed to identify patterns.

2.3 Results and Discussion

The CAPHYRA corpus had 14,954 word types and 386,966 tokens. The ten most frequent words and the normalized frequency (NF) are shown in Table 1. The two most frequent words were *the* and *of*. Marco (2000, p. 67), who analyzed collocational frameworks in medical RAs, suggested that "[t]he most common frameworks in [their] corpus are: *the ...of* (e.g., *the number of*), *be ... to* (e.g., be similar to), *a ... of* (e.g., *a variety of*)...."

Table 1: The most frequent words in the CAPHYRA corpus: Normalized per 1000 words (1551 CArdiovascular PHYsiology Research Article Abstracts)

Rank	Word	NF
1	the	40.3
2	of	39.3
3	and	38.7
4	in	33.7
5	to	16.5
6	a	15.8
7	with	10.3
8	that	9.8
9	by	8.1
10	was	7.8

The ten most frequent 4-grams are shown in Table 2. Among them, *vascular endothelial growth factor*, *endothelial growth factor vegf*, *human umbilical vein endothelial*, *umbilical vein endothelial cells*, and *of vascular endothelial growth* were highly specific technical expressions related to "the cell layer that lines the blood vessels" (Pocock, Richards, & Richards, 2006, p. 57). Two of the remaining five 4-grams were *in vitro and in* and *vitro and in vivo*. According to the American Medical

Association (AMA) Manual of Style (Iverson, Christiansen, Flanagin, & Fontanaroas. 2007, p. 925), the terms *in vitro* and *in vivo* are "considered to have become part of the English language" and "[i]talics are not used" in medical articles.

The remaining three were *in this study we*, *this study was to,* and *of this study was.* These are typical hint expressions which are likely to have been used intentionally by the writers of the abstracts to realize their rhetorical purposes (Tojo, Hayashi, & Noguchi, 2014; Mizumoto, Hamatani, & Imao, 2017).

Table 2: The most frequent 4-grams in the CAPHYRA corpus

Rank	Four-word expression	Frequency
1	vascular endothelial growth factor	293
2	endothelial growth factor vegf	173
3	in this study we	145
4	in vitro and in	118
5	vitro and in vivo	112
6	human umbilical vein endothelial	89
7	umbilical vein endothelial cells	77
8	this study was to	76
9	of this study was	72
10	of vascular endothelial growth	72

2.4 Pedagogical Implication

The concordance lines showed that the phrase *vascular endothelial growth factor* has no article in the third, fourth, and sixth sentences (Figure 1) and is followed by the abbreviation in all sentences. The abbreviation *VEGF* and its expanded form *vascular endothelial growth factor* appear in "the list of clinical, technical, and other common terms" in the AMA Manual of Style (Iverson et al., 2007, p. 519). Iverson et al. (2007,

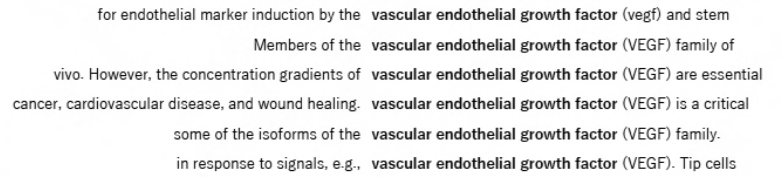

Figure 1: Example of concordance lines for *vascular endothelial growth factor*

p. 501) suggest that "Use common sense in deciding whether to abbreviate the terms." In the DDL setting, instructors may guide the learners to become aware of the usage and form of such terms and their abbreviations.

The first letter of the phrase *In the study, we* was capitalized in all 145 sentences of the concordance (Figure 2). The phrase was followed by verbs describing actions such as *investigate* and *have developed* or reporting verbs such as *demonstrate*. In classrooms, learners can be guided to notice that the use of "this study" indicates the research being reported in the paper itself. They can also learn about the types and tense of verbs that follow the phrase *In this study, we*.

blood vessel growth, and cancer invasion.	**In this study, we** investigate the influence
of the proteins ERK1 and 2 (ERK1/2).	**In this study, we** have developed a
genetic determinants are largely unidentified.	**In this study, we** sought to determine
receptors and intracellular signaling pathways.	**In this study, we** generated an α5
(NFs) into CAFs is largely unknown.	**In this study, we** determined the contribution
from human monocytes. Methodology and Results:	**In this study, we** demonstrate the molecular

Figure 2: Example of concordance lines for *in this study, we*

Word profiles of the corpus suggested that the the CAPHYRA corpus might have lexical features indicative of medical RAs. The most frequent four-word expressions were classified into three types: highly specific technical expressions in the cardiovascular field, technical words of Latin origin, and hint expressions which were used for rhetorical purposes. The concordance lines with these expressions should be useful for activities in the classroom.

3. Examination of a larger corpus

3.1 Compilation of a medical research article corpus—The MEDRA corpus

The CAPHYRA corpus included only the RA abstracts of cardiovascular physiology papers. Building a larger corpus with AntCorGen was, therefore, studied from the viewpoint of pedagogical application.

Four corpora comprising 4,863 cardiology, 5,552 gastroenterology, 5,524 pulmonology, and 4,821 cancer RA texts were built successfully by the some procedure as that for the CAPHYRA corpus. The collection of cancer RA texts was limited to

"clinical oncology," "oncology agents," "cancer risk factors," and "cancer detection and diagnoses" to obtain approximately 5,000 RAs. The disciplines were selected based on consultation with an informant specializing in histopathology and gynecology.

From the four corpora, 100 RAs each were "randomly sampled" (L. Anthony, personal communication, June 13, 2017) to obtain a medical RA (MEDRA) corpus. The MEDRA corpus, consisting of 400 RA body texts from the four disciplines, was examined using AntConc (Anthony, 2014) and CasualConc (Imao, 2017).

3.2 Findings—Profiles of the MEDRA corpus

The corpus had 33,734 types and 1,778,417 tokens. Each portion of the four areas, including cardiology, gastroenterology, pulmonology, and cancer, had approximately 15,000 types and 450,000 tokens (Table 3). The 30 most frequent words in the MEDRA corpus and the normalized frequency (NF) are shown in Table 4. Ten texts each from the four portions were sampled and compared with 10 texts from all 400 texts. The 30 most significant words according to "log-likelihood ratio" (LLR; Dunning, 1993, p. 68) were identified using CasualConc (Imao, 2017) and shown in Table 5.[1] In the table, the words having "document frequency" (DF) of at least five (Tabata, 2012, p. 3) are in bold letters.[1]

In the MEDRA corpus as a whole (Table 4), the words *patients* and *cells* occurred 8,150 times (4.6 per 1000 words) and 7,578 times (4.3 per 1000 words), respectively, in 318 and 213 files. The letter *p* appeared times (5.0 per 1000 words) in 380 files and was mostly used to express *p* values.

(1) **Patients** with early-stage or minimal residual disease usually have lower levels of ctDNA, making it difficult to precisely detect specific alterations.
(Fan, Zhang, Yang, Ding, Wang, & Li, 2017, p. 2 [emphasis added])

(2) No single gene responsible for the commitment of mesenchymal **cells** to the angioblast cell fate has been identified as yet.
(Sumanas & Lin, 2006, p. 60 [emphasis added])

(3) In the combined group of stages II and III, serum mSST remained as a significant independent predictor of worse OS (HR = 2.797, 95% CI, 1.34–5.84; ***P*** = 0.006). (Fan et al., 2017, p. 12 [emphasis added])

Table 3: Word profiles of the MEDRA corpus as a whole and for each portion

	The MEDRA corpus as a whole	Cardiology portion	Pulmonology portion	Gastroenterology portion	Cancer portion
Type	49,542	15,656	16,588	17,298	14,666
Token	1,356,539	434,818	459,558	462,163	421,878

Table 4: The 30 most frequent words in the MEDRA corpus
(Normalized per 1000 words)

Rank	Word	NF	Rank	Word	NF	Rank	Word	NF	Rank	Word	NF	Rank	Word	NF
1	the	47.6	7	with	11.8	13	as	6.2	19	at	4.3	25	c	3.5
2	of	37.0	8	for	10.7	14	is	5.8	20	cells	4.3	26	study	3.3
3	and	30.9	9	were	10.4	15	or	5.3	21	on	4.3	27	not	3.3
4	in	26.4	10	was	9.6	16	p	5.0	22	are	3.6	28	s	3.2
5	to	17.1	11	that	6.6	17	from	4.8	23	we	3.5	29	be	3.1
6	a	15.8	12	by	6.5	18	patients	4.6	24	this	3.5	30	n	2.9

Table 5: The 30 most significant key words with ten texts from each portion
(Sorted according to LLR)

Rank	Cardiology			Pulmonology			Gastroenterology			Cancer		
	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR
1	**artery**	**7**	**132.7**	infants	4	180.5	pain	3	124.0	**cancer**	**10**	**227.4**
2	pathways	1	105.5	neonatal	4	105.5	uc	2	117.1	oral	2	106.7
3	cc	2	93.9	deaths	4	103.0	asa	1	85.3	aging	2	81.4
4	vascular	3	62.4	**mortality**	**5**	**87.8**	liver	4	84.2	ga	1	73.7
5	air	2	59.0	**infection**	**7**	**83.1**	tnbs	1	70.0	**ci**	**7**	**71.3**
6	ci	4	54.0	mutation	2	76.4	**no**	**10**	**69.6**	occult	2	71.2
7	**cardiac**	**7**	**53.4**	cd117	2	74.1	mortality	3	57.8	**d**	**5**	**65.4**
8	**dysfunction**	**5**	**49.0**	resistance	4	58.6	coverage	2	57.7	relapse	2	62.2
9	**coronary**	**6**	**47.6**	**children**	**6**	**52.1**	duration	4	54.8	**vs**	**5**	**61.9**
10	**rate**	**8**	**47.2**	**respiratory**	**7**	**52.1**	us	3	47.3	screening	4	61.4
11	**m**	**6**	**42.8**	x	4	51.9	children	3	47.0	**growth**	**5**	**60.0**
12	**vs**	**8**	**41.3**	**subject**	**6**	**51.0**	colonic	3	46.2	**diagnosis**	**6**	**56.4**
13	access	3	39.5	**death**	**6**	**50.1**	lamina	2	40.7	prognostic	3	48.3
14	origin	1	39.3	**causes**	**5**	**40.8**	propria	2	40.7	kg	2	47.6
15	wild	2	36.7	mca	1	39.9	improved	3	39.6	**response**	**7**	**44.0**
16	**post**	**6**	**34.8**	**rates**	**6**	**39.9**	predictor	2	39.6	survival	3	43.2
17	**baseline**	**5**	**34.7**	**n**	**9**	**37.5**	epithelium	1	36.4	liver	3	41.7
18	yes	1	33.8	month	3	36.8	**severity**	**5**	**33.4**	**status**	**7**	**40.5**
19	**reported**	**10**	**33.4**	**kit**	**5**	**35.4**	costs	1	33.1	symptoms	3	39.3
20	**tests**	**5**	**32.4**	born	3	35.1	apoptosis	3	32.5	**et**	**8**	**37.9**
21	monocytes	1	31.4	individuals	3	30.3	genotype	2	31.4	apoptosis	4	36.9
22	san	3	31.4	**culture**	**6**	**30.0**	**model**	**7**	**30.2**	**al**	**8**	**36.9**
23	**classification**	**5**	**30.4**	**rate**	**6**	**27.9**	health	4	29.8	**reference**	**5**	**36.0**
24	associations	2	30.1	**incidence**	**5**	**26.1**	drug	1	28.5	score	4	34.0
25	**studies**	**9**	**29.6**	**due**	**9**	**23.3**	**effects**	**8**	**28.2**	**diagnosed**	**7**	**33.2**
26	**during**	**10**	**29.2**	delivery	2	23.2	**genes**	**5**	**26.8**	**patients**	**10**	**32.9**
27	**no**	**10**	**29.1**	**risk**	**6**	**23.0**	**tests**	**5**	**26.6**	months	4	32.9
28	constant	2	27.8	responses	3	20.9	**al**	**9**	**26.1**	**table**	**9**	**32.5**
29	course	4	27.8	hospitalization	3	20.4	penetration	1	24.8	**median**	**6**	**32.3**
30	**type**	**9**	**27.3**	cd44	1	20.0	processes	2	24.8	**baseline**	**5**	**32.0**

The DF values in Table 5 indicate that many of the most significant words occurred in less than half of the texts, suggesting that each portion of the MEDRA corpus might be diverse in their word profiles. Some of the words, including *ci*, *rate*, *vs*, *baseline*, *tests*, and *no*, occurred at least five times in the texts sampled from two different portions.

3.3 Classification of the MEDRA corpus by text-type

The examination suggested that each portion of the MEDRA corpus may have various types of texts. Swales (1990, p. 19) supports the idea of Geertz (1983), who indicated that "Grand rubrics like 'Natural Science' ... and 'The Humanities' ... merely block from view what is really going on out there." Salager-Meyer (2014, p. 50) regards genres as "text-types" and raises the examples of medical genres such as "research articles," "case reports," and "review articles." Williams (1996, p. 175) defines "two types of research articles" as "clinical" and "experimental" in his contextual study of verbs in medical RAs.

According to the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH), which is an international non-profit association for ensuring "development and registration of safe, effective, and high quality medicine" (ICH, *n.d.*), the regulatory format for application of new drugs has separate sections for clinical and non-clinical texts. The informant who participated in this study commented that "[o]ne of the major factors that influence RA writing may be the degree of homogeneity in the methods used in studies. Studies with human subjects may not be as homogeneously controlled as those using animals or cells. For example, a clinical study may enroll participants with different dietary habits. ... Therefore, the rationale of study design such as experimental systems and inclusion criteria should be discussed in RAs."

Based on these considerations, the RAs in the MEDRA corpus were classified into five text-types (National Center for Global Health and Medicine, 2009): in vitro studies, in vivo animal studies, reviews, case and cohort studies, and meta-analysis based on discussion with the informant.[2]

The corpus texts were classified by visiting the original RA websites, observing the hint expressions in the title, abstract, and the methods section, if necessary. The following text was classified as an in vivo animal study:

(4) **Weekly** Doxorubicin **Increases** Coronary Arteriolar Wall and Adventitial Thickness

... Doxorubicin (DOX) **is associated with** premature cardiovascular events **including** myocardial infarction. **This study was performed to determine if** the **weekly administration of** DOX **influenced** coronary arteriolar medial and/or adventitial wall thickening. ... Thirty-two **male** Sprague-Dawley **rats aged** 25.1 ± 2.4 **weeks were randomly divided into** three **groups and received weekly** intraperitoneal **injections of...high** (2.5 mg/kg, n=11) **doses of** DOX. (Eckman et al., 2013, p. 1 [emphasis added])

The hint expressions in the example text included *weekly administration of* and *male...rats,* and the text was classified as an in vivo animal study using rats. Of the 400 texts, 14 were classified as in vitro studies; 36 as in vivo animal studies, 204 as case and cohort studies, 23 as meta-analysis, and 10 as reviews.[3] Ten texts each from the five text-types were sampled for comparison with 10 samples of the 400 texts which were different from those used for the data in Table 5. Table 6 shows the 30 most significant words according to LLR obtained by CasualConc (Imao, 2017).[1] The words in bold letters occurred in at least five (50%) of the sampled texts.

3.4 Pedagogical implication

Concordance lines revealed that the word *cell* in the in vitro study portion was frequently used to modify other nouns in phrases such as ***cell*** *culture* and ***cell*** *line.* The word *cells* observed in the animal study portion was, however, used as a noun in phrases such as *cancer* ***cells*** and *endothelial* ***cells***. The word *ci* in the case cohort study portion appeared solely in the phrase *95%* ***CI*** and occurred ten times more than its expanded form *confidence intervals*. These features are noteworthy for pedagogy.

In classrooms, the MEDRA corpus could be used for learners from various disciplines in medicine. The concordance lines with a node word of learner's interest, such as "cells" could be presented, and the usage of the word in paragraphs or sections could be presented to identify the text-type or genre. Scott and Tribble (2006, p. 109) indicate that combinations of "KW [keyword] analysis and discourse analysis" are useful for understanding "how language is used." Further study is needed to explore more approaches for pedagogy.

Table 6: The 30 most significant key words with ten texts from five text-type portions (Sorted according to LLR; the MEDRA corpus)

Rank	In vitro			Animal			Case cohort			Meta analysis			Review		
	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LLR	Word	DF	LL
1	**cells**	**9**	**390.4**	**mice**	**5**	**632.5**	**cases**	**5**	**145.1**	**survival**	**6**	**127.8**	**vs**	**8**	**291.8**
2	**expression**	**9**	**245.0**	**ca**	**7**	**196.5**	**cancer**	**5**	**110.5**	**pooled**	**7**	**92.7**	**months**	**6**	**169.4**
3	**mice**	**7**	**245.0**	**group**	**10**	**141.3**	crt	1	110.4	**meta**	**8**	**74.2**	**cohort**	**7**	**151.7**
4	**genes**	**6**	**206.8**	**rats**	**5**	**128.8**	**positive**	**6**	**104.2**	**analysis**	**9**	**69.2**	fc	1	129.4
5	monocytes	2	**171.8**	ko	1	114.3	df	1	101.2	**studies**	**10**	**65.5**	**studies**	**10**	**125.7**
6	radiation	3	**152.9**	**mouse**	**6**	**113.9**	hbv	1	99.4	surgery	3	58.3	water	3	104.1
7	**gene**	**6**	**125.6**	glucose	2	113.0	prevalence	4	96.0	**heterogeneity**	**8**	**57.2**	breast	4	94.8
8	**figure**	**8**	**118.8**	**b**	**10**	**110.2**	leakage	1	91.5	**methods**	**10**	**50.1**	activities	2	91.0
9	**cell**	**10**	**92.0**	**weight**	**9**	**109.8**	survival	4	77.5	**dose**	**7**	**48.6**	rt	1	84.0
10	**fold**	**6**	**90.2**	**induced**	**9**	**86.6**	activity	4	61.7	method	4	44.9	v	3	82.1
11	**protein**	**8**	**85.2**	**increased**	**10**	**84.0**	breast	1	57.0	hcc	1	42.6	**systematic**	**9**	**80.5**
12	**ca**	**8**	**78.8**	**mg**	**9**	**74.9**	metastasis	3	55.1	incidence	4	42.4	**cancer**	**7**	**78.4**
13	**liver**	**5**	**78.8**	**expression**	**8**	**74.5**	limitation	3	52.8	**subgroup**	**6**	**39.3**	**review**	**10**	**76.9**
14	**f**	**6**	**77.5**	**stimulation**	**5**	**73.8**	copd	2	51.3	af	1	37.7	**reported**	**10**	**69.5**
15	**human**	**8**	**68.2**	**changes**	**9**	**71.5**	lymph	3	48.9	**cancer**	**6**	**37.1**	**level**	**9**	**67.5**
16	**induced**	**8**	**62.3**	**normal**	**9**	**69.0**	nodes	2	48.3	supplementation	1	36.8	**study**	**10**	**66.8**
17	**controls**	**7**	**61.5**	**pressure**	**6**	**65.5**	**bmi**	**5**	**46.2**	**difference**	**8**	**36.0**	comorbidities	1	64.3
18	**c**	**10**	**60.6**	**body**	**8**	**63.8**	plasma	4	45.7	**effect**	**8**	**33.6**	**or**	**10**	**55.5**
19	nuclear	2	**60.3**	infected	2	63.3	**subjects**	**7**	**45.6**	estimated	4	31.4	publications	3	52.3
20	**using**	**10**	**60.2**	infection	3	63.3	infection	2	44.9	**disease**	**10**	**31.2**	**strength**	**5**	**50.6**
21	hcc	3	**52.9**	transmission	1	58.3	scan	3	44.6	nd	2	30.0	radiotherapy	3	49.0
22	infection	2	**51.8**	**significantly**	**9**	**56.6**	**variables**	**10**	**43.8**	yes	2	28.6	**regarding**	**8**	**47.3**
23	**significantly**	**10**	**48.8**	**figure**	**7**	**56.5**	**age**	**10**	**43.2**	**tumors**	**5**	**28.4**	**one**	**10**	**46.2**
24	**tumor**	**7**	**48.7**	**protein**	**7**	**55.8**	pressure	4	42.6	**plot**	**7**	**27.7**	alterations	2	44.1
25	μm	4	**48.5**	**c**	**10**	**53.9**	**index**	**5**	**40.4**	rt	2	27.7	**criteria**	**9**	**40.7**
26	**by**	**10**	**46.7**	**function**	**8**	**53.7**	wall	2	40.1	**forest**	**7**	**26.9**	**exclusion**	**7**	**39.0**
27	macrophages	1	46.2	transient	4	**52.9**	controls	3	39.1	restricted	2	26.6	indicator	3	38.9
28	**pathway**	**7**	**46.2**	**d**	**8**	**49.2**	females	4	37.3	**bias**	**8**	**25.6**	**described**	**10**	**37.7**
29	**change**	**8**	**45.7**	**weeks**	**8**	**48.0**	**cohort**	**7**	**37.1**	**included**	**10**	**25.4**	nr	2	36.9
30	acid	4	45.2	af	2	**47.6**	**factors**	**10**	**35.8**	oral	3	24.8	cc	2	34.2

4. Further research directions

The word profiles of the two corpora were identified in this study; the next step will be to use the corpora in writing classrooms. Handford (2013) argues that it is essential to understand "the context in which the text is produced, and the constraints under which the writer is working" (p. 260) and that "such understanding can then be transferred to academic writing requirements the students may have" (p. 261). Activities such as exploring lexico-grammatical patterns could be attempted. Concordance lines for *cancer* in MEDRA corpus, for example, show that it is frequently used to modify another noun, as in *a* ***cancer*** *patient*; it is often modified by other nouns, as in *breast* ***cancer*** and *colon* ***cancer***; and it is also modified using an adjective, as in *esophageal* ***cancer*** and *pancreatic* ***cancer***. This suggests patterned usage in the discourse community,

which is worth examining in class. Activities could also be attempted from the genre-based approach. The phrase *in this study, we* in the MEDRA corpus occurred in the discussion section of approximately 80 texts and was seen in the introduction section of about 40 texts. It should be suggested that "[s]tudents run a concordance" (Stevens, 1991, p. 45) on the phrase *in this study, we* in the corpus and learn the rhetorical patterns. The combination of the genre-based framework and corpus-based study may allow us to learn specific usages of words and phrases that help students develop their proficiency in discipline-specific writing.

5. Conclusion

The corpora were built successfully with AntCorGen. The CAPHYRA corpus was considered to have linguistic features of medical RAs of the field. The MEDRA corpus should be useful for pedagogical applications for students who are or will be conducting various types of studies in medicine.

Corpora built using AntCorGen, with file names enabling access to the original texts, can be used to identify high frequency phrases in the entire corpus or a portion of it to identify the communicative purposes (genre or text-types). Corpora can also be used to accumulate the knowledge that needs to be presented for the genre in classroom settings.

Acknowledgments

The author would like to express sincere gratitude to Tomoji Tabata, Hisashi Iwane, Yasuhiro Imao, Maki Miyake, Hodošček Bor, Nobuyuki Hino, Judy Noguchi, Tomoko Wakasa, and three anonymous reviewers for their invaluable comments on the manuscript. The author also wishes to thank the editor of this journal. The author is also grateful to audience at the 43rd Annual Conference of JAECS at Kwansei Gakuin University in 2017 for their helpful comments. Any remaining errors are the author's responsibility.

Notes

1. The data are available upon request.
2. Several types of studies (National Center for Global Health and Medicine, 2009) were

combined for categorizing purposes.

3. The remaining texts were classified as *Other* because they were unclassifiable, but they were included in the analysis.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2017a). AntCorGen (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2017b). Automating the construction of individualized discipline-specific corpora for data-driven learning. In I. Holliday, K. Hyland, & L. L. C. Wong (Eds.), *Conference Programme Book for the Center for Applied English Studies (CAES) International Conference: FACES of ENGLISH 2, Teaching and Researching Academic and Professional English* (p. 71). Hong Kong.

Caze, A. L. (2011) The role of basic science in evidence-based medicine. *Biology & Physiology, 26(1),* 81–98. DOI 10.1007/s10539–010–9231–5

Coates, R., Sturgeon, B., Bohannan, J., & Pasini, E. (2002). Language and publication in Cardio-vascular Research articles. *Cardiovascular Research*, *53*, 279–285.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19(1),* 61–74.

Eckman, D. M., Stacey, R. B., Rowe, R., D'Agostino Jr., R, Kock, N. D., Sane, D. C., ... Hundley, W. G. (2013). Weekly doxorubicin increases coronary arteriolar wall and adventitial thickness. *PLOS ONE*, *8(2)E57554*, 1–6.

Fan G, Zhang K, Yang X, Ding J, Wang Z, & Li, Z. (2017). Prognostic value of circulating tumor DNA in patients with colon cancer: Systematic review. *PLOS ONE, 12(2)*, 1–17

Feak, C. B., & Swales, J M. (2012). *Academic Writing for Graduate Students, Essential Tasks and Skills* (3rd ed.). MI, USA: University of Michigan Press.

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing. *Applied Linguistics, 28(3)*, 440–465.

Flowerdew, J., & Wang, S. H. (2017). Teaching English for research publication purposes with a focus on genre, register, textual mentors and language re-use: a case study. In J. Flowerdew & T. Costley (Eds.): *Discipline-Specific Writing: Theory into Practice* (pp. 144–161). Oxon, UK: Routledge.

Geertz, C. (1983). *Local Knowledge: Further Essays in Interpretive Anthropology*. New York: Basic Books.

Handford, M. (2013). What can a corpus tell us about specialist genres? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 255–269). Oxford, UK: Routledge.

Hunston, S. (2013). How can a corpus be used to explore patterns? In A. O'Keeffe & M.

McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 152−166). Oxford, UK: Routledge.

ICH. (*n.d.*). M4: The Common Technical Document. URL: http://www.ich.org/products/ctd.html

Imao, Y. (2017). CasualConc (Version 2.0.7) [Computer Software]. Osaka, Japan: Osaka University. Available from https://sites.google.com/site/casualconc/

Iverson, C., Christiansen, S., Flanagin, A., & Fontanaroas, P. B. (2007). *American Medical Association Manual of Style* (10th ed.). New York: Oxford University Press.

Johns, T. (1991). From print out to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *English Language Research Journal, 4,* 27−45.

Jones, M., & Dimant, P. (2013). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 387−400). Oxford, UK: Routledge.

Lee, D. (2008). Corpora and discourse analysis: New ways of doing old things. In V. Bhatia, J. Flowerdew, & H. Jones (Eds.), *Advances in Discourse Studies* (pp. 86−99). London: Routledge.

Lee, D. & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, *25*, 56−75.

Maher, J. (1986). The development of English as an international language of medicine. *Applied Linguistics*, *7(2)*, 206−220.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, *19*, 63−86.

Maswana, S., Kanamaru, T., & Tajino, A. (2015). Move analysis of research articles across five engineering fields: What they share and what they do not. *Ampersand, 2*, 1−11.

Mauranen, A., (2017). Academically speaking: English as the lingua franca. In I. Holliday, K. Hyland, & L. L. C. Wong (Eds.), *Conference Programme Book for the Center for Applied English Studies (CAES) International Conference: FACES of ENGLISH 2, Teaching and Researching Academic and Professional English* (p. 22). Hong Kong.

Mizumoto, A., Hamatani, S., & Imao, Y. (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.

Nakatani, Y., & Bucsis, C. (Ed.). (2016). *Daigakusei no Tame no Academic Eibun Writing* [Academic Writing Strategies for University Students]. Tokyo: Taishukan Publishing Co., Ltd.

National Center for Global Health and Medicine. (2009). *Shoki Rinsho de Minitsuketai Rinsho Kenkyu no Essence* [Essence for Early Phase Clinical Research]. URL: http://www.imcj-gdt.jp/topics_04.pdf

Nesi, H. (2013). ESP and corpus studies. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 407−426). West Sussex, UK: Wiley Blackwell.

Noguchi, J., (2004). A genre analysis and mini-corpora approach to support professional writing by nonnative English speakers. *English Corpus Studies, 11*, 101–110.

Noguchi, J., Matsuura, K., & Haruta, S. (2015). *Judy Sensei no Eigo Kagaku Ronbun no Kakikata Zoho KaiteiBan* [An Efficient Approach to Writing Up Your Research]. Tokyo, Japan: Kodansha.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for Specific Purposes*, *16(2)*, 119–138.

Paltridge, B. (2013). ESP and pedagogy. In B. Paltridge & S. Starfield (Eds.), *The Handbook of English for Specific Purposes* (pp. 347–366). West Sussex, UK: Wiley Blackwell.

Pocock, G., Richards, C. D., & Richards, D. A. (2006). *Human Physiology*. UK: Oxford University Press.

Salager-Meyer, F. (1989). Principal component analysis and medical English discourse: An investigation into genre analysis. *System, 17(1)*, 21–34.

Salager-Meyer, F. (2014). Origin and development of English for medical purposes. Part I: Research on written medical discourse. *Medical Writing*, *23(1)*, 49–51.

Scott, M., & Tribble, C. (2006). *Textual Patterns*. Amsterdam: John Benjamins.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant authentic text. *English for Specific Purposes, 10,* 35–46.

Sumanas, S., & Lin, S. (2006). Ets1-related protein is a key regulator of vasculogenesis in zebrafish. *PLOS Biology*, *4(1)*, 0060–0069.

Swales, J. M. (1990) *Genre Analysis: English in Academic and Research Setting*. UK: Cambridge University Press.

Tabata, T. (2012). Dickens to Collins no kyocho sakuhin heno buntai tokeigakuteki approach. *IPSJ SIG Technical Report, 2012-CH-93(3),* Information Processing Society of Japan.

Tojo, K., Hayashi, H., & Noguchi J. (2014). Linguistic dimensions of hint expressions in science and engineering research presentations. *JACET International Convention Selected Papers*. *1*, 131–163.

Williams, I. A. (1996). A contextual study of lexical verbs in two types of medical research report: Clinical and experimental. *English for Specific Purposes, (15)3*, 175–197.

（浅野　元子　大阪大学大学院　Email: u029294j@ecs.osaka-u.ac.jp）