Lexical bundles in Wikipedia articles and related texts: exploring disciplinary variation

0. Abstract

Wikipedia is widely used by academics and students in higher education, but research on the linguistic characteristics of this genre is scarce (Kuteeva 2016). This paper explores the usefulness of lexical bundles as an analytical tool to describe disciplinary variation *within* Wikipedia articles, and to contrast Wikipedia writing with two neighbouring genres, student essays and research articles. The results indicate that the occurrence of lexical bundles in Wikipedia varies between disciplines, which is in broad agreement with previous studies on other academic genres. The analysis of bundles also suggests that a credible authorial persona is less crucial to Wikipedia articles. Indicative of this is the low frequency of bundles indicating stance and engagement, which are characteristic of professional academic writing (e.g. Hyland 2008a).

Keywords: Wikipedia, disciplinary differences, academic discourse, lexical bundles

1. Introduction

Since its establishment in 2001, the growth of Wikipedia has been impressive. This collaboratively edited, multilingual, free Internet encyclopaedia (https://en.wikipedia.org/wiki/Wikipedia) currently includes over 5 million articles, and holds seventh position in *Alexa Traffic Ranks*, which is a measure of a site's popularity based on the number of visitors globally.[1]

Wikipedia has gradually gained a foothold within higher education, although its status remains somewhat controversial. On the one hand, based on several reports, it is widely used not only by students but also instructors, in different ways and for a variety of purposes (see e.g. McCloud 2007, Eijkman 2010 and Konieczny 2014). In a recent article, Jemelniak (2014) calls Wikipedia "a professor's best friend", and several studies suggest that making use of Wikipedia and *wikis* has great potential for positive learning outcomes (Barton 2005, Barton and Cummings eds. 2008, Tapscott and Williams 2010). Moreover, Kuteeva (2016: 437) suggests that "Wikipedia is often treated as a credible source by scientists". On the other hand, many instructors have expressed concerns about the quality of the information in Wikipedia articles, as well as about students overusing them at the expense of textbooks and peer-reviewed articles (for an overview of this criticism, see, e.g., Myers 2010 and Kuteeva 2016).[2] For example, in a widely cited article entitled "Why You Can't Cite Wikipedia in My Class", historian Neil Waters questions the idea that "the vector-sum products of tens or hundreds of anonymous collaborators could have much value" (2007: 15) for academic historians. He recounts his own frustration about Wikipedia in a higher education context as follows:

> While grading a set of final examinations from my "History of Early Japan" class, I noticed that a half-dozen students had provided incorrect information about two topics [...] on which they were to write brief essays.

[1] Data from January, 2016, see further http://www.alexa.com/siteinfo/wikipedia.org.

> Moreover, they used virtually identical language in doing so. A quick check on Google propelled me via popularity-driven algorithms to the Wikipedia entries on them, and there, quite plainly, was the erroneous information. (Waters 2007, 15)

Waters' experience highlights two problems associated with Wikipedia. The first of them, quality of information, has been a major concern in the debate on Wikipedia.[3] By contrast, the second issue, namely what the appropriate way of using information from Wikipedia in higher education (HE) assignments would be, has received less attention. Yet the second issue is also important in today's HE, where the use of Wikipedia has become so widespread that some scholars (e.g. Myers 2010 and Kuteeva 2011) even refer to the Wikipedia article as "a new academic genre" (Kuteeva 2011: 46). Wikipedia articles thus form an increasingly important part of "writers' experiences of previous discourse" (Hyland 2000, 145), through which writing is mediated and produced. From this it follows that English for Academic Purposes (EAP) teachers in the 21th century need to consider what to tell students about Wikipedia: whether it can be consulted as part of learning tasks and assignments, and if so, in what way. This is an issue not only in academic writing courses offered to ESL students, but also content classes, where students may come across Wikipedia articles when they do research for their coursework.

In addition, it is well known that discipline is one of the main determinants of linguistic variation in academic writing, both as far as published writing and student writing are concerned (e.g. Hyland 2000, Fløttum et al 2006, Durrant 2014 and 2015). This being

[3] See for example Myers (2010: 129–30; 143–4).

the case, the issue of disciplinary variation in Wikipedia is obviously relevant. To provide support for these assessments, up-to-date empirical descriptions of language and discourse of Wikipedia are clearly needed, especially in relation to the kind of discourses that students are ultimately expected to master – competent *expert performances* (cf. Bazerman 1994) in their field of study.

To this end, the present chapter offers a data-driven investigation into textual characteristics of Wikipedia writing. As the focus is on higher education settings, specific emphasis is placed on features that might set Wikipedia apart from more established academic genres with which students routinely engage. Specifically, Wikipedia articles are compared to a baseline of two other kinds of writing: research articles and student essays. These genres differ from each other in important ways. The research article (RA) is a key academic genre both in terms of prestige and number, and is one of the preferred types of reference literature for student writing, on a par with academic textbooks. As textual expert performances, they can thus reasonably be taken to represent the kind of writing that serves as a model for student writers (cf. Tribble 2011:88). The latter category, student essay, contains texts representing several assessment-related genres in university settings.

The linguistic and textual differences between these text categories are investigated through an analysis of *lexical bundles*, recurrent sequences of word forms in discourse. After Biber et al.'s comprehensive treatment in the *Longman Grammar* (1999), lexical bundles have been analysed in a variety of genres.[4] The present study investigates their role in corpora that represent Wikipedia articles and related texts, with a specific focus

[4] Various terms are used in literature, which fully or partially overlapping denotations: including clusters (Scott 2006), prefabs (Granger 1998) and n-grams. For an overview of terminology in the study of formulaic language, see Wray 2002.

on variation between articles representing different areas of inquiry. As I shall argue, this "radical corpus-driven approach" (Biber 2009: 281) provides a low-effort yet powerful method for describing how Wikipedia writing differs from related texts: a frequency-driven approach with minimal theoretical assumptions is suited for the exploratory goals of the present study. Moreover, as will be shown, a quantitative analysis of bundles is able to draw attention to frequent language patterns and highlight differences between Wikipedia writing and other specialised corpora. In turn, these patterns can generate more specific hypotheses for a quantitative analysis, or they can be interpreted using other methods, such as genre analysis (Swales 1990).

The analysis identifies the most characteristic bundles of three and four words in Wikipedia articles representing three different fields: medicine, economics and literary studies. To interpret the quantitative findings, I consider the role of contextual factors which may account for the observed variation, and finally, consider their implications for English for Academic Purposes.

2. Methodology

In studies carried out after the publication of *Longman Grammar* (1999), lexical bundles are usually understood as word sequences with a statistical tendency to co-occur, but the precise operational definitions vary. For Biber et al. (1999: 992–993), any potential lexical bundle should occur at least ten times per one million words (pmw), and these occurrences should be spread across at least five texts.[5] Other studies in the EAP framework use uses more stringent criteria: Hyland (2008a: 9) requires that

[5] The latter criterion is introduced to distinguish lexical bundles from "local repetitions", which are characteristic of individual speakers and texts and not representative of general patterns of langauge use (Biber et al. 1999: 991). Salazar notes that the threshold frequencies are "somewhat arbitrary" (2014: 13).

potential bundles would occur 20 times per 1 million words, and across 10% of texts, Ädel and Erman (2012) and Baker and Chen (2010) apply a threshold of 25 pmw, Tribble's (2011) threshold is 30 pmw, and in Biber (2006) we find 40 pmw. In this paper, the criteria 20 pmw and 10% are used.[6]

It is often stated that lexical bundles play an important role in creating a coherent and idiomatic text, but previous research has also shown that their use is highly context-bound, in that depending on the genre of the text, different bundles tend to emerge as prominent (e.g. Biber 2006, 2009). Studies on lexical bundles in written academic language have been carried out for different purposes, and using a variety of research designs. A particular area of interest has been the use of bundles as determined by different contextual variables, including genre (e.g. Biber et al 2004; Biber and Barbieri 2006), writers' expertise (Tribble 2011), and native language (Chen and Baker 2010, Ädel and Erman 2012, Perez-Llantada 2014, Salazar 2014). These studies typically discuss pedagogic applications, either in the form of suggested activities or lists of purportedly useful common-core bundles, which have been identified empirically (e.g. Simpson-Vlach and Ellis 2010).

The emphasis on pedagogic concerns is hardly surprising, given that lexical bundles are commonly identified as a potential problem area in apprentice texts. Cortes (2004), for instance, found that compared to published texts, student papers in history and biology contain few bundles, and those that are found are often used repeatedly, which makes the discourse sound "redundant" (2004: 415). On the other hand, Hyland's (2008b) data on undergraduate dissertations demonstrates a heavy reliance on certain

[6] An example of a more complex operationalisation of these recurrent sequences of is Simpson-Vlach and Ellis' *Academic Formulas List* (2010), which is derived using a combination of corpus statistics, psycholinguistic metrics and instructors' assessment of utility.

bundles which are not characteristic of more proficient academic prose (PhD theses and RAs). Hyland offers two alternative explanations for these findings (which need not exclude one another): student genres may be more formulaic in nature, or students "need to display a more conciliatory approach to arguments and to demonstrate that alternative points of view have been considered (Hyland 2008b, 50).

Most lexical bundle studies concur that compared to competent expert writing by native speakers of English, apprentice writing both in L1 and L2 English exhibits a more restricted repertoire of lexical bundles (Granger 1998; Baker & Chen 2010; Ädel and Erman 2012, Salazar 2014).[7] This seems to apply to professional L2 academic writing as well: Pérez-Llantada suggests that "(proficient) L2 is partly, but not fully, native-like" (2014: 93), as far as the use of lexical bundles is concerned.[8] However, due to differences in the research designs, Baker and Chen (2010: 44) conclude that "it is still not conclusive as to whether there is a relationship between proficiency and the number of formulaic expressions used".[9]

The main variable in focus in this study, discipline, has been established as one of the major determinants of variation in academic writing in general (e.g. Bazerman 1981, Becher & Trowler 1998, Hyland 2000, Fløttum et al 2006, Groom 2009, Hiltunen 2010). Hyland (2008a) has further shown that bundle repertoires vary considerably between different disciplines in published academic writing, and more recently Durrant (2015) has demonstrated that the use of lexical bundles varies across disciplinary

[7] Baker and Chen (2010) considered L1 speakers of Chinese, and Ädel and Erman (2012) L1 speakers of Swedish.

[8] Based on data from L1 speakers of Spanish.

[9] A similar point is made by Pacquot and Granger (2012).

divisions. The results of these studies suggest that the discipline variable is potentially useful in describing variation in Wikipedia writing.

3. Corpus and context

There are several ways to extract text from Wikipedia. Articles can be downloaded directly from the site (e.g. Myers 2010), or using a web-based tool that extracts a sample of articles "on the fly" based on seed words (Davies 2015, cf. Baroni and Bernardini 2004 and Kilgarriff et al. 2010). Alternatively, researchers can rely on a previously released Wikipedia corpus, created from a "data dump" made available by Wikipedia itself.[10] The three subcorpora representing Wikipedia articles were extracted from one such Wikipedia corpus, the *Westbury Lab Wikipedia Corpus* (WLWC) (Shaoul and Westbury 2010), a 990-million-word snapshot of the English Wikipedia taken in April 2010.

To analyse disciplinary variation in Wikipedia writing, it is necessary to extract samples representing different areas of inquiry. This extraction cannot be fully automated, as the WLWC is released as a simple plain text corpus, and as such contains no information about the hierarchical relationships of articles or the links between them. The samples were therefore extracted using the titles of the articles, based on the assumption that for the most part, the main page of an academic discipline links to other articles relevant to the discipline and can therefore considered to represent the same field in a broad sense.[11] Accordingly, each sample includes all the articles which are

[10] See https://en.wikipedia.org/wiki/Wikipedia:Database_download and https://dumps.wikimedia.org/. For a list of other Wikipedia corpora, see for example Hiltunen (2014: section 3).

[11] For example, the Wikipedia page *Medicine* links to articles on different terms (e.g. *diagnosis, treatment, musculosceletal*), body parts (e.g. *abdomen, blood vessels*) and illnesses, (*diabetic ketoacidosis, heat failure*), but also to articles on the history of medicine and medical institutions (e.g. *Greek medicine, Edward Jenner, hospital*).

linked to from the main pages of the subject areas: *Economics*,[12] *Medicine*,[13] and *Literary criticism*[14] (accessed 25 July 2014). The *Econ-Wiki* and *Med-Wiki* subcorpora are roughly the same size, and nearly twice as large as LC-Wiki (see Table 1).

Subcorpus Name	**Discipline**	**Number of texts**	**Number of words**
Econ-Wiki	Economics	470	856,272
Med-Wiki	Medicine	439	837,269
LC-Wiki	Literary studies	182	473,604

Table 1. Subcorpora extracted from the WLWC: number of texts and words.

These corpora are then compared to reference corpora representing two other genres: peer-reviewed research articles and student essays. These reference corpora, summarised below in Table 2, consist of existing corpora. The three subcorpora of research articles (Econ-PUB, Med-PUB, and LC-PUB) comprise texts published in influential journals and represent different sub-disciplinary specialisms.[15] The subcorpora of student essays (Econ-BAWE, Med-BAWE and LC-BAWE) come from the *British Academic Written English* (BAWE) corpus (Nesi 2008). Given the

[12] https://en.wikipedia.org/wiki/Economics .

[13] https://en.wikipedia.org/wiki/Medicine.

[14] https://en.wikipedia.org/wiki/Literary_criticism.

[15] The journals include the *Journal of Financial Economics* and *Journal of Economic* Literature (economics), *Spine* and *Journal of Orthopedic* Research (medicine) and *American Literature* and *Comparative Literature Studies* (literary criticism). For a full description of Med-PUB and LC-PUB, see Hiltunen 2010. For a description of Econ-PUB; see Hiltunen and Mäkinen (2014) and Mäkinen and Hiltunen (2016).

exploratory nature of this paper, no attempt was made to limit the variety of student writing to be considered; all texts representing the three disciplines in focus were included in the analysis irrespective of their "genre family" in BAWE.[16]

Subcorpus Name	Discipline	Number of texts	Number of words
Econ-BAWE	Economics	96	221,821
Econ-PUB	Economics	50	550,633
Med-BAWE	Medicine	80	214,226
Med-PUB	Medicine	64	248,693
LC-BAWE	Literary studies	106	241,516
LC-PUB	Literary studies	64	516,242

Table 2. Reference corpora used in this study.

4. Data preparation and analysis

All bundles consisting of 3 or 4 words were retrieved in all subcorpora, and their type and token frequencies determined. After initial exploration of the data, 4-word bundles were taken under closer investigation, as they appeared to provide the highest "signal-to-noise" ratio for describing genre-based differences and disciplinary variation (cf. Hyland 2008: 22).[17] The identification of bundles was case-insensitive, so that *at the same time* and *At the same time* were treated as instances of the same bundle.[18]

[16] BAWE distinguishes between 13 ”genre families”, each of which contains a number of variously labelled genres which share some ”functional and structural properties” (BAWE manual, p. 7).

[17] As observed in many previous studies, three-word bundles are often subsumed under 4-word bundles. 5-word bundles are relative infrequent, and have little to contribute to the present analysis.

[18] Lowercase forms of bundles are used consistently in this paper, even if they include proper nouns.

Consistent with the data-driven approach adopted, data manipulation were kept minimal: overlapping bundles were not merged, bundles containing numbers were excluded from the analysis, but content bundles were kept in the data set.

Following the automatic retrieval of bundles, the subcorpora were compared in terms of overall frequencies and the degree of sharedness of bundle types across genres and disciplines. To assist these comparisons, following Hyland (2008a; 2008b) and Tribble (2011), individual bundles were also categorised in terms of their function in discourse into *research oriented*, *text-oriented,* and *participant-oriented* bundles.[19]

5. Discussion of findings

5.1 Frequency of bundles across genres and disciplines

The main finding emerging from the analysis of frequencies is that Wikipedia articles stand out from the other two genres, as far as the overall frequency of lexical bundles is concerned. Table 3 lists the number of bundle types and tokens, as well as the percentage of text found within bundles for each of the nine subcorpora. As can be seen, Wikipedia articles in all three disciplines contain many fewer 3-word bundles than the reference corpora, with economics displaying a particularly large difference across genres: compared to Econ-Wiki, Econ-PUB, is over one-third smaller but contains over ten times more bundle types. Although less frequent overall, the data for 4-word provides a similar picture, as shown in Table 4.

[19] Other functional classifications are found in the literature; for example, Biber et al (2004) use the terms *stance expressions*, *discourse organizers*, and *referential expressions*, and Pérez-Llantada (2014) the terms *referential*, *text-organising* and *stance* bundles.

Subcorpus	Bd Types	Bd Tokens	% of text in Bds
Econ-Wiki	112	12,054	4.2
Econ-BAWE	308	7,542	10.2
Econ-PUB	1,384	33,990	18.5
Med-Wiki	99	10,616	3.80
Med-BAWE	727	19,666	27.5
Med-PUB	551	49,700	12.0
LC-Wiki	421	10,405	6.6
LC-BAWE	260	6,756	8.4
LC-PUB	836	17,988	10.4

Table 3. Frequency of 3-word bundles (Bd) by DISCIPLINE and GENRE. (Wikipedia highlighted in grey.)

Subcorpus	Bd Types	Bd Tokens	% of text in Bds
Econ-Wiki	9	840	0.39
Econ-BAWE	26	569	1.03
Econ-PUB	172	3,510	2.54
Med-Wiki	10	954	0.45
Med-BAWE	341	12,010	22.42
Med-PUB	88	1,195	1.90
LC-Wiki	40	968	0.81
LC-BAWE	26	645	1.07
LC-PUB	83	1,583	1.22

Table 4. Frequency of 4-word bundles by DISCIPLINE and GENRE.

Table 3 and Table 4 point to clear differences, which appear to be primarily motivated by GENRE rather than DISCIPLINE: bundles are consistently more frequently used in published RAs than in Wikipedia articles, as shown by the fact that a larger proportion of text is found *within* bundles. As an appropriately varied repertoire of bundles is characteristic of competent academic writing (e.g. Cortes 2004, Hyland 2008b), this potentially indicates that Wikipedia writing is not a good a model for it. That said, some of the variation is clearly due to DISCIPLINE, especially when frequencies in student writing are taken into account. For example, student texts in medicine contain far more bundles than Wikipedia articles, but in the other two disciplines do not follow this pattern. To obtain a better understanding of the nature of differences, we shall next look at both individual bundles and the degree of sharedness between genres and disciplines in more detail.

5.2. Disciplinary variation in Wikipedia articles

A comparison of the frequency-ranked lists of lexical bundles shows that the three samples from Wikipedia (Econ-Wiki, Med-Wiki, and Lit-Wiki) share a great deal of common ground. As shown in Table 5, nearly all 4-word bundles (i.e., 8/9) are shared between Econ-Wiki and Med-Wiki, and these bundles also occur in Lit-Wiki. These core bundles (highlighted by black shading) include indicators of place (*in the united state*s, *at the university of*), as well as text-oriented indicators of comparison, contrast and result (*on the other hand*, *as well as the,* and *as a result of*).

	Economics		**Medicine**		**Literary criticism**	
	Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
1	in the united states	205	in the united states	325	at the university of	70

2	at the university of	129	as well as the	79	the end of the	61
3	on the other hand	80	at the university of	74	one of the most	50
4	one of the most	79	as a result of	72	on the other hand	39
5	the end of the	79	the end of the	72	in the united states	38
6	in the form of	73	one of the most	68	in the form of	22
7	as a result of	69	in the form of	59	as well as the	21
8	as well as the	67	on the other hand	50	as a result of	20
9	at the same time	59	is one of the	74	at the same time	40
10			in the united kingdom	81	is one of the	24
11					at the age of	49
12					was one of the	43
13					at the end of	38
14					one of the first	31
15					the beginning of the	29
16					as one of the	27
17					in the history of	23
18					for the first time	17
19					one of the greatest	17
20					the history of the	16

Table 5. 4-word bundles in Wikipedia articles (black shading = bundle occurs in all three subcorpora, light grey shading = bundle occurs in two subcorpora, no shading = bundle occurs in one subcorpus)

In addition, Table 5 also provides an indication of a usage specific to LC-Wiki, which contains 8 bundles unique to this subcorpus. Most of these discipline-specific bundles fall under *research-oriented bundles* in Hyland's (2008a) classification: they deal with the writer's experience of the real world, and their prominence clearly reflects the fact that LC-Wiki contains articles authors, critics, and literary figures, which are characteristically biographic or historical in their orientation, as illustrated by examples (1) and (2).

(1) **At the age of** 65, Blake began to work on illustrations for the "Book of Job". (*LC-Wiki*: *William Blake*)

(2) Goethe **was one of the** key figures of German literature and the movement of Weimar Classicism in the late 18th and early 19th centuries …(*LC-Wiki*: *Johann Wolfgang von Goethe*)

Articles with a similar orientation are certainly also found in the other two subcorpora – Econ-Wiki contains articles on economists like Adam Smith, J. M. Keynes, and Amartya Sen, and Med-Wiki articles on important figures in medical history like Vesalius, Pasteur, and Paul Broca. However, the proportion of biographically oriented articles is much smaller than in LC-Wiki, and sequences characteristic of this type of writing are therefore less likely to occur in a sufficiently large number of texts to be classified as lexical bundles. The prominence of bundles illustrated in (1) and (2) highlights the fact that many Wikipedia articles are have concerns that are not shared by much of research writing.

Three-word bundles tell a similar story: there is a large common core of bundles shared between the tree subcorpora, in addition to which LC-Wiki contains a considerable number of bundles which are unique to it. To illustrate, Table 6 shows 40 most frequently occurring bundles in each discipline, of which only one does not occur in all three subcorpora (*the theory of*); the rest occur in all three, as indicated by black shading.

	Economics		**Medicine**		**Literary criticism**	
	Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
1	one of the	382	as well as	436	as well as	143
2	as well as	339	in the united	411	the university of	119
3	the study of	273	one of the	399	the end of	106
4	such as the	268	such as the	283	a number of	102
5	in the united	239	the use of	257	in order to	93
6	the university of	236	the study of	229	at the time	83
7	in order to	212	part of the	209	the history of	83
8	part of the	206	the development of	206	at the university	80

9	a number of	201	the university of	196	such as the	75
10	the use of	197	a number of	190	the united states	73
11	the development of	172	in order to	189	of the most	70
12	the end of	155	according to the	152	the work of	69
13	at the university	147	referred to as	150	in the early	68
14	in terms of	146	known as the	141	part of the	67
15	referred to as	146	as a result	132	end of the	65
16	there is a	142	a variety of	130	a series of	61
17	known as the	139	the end of	129	the development of	61
18	as a result	137	the field of	125	of the first	59
19	the concept of	128	due to the	120	some of the	59
20	of the world	124	in addition to	116	the use of	58
21	the work of	123	there is a	116	the works of	54
22	some of the	116	based on the	108	was born in	54
23	there is no	111	in the world	105	many of the	52
24	in the late	110	in the early	98	to be a	51
25	due to the	109	some of the	97	was one of	51
26	on the other	108	was the first	96	at the age	49
27	the form of	107	of the most	94	the fact that	49
28	of the most	103	end of the	89	the beginning of	48
29	based on the	100	is one of	87	the study of	48
30	the field of	100	most of the	87	was published in	48
31	a variety of	99	a result of	84	in which he	47
32	in the world	94	in terms of	83	of his life	47
33	the existence of	94	because of the	81	on the other	47
34	in the early	93	as part of	80	to be the	47
35	a result of	91	many of the	79	he did not	46
36	in which the	89	the form of	79	of the world	46
37	the theory of	165	well as the	79	the concept of	45

Table 6. 3-word bundles in Wikipedia articles (black shading = bundle occurs in all three subcorpora, no shading = bundle occurs in one subcorpus)

Another noteworthy finding in Table 5 and Table 6 is the low frequency of text-oriented bundles, which are especially characteristic of the discursive rhetoric of the soft fields of inquiry (Hyland 2008a: 16); we shall come to this question in the next section, where we compare different genres within the same discipline.

5.3 Genre-based variation within disciplines

5.3.1 Economics

As previously noted, Econ-Wiki contains many fewer bundle types than expert writing (Econ-PUB); this is true for both 3- and 4-word bundles. For example, as shown in Table 7, Econ-Wiki contains merely nine 4-word bundles, five of which are core bundles found in the other economics genres: *on the other hand, in the form of, as a result of, as well as the*, and *at the same time.*

	Econ-Wiki		**Econ-BAWE**		**Econ-Pub**	
	Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
1	on the other hand	80	on the other hand	50	on the other hand	74
2	in the form of	73	as a result of	34	as well as the	44
3	as a result of	69	in the form of	27	at the same time	39
4	as well as the	67	as well as the	22	in the form of	23
5	at the same time	59	at the same time	22	as a result of	21
6	in the united states	205	an increase in the	29	in the united states	74
7	the end of the	79	in the case of	28	the extent to which	71
8	at the university of	129	it is possible to	28	in the case of	45
9	one of the most	79	that there is a	23	an increase in the	32
10			the extent to which	23	the fact that the	30
11			the value of the	18	the end of the	29
12			the fact that the	15	in terms of the	27
13			the size of the	15	the difference between the	26
14			in terms of the	13	the size of the	26
15			it is difficult to	13	it is difficult to	23
16			the difference between the	12	that there is a	22
17			in the long run	57	the value of the	22
18			in the short run	26	it is possible to	17
19			we can see that	23	are more likely to	93
20			in this case the	19	to the extent that	58
21			that there is no	16	in the context of	52
22			is known as the	13	on the basis of	52
23			take into account the	12	is consistent with the	45
24			to take into account	11	more likely to be	38
25			be explained by the	10	we find that the	37
26			can be explained by	10	it is important to	36

Table 7. 4-word bundles in three economics subcorpora, ranked by frequency (black shading = bundle occurs in all three subcorpora, light grey shading = bundle occurs in two subcorpora, no shading = bundle occurs in one subcorpus)

A comparison between the lists shows that many bundles frequently used in Econ-PUB are missing altogether from Econ-Wiki. This finding highlights the fact that Wikipedia articles differ dramatically from published RAs in terms of their purpose: many of these bundles serve important functions in research writing, as illustrated below in examples (3)–(7); in particular, they make connections between different units of discourse (*the extent to which, in the context of, in terms of the*), and make explicit the writer's epistemic stance or judgement of necessity (*are more likely to, it is important to*).

(3) Several recent studies question **the extent to which** market integration alone can explain the cross-listing effects. (Econ-PUB)

(4) Although these studies were **in the context of** foreign investment, it is plausible that similar arguments extend to exporters. (Econ-PUB)

(5) **In terms of the** structure of our model, we borrow some tools from the sizeable literature on search-theoretic approaches to the analysis of labor markets. (Econ-PUB)

(6) At the macro level, short-run fluctuations in disposable income **are more likely to** be dominated by the variance of temporary shocks that would be averaged out in the long run. (Econ-PUB)

(7) …but **it is important to** learn what types of investments are more effective.) (Econ-PUB)

These, and other similar bundles are clearly useful in original research reports, whose rhetorical task is to establish the value of the new research findings as scientific facts. In this process, the writer needs to signal their view on the status of the information being presented, by clearly marking speculation as such and expressing their reservations where necessary.

It is therefore not surprising that these bundles do not often find their way into Wikipedia articles, whose perspective on new discoveries is radically different: they are out of place in an online encyclopaedia and should be announced on some other forum. This perspective is articulated in the *No original research* policy, which is one of Wikipedia's core content policies. According to the policy, articles should only present material that is attributable to an existing published source. The content policy further cautions against improper syntheses as follows:

> If one reliable source says A, and another reliable source says B, do not join A and B together to imply a conclusion C that is not mentioned by either of the sources. This would be improper editorial *synthesis* of published material to imply a new conclusion, which is **original research** performed by an editor here. ("No original research")

Accordingly, the purpose of Wikipedia articles is to present factual information, on which there is a high degree of consensus, and steer clear of controversy and speculation. The need to negotiate the status of claims is therefore less acute, which in turn is clearly reflected in the relative absence of text-oriented and stance-oriented bundles. In this sense, too, the discourse of Wikipedia articles is different from research genres, which limits their usefulness in higher education settings.

The situation is similar for 3-word bundles: Econ-PUB contains nearly 1,300 three-word bundles which are not used in Econ-Wiki (data not shown). Many of these are subsumed under the four-word bundles introduced above (e.g. *more likely to*), but they also include other participant-oriented bundles (e.g. *is consistent with*, *the presence of*, and *the degree of*).

5.3.2 Medicine

Medicine offers a similar picture to economics. As shown in Table 8, 4-word bundles are infrequent in Med-Wiki compared to Med-PUB, and the few bundles that do occur in this subcorpus are common across the board – *as a result of, as well as the, the end of the*, and *on the other hand*.

	Med-Wiki		**Med-BAWE**		**Med-Pub**	
	Bundle	**Freq**	**Bundle**	**Fre q**	**Bundle**	**Freq**
1	as a result of	72	as a result of	65	as a result of	8
2	as well as the	79	it is important to	56	at the time of	85
3	the end of the	72	at the time of	41	as well as the	18
4	in the form of	59	in the case of	17	in the case of	16
5	on the other hand	50	has been shown to	14	has been shown to	15
6	at the university of	74	in the form of	12	the end of the	14
7	one of the most	68	in the treatment of	12	it is difficult to	13
8	in the united states	325	it is difficult to	9	on the other hand	12
9	in the united kingdom	81	of the patient’s problem	155	at the university of	9
10	is one of the	74	the patient’s problem s	155	in the treatment of	9
11			the nature of the	60	it is important to	8
12			analysis of history and	55	one of the most	7
13			formulation of the patient’s	55	on the basis of	37
14			the cause of the	53	in the current study	33
15			evidence based care and	52	in the control group	29
16			based care and issues	51	of the patients in	28
17			care and issues for	51	in the presence of	25
18			nature of the problem	51	the patients in the	22
19			a summary of key	50	the time of the	21
20			about the presenting illness	50	in the present study	20
21			all relevant information gathered	50	of this study was	20
22			and a summary of	50	this study was to	20
23			and the social and	50	for the treatment of	18
24			and their expectations for	50	are summarized in table	17

Table 8. Top 4-word bundles in three medicine subcorpora (black shading = bundle occurs in all three subcorpora, light grey shading = bundle occurs in two subcorpora, no shading = bundle occurs in one subcorpus).

However, even though published medical articles follow closely the IMRD structure with little rhetorical flourish, the analysis of lexical bundles suggests that Wikipedia articles are rather poor models for them. This can be seen in Table 8, which shows that many commonly used bundles in medical research writing are missing from Med-Wiki. These include various bundles associated with reporting details of the empirical research: *at the time of* (example 8; by far the most commonly occurring bundle in Med-Pub), *in the control group*, *of the patients in,* and *at the site of*. Many of these bundles are frequent in Med-BAWE, which indicates that student writing shares similar rhetorical concerns with published RAs. We also find a few standard text-organising formulas and location elements (*in the present study*, *of this study was*, *are summarised in table*) – but few signals of stance and engagement (cf. Hyland 2005). The ones that do occur are only moderately frequent; for instance, the bundle *it is important to* occurs only eight times (example 9).

(8) **At the time of** the latest follow-up, forty-nine patients had no pain or slight, intermittent pain. (Med-RA)

(9) **It is important to** note, however, that the magnitude of HDAC inhibitor-mediated induction of apoptosis is dependent on the drug concentrations […] (Med-RA)

The analysis of lexical bundles also highlights another genre-related difference. As previously noted, student essays in Medicine are particularly rich in 3- and 4-word bundles, a finding that initially appears unexpected, especially as many of these bundles are specific to Med-BAWE. However, this finding is explained by the fact that most texts in Med-BAWE (66/80) are case studies, which is a different genre from the research report. The purpose of case studies is “to gain an understanding of professional practice through the analysis of a single exemplar” (*The BAWE Corpus Manual*, p. 46), and they are written following a template, which defines standardised section headings to be used, and this in part accounts for the large number of bundles found. For instance, the bundles *of the patient’s problem* and *evidence based care and* clearly come from the assigned section heading, (in boldface in example (10))

(10) **Evidence based care and issues for research. A brief consideration of the evidence base required for the diagnosis and management of the patient’s problem(s).** As the single most likely explanation for the presenting symptoms in this case, I have chosen to present evidence for the management of pulmonary embolism and also generalised thromboembolic disease. (Med-BAWE)

This heading is reproduced verbatim in nearly all case studies in this subcorpus.

5.3.3. Literary criticism

As an academic discipline, literary criticism is divergent, characterised by different paradigms of enquiry (Sosnoski 1994), and RAs in this discipline are typically different from scientific RAs in that they allow writers to organise their texts more freely. Against this background, it is interesting that compared to the other disciplines, literary critical and RAs and Wikipedia articles on literary topics exhibit smaller differences, as far as the use of lexical bundles is concerned. Indicative of this is the fact that up to seven of the bundles are shared between the three subcorpora, as shown in Table 9. At the same time, the bundles used in LC-PUB are more varied and less tied to standard rhetorical moves in research articles, which makes it more difficult to determine what may be missing from Wikipedia articles, and why.

	LC-WIKI		**LC-BAWE**		**LC-PUB**	
	Bundle	**Freq**	**Bundle**	**Freq**	**Bundle**	**Freq**
1	the end of the	61	at the end of	49	the end of the	70
2	at the same time	40	the end of the	47	at the end of	68
3	on the other hand	39	on the other hand	35	at the same time	66
4	at the end of	38	the beginning of the	25	on the other hand	49
5	the beginning of the	29	at the same time	18	in the form of	29
6	in the form of	22	in the form of	14	the beginning of the	16
7	one of the most	50	the way in which	56	in the united states	59
8	in the united states	38	at the beginning of	29	in the case of	35
9	is one of the	24	the fact that the	23	the fact that the	26
10	in the history of	23	the rest of the	22	as well as the	24
11	as well as the	21	the image of the	16	in the context of	22
12	as a result of	20	to the fact that	16	as a result of	21
13	for the first time	17	that there is no	15	the way in which	20
14	on the basis of	15	through the use of	22	is one of the	19
15	in the case of	14	it could be argued	21	one of the most	18
16	the end of his	14	could be argued that	19	at the beginning of	17
17	in the context of	12	way in which the	17	for the first time	14
18	at the university of	70	the repetition of the	15	that there is no	14
19	at the age of	49	allows the reader to	14	on the basis of	12
20	was one of the	43	by the use of	14	the image of the	12
21	a member of the	33	the importance of the	14	the rest of the	12
22	one of the first	31	the role of the	12	in the history of	11
23	as one of the	27			the end of his	11
24	one of the greatest	17			to the fact that	11
25	the history of the	16			of the united states	36

Table 9. Top 4-word bundles in three literary studies subcorpora. (black shading = bundle occurs in all three subcorpora, light grey shading = bundle occurs in two subcorpora, no shading = bundle occurs in one subcorpus).

A number of relatively frequent bundles are specific to LC-Pub, including framing signals like *in the face of* (11) and *to the extent that* (cf. example 3 above), and *the ways in which* (12).[20] Another such bundle is *as a kind of* (13), which emerges from the use of the so-called *as*-predicative construction, which is a useful rhetorical resource for presenting and reporting interpretive claims in literary critical writing (Hiltunen 2010).

(11) Molina's retellings of the Hollywood movies speak of the power of imagination and narrative, even – or, perhaps, especially – **in the face of** the violence and repression of dictatorship. (LC-Pub)

(12) These critics, however, neglect **the ways in which** the Gregorian reform powerfully informed the writing of the eclogues. (LC-Pub)

(13) Rather than extolling its benefits, Austen here characterizes fashionable travel **as a kind of** willful errancy in which even those who profess to be traveling on business stray from established routes, risk bodily injury, and neglect domestic responsibilities. (LC-Pub)

In addition, the bundle *the fact that the* is used both in LC-Pub and Lit-BAWE, but not in LC-Wiki. This structure is useful for discussing sophisticated ideas, because, as Schmid (2000: 361–2) observes, shell nouns like *fact* enable the writer to encapsulate complex abstract relations into simple concepts and present them in a way that helps the reader make sense of them.

[20] LC-BAWE has 56 instances of the related bundle *the way in which*, but no instance of the plural variant.

(14) However, **the fact that the** German text is a translation of a translation confounds our attempts at direct symbolic assignations. (LC-Pub)

The occurrence of these bundles can thus be linked to the evaluative function of literary critical writing. Their absence in Lit-Wiki in turn points to different rhetorical concerns of the genres, and lends support to the idea that Wikipedia articles do not expose readers to the full range of discourse structures and argumentative strategies of literary critical research writing. As previously noted, Lit-Wiki is rich in bundles referring to various biographical details of authors and critics, illustrated below in examples (15)–(17).

(15) Jonathan Swift **was one of the** greatest of Anglo-Irish satirists, and one of the first to practise modern journalistic satire. (LC-Wiki)

(16) Tzara had enrolled **at the University of** Bucharest in 1914, studying Mathematics and Philosophy, but did not graduate. (LC-Wiki)

(17) Foucault was **a member of the** French Communist Party from 1950 to 1953. (LC-Wiki)

The high frequency of these bundles underlines the status of Wikipedia as a reference work, which aims to present factual information accurately and concisely. The kind of language associated with this communicative purpose is very different from literary critical discourse, which is expected to present well-argued evaluations and coherent interpretations of literary texts.

6. Conclusions

Taken together, the analyses presented here demonstrate that the occurrence of lexical bundles varies across both genres and disciplines. Such variation is not unexpected, given that earlier studies have shown bundles to vary depending on many contextual variables, including DISCIPLINE (cf. Hyland 2008a). It is, however, interesting that the GENRE variable seems to have more explanatory power: comparisons across subcorpora indicate that Wikipedia articles typically make limited use of a few core bundles (e.g. *on the other hand*), and contain hardly any instances of many bundles that previous studies have flagged as prominent resources for writers of academic papers.

What precisely is missing from Wikipedia articles is highlighted by the cross-genre comparison of lexical bundles within disciplines (section 5.2). The most important absences identified in this study are references to the research process, framing signals, and explicit indications of writer's stance. Instead, Wikipedia articles contain bundles that are used for stating indisputable facts as opposed to interpretations. This finding underlines the principal contrast between Wikipedia articles and RAs: they are different genres, with different communicative purposes and associated rhetorical strategies. The former are encyclopaedia texts, whose aim is to provide a comprehensive summary of a topic, while the latter typically present original research results. In addition, while the quality of RAs depends not only on the results themselves, but also on the writer's ability to establish a competent disciplinary identity, Wikipedia articles also require less interaction with readers. In this sense, they are similar to learner texts, which tend to have a low frequency of participant-oriented bundles (cf. Salazar 2014: 180)

These genre-based differences also have pedagogical implications. Given the ubiquity of Wikipedia, it is important to acknowledge the fact that linguistically it does not represent the full range of academic discourse and lacks many features of professional research writing. Recent research in academic writing has shown that to craft persuasive arguments, the writer's appropriate positioning is, and lexical bundles indicating stance and engagement are resources that commonly contribute to this goal crucial (e.g. Hyland 2005, 2008). As shown in this paper, these are comparatively rare in Wikipedia articles across the board. Wikipedia, together with compilations, anthologies, handbooks, traditional encyclopaedias, is accordingly best used as a 'first port of call' in a research project, and not as primary source of secondary literature to be cited in essays or papers (see e.g. Waters 2007) – which is a point that Jimmy Wales, the founder of Wikipedia, has also made (Young 2006). The findings of this study suggest that while obtaining reliable factual information from Wikipedia may be possible, Wikipedia articles do not expose readers to the variety of argumentation patterns writing and styles which characterise professional academic writing. Against this background, EAP content courses would clearly benefit from a discussion of Wikipedia's communicative purpose vis-à-vis traditional academic genres, which could involve practical activities, such as comparing Wikipedia texts to research genres with the help of concordances. Such activities could focus on lexical bundles, as they are easy to retrieve from a corpus and, as shown in this chapter, are often linked to specific rhetorical functions. These activities could promote students' rhetorical consciousness and enhance their understanding of the process of constructing academic knowledge.

Acknowledgements

Part of the data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the

Centre for Applied Linguistics, Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (School of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

References

Ädel, Annelie & Erman, Britt. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31 (2): 81–92.

Brook O'Donnell, Matthew, Römer, Ute & Ellis, Nick C. 2013. The Development of Formulaic Sequences in First and Second Language Acquisition: Investigating effects of frequency, association and native norm. *International Journal of Corpus Linguistics* 18(1): 83–108.

Baker, Paul & Chen, Yu-Hua. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2): 30–49.

Baroni, Marco & Bernardini, Silvia. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, 1313–16. <http://clic.cimec.unitn.it/marco/publications/lrec2004/bootcat_lrec_2004.pdf>

Bazerman, Charles. 1981. "What Written Knowledge Does: Three Examples of Academic Discourse". *Philosophy of the Social Sciences* 11(3), 361–387.

Becher, Tony and Paul Trowler. 2001. *Academic tribes and territories: intellectual enquiry and the culture of disciplines*. Buckingham: Society for Research into Higher Education & Open University Press.

Biber, Douglas. 2006. *University language: a corpus-based study of spoken and written registers*. Amsterdam: Benjamins.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3): 275-311.

Biber, Douglas & Barbieri, Federica. 2007. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes* 26: 263–286.

Biber, Douglas, Conrad, Susan & Cortes, Viviana. 2004. *If you look at…*: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25(3): 371–405.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman grammar of spoken and written English*. London: Longman.

Cortes, Viviana. 2004. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific* Purposes 23: 397–423.

Davies, Mark. 2015. *The Wikipedia Corpus.* <http://corpus.byu.edu/wiki/>

Durrant, Philip. 2013. Discipline and Level Specificity in University Students' Written Vocabulary. *Applied Linguistics* 35(3): 328–256.

Durrant, Philip. 2015. Lexical bundles and disciplinary variation in university students' writing: mapping the territories. *Applied Linguistics*: 1–30

Eijkman, Henk. 2010. Academics and Wikipedia: reframing Web 2.0+ as a disruptor of traditional academic power-knowledge arrangements. *Campus-Wide Information Systems* 27(3): 173–185.

Fløttum, Kjersti, Dahl, Trine & Kinn, Torodd. 2006. *Academic voices: across languages and disciplines*. Amsterdam: John Benjamins Publishing Company.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In *Phraseology: Theory, Analysis, and Applications*, A. P. Cowie (ed), 145–160. Oxford: Clarendon Press.

Groom, Nicholas. 2009. Phraseology and epistemology in academic book reviews: a corpus-driven analysis of two humanities disciplines. In *Academic Evaluation. Review Genres in University Settings*. Ken Hyland and Giuliana Diani. London: Palgrave Macmillan, 122–139.

Hiltunen, Turo. 2010. Grammar and disciplinary culture: a corpus-based study. PhD dissertation, University of Helsinki. <http://urn.fi/URN:ISBN:978-952-10-6464-7>.

Hiltunen, Turo & Mäkinen, Martti. 2014. Formulaic language in L2 academic writing for business studies and economics. In *Corpus Analysis for Descriptive and Pedagogic Purposes: English Specialised Discourse* [Linguistic Insights 200], Maurizio Gotti & Davide Giannoni (eds), 347–360. Berlin: Peter Lang.

Hyland, Ken. 2000. *Disciplinary discourses: social interactions in academic writing*. Harlow: Pearson Education.

Hyland, Ken. 2005. Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7(2): 173–192.

Hyland, Ken. 2008a. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1): 4–21.

Hyland, Ken. 2008b. Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1): 41–62.

Jemelniak, Dariusz. 2014. Wikipedia, a Professor's best Friend. *Chronicle of Higher Education*, 13 Oct 2014.

Kilgarriff, Adam, Reddy, Siva, Pomikálek, Jan & PVS, Avinesh. 2010. A Corpus Factory for Many Languages. In *Proceedings of the Seventh International Conference of Language Resources and Evaluation (LREC'10)*, 904–910.

Kuteeva, Maria. 2011. Wikis and academic writing: Changing the writer–reader relationship. *English for Specific Purposes* 30(1): 44–57.

Kuteeva, Maria. 2016. Research blogs, Wikis and Tweets. In *The Routledge Handbook of English for Academic Purposes*, Phillip Shaw & Ken Hyland (eds.), 431–443.

Mäkinen, Martti & Hiltunen, Turo. 2016. Creating a corpus of student writing in economics: structure and representativeness. In *Corpus linguistics on the move: Exploring and understanding English through corpora*. [Language and computers: Studies in Practical Linguistics], Maria José López Couso, Bélen Méndez Naya, Paloma Núñez Pertejo & Ignacio Palacios Martínez (eds), 41–58. Amsterdam: Rodopi.

Miller, Julia. 2012. Building academic literacy and research skills by contributing to Wikipedia: A case study at an Australian university. *Journal of Academic Language and Learning* 8(2): A72–A86

Myers, Greg. 2010. *Discourse of blogs and wikis*. New York: Continuum.

Nesi, Hilary. 2008. BAWE: an introduction to a new resource. In *Proceedings of the 8th Teaching and Language Corpora Conference*, Ana Frankenberg-Garcia, Tawfig Rkibi, Maria do Rosario Braga da Cruz, Ricardo. Carvalho, Cristina Direito & Diogo Santos-Rosa, 239–246. Lisbon: Instituto Superior de Línguas e Administração.

“No original research”. *Wikipedia*, accessed 15 April 2016, <https://en.wikipedia.org/wiki/Wikipedia:No_original_research#Synthesis_of_published_material>

Paquot, Magali & Sylviane Granger. 2012. Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics* 32: 130-149.

Pérez-Llantada, Carmen. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes* 14: 84–94.

Salazar, Danica. 2014. *Lexical bundles in Native and Nonnative Writing. Applying a Corpus-based study to Language Teaching* [Studies in Corpus Linguistics]. Amsterdam: Benjamins.

Schmid, H.–J. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: Mouton de Gruyter.

Shaoul, Cyrus & Chris Westbury. 2010. *The Westbury Lab Wikipedia Corpus (2010)*. Edmonton, AB: University of Alberta. <http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html>

Simpson-Vlach, Rita & Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4), 487-512.

Sosnoski, James J. 1994. *Token professionals and master critics: a critique of orthodoxy in literary studies*. Albany, NY: State University of New York Press.

Swales, John M. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tribble, Christopher. 2011. Revisiting Apprentice Texts: Using Lexical Bundles to Investigate Expert and Apprentice Performances in Academic Writing. In *A Taste for Corpora: In Honour of Sylviane Granger*, Fanny Meunier, Gaëtanelle Gilquin & Magali Paquot, 85–108. Amsterdam: Benjamins.

Waters, Neil L. 2007. Why You Can't Cite Wikipedia in My Class. *Communications of the ACM* 50(9): 15–17.

Young, Jeffrey R. 2006. “Wikipedia Founder Discourages Academic Use of His Creation.” *The Chronicle of Higher Education,* June 12, 2006. <http://chronicle.com/blogs/wiredcampus/wikipedia-founder-discourages-academic-use-of-his-creation/2305.>