Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals

Fan Pan a,*, Randi Reppen b, Douglas Biber b

a Huazhong University of Science and Technology, School of Foreign Languages, Wuhan, Hubei, 430074, PR China
b Northern Arizona University, Department of English, Applied Linguistics, Liberal Arts Building 18, PO Box 6032, Flagstaff, AZ 86011, USA

ARTICLE INFO

Article history:
Received 6 June 2015
Received in revised form 13 November 2015
Accepted 17 November 2015
Available online 24 December 2015

Keywords:
Lexical bundles
Structures
Functions
Academic articles

ABSTRACT

Corpus-driven research on phraseology has documented different types and functions of lexical bundles (recurrent formulaic sequences) in spoken versus written registers, providing a foundation for recent studies that explore patterns of language development in the use of these bundles. Some studies focus on changes in the phraseological patterns of novice/student writers compared to 'expert' writers, while others have focused on a comparison of L1-versus L2-English students. However, no study to date has directly compared the use of lexical bundles by L1-English versus L2-English academic professionals. The current study addresses this gap by examining the structural and functional types of lexical bundles employed by L1 English and L1 Chinese professionals writing for English medium Telecommunications journals. The findings show major structural differences (phrasal vs. clausal) between L1 and L2 writers. L2 writers mostly use bundles consisting of verbs and clause fragments (especially passive verb structures), while L1 writers use bundles consisting of noun and prepositional phrases. Results also demonstrate that L2 professionals use bundles that are functionally different from the L1 professionals, and even misuse certain bundles. In addition, the relationship between structural types and functional categories shows that L1 and L2 professionals employ bundles with different structural characteristics serving similar functions.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Over the last two decades, an extensive body of research has employed corpus-driven methods to explore the phraseological patterns of spoken and written registers (see Gray & Biber, 2015). Many of these studies have explored the use of lexical bundles: recurrent lexical sequences (e.g., take a look at, know what I mean) identified through corpus analysis that includes specific frequency thresholds and dispersion requirements. Studies have described lexical bundles in a range of spoken and written registers and have built on earlier research to explore patterns of language development in the use of lexical bundles. Lexical bundles have been used to explore developmental differences between novice/student and expert writers (e.g., Chen & Baker, 2010; Cortes, 2004), proficiency levels of L2-English writers (e.g., Staples, Egbert, Biber, & McClair, 2013), and between L1-English and L2-English student writers (e.g., Adel & Erman, 2012; Chen & Baker, 2010).
Scholars disagree on whether academic writers follow the same developmental progression in the use of lexical bundles, or whether L2-English writers differ from L1-English writers in the use of bundles. Results of previous research are unclear as to whether the novice/student versus expert distinction, or the L1 versus L2 distinction leads to the differences that have been found. For example, Chen and Baker (2010), and Adel and Erman (2012) both found consistent and interpretable differences between L1 and L2 student writers, while Cortes (2004) and Römer (2009) argue that the novice/student versus expert distinction is as important as the distinction between L1 and L2 writers. Surprisingly, though, most developmental studies have confounded the influence of expertise with the influence of L1. Perez Llantada (2014) is the only previous study that we found that compared the use of lexical bundles in academic journals by L1-English and L2-English professional scholars (and also L1 Spanish writing in Spanish). In this large scale study that looked at these different L1 authors of research articles from twelve disciplines, she finds that register was one of the most important predictors of bundle use.

It seems uncontroversial that all novice writers (L1 and L2 alike) must learn the discourse conventions of advanced academic writing, including the appropriate use of lexical bundles (see, e.g., Biber, Gray, & Pooneh, 2013; Cortes, 2004). But some previous research indicates that L1-English writers have a head start in this task, resulting in consistent differences between L1 and L2 groups of students who are otherwise matched for level (and discipline in the case of the Adel & Erman study). The question investigated in the current study is whether those differences disappear at the expert level, or if the differences between L1 and L2 continue for experts.

We investigate this question through a corpus-based comparison of published academic research articles by L1-English writers versus L1-Chinese writing in English from the same academic discipline. In the following section, we introduce the construct of lexical bundles as it is used in the current study. Then, we present a corpus-based investigation that compares the use of lexical bundles by English L1 and L2 professional scholars writing in Telecommunications journal articles.

1.1. Previous research on lexical bundles

Altenberg (1998) was probably the first to employ corpus analysis to investigate frequently recurring lexical phrases in English, identifying 470 3-word sequences that occurred at least 10 times in the London–Lund Corpus (spoken English). Around the same time, Biber, Johansson, Leech, Conrad, and Finegan (1999; Chapter 13) identified the most common ‘lexical bundles’ in conversation and academic writing, defined as sequences of words that occurred at least 10 times per million words in the target register, distributed across at least 5 different texts. These bundles were interpreted in structural/grammatical terms (e.g., main clause fragments: have a look at; noun phrase or prepositional phrase fragments: the end of the).

Many studies have employed a lexical bundle framework to describe expressions typical of different registers, focusing on variation across registers, and describing the discourse functions served by different types of lexical bundles. Biber, Conrad, and Cortes (2004) compare the distribution and functions of lexical bundles in four registers: conversation, university classroom teaching, university textbooks, and published academic research writing. That study found systematic differences at all levels. Overall, there were more frequent (i.e., tokens) and more different (i.e., types) bundles in speech than in writing. Conversation bundles consist mostly of verbs and clause fragments (including dependent clause fragments), while in academic writing lexical bundles consist mostly of noun phrase and prepositional phrase fragments.

Several researchers (Biber, 2006; Biber et al. 2004; Hyland, 2008a) have shown that lexical bundles also vary in their discourse functions (e.g., expressing stance, discourse organization, or referential meanings). It turns out that these functional differences are as important as structural differences for the description of register variation; for example, conversation tends to use bundles for stance functions, while academic writing tends to rely on referential bundles.

More recently, lexical bundle studies have explored issues related to language development. Several studies have investigated the use of lexical bundles by L2-English writers across proficiency levels (e.g., Staples et al., 2013), or compared the use of bundles by L1-English versus L2-English writers (e.g., Adel & Erman, 2012; Chen & Baker, 2010; De Cock, 2000; Nekrasova, 2009). Most of these studies conclude that L1-English writers use more lexical bundles, and more varied bundle types, than L2-English writers (Adel & Erman, 2012; Chen & Baker, 2010). However, Staples et al., (2013) found that lower proficiency L2-English writers use more lexical bundles overall (not distinguishing between clausal and phrasal types) than higher proficiency L2 writers.

Other studies have compared novice versus expert writers, not focusing on the L1 versus L2 distinction. For example, Cortes (2002, 2004) finds that novice writers (student writers) use bundles in ways that are functionally different from those of experts (published authors). In addition, Cortes (2004) also finds that many bundles used by experts are rarely used by undergraduate and graduate students.

Römer (2009) finds that some bundles common in expert writing occur much less frequently in either L1-English or L2-English student writing. Based on that finding, Römer argues that the novice versus expert distinction is more important than the L1/L2 distinction for understanding language development in the use of lexical bundles. However, Hyland (2008a) documents a somewhat different developmental progression, finding that postgraduate students tend to employ more lexical bundles (types) in their academic writing than professional academics, apparently as a way of displaying their competence in academic discourse. Similar to Römer (2009), Chen and Baker (2010) do a 3-way comparison: L1-English students versus L2-English students versus L1-English expert writing (published research articles). Chen and Baker find few differences between the L1 and L2 student writing. Most differences in their study were between student versus expert writing with students using more verb phrase based bundles and more discourse organizing bundles than the expert writers.
1.2. Structural and functional characteristics of lexical bundles

Many of the studies surveyed in the last section compare lists of specific lexical bundles representing different groups of writers or different registers. However, there are several methodological challenges for such analyses related to the procedures used to identify these lists of the most important lexical bundles in a corpus, and the comparability of the corpora used for the analysis (see, Adel & Erman, 2012, pp. 87–88). One approach to this problem is to explore the use of statistical techniques for identifying the important lexical sequences in a corpus (e.g., Gries, 2008; Gries & Mukherjee, 2010; O’Donnell, Römer, & Ellis, 2013), although no study to date has systematically explored the influence of corpus design and composition on the specific lists of lexical bundles identified across corpora (see Miller & Biber, 2015).

An alternative approach is to focus on the structural and functional characteristics of lexical bundles across corpora, which are more stable and generalizable than lists of specific lexical bundles. The structural classification of lexical bundles first proposed by Biber et al. (1999, Chapter 13) has been widely used in subsequent studies of lexical bundles. The primary distinction employed for this purpose is between clausal bundles (e.g., I don’t know what) and phrasal bundles (e.g., in the case of). Other distinctions can be made within each of these two major structural types. For example, clausal bundles include simple verb phrases (e.g., have a look at) as well as bundles that incorporate a main clause and the beginning of an embedded dependent clause (e.g., I don’t know how). Phrasal bundles can be composed of noun phrases with the start of an embedded prepositional phrase (e.g., the nature of the), noun phrases with the start of an embedded relative clause (the way in which), and prepositional phrases, which often include the start of an embedded prepositional phrase (e.g., as a result of).

There are systematic differences across registers and across groups of writers in their reliance on these different structural types of lexical bundles. For example, Biber et al. (1999, pp. 996–997) document a major difference between conversation and academic writing: conversation relies almost entirely on clausal bundles, while academic writing relies almost entirely on phrasal bundles. Biber et al. (2004) show that university textbooks are similar to academic research articles in their reliance on phrasal bundles. In addition, there is some evidence that groups of writers differ in their reliance on these structural types: expert writers use more phrasal bundles than novice writers, and L2-English writers use slightly more phrasal bundles than L1-English writers (see Chen & Baker, 2010, Table 5).

These phraseological patterns can be compared to recent research on grammatical complexity by Biber and Gray (2010, 2011, in press) and Biber, Gray, and Poonpon (2011, 2013). These studies show that conversation and other spoken registers tend to rely on clausal complexity features, while academic prose tends to rely on phrasal complexity features. These same patterns can be applied to the description of written language development. For example, Biber, Staples, and Gray (2014) show that high proficiency L2-English writers rely on phrasal structures to a greater extent than lower proficiency writers (but all groups of student writers use phrasal structures less than professional academic writers).

The second major approach that has been used to characterize lexical bundles is to investigate their discourse functions. Biber et al. (2004) develop a functional taxonomy with three major categories: stance expressions (e.g., it is possible to), discourse organizers (e.g., as a result of), and referential expressions (e.g., as shown in figure). They find dramatic differences between spoken and written registers in their reliance on these functional types of bundles. Conversation mainly employs stance bundles, while academic writing mostly uses referential bundles. Discourse organizing bundles are less common.

Hyland (2008b, p. 13) proposes a modified version of Biber et al.’s (2004) categories, with different labels that are relevant to the domain of academic writing (instead of the range of registers considered by Biber et al. 2004). Hyland’s threefold distinctions categorize bundles functionally as research-oriented (corresponding to ‘referential’ bundles in the Biber et al. framework), text-oriented (corresponding to ‘discourse-organizing’), or participant-oriented (corresponding to ‘stance’ bundles), focusing on the role of the writer or the reader as related to academic writing. For example, research-oriented bundles (e.g., the purpose of the, the role of the) are used by writers to structure their experiences, while text-oriented bundles are used to organize the text.

1.3. Overview of the present study

Previous research has shown that there are meaningful differences in the use of lexical bundles between novice versus expert writers, as well as differences between L1-English and L2-English student writers at the same level of university study. However, no study to date has compared the patterns of use for L1 and L2 English professional writing to explore if differences become leveled at highly advanced stages of proficiency and expertise.

To address this question, we carried out a comparison of Telecommunications research articles written by L1-English and L2-English professionals. To avoid possible confounding influences, we controlled for three factors: we only included texts from a single academic discipline (Telecommunications); and a single register (published academic research articles); and all L2-English authors have the same native language (Chinese). In contrast, some previous studies have confounded register/discipline differences with the difference between groups of writers (e.g., comparing general essays written by students to research articles written by professionals) and some have included L2-English students from multiple first languages. By controlling these factors, we hope to identify similarities and differences in language use between writers at an advanced professional level in the same discipline and with the L2 writers sharing the same L1 (Chinese).

As described in the previous section, lexical bundle use is described much more reliably when the focus is on the structural and functional types of bundles (rather than comparing lists of specific bundles). Thus, we investigate the following research questions:
1. What differences exist in the structural types of lexical bundles used by L1-English versus L1-Chinese scholars in published Telecommunications research articles?

2. What differences exist in the discourse functions of lexical bundles used by L1-English versus L1-Chinese scholars in published Telecommunications research articles?

2. Corpora and methods

2.1. Corpus collection

The corpora analyzed for this study are Telecommunications research articles published between 2007 and 2011 by L1-English writers (TELE-EN) and by L1-Chinese writers (TELE-CH). The TELE-EN corpus research articles were selected from the world’s leading scholarly Telecommunications research journals (as reflected by high impact factors). The TELE-CH corpus articles appeared in major Telecommunications journals published in China (see Appendix A). The two corpora are closely matched for total number of words (c. 500,000 in each), but not for number of texts (see Table 1). The TELE-CH articles are shorter than TELE-EN articles; therefore, there are more articles in the TELE-CH corpus.

There is of course no exact way to ascertain the first language of a writer, especially for published texts. Therefore, we followed the methods proposed by Wood (2001) that operationally defined ‘L1-English’ writers to be any author affiliated with an institution in a country where English is spoken as the first language who also has a first and last name that can be considered native to English-speaking countries. (We erred on the side of excluding authors with questionable names.) ‘L1-Chinese’ (and thus L2-English) writers were operationally defined as any author affiliated with an institution in China who also has a first and last name that could be considered native to China.

The journals selected for the TELE-CH corpus are viewed as somewhat comparable to the TELE-EN corpus based on these considerations. The journals in the TELE-CH corpus are published by top universities in China. Second, these journals state that they follow the academic conventions of international journals and are peer-reviewed. The editorial boards include international members and target Chinese as well as international readers. In addition, by selecting L2 articles from journals published in China, we are more likely to reflect the language produced by Chinese writers in a Chinese context for an international audience. Thus, a comparison of these two corpora is taken to represent the similarities and differences in the written discourse of L1-English versus L1-Chinese academic professionals.

2.2. Identification of lexical bundles

Lexical bundle research begins by identifying the important lexical sequences typically used in a variety. To achieve this goal, lexical bundle studies consider two criteria: frequency and dispersion.

Frequency thresholds used in lexical bundle studies range from 20 occurrences per million words to 40 occurrences per million words. Of course, many lexical bundles occur much more frequently (even 200 times per million words). But the main goal of setting a frequency threshold is to identify bundles that recur often enough to be regarded as typical of the target register.

Dispersion thresholds are important to ensure that bundles are not restricted to a few texts or authors. A dispersion requirement ensures that the bundles are typical of the entire corpus, not just a few texts. Early lexical bundle studies required that bundles occur in at least five different texts of the corpus (e.g., Biber et al., 1999), while studies of student writing have sometimes employed a lower dispersion requirement of three texts (e.g., Adel & Erman, 2012; Chen & Baker, 2010). One methodological problem is that corpora differ in average text lengths, and therefore in the number of texts, making it desirable to set different dispersion thresholds for different corpora (Hyland, 2008a, b).

Another decision is the length of sequences to be included in the study. Following the practice of previous research, the present study focuses on 4-word sequences, “because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles” (Hyland, 2008b, p.8).

For our study, we chose the moderately high frequency threshold of 40 occurrences per million words (the standard in most previous lexical bundle studies), with different dispersion requirements for the two sup-corpora: 5 texts in TELE-EN and 10 texts in TELE-CH (since there are roughly twice as many texts in the TELE-CH corpus as in the TELE-EN corpus). Using WordSmith 4.0, these criteria resulted in 57 bundles from TELE-EN and 72 bundles from TELE-CH. Three lexical sequences were composed entirely of discipline specific content words (i.e., the transmitter and receiver (TELE-EN), signal to noise ratio (in both corpora)); these were eliminated from further consideration, giving a total of 55 lexical bundles from TELE-EN and 71 bundles from TELE-CH.

<table>
<thead>
<tr>
<th>Table 1 Composition of the corpora.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpora</td>
</tr>
<tr>
<td>Number of texts</td>
</tr>
<tr>
<td>Mean length of texts</td>
</tr>
<tr>
<td>Total corpus size</td>
</tr>
</tbody>
</table>
As we noted in Section 1.2, there is no ‘correct’ list of the important lexical bundles in a register, because the identification process is highly influenced by the corpus design/composition and by the identification procedures. As a result, it can be problematic to compare the specific bundles identified in different corpora. In the present study, we focus instead on the typical structural characteristics of bundles used by the two groups of writers (Section 3.2) and the typical discourse functions served by bundles used by the two groups (Section 3.3). Comparisons at these levels are more robust and less influenced by the specific corpus designs and identification procedures than direct comparisons of specific bundle lists. (However, for the sake of reference, we present the actual lists in Appendix B.)

3. Results and discussion

3.1. Lists of the specific lexical bundles in TELE-EN versus TELE-CH

Appendix B lists the lexical bundles extracted from the two corpora, including bundles that are shared by both groups and those specific to one or the other group. The criteria described above yielded 55 4-word bundles in TELE-EN and 71 4-word bundles in TELE-CH, with 24 bundles used by both groups (43% of the L1 bundles and 34% of the L2 bundles).

The bundles that met the inclusion criteria (frequency and dispersion) for only one corpus were checked to see if they occurred at all in the other corpus. Interestingly, bundles in the TELE-EN list were used by TELE-CH authors, but with lower frequency or restricted dispersion. In contrast, three bundles from the TELE-CH list (with the increase of, the reason is that, it is obvious that) did not occur in the TELE-EN corpus. These bundles are a translation from Chinese to English.

The fact that there are a greater number of different 4-word bundles in the L2-English corpus is a preliminary indication that these L2-English professional writers rely on lexical bundles to a greater extent than L1-English writers. In the following sections, we explore the structural and functional associations of those bundles.

3.2. Comparison of the structural types of lexical bundles

Using the structural classification of lexical bundles proposed by Biber et al. (1999; Section 1.2), the bundles from the two corpora were categorized according to structural correlates. Similar to Chen and Baker (2010), we grouped the lexical bundles under three broader categories: “NP-based”, “PP-based” and “VP-based” (see Table 2). NP-based bundles and PP-based bundles include noun phrases and prepositional phrases, while VP-based bundles refer to word combinations with a verb component (Chen & Baker, 2010, p.35). Table 2 presents the types (different bundles) and tokens (total bundles) of each subcategory. Log-likelihood tests were performed to identify significant differences across the two corpora.

3.2.1. Comparison of the distribution of structural categories

As Table 2 shows, L1 writers and L2 writers in the same discipline rely on different grammatical types of bundles in writing research articles. Log-likelihood tests comparing the tokens for L1-English show that TELE-EN writers use significantly more NP-based and PP-based bundle tokens than TELE-CH writers (the numbers showing higher use are bolded in Table 2). In contrast, TELE-CH writers use significantly more VP-based bundle tokens than TELE-EN writers in most VP-based subcategories.

Table 3 presents a comparison of the percentages of the main structural categories across the two corpora. The primary distinction here is between clausal bundles and phrasal bundles. Clausal bundles mainly include simple verb phrases as well as bundles that incorporate a main clause; phrasal bundles are mostly composed of noun phrases and prepositional phrases (Biber et al., 1999 section 13.2). We find that TELE-EN writers primarily use phrasal bundles (NP or PP), accounting for 69% of the bundle types and approximately 67% of the bundle tokens. This finding is consistent with previous studies that show lexical bundles in written academic prose are predominantly phrasal rather than clausal (Biber & Conrad, 1999; Biber et al.,

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Distribution of structural subcategories.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural subcategories</td>
<td>Types</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>NP-based</td>
<td>Noun phrase with of –phrase fragment (the size of the)</td>
</tr>
<tr>
<td></td>
<td>Noun phrase with other post-modifier fragment (the difference between the)</td>
</tr>
<tr>
<td>PP-based</td>
<td>Prepositional phrase with embedded of –phrase (in the case of)</td>
</tr>
<tr>
<td></td>
<td>Other prepositional phrase fragment (at the same time)</td>
</tr>
<tr>
<td>VP-based</td>
<td>Copula be + noun phrase/adjunctive phrase (is due to the)</td>
</tr>
<tr>
<td></td>
<td>(Verb phrase)+ with active verb (we can get the)</td>
</tr>
<tr>
<td></td>
<td>Anticipatory it + verb phrase/adjunctive phrase (it is easy to)</td>
</tr>
<tr>
<td></td>
<td>Passive verb + prepositional phrase fragment (can be seen in)</td>
</tr>
<tr>
<td></td>
<td>(Verb phrase)+that clause fragment (we assume that the)</td>
</tr>
<tr>
<td>Others</td>
<td>(as well as the, than that of the)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

* = significant at $p < .05$ level ** = significant at $p < .01$ level.
1999, 2004, 2011, 2013; Biber & Gray, 2010, 2011; Byrd & Coxhead, 2010). In contrast, the clausal nature of TELE-CH is surprising: TELE-CH writers rely heavily on VP-based bundles, accounting for almost 58% of the bundle types and 56% of the bundle tokens.

Phrasal features in written academic prose are associated with its high informational focus. Careful integration of information in academic prose requires the use of noun phrases and prepositional phrases, which leads to a shift from clausal style to phrasal style in academic prose. Two examples from Halliday (1989, p.61, emphasis added) illustrate this point. Compared with two bolded verb clauses in example 1, the bolded noun phrase and prepositional phrases in example 2 make the sentence more informationally efficient.

1. **If you invest in a rail facility**, this implies that **you are going to be committed for a long term**.
2. **Investment in a rail facility** implies **a long term commitment**.

Similar differences characterize TELE-EN versus TELE-CH discourse. For example consider the text examples below (phrase and clause bundles in bold).

Text sample 1: TELE-EN

Fig. 5 shows a performance comparison with different initialization schemes as a function of the number of fitness evaluations with $N_u = 8$, $N_r = 5$, $S = 80$, and $P = 1024$. When soft biasing the initial population with the output of an mmse receiver that has been optimized and biased as in Section II-B, performance is almost identical to that with ZF soft biasing. Thus, in the remainder of this paper, the ZF receiver soft output is used. In the most common form of biased initialization [8], [19], the hard output of a linear receiver is used to seed the initial population through mutation.

Text sample 2: TELE-CH

We can see that the proposed technique outperformed other algorithms in the test. The reason is that the curvelet—wavelet—fractal technique makes use of facial similarities that exist in the detailed images. Furthermore, in the sparse components of an ICA transformation, only a few of the components are active (i.e., significantly nonzero) simultaneously, thus one may assume that the components with small absolute values (smaller than the noise level) are pure noise. By setting them to zero we can retain just a few components with large activities. This method is closely related to the wavelet shrinkage method [6], but the orthogonal sparse wavelet-like basis is adaptive to the data. Thus, we can obtain the detailed edge of the image and remove most noise.

To further explore the clausal vs. phrasal nature of L2 research writing, we compared our results to Chen and Baker (2010). Chen and Baker included four comparison corpora in their study: FLOB-J, BAWE-EN, BAWE-CH, and the academic writing sub-

<table>
<thead>
<tr>
<th>Structural categories</th>
<th>Types (%)</th>
<th>Tokens (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TELE-EN</td>
<td>TELE-CH</td>
</tr>
<tr>
<td>NP-based – Phrasal</td>
<td>36.4</td>
<td>21.1</td>
</tr>
<tr>
<td>PP-based – Phrasal</td>
<td>32.6</td>
<td>12.7</td>
</tr>
<tr>
<td>VP-based – Clausal</td>
<td>25.5</td>
<td>57.7</td>
</tr>
<tr>
<td>Others</td>
<td>5.5</td>
<td>8.5</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 3**

Distribution of main structural categories.

![Fig. 1. Comparison to previous research of the distribution of structural types of lexical bundles.](image-url)
The proportional distribution of lexical bundle types in each corpus is presented in Fig. 1 (based on Table 5 from Chen & Baker, 2010, p. 35, compared to the results from our study). Although Chen and Baker (2010) put “Others” in the VP-based category, we listed it separately in Fig. 1, because bundles such as “as well as the, than that of the, better than that of” do not fit in the VP-based category. In Fig. 1, the first 3 bars on the left represent L1-English expert writing (ACD (LGSWE), FLOB-J and TELE-EN). The three bars on the right represent three types of novice or L2 writing (BAWE-EN (L1 learners), BAWE-CH (L2 learners) and TELE-CH (L2 professionals)).

Fig. 1 shows a sharp contrast between the L1-expert groups on the left and the novice and/or L2 groups on the right. The L1 expert writing corpora on the left all exhibit a strong preference for phrasal bundles, with PP and NP making up roughly two thirds of all the bundles (63%, 69%, and 69%, respectively). In these corpora, clausal bundles make up one third or less of all bundles. However, the novice and/or L2 corpora on the right show a high use of VP or clausal bundles (53%, 48%, and 58%, respectively) and a much lower use of phrasal bundles. This difference is most pronounced in the L2-expert TELE-CH corpus.

This structural difference between the expert and the novice and/or L2 corpora may reflect an important developmental stage of academic writing: the phrasal style of expert writers, and the clausal style of less proficient writers that includes both L1 and L2 writers. This finding is consistent with Biber et al. (2014) and Parkinson and Musgrave (2014), who find that high proficiency L2-English writers rely on phrasal structures to a greater extent than lower proficiency writers, and that all groups of student writers use phrasal structures less than proficient academic writers. Taken together, this supports Biber et al.’s (2011, pp. 30–31) hypothesis that both L1 and L2 academic writers follow a similar series of developmental stages from clausal style to phrasal style. That is, these findings support the claim that both L1 and L2 writers acquire a clausal style of discourse at an early stage and move to a phrasal style of academic writing at later stages. Many types of complex phrasal embedding are not acquired naturally, and many native speakers of English rarely (or never) produce language of this type (Biber et al. 2011, p.29). Thus, the progression from clausal to phrasal style is challenging for both L1-novice and L2 writers. In our study, L2 English academic professionals demonstrate a reliance on clausal features, similar to L1 and L2 student writers, indicating that these academic professionals have not acquired the consistent use of phrasal bundles like their L1-English counterparts.

3.2.2. Comparison of frequent frames in the main structural categories

Similar to Chen and Baker (2010), we chose the most frequent and productive lexical frame in each main category for more detailed consideration, comparing the use of these frames across the two corpora. Three frames are discussed in the present section: ‘the + Noun + of the/a’ in NP-based category, ‘in the + Noun + of’ in the PP-based category, and ‘Passive Verb + Prepositional Phrase’ in the VP-based category.

Table 4 shows that compared with TELE-CH writers, TELE-EN writers only use a slightly wider range of nouns (both types and tokens) that collocate with the NP frame. However, in Chen and Baker’s (2010) study, L1 expert writers differ greatly from the L2 student group in both types and tokens with twice the number of types (16:8) and almost three times as many tokens (100:37).

Table 5 shows that TELE-EN writers and TELE-CH writers are similar in the number of filler types used in the frame in the + Noun + of, but TELE-EN writers use almost two times as many tokens as TELE-CH writers. However, in Chen and Baker’s (2010) study, L1 expert writers differ greatly from the L2 student group in both types and tokens, with over three times the number of types (10:3) and over four times the number of tokens (87:19).

As Table 6 shows, TELE-CH writers use many more types and tokens of the frame Passive Verb + Prepositional Phrase than TELE-EN writers: over three times the number of the types (17:5) and over four times for tokens (755:181). However, in Chen and Baker’s study, L2 student writers used fewer passive patterns than L1 expert writers, less than half the types (4:7) and tokens (19:34). Besides, Table 6 shows that TELE-CH writers display a strong preference for the frame ‘can be + passive verb + complement’ (e.g., can be expressed as) using six different verbs (e.g., describe, divide, express, obtain, use, write), followed by different prepositions, which fit the slot. TELE-CH writers use six different bundles with this frame while TELE-EN writers use only one such bundle. Both L1 and L2 student groups in Chen and Baker’s (2010) study also demonstrate the same preference (p. 37, Table 10).

The above comparisons demonstrate that the differences between L1 and L2 found in previous research do not disappear at the expert level, however we find these differences to a lesser degree than those found in previous studies.

3.3. Comparison of the function types and tokens of lexical bundles

Finally, the lexical bundles used by L1-English versus L2-English professionals are compared for their typical discourse functions, based on the framework developed by Biber et al. (2004) and modified by Hyland (2008b, pp. 13–14) (see Table 4).
We adopt Hyland’s labels of ‘research-oriented’ bundles and ‘text-oriented’ bundles for the first two categories, corresponding to the categories of ‘referential’ bundles and ‘discourse-organizing’ bundles in Biber et al. (2004). However, we find Hyland’s label of ‘participant-oriented’ bundles problematic when applied to academic writing, so we have adopted instead the label ‘stance-oriented’ bundles for the third category (building on the original label of ‘stance’ bundles employed in the Biber et al. study).

Based on this functional framework, the bundles from the TELE-EN and TELE-CH were classified by two raters. The two raters first classified the bundles independently and reached 74% agreement. Then, the discrepancies were discussed to reach 100% agreement (see Appendix C). Multi-functional bundles (e.g., at the same time) are categorized according to their primary function.

As seen in Fig. 2, the two corpora display similar proportions of the three main functional categories. Text-oriented bundles (types) rank as the largest category in both TELE-EN and TELE-CH, having similar proportions at 49% and 45% respectively, whereas Stance-oriented bundles constitute the smallest proportion, with these having the higher percentage of use in TELE-CH (17%) than in TELE-EN (9%).

Table 7 presents the results of a Chi-square test, showing a significant difference in the functional distribution of bundle types, between TELE-EN and TELE-CH. However, none of the cells has an absolute value of $R$ greater than 1.96, which suggests that TELE-EN writers and TELE-CH writers do not differ much in the functional types of bundles they use.

The token distribution of functions in the two corpora is similar to the type distribution. As seen in Fig. 3, text-oriented bundles comprise the largest proportion, research bundles are also important, while stance-oriented bundles occur much less frequently.
Table 7
Standardized residuals in a Chi-square contingency table for functional distribution (types).

<table>
<thead>
<tr>
<th></th>
<th>Research-oriented</th>
<th>Text-oriented</th>
<th>Stance-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>TELE-EN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed Count</td>
<td>23</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>Expected Count</td>
<td>20.4</td>
<td>24</td>
<td>10.6</td>
</tr>
<tr>
<td>R</td>
<td>0.6</td>
<td>0.6</td>
<td>−1.7</td>
</tr>
<tr>
<td>TELE-CH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed Count</td>
<td>27</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>Expected Count</td>
<td>29.6</td>
<td>35</td>
<td>15.4</td>
</tr>
<tr>
<td>R</td>
<td>−0.5</td>
<td>−0.5</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Fig. 3. Functional distribution (tokens).

A Chi-square test of these frequencies (Table 8) shows that there are significant differences in the functional distribution of bundle tokens. The standardized residuals (R) show that TELE-EN writers use significantly more research bundles and fewer stance bundles than expected. In contrast, TELE-CH writers use significantly more stance bundles than expected.

3.3.1. Comparison of specific functional subcategories

Bundles can also be classified for their specific functional subcategories (Hyland, 2008a, p. 49). Table 9 shows significant differences in bundle tokens across the two corpora in almost all functional subcategories. TELE-CH writers use lexical bundles significantly more frequently than TELE-EN writers in most functional subcategories (description, transition, structuring and all stance-oriented bundles), but significantly less frequently than TELE-EN writers in the categories of quantification and framing subcategories.

Among subcategories in research-oriented bundles, quantification is the only subcategory that L2 writers use bundle tokens significantly less frequently than L1 writers do. As seen in Appendix C, this subcategory consists mainly of NP-based bundles. L1 writers use more types and significantly more tokens of quantification bundles than L2 writers. L1 writers use 14 types of bundles (e.g., a large number of, a wide range of), while L2 writers use only 8 types of bundles and mainly rely on the number bundles (e.g., is the number of, the average number of). Cortes (2004) also notices the absence of quantifying bundles in the student academic writing in biology (p. 415). It seems that novice writers pay less attention to the use of quantification expressions in their academic writing.

Text-oriented bundles serve the purpose of organizing the text and establishing textual cohesion, which include four subcategories: transition, resultative, structuring and framing. L1 and L2 writers are similar in their choices of the most frequently used transition bundles (on the other hand, as well as the), but L2 writers may use the bundles to convey meanings or functions different from those used by L1 writers. For example, as well as the, is misused as a conjunction in some cases, as shown in Text Sample 3.

Text sample 3: TELE-CH

Complicated focusing function calculation for squint data is saved as well as the computation efficiency improves.

In the structuring subcategory, both groups use structuring bundles to introduce the organization of the paper (e.g., paper is organized as) or refer to a certain Table or Figure as the source of the data (e.g., as shown in Figure, are shown in Table). However, three ‘section’ bundles (e.g., in the next section, in the previous section, in this section we) frequently used by L1 writers were rarely used by L2 writers. None of the ‘section’ bundles reaches the frequency criterion for inclusion. This may indicate that L2 writers do not have as strong audience orientation, as L1 writers. L2 writers seem to focus less on directing reader’s attention to other parts of the texts to facilitate understanding. However, this pattern could be related to text length since the average TELE-EN text is over twice the length of the average TELE-CH text. Longer texts typically use more signals to guide the reader through the text and thus use more ‘section’ bundles.

Framing is the only subcategory in text-oriented bundles where L2 writers use bundle tokens significantly less frequently than L1 writers do. This subcategory consists mainly of PP-based bundles (See Appendix C). L1 writers use a wider variety of framing bundles to focus readers on a given case (in the case of) or to emphasize aspects of an argument (in terms of), or specify the conditions (in the context of). In contrast, L2 writers rely heavily on a limited number of framing signals (e.g., in the
case of, with respect to the) to serve similar purposes. It is interesting to notice that, 4 out of 5 framing bundles used by L2 writers are also frequently used by L1 writers. However, the second most frequently used bundle (in the context of) by L1 writers is rarely used by L2 writers. The underuse of this bundle by novice writers also has been noticed in previous research (e.g., Römer, 2009; Chen & Baker, 2010). It is still unclear why this frequent and useful phrase in academic writing is generally neglected by novice writers. Future research is needed to determine why.

The Stance-oriented category is an area that L2 writers are reported to struggle. Hyland and Milton (1997) find that L1 Chinese student writers differ significantly from L1-English writers, relying on a more limited range of items, offering stronger commitments, and exhibiting greater problems in conveying an appropriate degree of doubt and certainty (p.183). In our study, we find that L2 professionals use a wider range of stance bundles and demonstrate better control expressing the degree of doubt and certainty. However, they still use two certainty bundles (it is obvious that, it is clear that) to make strong assertions without providing sound proof, as in Text Sample 4. On the other hand, L2 professionals also use tentative bundles (is assumed to be, we assume that the) to demonstrate control of the degree of their certainty. Though L2 writers still have certain difficulty in presenting their propositions in appropriate ways, they are at a more advanced level than L2 student writers. It is interesting to note that L2 professionals tend to use evaluative bundles (it is difficult to, it is easy to) to present their personal opinion on a proposition (see Text Sample 5), however, the use of subjective adjectives in academic prose may diminish the author’s credibility. These differences may make L2 writers vulnerable to violating discourse norms by making their writing appear too personal.

Text sample 4: TELE-CH

It is obvious that the proposed method is of strong robustness.

Text sample 5: TELE-CH

First, it is easy to build power line network because no new wire is needed.

3.4. Comparison of the relationship between structural and functional categories

Biber et al. (2004) find that there is a strong relationship between structural type and discourse function for lexical bundles. In particular, most stance bundles are composed of dependent clause fragments, while most referential bundles are composed of noun phrase or prepositional phrase fragments (p.39). In our study, we compare the numbers of bundle types with different structures in each functional category to find out whether such a relationship is found in L1 and L2 research writing.

In Fig. 4, we see that most TELE-EN research bundles are composed of noun phrase fragments, while the TELE-CH research bundles are composed a combination of noun phrase fragments and verb phrase fragments. The majority of the text-oriented bundles in TELE-EN are prepositional phrase fragments, while the majority of the text-oriented bundles in TELE-CH are verb phrase fragments. The Stance-oriented category in both corpora is dominated by one structural category - verb phrase fragments. L1 writers and L2 writers employ bundles of different structures to serve similar functions in their research writing. L2 writers tend to use more VP-based bundles especially passive bundles to describe research. In particular, they seem to focus more on definitions, concepts and classifications (can be defined/described/expressed/written as). They also rely on VP-based bundles to organize texts and develop arguments (e.g., are shown in figure, are shown in table, are the same as, is organized as follows). The variation between L1 and L2 use of formulaic language in academic writing is an area in which L2 writers sometimes deviate from that of more proficient L1 writers.

4. Conclusion and implications

Our study investigated the structural and functional patterns of lexical bundles of L1 versus L2 professionals in published research articles from a single academic discipline (Telecommunications). The study shows that there are major structural differences (phrasal vs. clausal) between L1 and L2 writers who are experts in their academic discipline even at the professional level. L2 professionals use bundles consisting mostly of verbs and clause fragments (especially passive verb structures), while L1 professionals use bundles consisting of noun phrase and prepositional phrase fragments. Comparing our results to Biber et al. (1999), and Chen and Baker (2010) indicated that such phrasal-clausal differences exist between L1 expert academic writers and all other groups: L1 novice writers, L2 novice writers, and L2 expert academic writers. This

Table 8

<table>
<thead>
<tr>
<th></th>
<th>Research-oriented</th>
<th>Text-oriented</th>
<th>Stance-oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TELE-EN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed Count</td>
<td>800</td>
<td>895</td>
<td>150</td>
</tr>
<tr>
<td>Expected Count</td>
<td>739.4</td>
<td>899.9</td>
<td>205.7</td>
</tr>
<tr>
<td>R</td>
<td>2.2</td>
<td>−0.2</td>
<td>−3.9</td>
</tr>
<tr>
<td><strong>TELE-CH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed Count</td>
<td>1001</td>
<td>1297</td>
<td>351</td>
</tr>
<tr>
<td>Expected Count</td>
<td>1061.6</td>
<td>1292.1</td>
<td>295.3</td>
</tr>
<tr>
<td>R</td>
<td>−1.9</td>
<td>0.1</td>
<td>3.2</td>
</tr>
</tbody>
</table>
finding lends support to Biber et al.’s (2011) hypothesis that both L1 and L2 academic writers may follow a developmental progression from clausal to phrasal styles. The present study indicates that L1 professionals have a lead over L2 professionals in the transition to expert writers as measured by phrasal features.

L1 and L2 professionals do not differ much in the proportions of the functional distribution of bundle types and tokens. However, L2 professionals use significantly fewer research-oriented bundles and more stance-oriented bundles than L1 professionals. This study also demonstrates that L2 professionals use specific bundles in ways that are functionally different from those of the L1 professionals and even misuse certain bundles. Our study indicates the neither dichotomy (L1 versus L2, or novice versus expert) is adequate to explain the differences found. It is not simply a matter of L1 vs L2 and not simply a novice vs expert distinction, but the interplay of L1 and expertize as we show differences at the expert level. It seems that English L1 apparently has an important difference in ability to develop this discourse style.

Based on this finding, our study has potential pedagogical implications for teaching English for academic writing for professionals. Teachers need to pay attention to the structural patterns of academic writing, to help students and L2 professionals transition from a clausal style to a phrasal style in order to better integrate information in their academic writing. Teachers may also focus on the use of specific bundles, through contextualized examples.

Acknowledgments

We thank the anonymous reviewers and JEAP editors for their helpful comments. This research is supported by National Social Science Foundation of China (No. 14BYY148).

1. We used Paul Rayson’s Log likelihood calculator from http://ucrel.lancs.ac.uk/llwizard.html.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jeap.2015.11.003.

References

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. University of Sydney Papers in TESOL.

Fan Pan is an associate professor of Applied Linguistics at School of Foreign Languages, Huazhong University of Science and Technology, P.R. China. Her research interest includes corpus linguistics and second language academic writing.

Randi Reppen is Professor of Applied Linguistics and TESL at Northern Arizona University where she teaches in the MA TESL and Applied Linguistics Ph.D. programs. Her work revolves around the use of corpus linguistics to gain knowledge about language use and to inform language teaching.

Douglas Biber is Regents’ Professor of English (Applied Linguistics) at Northern Arizona University. His research on corpus linguistics, English grammar, and register variation has resulted in over 200 research articles, 8 edited books, and 14 authored books and monographs.