Journal of English for Academic Purposes 54 (2021) 101054

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Profiling figure legends in scientific research articles: A corpus-driven approach

Zhongquan Du, Feng Jiang (Kevin)*, Luda Liu

School of Foreign Language Education, Jilin University, China

ARTICLE INFO

Keywords:
Figure legends
Keywords
Lexical bundles
Corpus-driven approach

ABSTRACT

Figure legends, a descriptive statement accompanying a graph, is an essential component of science writing, and novice writers often find it difficult to write an adequate and plausible one. However, almost no discourse analysis has been done of this short yet important part of research articles. Taking a corpus-driven approach, this study explores the use of keywords and lexical bundles in figure legends in order to understand how the discourse is constructed by the salient recurrent single and multiple lexical items. Results show that in figure legends writers are most concerned with the textual descriptions of *blue* and *red* as markers of contrasting scientific items and *bars* and *images* as representative of graphic shapes. Additionally, writers also frequently use *shown* and *indicated* to introduce the demonstrative relationship between graphic elements and scientific information. Further, noun/prepositional bundles account for the overwhelming proportion, followed by verb-related and then clause-related forms, while research-oriented functions of bundles are more frequently used than text and participant-oriented functions. The results offer useful pedagogical input and help novice and L2 writers come to grips with this part-genre.

1. Introduction

Effective academic communication is not only based on well-constructed and persuasive arguments but also depends on a sufficient and plausible presentation of data. Almost a half of scientific data are represented visually (Hyland, 2006, p. 53), so visual representations are not mere add-ons or ways to popularise a complex reasoning, but are an essential part of academic discourse (Moghaddasi et al., 2019; Pauwels, 2006). As Luna (2013) argues, figures in particular serve as the “spine of scientific story” and also “the backbone of a manuscript” (p. 62). Furthermore, many scientists may read only the abstract, figures and conclusions of a paper, expecting figures to be well labelled and well documented (Armer & Day, 2011; Cargill & O’Connor, 2013). Therefore, writing figure legends (i.e. *figure captions*), a descriptive statement accompanying figures, is an important skill in academic communication, since a small change in figure legends “can improve the communication of the main message” (Cargill & O’Connor, 2013, p. 27). However, this short yet important part-genre of research articles has almost escaped scrutiny in the EAP literature.

This paucity of related research is crucial in practice. Kroen (2004) reports that figure legends present enormous difficulty for students, attributing it not only to students’ unfamiliarity with this piece of writing but the unavailability of legend models and the linguistic resources typically used. Even submissions to *Nature* journals are also reported to contain “convoluted figure legends” (Nature research, 2019, p. 104). As EAP teachers, in addition, we also find students grapple with this aspect of research science writing,

* Corresponding author.
E-mail address: kevinjiang@jlu.edu.cn (F. Jiang).

https://doi.org/10.1016/j.jeap.2021.101054
Received 8 January 2021; Received in revised form 10 September 2021; Accepted 10 September 2021
Available online 21 September 2021

and as academic discourse analysts, we aim to explore figure legends with the methods that "make sense in environments where we also have to teach, develop pedagogy, produce course materials and the like" (Swales, 2001, p. 45).

Following Grabowski (2015), we believe that a corpus-driven description of the use and functions of key vocabulary complemented by a similar description of lexical bundles is particularly useful for the teaching and research purposes. The information given by these recurrent single-word and multi-word units and their corpus operation can be easily accessed by students (Jiang, 2019). By examining keywords and lexical bundles used in figure legends of research articles in physical and life sciences, we hope to extend our knowledge about this particular part-genre and assist students in this aspect of research writing.

2. Figure legends and visual presentation of scientific knowledge

The issue of representation, both textual and visual, touches on the lifeblood of scientific activities, and what is communicated as science is the result of a series of representational practices (Graves, 2014; Pauwels, 2006). Visual representations, in particular, are considered as part of academic persuasion, buttressing arguments and signalling the importance of research articles (Hyland, 2006). In biology, for example, Miller (1998) shows that visuals in research articles both "prove" and "clarify" academic arguments by bolstering up knowledge claims and condensing new information for the "informed and potentially skeptical reader" (p. 43). Similarly, focusing on nanotechnology as an interdisciplinary field, Graves (2014) examined the relationship between visuals and texts, and demonstrates that images and data displays contribute to the argument in experimental work to create knowledge in science. Therefore, visual figures are accorded special prominence in scientific texts, not simply accompanying academic texts but actively constructing meanings (Lemke, 1998; Moghaddasi et al., 2019). This should align with the fragmented accretion of knowledge in sciences which is assembled in relatively small pieces (Becher & Trowler, 2001; Hyland, 2004).

However, figures are difficult to comprehend without an associated text, commonly referred to as figure legend or figure caption. Fig. 1 illustrates a typical figure and the figure legend.

Clearly, a figure legend is essential to ensure the intelligibility of the figure without reference to the whole paper, explaining what is shown in the figure and offering readers all pertinent facts to interpret it (Aliotta, 2019; Skern, 2011). Most of what we know about figure legends comes from journal guidelines to authors, like the following from *Nature Materials*:

Figure legends begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used, focusing on describing what is shown in the figure and de-emphasizing methodological details. The meaning of all error bars and how they were calculated should be described. Each legend should total no more than 250 words[2].

Additionally, some writing handbooks offer stylistic advice on font design and referential content of legends. For example, Cargill and O'Connor (2013) suggest that in science writing figure legends have a general form with five parts which usually occur in sequence (p. 31):

(1) A title which summarizes what the figure is about;
(2) Details of results or models shown in the figure or supplementary to the figure;
(3) Additional explanation of the components of the figure, methods used, or essential details of the figure's contribution to the results story;
(4) Description of the units or statistical notation included;
(5) Explanation of any other symbols or notation used.

While the above sequence describes legends as a meaningful narrative, their linguistic features are rarely clarified, and little is known about the salient lexical items and strings science writers recurrently use to build up these descriptive components. The aspects of linguistic information are crucial, especially to L2 and student writers given the case studies of how students struggle with this aspect of research writing (Carolina & Carlino, 2017; Tardy, 2009; Zimbardi et al., 2013). Skern (2011) even concludes that almost all scientists, including experienced ones, find it difficult to compose this aspect of manuscript writing (p. 104).

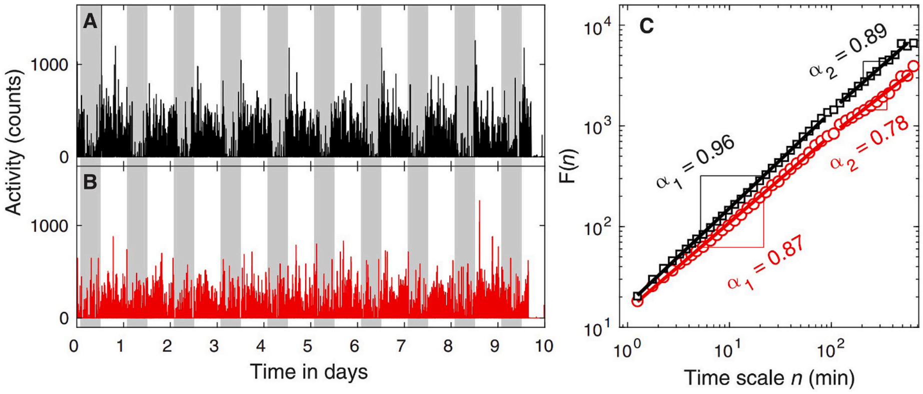

Fig. 1. A screenshot example of figure and the legend.[1].

Therefore, taking a corpus-driven approach, this study aims to fill the gap by an empirical corpus analysis of frequency distributions of keywords and lexical bundles that are typically used in figure legends in published research articles.

3. Keywords and lexical bundles in academic discourse

Computer-aided corpus linguistics has made it possible to statistically identify recurrent single or multiple word combinations in discourse. Different from a corpus-based approach, which works within a specific theoretical framework of linguistic description, the corpus-driven analysis assumes that "only the existence of words, co-occurrence patterns among words, discovered from the corpus analysis, are the basis for subsequent linguistic descriptions" (Biber, 2009, p. 276). Keywords and lexical bundles, which are typically identified by this corpus-driven approach as the recurrent vocabulary and prefabricated strings of word forms, are notably illustrative of a specialized text (Cortes, 2004, 2013; Grabowski, 2015; Hyland & Jiang, 2020), especially when a priori linguistic knowledge about the discourse is not available.

Keywords are those words whose frequencies are unusually high "in a corpus of specialized texts against the frequencies of the same wordforms in a larger and more general reference corpus" (Groom, 2010, p. 60). They help to not only reveal the 'aboutness' (Scott & Tribble, 2006) of texts but specify the salient lexical items which are functionally related to the discourse (Gilmore & Millar, 2018; Grabowski, 2015; Hyland & Jiang, 2020). While suggesting the topics of discourse, keywords are single words. A more informed interpretation of rhetorical resources and functional purposes entails an extension of single lexical items to multi-word combinations.

Lexical bundles, also referred to as clusters and chunks, are one of the most noticeable multi-word combinations of academic discourse (Cortes, 2004; Hyland, 2008; Chen & Baker, 2010). Simply, they are statistically the most frequent "recurring sequences" of words in a collection of texts: extended collocations which appear more repeatedly than expected by chance across a given range of texts (Biber et al., 1999, p. 990). Lexical bundles are routinely used to assist pragmatically efficient communication, and in academic discourse help to reduce readers' processing time because such familiar patterns structure a discourse, guiding readers through it (*in this figure*, *shown in fig*) or linking ideas (*with respect to*, *are based on*) (Biber et al., 2004; Cortes, 2013; Hyland, 2004). In addition, by signaling appropriate use of a disciplinary code, they allow writers to display solidarity with colleagues (Hyland & Jiang, 2018) and to construct a disciplinary competent voice (Biber, 2009; Hyland & Jiang, 2019). But while we know much about lexical bundles in main body texts of research articles, almost nothing has been written about their role in the composition of figure legends.

From the above, therefore, this study sets out to explore what figure legends are mainly concerned with and how lexical bundles are typically used in their presentation. We first describe our corpus and methods, before discussing the most prominent single and multiple lexical units which characterise figure legends. In the end, implications and limitations are raised about the study.

4. Corpus and methodology

4.1. Corpus description

Considering that figure legends feature research writing in hard sciences (Pauwels, 2006), we focus on four disciplines, *biology* and *medicine* as life sciences and *physics* and *engineering* as natural sciences. We aimed at five journals in each discipline which had achieved the top ranking in their field according to the 5-year impact factor published by Thomson Reuters's *Web of Knowledge ISI* in 2019 (Appendix 1). Their representativeness was also confirmed by consulting professors in each discipline. After having scrutinised sampled articles, we finally included ten papers in each of the 20 journals in 2019. Therefore, the corpus we compiled, comprising 200 research articles, includes 1190 figure legends which total 117,014 words. Table 1 summarizes the corpus information.

4.2. Keyword calculation

The keywords were calculated by *AntConc* (Anthony, 2019), which statistically compared the word frequencies in our target corpus against a 16 million-word reference corpus that comprises academic genres of the British National Corpus (BNC). This comparison gives a better characterisation of the differences between two corpora than simple frequency counts as it identifies items which are most prominent and not just common in figure legends.

4.3. Bundle identification

Given the length of figure legends, we decided on three-word bundles because they are more common than four and five-word strings and offer a clearer range of structures and functions than two-word bundles. While "somewhat arbitrary" (Biber & Barbieri,

Table 1
Corpus characteristics.

Discipline	Number of articles	Number of figure legends	Total number of figure legend words
Biology	50	274	41,574
Medicine	50	186	17,588
Physics	50	287	27,308
Engineering	50	443	30,544

2007, p. 267), different frequency and distribution criteria have been used in the literature regarding how to identify lexical bundles. The cut-off frequency, which determines the number of bundles to be included in the list, has ranged from 20 (Cortes, 2004; Hyland, 2008) to 40 times per million words (Biber et al., 2004) and even raw frequencies (Chen & Baker, 2010). A second criterion is that combinations have to occur in different texts, usually in at least three to five texts (e.g. Biber & Barbieri, 2007) or 10% of texts (Hyland & Jiang, 2018) to avoid idiosyncrasies of individual writers.

After repeated trials with the corpus data, we set the frequency threshold to five times, which is shown to work agreeably with corpora smaller than one million words (see Cortes, 2012), and included those bundles occurring in at least five texts. This resulted in an "optimum" number of bundles, which can sufficiently represent the corpus under investigation. After an automatic search for three-word bundles by *AntConc*, we manually excluded bundles with text-dependent noun phrases (e.g. *two independent experiments*), but retained overlapping word sequences where two three-word bundles are perhaps part of a four-word string (e.g. *are based on* and *based on the*).

The results were then transferred into an Excel file where we coded each example for its grammatical structure and function (see Table 3 and Table 5 respectively). The three authors worked independently, and achieved an inter-rater reliability of 99% on structure and 97% on function before resolving disagreements. Because of the highly compact nature of scientific discourse that values concision, a single bundle may have multiple functions even in a single occurrence. For example, bundles such as *are used to* can function to refer to an analytical procedure (*successive compression and expansion (C/E) cycles are used to achieve refrigeration in a particular subsystem*) or guide readers to the organisation of the discourse (*coloured asterisks are used to uniquely identify the cells*). We examined the context to determine a primary function so that each of 18 potentially multi-functional bundles was coded to one specific function.

5. Keywords in figure legends: what are they about?

The basic idea of keyness is that a word form or cluster of words which are statistically more frequent in a corpus than might be expected by chance are *key* to it: they are what the texts are 'about' or "what it boils down to … once we have steamed off the verbiage, the adornment, the blah blah blah" (Scott & Tribble, 2006, p. 56). Keywords can also "specify the salient features which are functionally related to the genre" (Grabowski, 2015, p. 24). Our comparison of the legend corpus with the BNC academic genres found 385 keywords. Table 2 shows the 20 with the highest keyness value.

Among the 20 keywords, there are four relating to colours (*blue*, *red*, *gray* and *green*) and six describing graphic shapes (*bars*, *images*, *dashed*, *curves*, *lines* and *plots*). While other colours and images fell out of the top 20, this result shows that legends were mainly textual presentations of figures, making intelligible what graphic illustration means as in (1) and (2).

(1) each **red** arrow represents one paraoxon exposure. (Medicine)
(2) In all figures, the **dashed curves** represent the modal excitation distribution at the input … (Physics)

It is interesting, however, to see that *blue* and *red* had the highest keyness with a relatively high frequency. As seen in (3) and (4), the two contrasting colours were used to foreground the most prominent features that the writers have found in their research, and easily attracted the audience's attention to the prominence. The high keyness and frequency indicate that these striking distinctions are the most essential information writers seek to clarify in figure legends.

Table 2
The 20 keywords with the highest keyness value in the corpus.

keyword	frequency	keyness		effect size
blue	266	+	1631.49	0.0045
red	300	+	1555.39	0.0051
bar	160	+	1031.08	0.0027
image	205	+	1010.02	0.0035
dashed	109	+	948.98	0.0019
versus	161	+	938.41	0.0027
curve	137	+	843.18	0.0023
normalized	84	+	736.51	0.0014
line	198	+	650.25	0.0033
plot	98	+	598.40	0.0017
shown	273	+	569.46	0.0044
gray	86	+	565.12	0.0015
green	134	+	555.68	0.0023
data	334	+	538.71	0.0053
indicated	154	+	531.96	0.0026
value	219	+	531.60	0.0036
showing	139	+	523.43	0.0024
respectively	149	+	522.16	0.0025
indicate	154	+	507.14	0.0026
inset	80	+	492.53	0.0014

Table 3
Structural classification of three-word lexical bundles in academic writing.

verb phrase-related bundles	• passive verb (*are shown in*, *was used to*) • copular be (*are representative of*, *are available in*) • imperative (*note that the*)
clause-related bundles	• abstract subject (*error bars represent*, *the data are*) • *as*-fragments (*as in the, as well as*) • *conj*-fragments (*and in the*, *or absence of*) • *adj*-fragments (*relative to the*) • *wh*-fragments (*which shows a*)
noun-related bundles	• noun phrase with *of*-phrase fragment (*the number of*, *the presence of*) • noun phrase with other post-modifier fragment (the *relationship between*, *the references to*)
preposition-related bundles	• prepositional phrase expressions (*in terms of*, *with respect to*) • comparative expressions (*before and after*, with *and without*)

Table 4
Structural distribution of three-word lexical bundles (frequency & percentage).

structural categories	frequency	percentage
verb-related bundles	37	19.8
passive verb	34	18.2
copular be	2	1.1
imperative	1	0.5
clause-related bundles	**20**	**10.7**
abstract subject	14	7.5
as-fragments	2	1.1
conj-fragments	2	1.1
adj-fragments	1	0.5
wh-fragments	1	0.5
noun-related bundles	**83**	**44.4**
noun phrase with *of*-phrase fragment	72	38.5
noun phrase with other post-modifier fragment	11	5.9
preposition-related bundles	**45**	**24.0**
prepositional phrase expressions	38	20.3
comparative expressions	7	3.7

Table 5
Functional categories of three-word lexical bundles.

research-oriented bundles	• location – indicating time and place (*before and after*, *are available in*); • procedure (*was used to*, *the presentation of*); • quantification (*the number of the*, *one of the*); • description of tangible attributes (*the structure of*, *the size of*); • description of intangible attributes (*the stability of*, *the basis of*)
text-oriented bundles	• transition signals – establishing additive or contrastive links between elements (*as well as*, *compared with*); • resultative signals – mark inferential or causative relations between elements (*due to the*, *with respect to*); • structuring signals – text-reflexive markers which organise stretches of discourse or direct reader elsewhere in text (*shown in fig*, *shown in the*); • framing signals – situate arguments by specifying limiting conditions (*based on the*, *with respect to*).
participant-oriented bundles	• stance features – convey the writer's attitudes and evaluations (*bars indicate the*, *the probability of*); • engagement features – address readers directly (*note that the*, *reader is referred*).

(3) For clarity, 2-butene molecules are in **red** and isobutane molecules in **blue.** (Chemical engineering)
(4) **Red** indicates enhanced central accumulation relative to non-targeted signaling intermediate …, **blue** similarly indicates diminished central accumulation. (Biology)

Gray and *green* were less frequent options. *Gray* was typically used to portray the setting information against which more prominent features were observed (5) and (6). Conversely, *green* was not used as a background colour but often an addition to other commonly-used ones, especially red (7) and blue (8).

(5) Four shades of **gray** indicated level of IL-2 mRNA relative to non-targeted signaling intermediate … (Biology)
(6) SNPs in **gray** background are different from the reference genome (isolate P6497). (Biology)
(7) The red, **green** and blue balls represent the Nb, Si and Te atoms, respectively. (Physics)

(8) Red, **green** and blue lines correspond to uncoordinated PF6, coordinated PF6, FEC ring deformation, and total spectrum, respectively. (Physics)

Therefore, the colour keywords and what they represent in the above examples show us something about how legends portray and complement the visual presentations. Turning to graphic shapes, *bar* and *image* had the highest keyness to this short text of figure descriptions and thus were the two most commonly recounted in legends. As shown in (9) and (10), science writers sought to clarify either what bars refer to in figures or the relationship between elements marked by bars. Differently, images were often confined visually by a pre-modification, such as immunofluorescence (11) and TEM[3] (12) where the writers specified the experimental sources of the reported images.

(9) **Bars** to the right of the center indicate that, relative to contractile SMCs, modulated SMCs have shifted towards a given cell type. **Bars** to the left indicate that they have shifted away. (Medicine)
(10) Genes of the isolate IPO323 (light **bars**) classified as either accessory or singleton in the pangenome were analyzed for clustering in the genome (dark **bars**). (Biology)
(11) Representative fluorescence **images** of cells and quantified results are from three independent experiments. (Engineering)
(12) TEM **images** of five fields of view from two independent experiments are shown. (Physics)

Despite a lower keyness, *dashed*, *curve*, *line* and *plot* represent the shapes which were also frequently spelled out in legends. Among 109 occurrences of *dashed*, 79 cases collocated with *lines* in the explanation of what the lines differing from regular solid ones referred to (13). Curves were usually described in terms of the way they were obtained (14) rather than simply what they symbolised, whereas what lines represented were always introduced in legends (15). Plots were often mentioned when writers sought to specify plot shapes and their composition (16).

(13) **Dashed** lines indicate the theoretical predictions for the thermalized energy in each of the two species. (Physics)
(14) The black **curves** were obtained using the mean and variance of the data shown in Fig. 2. (Biology)
(15) Horizontal **lines** indicate corresponding 95% confidence intervals around hazard ratios. (Medicine)
(16) Box **plots** depict the IQR, with a white horizontal line representing the median. (Medicine)

In addition, *versus* and *normalized* are another two keywords worthy of comment. *Versus* establishes a comparison between variables or parameters. As shown in (17) and (18) science writers were concerned with such a comparison and attempted to make clear how it was marked in figures. *Normalized* typically expresses the adjusted values which are intended for a comparison with another set of values. This calculation was often reported in legends to help readers make sense of experimental data (19) and (20).

(17) Cell subsets are scattered by the difference between the Pearson correlation values of their individual-level slopes calculated either **versus** individuals' baseline positions in the PCA space or **versus** age (x-axis). (Medicine)
(18) Microscopy was subdivided into gametocyte **versus** asexual trophozoite positivity/negativity. (Medicine)
(19) Data are **normalized** to the pS19-MLC cell body ($n = 46$ cells from three independent experiments and $P < 0.0001$ as determined by a two-tailed Wilcoxon matched-pairs signed rank test). (Physics)
(20) Weight of the liver and spleen are **normalized** to total bodyweight of wildtype and Mfsd1 KO mice (age: 14 weeks).(Biology)

Furthermore, *shown*, *indicated*, *showing* and *indicate* were the four verb-related words that were most *key* to figure legends. It is interesting to note, however, that the present and past participle of *show* and *indicate* had a higher keyness than the basic forms. The use of the past participle indicates that compared with other academic textual segments such as research abstracts (Hyland & Tse, 2005), legends are more inclined to take a passive voice in the reporting of graphic components (21) and (22). *Showing* was also particularly used as a way of assisting science writers to compact much information in such a short piece of text as legends (23) and (24).

(21) Fits are **shown** along the x direction (a), y direction (b) and z direction (c). (Physics)
(22) Expression levels are **indicated** by scales in the lower left of each panel. (Medicine)
(23) The dashed lines are guides for the eye **showing** the β and β' bands. (Physics)
(24) The histograms **showing** the posterior distributions of population mean and standard deviation hyperparameters are given in Fig. 1.(Biology)

Additionally, *data* and *value* were also prominently used in figure legends compared with the BNC academic genres. However, to a larger extent *data* was mentioned to describe the source and component of experimental data by an intertextual link to supplemental materials (25) and (26). *Value* often referred to probability or mean values in legends, and writers sought to explain the calculation of probability values (27) and the figured presentation of mean values (28).

(25) The source **data** and computer code with instructions of implementation to generate Fig. 1 are fully publicly available at https://doi.org/10.26188/5cde4c26c8201. (Biology)
(26) Underlying **data** are available in S1 Data. (Biology)

(27) P **values** were determined with Student's t-test and corrected for multiple testing by the Benjamini–Hochberg procedure. (Medicine)
(28) Median **values** of scaled frequencies measured in young individuals are in left bar. (Medicine)

Moreover, *respectively* also appeared frequently in figure legends, making clear the connection between graphic shapes and what they concerned (29) and (30). Perhaps particular to scientific papers, *inset* was often used to refer to a small diagram inside a larger one (31 and 32).

(29) The deconvoluted profiles for CoCo and the core complexes are described by the dotted line and the dashed line, **respectively**, for CuFeCoCo, CuCoCoCo and CoFeCoCo. (Engineering)
(30) Error bars in c and f are derived from Poissonian statistics and range from 0.013 to 0.046 and 0.02 to 0.063, **respectively.** (Physics)
(31) The **inset** shows the same data on an expanded y axis. (Medicine)
(32) **Inset** is a schematic of equivalent circuit of the EH for measuring the voltage across a 33 F capacitor. (Physics)

In summary, the keywords above reveal the "aboutness" of figure legends. The most commonly used colours suggest that data visualization strategies inform the legend content and the keywords. We have also seen the graphic shapes science writers are most concerned with in the textual representation of figures. Additionally, verbal and connective keywords indicate the informational and textual characteristics of this short text of graphic description.

6. Lexical bundles in figure legends: building blocks of textual presentations

We identified 2047 three-word lexical bundles in the corpus using our criteria, amounting to 10.2 instances per figure legend. This total number comprises 187 different types, with the most frequent being *a function of* closely followed by *as a function*. This analysis shows that despite their limited length, figure legends are composed of these repeated lexical strings, which function as "important building blocks in discourse" (Biber & Barbieri, 2007, p. 270) and also facilitate pragmatically efficient communication of what scientific figures indicate to readers. Additionally, it was also found that lexical bundles in the legends overwhelmingly included parts of prepositional or noun phrases and that they mainly related to referential content rather than to the discourse itself or the participants. We discuss the structures and functions of the three-word lexical bundles found in figure legends in more detail below.

6.1. Structures of lexical bundles in legends

Structural properties of lexical bundles are a differentiating feature of academic discourse (Biber et al., 1999). In academic writing, bundles are typically prepositional phrases with -*of* fragments (*in terms of*) and noun phrase + *of* fragments (*the number of*) (e.g., Hyland, 2008) or passive-verb fragments (*is shown in*) (e.g., Chen & Baker, 2010). According to Biber et al. (1999), these three forms comprise over 70% of three-word patterns in academic prose (pp. 994–995) but rarely figure in conversation, where the majority of bundles contain a verb phrase, particularly 'personal pronoun + verb phrase' (e.g., *I don't know*) (pp. 1006–1007).

Based on Biber et al.'s (1999) 12 categories of written academic bundles and Hyland and Jiang's (2018) analysis of a diachronic corpus of disciplinary writing, we identified four distinct categories of formal realisation as shown in Table 3.

The analysis shows that noun-related forms accounted for the largest proportion of bundles, followed by preposition-related and then verb-related bundles. Table 4 presents a detailed structural distribution of three-word bundles in the corpus. It is noteworthy that the proportion of noun-related bundles (44.4%) was remarkably higher than that in full research articles (27.0%) as reported in Hyland and Jiang's (2018) study, whereas preposition-related bundles registered a much lower percentage (24.0% vs 49.5%). The differences may reveal the informational density of this part-genre, since as seen in Table 1 each legend comprised 102.6 words on average. Hence, the information is mainly compacted by nominal strings.

Additionally, these differences perhaps relate to the particular communicative purpose of figure legends. A sufficient explanation is always intended in figure legends to ensure that the figures and graphs stand independent of the remainder of a manuscript (Alley, 2018), so science writers endeavoured to either define symbols (33) and measurement units (34) or explain the probability of statistically significant values marked in figures (35). As we can see, these explanatory acts increased the proportional use of verb-related bundles in the legends.

(33) Those SNPs **are representative of** a cluster of SNPs defining a haplotype. (Biology)
(34) Statistical analysis **was performed using** a logl-rank (Mantel-Cox) test. (Engineering)
(35) Genes in candidate PTB taxa identified in 16S rRNA analyses that differ significantly at Padj <0.05 with a two-sided Wald test **are shown in** red and those that were not statistically significant **are shown in** pink. (Medicine)

Furthermore, passive-verb bundles were more often used than copular-verb strings. This provides evidential support to the stylistic advice Armer and Day (2011) offer to students on using passive voices in figure legends. As mentioned above, in legends it is essential to define and describe concisely what science figures are supposed to show, so writers rely on the recurrent verbal strings to achieve this communicative purpose. The rhetorical function can be also seen in the passive verbs used in the strings: *shown* (10), *used* (5), *indicated* (3), *based* (2), *compared* (2), *plotted* (2), *referred to* (2), *associated with* (1), *calculated* (1), *found* (1), *given* (1), *normalized* (1), *performed* (1), *presented* (1) and *treated* (1)[4].

These verbal choices fall into three groups: representative, including *shown*, *indicated*, *plotted*, *referred to*, *found*, *given* and *presented*; methodological, including *used*, *based*, *calculated*, *normalized*, *performed* and *treated*; relational, including *compared to/with* and *associated with*.

The representative group was the most frequent, introducing the scientific information that is graphically illustrated (36) or that readers may find in supplementary materials (37). In addition, the methodological group helped to create an interpretative context for the results reported in figures, clarifying either experimental details (38) or statistical treatments applied (39). Moreover, the relational group was also indispensable for spelling out the relationship displayed in graphs in order to establish an interpretative context (Anderson, 2014), so we also found *compared to* (40) and *associated with* (41) in the corpus.

(36) Expression levels **are indicated by** scales in the lower left of each panel. (Engineering)
(37) P values for homogeneity (all >0.05) **are shown in** supplementary materials. (Biology)
(38) Successive compression and expansion (C/E) cycles **are used to** achieve refrigeration in a particular subsystem (cold subsystem).(Physics)
(39) *r* values were **calculated from the** z-statistic of the Mann–Whitney–Wilcoxon test. (Medicine)
(40) Position 5 furthermore exhibits different drying dynamics as **compared to the** four other positions because of irregular movements of the drying front. (Physics)
(41) The black line depicts the proportion of letters **associated with the** favourite colour (y-axis), for the first initial versus all other letters (x-axis) for English-speaking non-synaesthetes. (Biology)

While only two copular *be* three-word bundles, *are available in* and *are representative of*, were identified, they occurred frequently and across a wide range of texts. The former was used to direct readers to supplementary materials (42), and the latter functioned to introduce what images and variants represented (43).

(42) Enlarged graphs of Mulliken charge distribution of atoms **are available in** the Supporting Information. (Physics)
(43) The images **are representative of** the presence or absence of a large lipid-rich plaque detected by near-infraredspectroscopy at baseline. (Medicine)

Turning to noun and preposition-related bundles, they comprised the major categories, especially noun phrase with *of*-phrase fragments (38.5%) and prepositional phrase expressions (20.3%). *A function of*, *the number of*, *the presence of*, *diagram of the* and *illustration of the* were the most common five noun-related bundles with *of*-phrase fragments. This structural pattern normally covers a range of meanings in academic discourse and is often used to specify the attributes of what is being discussed (Biber et al., 2004; Hyland, 2008). In figure legends, however, this form of three-word bundles was typically used to mark a representative relationship between images and scientific information (44) and a formative relationship between dependent and independent variables (45).

(44) The X-axis represents **the number of** cells in which specimens of an OTU occurred, and the Y-axis represents **the number of** cells in which they were projected to occur after modeling. (Biology)
(45) FM is due to **the effect of** the applied magnetic field due to PM1.(Engineering)

In addition, three prepositional phrase bundles were identified, including *as a function*, *in the presence* and *of the two*, and they were used frequently and widely. As shown in (46) and (47), writers sought to set up an interactive relationship between experimental variables by this form of bundles. Additionally, the use of *the two* typically corresponded to the complex interplay between material information in the investigation of scientific world.

(46) Mean uptake rate by bacterial cells are mathematically simulated **as a function** of initial cell density within a spherical aggregate of diameter 20 μm, 20 h after inoculation. (Biology)
(47) TTP399 was tested **in the presence** of 3 mM glucose. (Medicine)
(48) The combination **of the two** layers forms the QPGM shown on the right. (Physics)

Rather than being dominated by anticipatory *it* structures (*it is believed*, *it is necessary*) as Hyland and Jiang (2019) found in research articles, clause related bundles in figure legends were principally composed of abstract-subject fragments. A sweep of the concordance lines shows that they were used to introduce what graphic shapes illustrated in figures (49) and (50).

(49) **Error bars represent** standard deviations of the estimated values.(Physics)
(50) Solid **lines indicate the** relative energy input necessary to maintain a body temperature of 40 C for different air temperatures. (Biology)

As seen above, the building blocks of figure legends primarily comprise noun-related bundles, which points to the informational density of this short piece of text. However, to increase its intelligibility, writers also frequently employ verb-related strings, clarifying explanatory relationships between the items presented in the visuals.

6.2. Functions of lexical bundles in legends

For a better interpretation of the bundle uses, we grouped bundles into functional categories, and our categorisation of the functions seeks to avoid the proliferation of types and sub-types found in the work of Nattinger and DeCarrico (1992), aiming to distil the data into a compact model. Here we followed Hyland and Jiang (2018) in grouping bundles into three main functional groups: research-oriented, dealing with referential functions in the real world; text-oriented, concerned with the organisation of the discourse; and participant-oriented, concerned with authorial stance and reader engagement. However, we expanded the description sub-category to include tangible and intangible attributes with reference to the categorisation suggested by Biber et al. (2004), and the modified functional categories are shown in Table 5.

The data in Table 6 show that research-oriented functions were much more frequently used than the other two categories, referring to different aspects of experimental details and scientific results. Among these referential descriptions, tangible features were most often represented. We see this as a consequence of the mathematization of research findings (Graves, 2014), in which *a schematic of* and *the ratio of* were used to create the impression that the objects or relations they represented are inherently mathematical.

The high frequency of tangible features may also show a common interplay between material and methods section and figure legends (Skern, 2011). To a large extent, this section reports the basics of an experimental design and its methodological approach, while figure legends then specify the exact conditions used in the experiment. Therefore, science writers sought to be specific about the surface of electric cells in a physical experiment (51) or about the amount of electric current traveling per unit at the catalytic sites in an engineering project (52).

(51) Inset shows the result of the Li metal anode **surface of the** control Li-CS cell (78 cycles). (Physics)
(52) In both (C and D) frames, MOF_CNx_Ar + NH3 represents the contribution to the NC **current density of** the non-metallic ORR catalytic sites. (Engineering)

In addition, compared with full research articles (Hyland & Jiang, 2018), figure legends registered a higher proportion of text-oriented functions, and the difference comes from the use of lexical bundles as structuring and framing signals. As Cargill and O'Connor (2013) note, it is essential for figure legends to "explain what the data being presented are and highlight the key points of the part of the results story presented" in the visuals (p. 31). Therefore, text-related bundles were often applied to bridge the link between the presented data and results in a way that readers can easily recognise (53) and (54).

(53) the total number **is shown in** the center of the pie charts.(Medicine)
(54) Corresponding imaging data **are given in** the respective columns below: wild type or Itk-deficient 5C. (Biology)

Accounting for the highest proportion among the functional sub-categories, framing signals played a critical role in defining the conditions which either controlled methodological calculation (55) or restricted the interpretation of results (56), so ensured that readers can make sense of the arguments presented in the legends without reference to other parts of academic prose.

(55) The tip-Csample voltage-dependent dissipation taken at constant d = 5 nm shows a series of dissipation-enhanced features marked by arrows positioned symmetrically **with respect to** Vs = 0 V … (Engineering)
(56) Tc values are clustered **according to the** underlying mechanism (labelled), which controls the magnitude and sign of Tc. (Physics)

Furthermore, it may not be surprising to find that participant-oriented functions were the least often used among the functional categories in figure legends because this account needs to be concise and conveys essential information about the content of a figure and where any data come from (Alley, 2018; Cargill & O'Connor, 2013). Writers, therefore, are more concerned with a succinct

Table 6
Functional distribution of three-word lexical bundles (frequency & percentage).

functions of bundles	frequency	percentage
research-oriented bundles	100	53.5
location	22	11.8
procedure	23	12.3
quantification	16	8.6
description of tangible attributes	34	18.2
description of intangible attributes	5	2.7
text-oriented bundles	**70**	**37.4**
transition signals	7	3.7
resultative signals	5	2.7
structuring signals	21	11.2
framing signals	37	19.8
participant-oriented bundles	**17**	**9.1**
stance features	13	7.0
engagement features	4	2.1

reporting of what is shown in figures than with establishing an interactional relationship with readers. While we found cases of writers commenting on the possibility of knowledge (57) and directly bringing readers in discourse (58), they were not frequent.

(57) The vertical axis in B represents the logarithm of the cumulative number of values exceeding the current velocity, i.e. **the probability of** finding a current with greater velocity than that plotted on the horizontal axis. (Biology)

(58) **Note that the** component 1 axis for the TB sCCA (left) has been reversed for effective visual comparison with PTB sCCA. (Medicine)

In summary, research-oriented bundles outnumbered both text-oriented and participant-oriented strings as a consequence of mathematical framing of research results, so that aspects of methodological details and experimental conditions are always specified in the legends by these referential expressions. Text-related bundles assist in the textual specification and make figure legends easily accessible to readers.

7. Conclusion

Figure legends are an essential component of science research writing. They are built concisely on referential content as journal and stylistic guidelines suggest, and more importantly on a plausible choice of key vocabulary and lexical strings. Keywords not only are a summary of the salient lexical items distinguishing figure legends from other types of academic discourse, but also reveal what science writers are most concerned with in both visual designs and textual descriptions. Among the 20 keywords, *blue* and *red* are the colours used most prominently to mark contrasting features shown in figures, while *bars* and *images* appear dominant to describe graphic shapes. In addition, *shown* and *indicated* are frequent verb choices to introduce the demonstrative relationship between graphic elements and scientific information.

Lexical bundles are recurrent lexical sequences regularly applied to assist academic communication so that readers find such communication efficient and familiar. The structural distribution of commonly used three-word bundles in legends shows that nominal and prepositional forms account for the overwhelming proportion, followed by verb-related and then clause-related bundles. This may relate to the way that writers seek to clarify either graphic symbols or statistical measurement in legends. In addition, research-oriented functions are more frequently used than text and participant-oriented functions, referring to different aspects of experimental details and scientific results. This functional use of lexical strings can be a consequence of mathematization of research findings, and may pertain to the typical interplay between methods section and figure legends.

The results indicate that the legends in the corpus are largely in line with journal and stylistic guidelines. More importantly, however, we see that writers employ what these single and multi-word sequences afford to formulate the short descriptive texts in a way readers may find accessible and meaningful. In addition, a methodological implication from the study is that a corpus-driven approach is productive if there is little linguistic knowledge about an unfamiliar academic textual component, and keywords and lexical bundles can be a starting point of corpus analysis by EAP practitioners. Nevertheless, as future lines of research, genre approaches are complementary and can help to unpack the interplay between legends and the figures and how the two semiotic systems collaboratively contribute to the communicative purpose of research articles.

Legends are succinct and work to ensure the intelligibility of science figures, but more complex is the connection between this rhetorical end and the single and multiple lexical units that are repeatedly used. This may account for the difficulties novice and L2 writers have in writing this particular part-genre, so the findings in this study can be translated into teaching materials. For one thing, text placement activities can be developed to sensitise students to particular forms and functions of key vocabulary and lexical sequences in figure legends. For another, explicit instruction is also necessary to show students the link between stylistic advice given by journals and handbooks and the recurrent word combinations in illustrative examples of figure legends as used in published research articles.

Notes

1. Adapted from Li, P., Lim, A., Gao, L., Hu, C., Yu, L., Bennett, D., Buchman, A. and Hu, K. (2019). More random motor activity fluctuations predict incident frailty, disability, and mortality. *Science Translational Medicine*, 11 (516), p.2.
2. https://www.nature.com/nmat/for-authors/preparing-your-submission
3. TEM is short for Transmission Electron Microscope.
4. The number in parenthesis presents the frequency of each passive verb.

Author statement

Zhongquan Du: Writing – review & editing, Data curation, Validation, Feng Kevin Jiang: Conceptualization, Methodology, Writing – review & editing, Funding acquisition, Luda Liu: Formal analysis, Data curation, Validation.

Acknowledgement

This study was supported by the Research Project of Social Sciences in Jilin Province (2019B163), the Grant for Young Researchers in Social Sciences by Jilin University (2019FRLX04) and also by the Research Project of Humanities and Social Sciences administered by Education Department of Jilin Province (JJKH20211252SK).

Appendix 1. Selected journals in the corpus

Biology.
Biological Review.
BMC Biology
eLife
Philosophical Transactions of the Royal Society B Biological Sciences.
PLOS Biology.
Medicine.
British Medical Journal.
Nature Medicine.
Science Translational Medicine.
The Lancet.
The New England Journal of Medicine.
Physics.
Advanced Energy Materials.
Advanced Functional Materials.
Nano Energy.
Nature Materials.
Nature Photonics.
Engineering.
Applied Catalysis B-Environmental.
Applied Energy.
Chemical Engineering Journal.
Energy & Environmental Science.
Journal of Catalysis.

Appendix 2. The 50 keywords with the highest keyness value in the corpus of figure legends

keyword	frequency	keyness		effect
blue	266	+	1631.49	0.0045
red	300	+	1555.39	0.0051
spectra	156	+	1043.33	0.0027
bars	160	+	1031.08	0.0027
images	205	+	1010.02	0.0035
dashed	109	+	948.98	0.0019
versus	161	+	938.41	0.0027
curves	137	+	843.18	0.0023
normalized	84	+	736.51	0.0014
lines	198	+	650.25	0.0033
plots	98	+	598.40	0.0017
shown	273	+	569.46	0.0044
gray	86	+	565.12	0.0015
green	134	+	555.68	0.0023
data	334	+	538.71	0.0053
indicated	154	+	531.96	0.0026
values	219	+	531.60	0.0036
showing	139	+	523.43	0.0024
respectively	149	+	522.16	0.0025
indicate	154	+	507.14	0.0026
inset	80	+	492.53	0.0014
test	217	+	492.24	0.0036
layer	109	+	491.75	0.0019
representative	129	+	487.21	0.0022
density	116	+	471.99	0.002
plotted	74	+	448.80	0.0013
expression	169	+	443.76	0.0028
bar	113	+	499.97	0.0019
line	187	+	441.62	0.0031
image	142	+	436.94	0.0024
scale	178	+	427.30	0.003

(*continued on next page*)

(*continued*)

keyword	frequency	keyness		effect
represent	138	+	423.95	0.0023
samples	133	+	420.95	0.0022
axis	91	+	419.66	0.0015
profiles	74	+	410.06	0.0013
plot	87	+	403.43	0.0015
experiments	124	+	389.37	0.0021
dotted	56	+	377.47	0.001
purple	52	+	371.81	0.0009
corresponding	112	+	362.85	0.0019
orange	58	+	361.75	0.001
arrowheads	38	+	348.11	0.0006
black	150	+	346.51	0.0025
bottom	90	+	342.58	0.0015
with	1351	+	339.76	0.012
represents	102	+	325.69	0.0017
comparison	117	+	320.77	0.002
dots	54	+	313.38	0.0009

Appendix 3. Three-word lexical bundles in the corpus of figure legends

frequency	range	3-word bundles
58	32	a function of
56	31	as a function
33	19	the number of
30	13	in the presence
28	19	are shown in
26	16	the presence of
25	17	as well as
25	17	diagram of the
25	19	illustration of the
24	12	representation of the
23	18	schematic illustration of
22	16	of the two
22	14	schematic diagram of
21	12	before and after
21	14	each of the
21	13	image of the
21	7	photograph of the
20	12	a schematic of
20	5	representative images of
20	6	was used to
19	5	are available in
19	14	relative to the
19	13	schematic of the
19	16	with respect to
18	14	comparison of the
18	11	the effect of
18	7	version in colour
17	10	compared with the
17	6	current density of
17	15	images of the
17	9	in the absence
17	15	shown in the
16	13	according to the
16	5	are representative of
15	5	data are presented
15	12	images of a
15	8	normalized to the
15	10	schematic representation of
15	11	view of the
14	8	error bars represent
14	10	is shown in
14	8	the position of
13	8	as in a
13	7	based on the
13	7	in colour figure
13	9	indicated by the

(*continued on next page*)

(*continued*)

frequency	range	3-word bundles
13	8	position of the
13	7	with or without
12	7	are indicated by
12	5	are presented as
12	6	bars indicate the
12	5	calculated from the
12	9	due to the
12	7	image of a
12	8	on the left
12	7	the distribution of
12	5	the percentage of
12	8	the top of
11	5	a schematic diagram
11	6	are based on
11	9	function of the
11	6	of a single
11	9	respect to the
11	5	was performed using
11	5	were treated with
10	9	and in the
10	5	are given in
10	7	are shown for
10	6	are shown with
10	8	curves of the
10	8	distribution of the
10	8	is indicated by
10	7	line represents the
10	7	lines represent the
10	7	of the same
10	6	open circuit voltage
10	6	or absence of
10	6	presence or absence
10	6	the presence or
10	7	the proportion of
10	9	the surface of
10	7	top of the
10	6	with and without
9	7	a comparison of
9	5	in response to
9	6	lines indicate the
9	8	on the right
9	7	relationship between the
9	8	shown on the
9	8	the absence of
9	7	the dashed lines
9	7	value of the
8	5	are plotted as
8	6	are shown as
8	5	are used to
8	5	compared to the
8	8	for each of
8	6	note that the
8	5	of the cell
8	7	patterns of the
8	5	plot of the
8	5	plotted as a
8	7	section of the
8	6	surface of the
8	5	test of the
8	8	the dashed line
8	6	the data are
8	8	the ratio of
8	5	total number of
8	7	values of the
7	7	a schematic illustration
7	7	at least one
7	5	at the top
7	5	between the two
7	6	cross section of
7	5	dashed lines indicate

(*continued on next page*)

(*continued*)

frequency	range	3-word bundles
7	5	effect of the
7	6	error bars indicate
7	7	in the same
7	6	in the top
7	6	in the two
7	5	is used to
7	6	left to right
7	5	of the experimental
7	6	on the surface
7	5	shown in fig
7	5	side view of
7	6	stability of the
7	6	the case of
7	5	the dotted line
7	7	the end of
7	7	the formation of
7	5	the probability of
7	5	the time of
7	6	the total number
7	5	views of the
7	6	well as the
6	6	at the end
6	5	correlation between the
6	6	difference between the
6	5	found in the
6	5	in the lower
6	6	in this figure
6	5	in which the
6	5	is shown on
6	5	line indicates the
6	5	mapping images of
6	5	mean and standard
6	5	of the cross
6	6	of the first
6	5	one of the
6	6	ratio of the
6	5	the amount of
6	5	the data points
6	6	the location of
6	6	the results of
6	6	the size of
6	5	to the same
6	5	up of the
6	5	were used as
5	5	are also shown
5	5	associated with the
5	5	characterization of the
5	5	colour in this
5	5	curves of a
5	5	data are shown
5	5	end of the
5	5	for interpretation of
5	5	for the two
5	5	is referred to
5	5	of the different
5	5	of the samples
5	5	of this article
5	5	on the basis
5	5	orientation of the
5	5	overview of the
5	5	performance of the
5	5	reader is referred
5	5	referred to the
5	5	the basis of
5	5	the inset is
5	5	the reader is
5	5	the right of
5	5	this figure legend
5	5	to colour in
5	5	to the right
5	5	used for the

References

Aliotta, M. (2019). *Mastering academic writing in the sciences: A step-by-step guide*. Boca Raton: CRC Press Taylor & Francis Group.

Alley, M. (2018). *The craft of scientific writing*. New York: Springer.

Anderson, G. (2014). *How to write a paper in scientific journal: Style and Format*. Bates College: Lewiston: Department of Biology.

Anthony, L. (2019). *AntConc*. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/.

Armer, T., & Day, J. (2011). *Cambridge English for scientists. Student's book//Cambridge English for scientists*. Cambridge: Cambridge University Press.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines* (2nd ed.). Philadelphia PA: Open University Press.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics, 14*(3), 275–311.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Cargill, M., & O'Connor, P. (2013). *Writing scientific research articles: Strategy and steps*. Oxford: Wiley-Blackwell.

Carolina, R., & Carlino, P. (2017). Reading to write in science classrooms: teacher's and students' joint action. In S. Plane, C. Bazerman, F. Rondelli, C. Donahue, A. N. Applebee, C. Boré, P. Carlino, M. M. Larruy, P. Rogers, & D. Russell (Eds.), *Research on writing: Multiple perspectives* (pp. 415–436). Colorado: The WAC Clearinghouse.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language, Learning and Technology, 14*(2), 30–49.

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

Cortes, V. (2012). Lexical bundles and technology. In C. A. Chapelle (Ed.), *The encyclopaedia of applied linguistics*. Oxford: Blackwell.

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

Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes, 51*, 1–17.

Grabowski, Ł. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes, 38*, 23–33.

Graves, H. (2014). The rhetoric of (interdisciplinary) science: Visuals and the construction of facts in nanotechnology. *Poroi, 10*(2), 1–19.

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi, & M. Scott (Eds.), *Keyness in texts* (pp. 59–78). Amsterdam: John Benjamins.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: University of Michigan Press.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. London: Routledge.

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

Hyland, K., & Jiang, F. K. (2018). Academic lexical bundles: How are they changing? *International Journal of Corpus Linguistics, 23*(4), 383–407.

Hyland, K., & Jiang, F. K. (2019). *Academic discourse and global publishing: Disciplinary persuasion in changing times*. London: Routledge.

Hyland, K., & Jiang, F. (2020). "This work is antithetical to the spirit of research": An anatomy of harsh peer reviews. *Journal of English for Academic Purposes, 46*, 1–13.

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

Jiang, F. (2019). *Corpora and EAP studies*. Beijing: Foreign Language Teaching and Research Press.

Kroen, W. (2004). Modeling the writing process: Using authentic data to teach students to write scientifically. *Journal of College Science Teaching, 34*(3), 50–53.

Lemke, J. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text. In J. R. Martin, & R. Veel (Eds.), *Reading science: Critical and functional perspectives on discourses of science* (pp. 87–113). London: Routledge.

Luna, R. E. (2013). *The art of scientific storytelling: Transform your research manuscript, using a step-by-step formula*. Lexington, Kentucky: Amado International.

Miller, T. (1998). Visual persuasion: A comparison of visuals in academic texts and the popular press. *English for Specific Purposes, 17*(1), 29–46.

Moghaddasi, S., Graves, H. A. B., Graves, R., & Gutierrez, X. (2019). "See figure 1": Visual moves in discrete mathematics research articles. *English for Specific Purposes, 56*, 50–67.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nature research. (2019). Writing for a nature journal. Retrieved from https://www.nature.com/nature-research/for-authors/write#how-to-write-a-scientific-paper.

Pauwels, L. (2006). *Visual cultures of science: Rethinking representational practices in knowledge building and science communication*. Hanover: University Press of New England.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Skern, T. (2011). *Writing scientific English: A workbook*. Wien: Facultas Universitätsverlag.

Swales, J. (2001). EAP-related linguistic research: An intellectual history. In J. Flowerdew, & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 42–54). Cambridge: Cambridge University Press.

Tardy, C. M. (2009). *Building genre knowledge*. West Lafayette: Parlor Press.

Zimbardi, K., Bugarcic, A., Colthorpe, K., Good, J. P., & Lluka, L. J. (2013). A set of vertically integrated inquiry-based practical curricula that develop scientific thinking skills for large cohorts of undergraduate students. *Advances in Physiology Education, 37*(4), 303–315.

Zhongquan Du is associate Professor in applied linguistics in the School of Foreign Language Education at Jilin University, China and teaches and researches in academic writing at both undergraduate and postgraduate levels.

Feng (Kevin) Jiang is Kuang Yaming Distinguished Professor in applied linguistics in the School of Foreign Language Education at Jilin University, China and gained his PhD under the supervision of Professor Ken Hyland at the Centre for Applied English Studies at the University of Hong Kong. His research interests include disciplinary discourse, corpus studies and academic writing, and his publications have appeared in most major applied linguistics journals.

Luda Liu is a PhD student in applied linguistics in the School of Foreign Language Education at Jilin University, China. His research interests include disciplinary writing, corpus analysis and genre studies.