Journal of English for Academic Purposes 52 (2021) 100999

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

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

# Is core vocabulary a friend or foe of academic writing? Single-word vs multi-word uses of THING

Sylviane Granger [a], Tove Larsson [a, b, *]

[a] Centre for English Corpus Linguistics, University of Louvain, Belgium
[b] Department of English, Uppsala University, Sweden

ARTICLE INFO

*Article history:*
Received 3 December 2020
Received in revised form 22 March 2021
Accepted 23 March 2021
Available online 28 March 2021

*Keywords:*
Core vocabulary
L2 writing
Phraseology
Multi-word units
THING

ABSTRACT

Core vocabulary items (e.g. *thing*, *way*) are often viewed as the enemy of effective academic writing, and style guides and textbooks often advise against using them. However, their bad reputation seems to stem from a single-word perspective that ignores the rich phraseological units that such items tend to figure in. In this study, we focus on the core vocabulary lemma THING to investigate the extent to which a phraseological approach can redeem its reputation. We look at learner essays from ten different first-language backgrounds from the *International Corpus of Learner English* and compare these to reference corpora from the endpoints of the informal-formal continuum: the Spoken BNC2014 and the Corpus of Academic Journal Articles. The results show that a phraseological approach indeed provides a more nuanced view of the core lemma THING: it is used in a wide variety of multi-word units, many of which common in academic writing. Although some signs of novice production are evident in the learners' writing, their use is closest to that of the expert academic writers. The paper concludes with a discussion of the role of phraseology in vocabulary lists used in teaching and assessment.


## 1. Introduction

The notion of core (or basic) vocabulary is widely used in applied language studies. The first core vocabulary list, West's *General Service List* (GSL), was compiled as early as 1953 and has since proved its usefulness for both teaching and testing purposes. The rationale that underpinned the GSL was that a lexical repertoire consisting of the most basic 2,000 words of English would be a good basis for learning English as a foreign language. Although the list is now dated, interest in core vocabulary has continued unabated and several new core vocabulary lists have been compiled, among them the *New General Service List* (NGSL) (Brezina & Gablasova 2015), which relies on the frequency of words in large electronic corpora of speech and writing. However, while lists of core words have proved their worth, they suffer from one major weakness: they include only single words. They thus disregard the wide range of productive multi-word units that these high-frequency words tend to generate, many of which "are as frequent as or more frequent than single items which everyone would agree must be taught" (O'Keeffe et al., 2007: 46).

Although core vocabulary is seen in a very positive light for general language purposes, it is commonly regarded as the enemy of effective academic writing. The acquisition of academic writing skills is seen as involving a major "vocabulary shift"

* Corresponding author. Department of English, Box 527, Uppsala University, SE-75120, Uppsala, Sweden.
*E-mail address:* tove.larsson@engelska.uu.se (T. Larsson).

https://doi.org/10.1016/j.jeap.2021.100999

(Swales & Feak 2004: 18), i.e. the replacement of high-frequency, informal words by less frequent, formal alternatives. Most academic vocabulary lists purposely exclude the 2,000 core words of the GSL. The bad reputation of core vocabulary in English for Academic Purposes (EAP) seems to relate to the single-word-based approach to core words and may not be justified in the case of an approach that takes multiword units into account.

Against this background, the main objective of our study is to investigate the extent to which the negative reputation of core words is still warranted when their phraseology is built into the analysis. To achieve this objective, we focus on the lemma THING, which epitomises the notion of coreness, as outlined further down.

The article is structured as follows. Section 2 defines and explores the notion of core vocabulary. After a description of the corpus data and methodology (Section 3), we examine the frequency of single vs multi-word uses of THING in learner writing and reference corpora of native speech and writing (Section 4), before investigating the discourse functions they serve (Section 5). Section 6 investigates similarities and differences in the use of THING phrases across different learner populations. Section 7 brings together the main threads of the study and provides some pedagogical recommendations.

## 2. Core vocabulary

Core vocabulary is elusive and difficult to define. Lee (2001) lists no fewer than seven conceptions of core vocabulary, each of which relies on one distinctive criterion (high frequency in the language as a whole, a particular medium or demographic grouping, genericity, saliency, range and definition value). While some linguists rely on definitions involving only one criterion (usually frequency), most resort to multi-criteria definitions. For example, Nation and Hwang's (1995: 35) notion of "general service vocabulary" relies on frequency and range: "General service vocabulary consists of words that are of high frequency in most uses of the language. It is the essential common core. (…) General service words occur frequently across a wide range of texts." Papp and Nicholson's (2011: 16) definition involves frequency, range and definition value. They define core words as "frequent words that are widely and relatively evenly distributed among texts of different kinds, and words that can be used to define other words."

The purpose of identifying core vocabulary is closely linked to teaching. The GSL has had – and to some extent, still has – a tremendous impact on textbook design and pedagogical lexicography. Core words and their counterparts (i.e. advanced/sophisticated words) also figure prominently in vocabulary assessment. Learners' level of lexical sophistication (or its opposite, basicness) is measured as the ratio of sophisticated (or basic) word types to the total number of word types (Wolfe-Quintero et al., 1998: 102). This measure can be automated with lexical profiling software such as Cobb's *VocabProfilers* (https://www.lextutor.ca/vp/), which breaks down the words used in texts into frequency bands (the first 1,000 most frequent words, the second 1,000 words, etc.) and computes the percentage of words in each band. Laufer's (1995) 'beyond 2000' measure relies on the difference between core (the first 2,000 words) and non-core (beyond 2,000) words. As observed by Read (2000: 204), "[t]he 'beyond 2000' percentage is in fact an alternative way of calculating lexical sophistication. Since the profile always adds up to 100 per cent, more words beyond the 2000-word level inevitably means a smaller proportion of the high-frequency words". A wide range of studies have shown that the more core words learners use, the less proficient they are. Conversely, a high proportion of non-core words has been found to be an indicator of more advanced proficiency (Crossley et al., 2013; Laufer & Nation, 1995). Comparisons of learner and native writers, such as Hasselgren's (1994) analysis of Norwegian learners' writing, show that "core words—learnt early, widely useable, and above all safe (…) are hugely overused, even among learners sufficiently advanced to have been weaned off them" (Hasselgren, 1994: 250). As such words tend to be particularly frequent in spoken production, they are often perceived as informal and as a sign of novice writing that needs to be remedied.

The need to wean learners off core vocabulary is particularly in evidence in the teaching of language for academic purposes. To help students and teachers of academic writing, several lists of academic words have been created, the most widely used being Coxhead's (2000) *Academic Word List* (AWL), which contains lexical items beyond the top 2,000 core words of the GSL that occur frequently across a wide range of academic material. Although exclusion of the GSL is justified to some extent, in that many core words are rarely used in EAP, Paquot (2007) has demonstrated that many core words hold great academic potential. Her *Academic Keyword List* (AKL) makes no a priori exclusion of GSL words and as a result includes many words that are absent from Coxhead's list (e.g. *reason, result, discuss, namely*). The same holds true for Gardner and Davies's (2014) *Academic Vocabulary List*.

Lists of both core words and academic words have proved their worth in teaching and assessment. However, as pointed out by Lindqvist et al. (2013: 122), the exclusion of multi-word units from vocabulary lists is likely to have a major impact on lexical assessment. The words contained in multi-word units may belong to different frequency bands, and "[t]reating these words separately means that the number of words categorised as highly frequent will rise, although this may not correspond to the frequency of the whole expression in the target language input".

Marx once said that religion was the opium of the masses. **Right now** – **at the end of** the twentieth century – we could replace religion with television. The church used to preach – and still preaches – dogmas and rules which no one thought about and which were accepted by practically everyone**. A lot of** people did not mind being in this situation. Now, **in a way**, this role has been **taken over** by television. Fundamentally, religion **as well as** television are good things. They were – and still are – necessary they could – and still can – enrich a person mind and personality. Unfortunately both have been misused by man.

Television **started out** as merely a **means of communication**, with a little entertainment **every now and again**. **Over the years** it has developed into something else now the emphasis seems to be on entertaining people, **rather than** on informing and educating them. Entertainment on TV is not necessarilly **a bad thing**: everyone needs to relax **from time to time.** But we have **come to the point of** accepting anything that is on. **A lot of** people **make a habit of turning on** the television set.

**Fig. 1.** Multi-word units in a learner text.

The text excerpt in Fig. 1[1] will serve to illustrate this point. The highlighted multiword units are examples of words that commonly co-occur.[2] Although several of them are clearly not "basic", the learner who wrote this text would not be rewarded for using them, as the vast majority of the words that compose them are counted as K1 words (i.e. the first 1,000-word frequency band) by single-word-based lexical profiling software such as Cobb's *VocabProfilers*.

Despite the fact that many vocabulary specialists are aware of this shortcoming (Lee, 2001; Cobb, 2013; Lindqvist et al., 2013, Brezina & Gablasova 2015), this awareness has not yet been translated into the inclusion of multi-word units in vocabulary lists (with the notable exception of the English Vocabulary Profile Wordlists, see Section 6). It has, however, led some linguists to compile phrasal lists, i.e. lists solely made up of multi-word units, such as Martinez and Schmitt's (2012) *Phrasal Expressions List*, which contains 505 frequent non-transparent multi-word expressions in English, intended in particular for receptive use. There have been several initiatives to produce lists of word combinations typical of academic discourse (Ackermann & Chen, 2013; Durrant, 2009; Simpson-Vlach & Ellis, 2010). The lists assembled by Durrant (2009) and Ackermann & Chen (A&C) (2013) both contain typical EAP collocations (e.g. *significant difference, vary widely*), but they have been compiled on the basis of very different criteria, and as a result show very little overlap (see Granger, 2017 for a more detailed comparison). Simpson-Vlach and Ellis's (2010) *Academic Formulas List* differs markedly from the preceding two in that it comprises lexical bundles, i.e. the most frequent recurrent sequences of contiguous words (e.g. *on the other hand, it should be noted*). It must be admitted, however, that compared with single-word academic lists such as the AWL, phrasal academic lists have not had a great impact on EAP teaching and assessment to date.

This is arguably unfortunate: a single-word approach may lead teachers and language assessors to disregard multi-word units present in students' texts, as illustrated in Fig. 1. Put another way, it is not unreasonable to assume that core vocabulary items have an undeservedly bad reputation in academic writing, which could be redeemed if their phraseology were taken into account. That is to say, if core vocabulary items are in fact part of multi-word units that occur in expert academic writing and/or speech, then dismissing these items as too basic when we look at them in isolation in novice writing would lead us to miss multi-word units that could be helpful for teaching and assessment. In this study, we test this assumption empirically by looking more closely at a core vocabulary item that can be said to have a particularly bad reputation in the academic context, namely THING, in order to see whether a phraseological approach could help vindicate its reputation. We focus on phraseological uses in second-language (L2) writing compared with written and spoken expert data and discuss some general implications for L2 teaching and assessment. THING stands out as an especially interesting core vocabulary item to investigate as it has been shown to be very frequent in student writing (Ringbom, 1998; Tåqvist, 2016), and is often mentioned in academic resources as an example of a word to be avoided (e.g. Bailey, 2011: 152).[3]

[1] This sample text was extracted from the Dutch subcorpus of the International Corpus of Learner English (Granger et al., 2020).

[2] The highlighted multiword units were identified manually. They represent a whole range of 'chunks of language', i.e. 'conventionalized form/function composites' that occur more frequently than newly minted word combinations (Nattinger & DeCarrico 1992: 1). This definition is vague, and our selection in Fig. 1 is therefore somewhat subjective. In our study we used a lexical bundle approach that identifies a category of multiword units in a more systematic manner.

[3] This warning is particularly in evidence on web-based academic resources, where the word THING is almost systematically listed as a word to be avoided at all costs (cf. e.g. https://www.academic-englishuk.com/academic-style and https://www.eapfoundation.com/writing/style/).

**Table 1**
Data used in the present study.

| Corpora | Words | Texts |
|---|---|---|
| CAJA | 83,544,346 | 13,116 |
| BNC2014 | 11,422,617 | 1,251 |
| ICLE-10 Total | 2,269,734 | 3,408 |
| *ICLE-IT* | *231,420* | *398* |
| *ICLE-SE* | *281,005* | *472* |
| *ICLE-RU* | *227,691* | *274* |
| *ICLE-PO* | *237,842* | *366* |
| *ICLE-NO* | *213,428* | *316* |
| *ICLE-GE* | *241,057* | *445* |
| *ICLE-FR* | *206,291* | *314* |
| *ICLE-FI* | *194,146* | *261* |
| *ICLE-DU* | *234,738* | *262* |
| *ICLE-BU* | *202,116* | *300* |
| **Total** | **97,236,697** | **17,775** |

## 3. Data and method

### 3.1. Data

The learner data come from the third version of the *International Corpus of Learner English* (ICLE; Granger et al., 2020), which is made up of argumentative texts written by students from a large number of different first-language (L1) backgrounds. Using the online interface (https://corpora.uclouvain.be/cecl/icle/), we extracted data from ten different L1 backgrounds where a majority of the texts rated were assessed as C1 or higher on the Common European Framework of Reference for languages (CEFR) scale (Granger et al., 2020: 12).[4] The L1 backgrounds included are Italian (IT), Swedish (SE), Russian (RU), Polish (PO), Norwegian (NO), German (GE), French (FR), Finnish (FI), Dutch (DU), and Bulgarian (BU). The inclusion of these particular L1s enabled us to study a typologically diverse set of languages, in that four different language families were represented in our subset. In total, approximately 2.3 million words from 3,400 texts were included. ICLE stood out as the best choice among available learner corpora for two main reasons: (i) the texts are homogeneous in terms of register (argumentative writing) and (ii) it allowed us to include a wide variety of L1 backgrounds from writers at comparable proficiency levels.

To situate the learners' usage on the informal-formal continuum (see, e.g., Larsson & Kaatari, 2019), two reference corpora were used to represent the endpoints on this continuum: the Spoken BNC2014 (BNC2014; Love et al., 2017) and the *Corpus of Academic Journal Articles* (CAJA; Kosem, 2010). BNC2014 was compiled between 2012 and 2015 (McEnery et al., 2017: 312), and contains spoken British English conversation data from a range of socio-economic and geographical backgrounds (Love et al., 2017). CAJA comprises approximately 83.5 million words (13,000 articles) from 28 different disciplines (Kosem, 2010: 100). The articles come from over 2,000 different high-ranking international journals (Kosem, 2010: 107–109). An overview of the three corpora and subcorpora included is provided in Table 1.

### 3.2. Method

Investigations of formulaic language and phraseology have been carried out using a wide variety of different techniques and approaches, focusing on several types of phraseological unit, including collocations, colligations, compounds, P-frames and lexical bundles. Academic language is known to be highly formulaic (e.g. Durrant & Mathews-Aydınlı, 2011), a characteristic that also extends to learner language (e.g. Granger, 2017). In the present study, we are particularly interested in formulaic, routine uses of learner language in an academic register. We therefore opted for an approach that would allow us to investigate longer multi-word units at the fixed end of the continuum, namely lexical bundles. Lexical bundles, defined as sequences of contiguous words that recur in a particular register (Biber et al., 1999: 990–1024), have been used extensively in investigations of learner writing to look at differences pertaining to L1 background and proficiency (e.g. Chen & Baker, 2010; De Cock, 2000). Their importance in academic writing in particular has been stressed, as a lexical bundle approach helps identify routine uses of language typical of academic writing (Gilquin et al., 2007).

Specifically, we focus on four-word bundles containing THING (in the singular or plural), as manual investigations of the data indicated that this bundle size would best enable us to balance frequency and "noise" (i.e. substantively irrelevant hits, such as *thing and that*). Whereas investigations of longer bundles tend to yield very low frequencies with very little noise, searches for short bundles conversely yield high frequencies with a high proportion of noise. Limiting the investigation to one type of bundle (as opposed to a size range) furthermore facilitates a frequency-based investigation of the data since the frequencies are not boosted by nested bundles (e.g. instances where a three-word bundle is fully contained in a four-word bundle). However, we still allowed for partial overlap between four-word bundles in our data, so as to enable nuances to remain

[4] Based on a sample of 20 randomly selected texts that were assessed as part of the ICLE compilation process (Granger et al., 2020: 12).

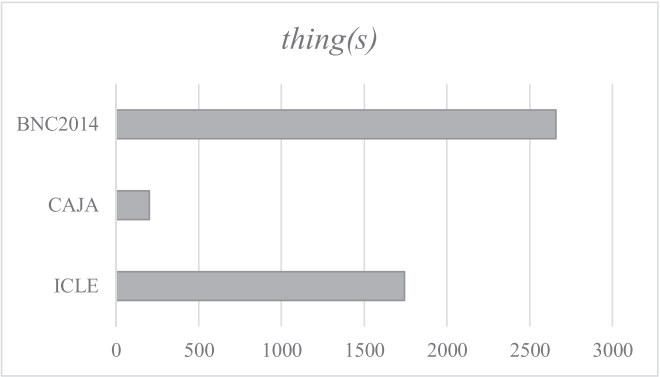


**Fig. 2.** Frequencies per million words of all uses of THING in the corpora.

visible, with the rationale that while it might be possible to merge some instances of specific bundles, this is not the case for all instances of these particular bundles. For example, it might seem as if the bundles *no such thing as* and *such thing as a* could be merged into one five-word bundle (*no such thing as a*); however, our data cautioned us against applying such an approach, as there were instances where these bundles could not be merged into a five-word bundle, as the examples in (1) and (2) show (emphasis has been added throughout).

(1) there is **no such thing as** absolute safety in a city (CAJA_Archit_18_2001_kitchen)
(2) it is difficult to believe that there was any **such thing as a** unified Burgundian identity (CAJA_Art_31_2008_rothstein)

Furthermore, as our corpora differed considerably in size and since the raw frequencies from the learner corpus were relatively small,[5] we opted for an approach where we looked at the highest-ranked (i.e. the top 15 most frequent) THING bundles in each corpus, rather than attempting to extract comparable lists from each corpus based on cut-off points for frequency and range (see, e.g., Cortes, 2004). We believe that this method enabled a fairer comparison between the corpora, especially since the study does not set out to provide a comprehensive account of all THING bundles but rather to look at the learners' use as compared with the use in the corpora representing the two endpoints of the (in)formality continuum.

The data and results were processed using R (R Core Team, 2020). Two different exploratory statistical techniques were applied in order to facilitate interpretation of the results: cluster analysis and correspondence analysis. Cluster analysis enables the elements that behave most similarly to cluster together in a tree-based structure most often displayed through a dendrogram; the higher up they appear in the graph, the more dissimilar the two branches are (e.g. Gries & Otani, 2010). Similarly, correspondence analysis, which is used on categorical data, detects and represents underlying structures in the data. It does so by representing the data points on a two-dimensional plane where proximity indicates similarity (e.g. Baayen, 2008; Glynn, 2014).

In the subsequent three sections, we present the results of our investigation of the role played by the core word THING and the numerous multi-word units it occurs in. We begin by giving an overview of what a frequency-based, single-word approach can tell us about where on the informal-formal (i.e. speech vs academic writing) continuum the learners are situated and then move on to give an account of what additional information a multi-word approach can provide on this matter. After that, we zoom in on the bundles in the written learner and expert data to investigate what functions the bundles serve in academic discourse. Finally, we look at the different L1 groups separately to compare their usage. Our research questions are as follows:

- What can a single-word vs multi-word approach tell us about where on the informal-formal continuum the learners' use is situated (Section 3)?
- What functions do the bundles identified through a multi-word approach serve in academic discourse (Section 4)?
- Are there differences in bundle use between the different L1 groups in the learner data (Section 5)?

## 4. A single-word vs multi-word approach

In terms of overall frequencies in the three corpora, the results show that THING is more strongly associated with the informal, spoken register than with the formal, written register: 2,657 instances per million words (pmw) were found in the

[5] The bundle with the lowest frequency that made it into the analysis, *the best thing to*, comes from the learner data and had a frequency of 14. It should also be mentioned that the bundles were well spread out across the texts; for example, *the best thing to* occurred in 14 different texts.

**Table 2**
Top 15 most frequent 4-word bundles with THING in the three corpora.

| | CAJA | | | BNC2014 | | | ICLE | | |
|---|---|---|---|---|---|---|---|---|---|
| Rank | Bundle | Raw | Per 10mw | Bundle | Raw | Per 10mw | Bundle | Raw | Per 10mw |
| 1 | other things being equal | 150 | 18 | that's the thing | 632 | 553 | the most important thing | 77 | 301 |
| 2 | no such thing as | 120 | 14 | and things like that | 418 | 366 | a lot of things | 45 | 176 |
| 3 | is no such thing | 96 | 11 | that sort of thing | 329 | 288 | the only thing that | 39 | 152 |
| 4 | the same thing as | 76 | 9 | that kind of thing | 314 | 275 | is a good thing | 25 | 98 |
| 5 | the only thing that | 75 | 9 | but the thing is | 289 | 253 | most important thing is | 23 | 90 |
| 6 | one of the things | 67 | 8 | one of the things | 159 | 139 | the first thing that | 19 | 74 |
| 7 | of the things that | 63 | 8 | thing isn't it | 156 | 137 | one of the things | 17 | 66 |
| 8 | it is one thing | 63 | 8 | a lot of things | 143 | 125 | of the things that | 16 | 63 |
| 9 | such a thing as | 58 | 7 | the thing is I | 142 | 124 | important thing is that | 16 | 63 |
| 10 | is one thing to | 54 | 6 | one of those things | 141 | 123 | no such thing as | 16 | 63 |
| 11 | is a good thing | 54 | 6 | the only thing I | 120 | 105 | important thing is to | 15 | 59 |
| 12 | such thing as a | 52 | 6 | well the thing is | 119 | 104 | thing to do is | 15 | 59 |
| 13 | right thing to do | 51 | 6 | sort of thing yeah | 111 | 97 | and the only thing | 14 | 55 |
| 14 | the right thing to | 49 | 6 | the thing is that | 111 | 97 | most important thing in | 14 | 55 |
| 15 | the kind of thing | 49 | 6 | 's the only thing | 109 | 95 | the best thing to | 14 | 55 |

spoken data, compared with 201 pmw in the data from the expert academic writers. While the frequencies in the learner data (1,745 pmw) are situated between the two, the learners' use is much closer to the spoken data than to the written data, as shown in Fig. 2. This tendency among the learners is in line with Tåqvist's (2016: 96) finding that learners (L1 Swedish in this case) tended to make very frequent use of *thing*, noting that in the learner data, it is "almost ten times more frequent than in the expert corpus".

Based solely on this approach, which does not take phraseology into consideration, we would have to conclude that THING is associated primarily with spoken conversation and therefore may deserve its somewhat tarnished reputation as a word that is best avoided in academic writing (e.g. Swales & Feak, 2012: 15). That said, it is not the case that THING is *never* used in expert academic writing, as style guides and textbooks may have us believe. Nonetheless, as the learners' frequencies are closer to those of spoken conversation, there would seem to be evidence to suggest that the learners are being overly informal and vague in their writing.

However, a closer look at the data through a phraseological lens suggests that this single-word approach does not tell the whole story. In fact, the results from the lexical bundles analysis show that THING occurs in a wide variety of high-frequency multi-word units in all three corpora, many of them with precise discourse functions, as we shall see in the next section. Examples of four-word bundles are shown in (3) and (4). An overview of the 15 most frequent bundles in each corpus can be found in Table 2. As is clear from the list, some of the bundles are partially overlapping (e.g. *no such thing as* and *is no such thing*; *and the only thing* and *the only thing that*).

(3) […] in other words, absence of evidence is not **the same thing as** evidence of absence. (CAJA_Educ_27_2006_mackey_Education)
(4) **Other things being equal**, the higher a vowel, the higher is its pitch. (CAJA_Ling_10_2006_plag)

As can be seen, there are clear differences between the registers. The expert academic writers make use of bundles such as *other things being equal* and *no such thing as*, whereas the spoken data contains bundles such as *that's the thing* and *things like that*. We also note that the learners' use is somewhat less varied than that of the other two groups: a third of the top 15 bundles in the learner data include *important thing*. Overreliance on so-called *phraseological teddy bears* is a commonly noted feature of learner language (Hasselgård, 2019) and seems to be present in our data as well.

With regard to the question of where on the informal-formal continuum the learners are situated, a Correspondence Analysis created using the R package *LanguageR* (Baayen & Shafaei-Bajestan, 2019) shows that the learners' use is in fact more similar to that of the academic experts than to the spoken data when a multi-word approach is applied (Fig. 3).

The graph shows further that the two corpora chosen to represent endpoints on the informal-formal continuum are indeed clearly different from one another when it comes to phraseological uses of THING,[6] which suggests that the bundles are in fact highly register-specific.

The patternings in the data summarised in the correspondence analysis graph can largely be explained by the fact that there is more overlap between the learners' and the expert academic writers' bundles than between the bundles in these two corpora and the spoken data. In fact, a third of the bundles (5/15) are shared between the learners and the expert academic

[6] This technique provides a visual overview of the bundle clusters and the relative degree of overlap found in the data (see Fig. 4 for an overview of the bundles). Functionally, the first dimension can be viewed as illustrating the formal−informal continuum, and the second dimension as primarily concerned with bundles used to stress importance.

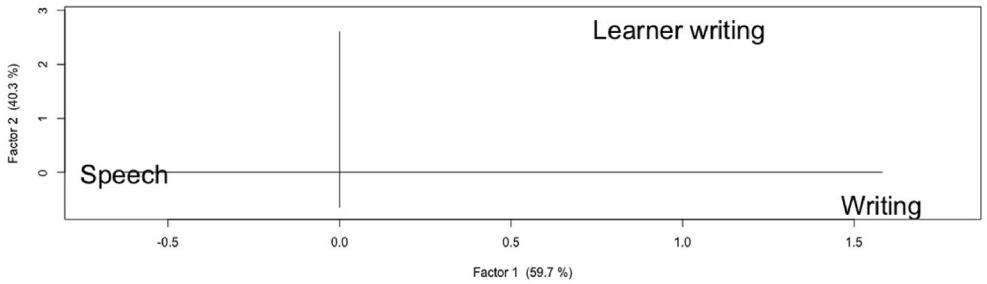


**Fig. 3.** Correspondence Analysis of the top 15 bundles in each corpus.

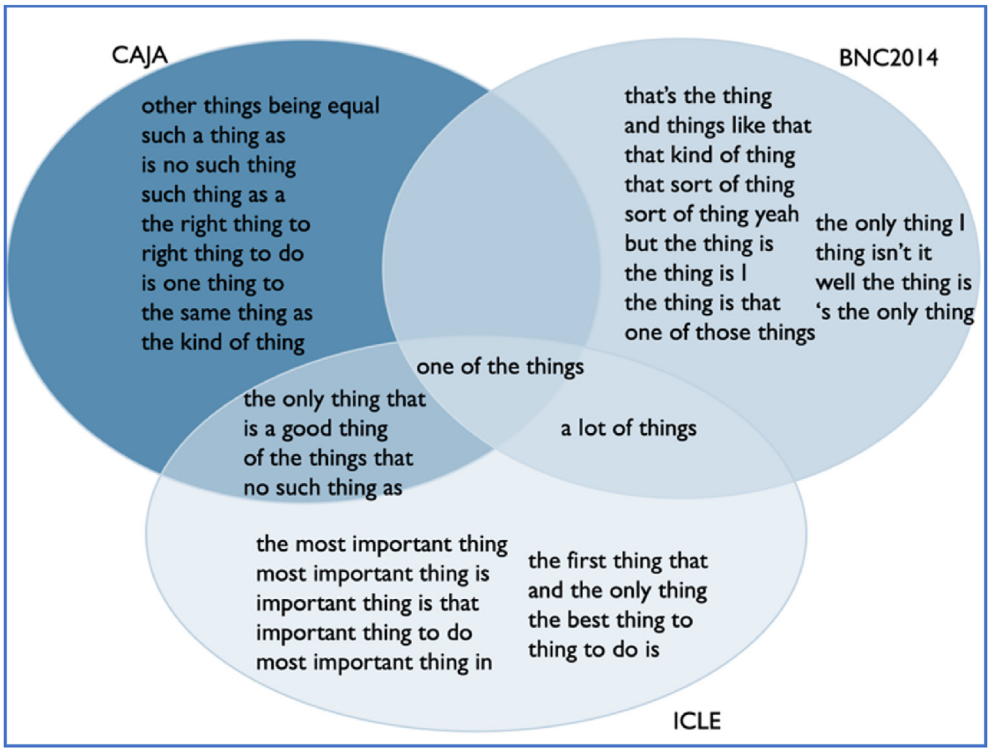


**Fig. 4.** Overlap between the three corpora.

writers, whereas 2 out of 15 are shared between the learner data and the spoken data. Only one bundle is shared between the academic writing data and the spoken data: *one of the things*. Fig. 4 summarises the overlap between the corpora.

Thus, while a single-word, frequency-based approach that did not take phraseology into account suggested that the learners' use of THING is more informal than formal, the phraseological approach indicated that there is more to the story than that, at least for the most fixed formulaic multi-word units. When we take phraseology into consideration, the learners' usage is actually more similar to that of the academic experts than to the spoken data. It thus appears that phraseology can help provide a more nuanced and fairer view of second-language use.

While the results of the present section have helped us debunk the myth that core vocabulary items tend to be avoided in expert academic writing, we have yet to address the criticism voiced against such items (and in particular THING) to the effect that they are overly vague and imprecise for academic writing. For this purpose, we investigated the THING bundles in the expert academic writing and the learner writing in order to explore what functions they serve in academic discourse.

## 5. Discourse functions

Lexical bundles have been found to serve a number of different discourse functions. Bundles are often formed around so-called *metadiscursive nouns* (e.g. *fact*, *belief*) and used as discourse organisers to help shape readers' understanding of a text

(cf. Jiang & Hyland, 2017; Tahara 2020). Similarly, THING bundles have been quoted as being used for discourse management in spoken academic English, linking "prior to immediate utterances" (e.g. *the thing is*) (Swales, 2001: 35).

In isolation, however, THING has a reputation for being vague and imprecise. Echoing Ringbom's (1998) description of *thing* as "a particularly vague noun", Tåqvist (2009: 96) states that "*thing* is arguably the least specific and the most neutral of all the DONs [discourse-organizing nouns]". Nonetheless, once again, our results show that a phraseological approach can be used to nuance the picture. Indeed, a closer look at the bundles shows that while THING in isolation may be neutral and vague, the phraseological units it occurs in serve a wide variety of precise discourse functions in academic writing, as discussed in Swales (2001) in relation to the use of THING in academic speech. Nonetheless, as not all bundles with THING are equally useful or precise, we focus here on four particularly widely used functions in the learner and expert academic writing: contrast/opposition marking, evaluation, comparison and emphasis.

The writing of both groups contained THING bundles used to mark contrast/opposition and evaluate claims. Examples of contrast/opposition bundles include *it is one thing* and *no such thing as*, as in (5) and (6). *It is one thing* gives the writer a chance to engage the reader in an on-the-one-hand/on-the-other-hand line of argumentation, whereas *no such thing as* helps the writer to take a strong stance against a concept or argument.

(5) **It is one thing** for an object to be beautiful, and another thing for it to express aesthetic ideas. (CAJA_Art_47_2007_murray)
(6) There is **no such thing as** a real world. (ICLE-NOUO1058)

Evaluation bundles include *is a good thing*, *right thing to do*, and *the best thing to*, as in (7)–(9), all of which express the writer's (positive) evaluative stance towards the proposition put forward. Similarly, Swales (2001: 48) in his analysis of academic speech noted that THING often has positive prosody.

(7) And this **is a good thing**, not a waste of resources that could have been avoided if they had settled their differences before the fact (CAJA_Hist_35_2006_dagostino)
(8) Moral behaviors must be freely chosen by the agent and chosen because the agent believes it is the **right thing to do.** (CAJA_Educ_1_2006_barrow)
(9) **The best thing to** do, if the government wanted to lower the consumption of alcohol and tobacco, would not be to raise the taxes on those articles. (ICLE-SWUL7047)

However, other functions, most notably comparison, were only found among the most frequent bundles in the expert writing.[7] Examples include *other things being equal* and *the same thing as*, as in (10) and (11). The bundle *other things being equal* helps writers to position themselves vis-à-vis a claim with the caveat that their stance is based on the assumption that all other factors are (and will remain) the same. Through the use of the bundle *the same thing as*, the writer can equate two concepts or ideas.

(10) Therefore, **other things being equal**, a firm operating in circumstances in which productivity is a function of the wage rate is more likely to give in to demands for wage increases. (Fin_38_2006_bhalotra)
(11) Since monomorphisms are **the same thing as** injective homomorphisms of NM-algebras, we are done. (CAJA_Comp_17_2007_aguzzolietal_Computer science)

In the learner data, THING bundles were very often used to place emphasis on a proposition. Bundles including *important* + *thing* were especially common in the learner data, in particular *the most important thing*,[8] as exemplified in (12) and (13). This is in line with Tahara's (2020) study in which the (L1 Japanese) learners were found to rely more heavily on this kind of bundle than their native-speaker peers. We will return to this bundle in the next section.

(12) Nowadays, it seems as if money is **the most important thing** on earth. (ICLE_DBAN2034)
(13) **the most important thing** is to keep the discussion alive (ICLE_SWUL9014)

In sum, as can be seen from these examples, THING bundles are often highly specified functionally. We are of course not suggesting that all THING bundles should be taught in English for Academic Purposes classes, but our results show that teaching students to *avoid* using core vocabulary items completely would be reductive and counterproductive. That is, if we ignore high-frequency core vocabulary items on the grounds that they are too basic or vague, we will miss the specified and often very useful multi-word units that they occur in.

[7] While bundles of this kind did not feature in the top 15 most frequent bundles in the learner data, *the same thing as* can be found further down the list, as the 35th most frequent bundle; however, *other things being equal* does not occur even once in the learner data.
[8] Although much less frequent in the expert data than in the learner data (0.51 times pmw vs 30.1 times pmw), *the most important thing* is also found in the expert data, albeit not among the 15 most frequent bundles.

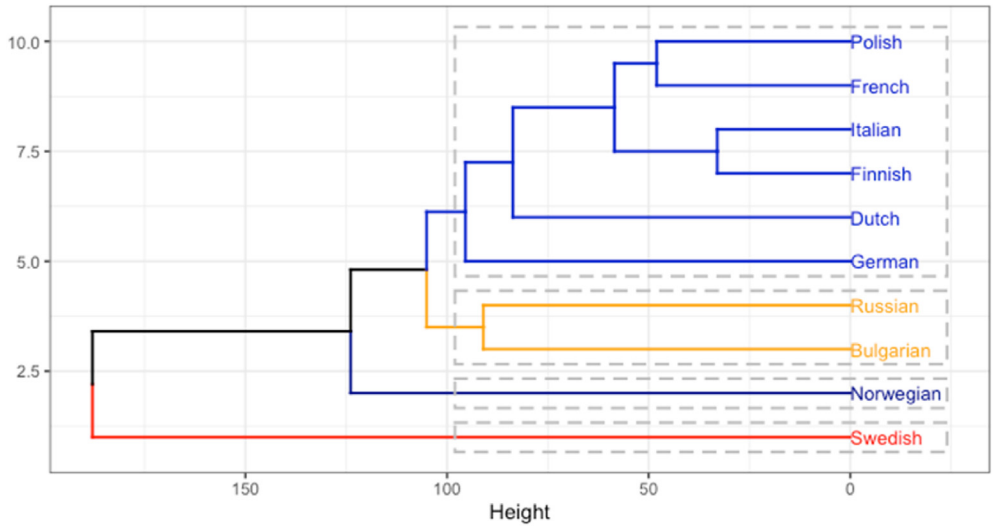


**Fig. 5.** Cluster dendrogram of THING bundles with a raw frequency higher than 3 across L1.

The next section zooms in further on the learner data. So far, we have treated the learners from the different L1 backgrounds as one group in order to attain more robust frequencies. However, as studies have shown that L2 learners in fact tend not to be a homogeneous group (e.g. Granger, 2017), we will now turn to an investigation of possible L1-specific uses of THING bundles.

## 6. Differences and similarities across the L1 groups

As the subcorpora of ICLE are rather small, our goal here is not to give a full-blown quantitative analysis of the data from each of the L1 backgrounds, but rather to highlight some similarities and differences between the L1 groups (see Section 3.1) in order to give a more accurate account of what the numbers presented in the previous sections are made up of.

To see what patterns emerged with regard to the L1 groups, a cluster analysis was carried out and illustrated through a dendrogram using the R package *factoextra* (Kassambara & Mundt, 2019). The dendrogram is shown in Fig. 5 (similarity measure: Manhattan; amalgamation rule: Ward's method). The L1 groups that behaved most similarly with respect to the THING bundles cluster together; the higher up in the graph the tree branches, the more dissimilar the two branches are.

As is shown, while the overall differences between L1s were relatively minor, the L1 Swedish and, to a lesser degree, the Norwegian students' usage stands out somewhat in terms of frequency and bundle type. We can also see that the L1 Russian and Bulgarian groups cluster together, as do the remaining L1 groups. Apart from the results for the L1 Russian and Bulgarian students, there is no clear evidence to suggest that L1s from the same language family behave similarly with regard to the bundles.

The reason why the Swedish students' usage stands out seems primarily to be that the most frequent bundle, *the most important thing* (18), takes the lion's share of the instances in the data, in that it is three times as frequent as the second most frequent bundle, *a lot of things* (6); in the other L1 data, the distribution is more even in that the frequency difference between the highest ranked and the second highest ranked bundle is much less pronounced. With regard to the Norwegian students, the main difference vis-à-vis the other L1s seems to be their relatively strong reliance on *right* + *thing* bundles (e.g. *the right thing to*): three of the bundles among the top 5 most frequent bundles are of this sort. However, a word of caution is warranted here: with frequencies this low (between 3 and 18 per bundle and L1), any quantitative findings are bound to be unstable, so these results would need to be confirmed in a more large-scale study. We therefore turn instead to some general trends.

The most striking similarity between the L1 groups was the fact that *the most important thing* was found among the top 3 bundles in all groups but two. The top bundles across L1 backgrounds can be found in Table 3 (*the most important thing* is bolded).[9]

The widespread reliance on this bundle may not be all that surprising if we consider its function: to emphasise importance. Multiple studies have noted a tendency for learners to make very frequent use of this function (e.g. Larsson, 2019; Lorenz 1998). However, there are of course many other patterns through which this function can be realised, most of which do not include THING (e.g. *the importance of*, *it is important to*, and *what is most important*), which suggests that we may also need to look beyond function to explain the use of this particular THING bundle.

[9] Bundles exhibiting the same frequency were placed in alphabetical order.

**Table 3**
Top 3 most frequent THING bundles in each L1 subcorpus.

| | 1 | 2 | 3 |
|---|---|---|---|
| SE | **the most important thing** | a lot of things | all the things that |
| IT | **the most important thing** | most important thing is | the first thing that |
| RU | **the most important thing** | the thing is that | the only thing that |
| PO | one thing is certain | **the most important thing** | important thing is to |
| NO | a lot of things | right thing to do | the right thing to |
| GE | such a thing as | the best thing to | **the most important thing** |
| FR | a lot of things | of looking at things | one thing is sure |
| FI | **the most important thing** | the only thing that | a good thing for |
| DU | is a good thing | **the most important thing** | a good thing to |
| BU | a lot of things | **the most important thing** | the first thing that |

In addition to a wish to emphasise importance, the frequency of this bundle could, at least in some cases, potentially be attributed to L1 transfer. For example, unlike English, languages such as Swedish and Norwegian allow for a headless noun phrase of the kind *the most important* + verb (e.g. *det viktigaste*/*viktigste* + verb), as exemplified in (14) and (15), along with its translations into (ungrammatical) English, a hypothesis also proposed by Tåqvist (2016: 99). These corpus examples are taken from the corpora *Akademiska texter – Humaniora* (https://spraakbanken.gu.se/resurser#corpora) and *Norsk aviskorpus* (http://avis.uib.no/), respectively. Having learned that headless noun phrases of this kind are not grammatical in English, the learners from such L1 backgrounds might therefore add *thing* as a place filler in order to be able to make a semantically similar statement.

(14) de säger att **det viktigaste är** att uttrycka musiken och sig själva som individer (Akademiska texter_PhD thesis_humanities) 'they say that **the most important is** to express the music and themselves as individuals'
(15) **Det viktigste er** at vi stoler på dem som er valgt til å styre oss (Norsk aviskorpus_SA200728_ http://avis.uib.no/). '**The most important is** that we trust those who have been elected to govern us'

However, while grammatical and idiomatic, the high frequencies of this bundle in the learner data as compared to the expert academic writing (see Section 3 above) seem to suggest that *the most important thing* is a bundle that is characteristic of learner language and thus a case where learners might benefit from being exposed to other semantically similar structures. Overall, the differences noted between the L1 groups were relatively minor, which suggests that the function and use of THING bundles is quite stable in the learner data, thus seemingly reflecting novice argumentative writing rather than differences resulting from individual L1s.

## 7. Discussion and conclusion

Our analysis of the lemma THING demonstrates that the two approaches to core words – single-word vs multi-word – result in quite different pictures. A single-word approach does indeed show that the frequency of THING use by L2 learners differs markedly from that exhibited by expert writers in being much closer to the spoken end of the informal/spoken-formal/written continuum. A multi-word approach tells a different story, however. A corpus-driven analysis of the top 15 most frequent four-word lexical bundles shows that, like many high-frequency words, THING enters into a wide range of multi-word units, many of which are typical of academic writing. Interestingly, contrary to the single-word-based findings, learners' frequency of multi-word use proves to be closer to the formal/written end of the continuum. This is testimony to the upper intermediate/advanced level of proficiency of the learners represented in the *International Corpus of Learner English*. However, an analysis of the actual bundle types used and the discourse functions they serve shows that there is only partial overlap between the learners and the expert writers. While an important function of the bundles in the expert data was comparing and contrasting (e.g. *the same thing as, it is one thing, other things being equal*), this function was absent in the learner data, where placing emphasis was the primary function. Particularly noticeable is the learner overuse of *the most important thing*, which ranks first in the learner corpus but is situated much further down the list (rank 18) in the expert writing. A comparison across the ten L1 populations represented in the learner corpus shows that learner behaviour is quite homogeneous, which suggests that differences from expert use are more likely to be developmental than L1-induced, although L1 transfer may also play a role.

Although we have only looked at one core vocabulary item, it is arguably emblematic enough in terms of its reputation in academic writing to merit generalizations beyond this particular item. What our findings plainly demonstrate is that a single-word approach is highly reductive. A much truer and fairer assessment of vocabulary use, in both general and academic language, needs to rely on both single-word and multi-word use. If the former approach alone is adopted, learners will not be

rewarded for using core words as part of phraseological units, some of which give evidence of a high level of proficiency on scales such as the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001).

As Nation (2016: xi) notes, "[w]ord lists lie at the heart of good vocabulary course design, the development of graded materials for extensive listening and extensive reading, research on vocabulary load, and vocabulary test development". However, it is essential to "phrase up" (Granger & Lefer, 2013) commonly used vocabulary lists, which are currently only made up of single words. Admittedly, there have been several recent efforts to provide phrasal lists, but to do justice to the ubiquity of phraseology in language, these units (or, at least, some of them) should be incorporated into vocabulary lists alongside single words. This would be a very concrete way of implementing Sinclair's (1991, 2004) phraseological view of language, which, as rightly observed by Gardner (2007: 255), "is perhaps the strongest indictment of traditional applied corpus-based vocabulary research that has relied heavily on frequency counts of individual word forms". We fully agree with Martinez and Schmitt (2012: 317) that "[i]t is mostly to the advantage of all interested parties that formulaic vocabulary be eventually seen as simply being 'vocabulary'".

The only lists that have actually achieved this are the *English Vocabulary Profile* (EVP) lists[10], which contain a very large number of multi-word units (phrases, phrasal verbs, collocations and idioms), graded according to the six levels of the CEFR scale. As stated by Capel (2012: 8–9), the rationale was that "even if learners know the top 2,000 words in English, the use of these words in phrases will not always be grasped, particularly when the meaning of the phrase as a whole is more figurative." As a result, even at the advanced C1 and C2 levels, the lists contain "many phrases formed from very frequent words." Although the EVP lists are not without their problems (for example, as regards the CEFR level assignments, see Negishi et al., 2012), they constitute a unique pedagogical resource which at long last gives multi-word units the place they deserve.

To make such lists maximally useful, however, it is necessary to go one step further and provide information on register. To return to the case of THING, the EVP lists include no less than 16 multi-word units distributed across levels B1 to C2. However, as shown by their respective frequencies in CAJA and BNC2014, some (the majority, in fact) are typical of informal spoken English (*things like that, the thing is, the next thing I knew*), while others are more frequent in academic writing (*among other things, all things considered, for one thing*). In addition, criteria for inclusion are not always obvious. For example, *one sure thing*, which is included as a C1 phrase, does not have a single occurrence in either CAJA or BNC2014. Any attempt at phrasing up vocabulary lists will be faced with the thorny issue of deciding not only which units to include but also how to integrate them into the lists. Lexical bundle extraction yields a mixed bag of quite different types of multi-word unit, and this bag needs to be unpacked in order to produce truly effective teaching and assessment resources. Overall, the most fixed, word-like units should be included in vocabulary lists as independent headphrases alongside headwords, while other less fixed units such as collocations should be listed under the relevant single word entry. However, lexical bundles are very numerous and it is neither possible nor desirable to include all of them in vocabulary lists. As shown by Khany & Malmir (2020), bundles also find a rightful place in learning and teaching resources for academic writing courses, where they can be grouped according to the rhetorical move they signal.

The results of our study show that multi-word units which include THING are surprisingly productive in expert academic writing, thus suggesting that teaching students to avoid using THING and other core words in their academic productions is reductive and potentially counterproductive. Instead, we should aim to raise learners' awareness of the stylistic preferences of multi-word units and move away from talking about "taboo words" that must be avoided towards a more nuanced view which acknowledges that high-frequency words occur in a wide range of constructions, some more formal than others.

All in all, our study adds to the growing list of those that highlight the benefits of a corpus-driven approach to phraseology based on lexical bundles. However, it has a number of limitations. First, we have only focused on one core vocabulary item; future studies are needed to further explore what gains a phraseological approach may have for other such items. Second, a lexical-bundle approach should ideally be based on very large corpora, and learner corpora (especially those that are made up of several L1-differentiated subcorpora) tend to be relatively small. Third, we only looked at one bundle size and therefore failed to extract some highly relevant bundles such as *among other things* or *all things considered*. Fourth, lexical bundles are sequences of contiguous words and are therefore not the ideal type of unit to bring out the variability that some units can display. As shown by Lu et al. (2021), discontinuous units such as phrase-frames which include an open slot that can be filled by different words provide useful information on the variability of multi-word expressions and the potential for creativity they offer (e.g. *play a(n) * in ⇒ play a(n) important/undisputed/central/crucial role in*). Phraseology is a very rich field which encompasses a wide range of multi-word units. If we are to cover the field in all its diversity, we need studies focused on each of these types of unit, but we must do this in full awareness that each study only lifts one small corner of the veil.

## Author statement

Sylviane Granger: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Validation; Visualization; Writing - original draft; Writing - review & editing.

Tove Larsson: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Validation; Visualization; Writing - original draft; Writing - review & editing.

[10] https://www.englishprofile.org/wordlists/evp.

## References


Ackermann, K., & Chen, Y.-H. (2013). Developing the academic collocation list (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235–247.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., & Shafaei-Bajestan, E. (2019). *languageR: Analyzing linguistic data. A practical introduction to statistics. R package version 1.5.0*. https://CRAN.R-project.org/package=languageR.

Bailey, S. (2011). *Academic writing: A handbook for international students* (3rd ed.). London & New York: Routledge.

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

Capel, A. (2012). Completing the *English vocabulary profile*: C1 and C2 vocabulary. *English Profile Journal, 3*(1), 1–14.

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

Cobb, T. (2013). Frequency 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 79–108). Eurosla Monographs Series, 2 http://www.eurosla.org/monographs/EM02/Cobb.pdf.

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

Council of Europe. (2001). *The common European Framework of reference for languages: Learning, teaching and assessment*. Cambridge: Cambridge University Press.

Crossley, S. A., Cobb, R., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*, 965–981.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair, & M. Hundt (Eds.), *Corpus Linguistics and linguistic theory* (pp. 51–68). Amsterdam: Rodopi.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*, 157–169.

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes, 30*, 58–72.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241–265.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*, 319–335.

Glynn, D. (2014). Correspondence analysis: An exploratory technique for identifying usage patterns. In D. Glynn, & J. A. Robinson (Eds.), *Corpus methods in cognitive semantics: Quantitative studies in polysemy and synonymy* (pp. 443–485). Amsterdam: John Benjamins.

Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. In R. Vatvedt Fjeld, K. Hagen, B. Henriksen, S. Johannson, S. Olsen, & J. Prentice (Eds.), *Vol. 9. Academic language in a nordic setting: Linguistic and educational perspectives* (pp. 9–27). Oslo Studies in Language (3).

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English*. Louvain-la-Neuve: Presses universitaires de Louvain. Version 3.

Granger, S., & Lefer, M.-A. (2013). Enriching the phraseological coverage of high-frequency adverbs in English-French bilingual dictionaries. In K. Aijmer, & B. Altenberg (Eds.), *Advances in corpus-based contrastive linguistics* (pp. 157–176). Amsterdam: Benjamins.

Gries, S. T., & Otani, N. (2010). Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal, 34*, 121–150.

Hasselgård, H. (2019). Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In M. Mahlberg, & V. Wiegand (Eds.), *Corpus Linguistics, context and culture* (pp. 339–362). Berlin: De Gruyter.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237–260.

Jiang, F., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes, 46*, 1–14.

Kassambara, A., & Mundt, F. (2019). *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.6 https://CRAN.R-project.org/package=factoextra.

Kosem, I. (2010). *Designing a model for a corpus-driven dictionary of academic English*. Aston University. PhD thesis.

Khany, R., & Malmir, B. (2020). A move-marker list: A study of rhetorical move-lexis linguistic realizations of research article abstracts in social and behavioural sciences. *RELC Journal, 51*(3), 381–396. https://doi.org/10.1177/0033688219833131.

Larsson, T. (2019). Grammatical stance marking in student and expert production: Revisiting the informal-formal dichotomy. *Register Studies, 1*(2), 243–268.

Larsson, T., & Kaatari, H. (2019). Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. *International Journal of Learner Corpus Research, 5*(1), 33–62.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In L. Eubank, L. Selinker, & M. Sharwood Smith (Eds.), *The current state of interlanguage: Studies in honor of William E. Rutherford* (pp. 265–272). Amsterdam: Benjamins.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307–322.

Lee, D. Y. W. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres. Evidence of a gradience of variation from the British National Corpus. *Journal of English Linguistics, 29*(3), 250–278.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 109–126). Eurosla Monographs Series, 2 http://www.eurosla.org/monographs/EM02/Lindqvist_etal.pdf.

Lorentz, G. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on computer* (pp. 53–66). Harlow: Longman.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319–344.

Lu, X., Yoon, J., & Kisselev, O. (2021). Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes, 61*, 63–83.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320.

McEnery, T., Love, R., & Brezina, V. (2017). Compiling and analysing the spoken British national corpus 2014. *International Journal of Corpus Linguistics, 22*(3), 311–318.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam & Philadelphia: Benjamins.

Nation, J., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35–41.

Negishi, M., Tono, Y., & Fujita, Y. (2012). A validation study of the CEFR levels of phrasal verbs in the English Vocabulary Profile. *English Profile Journal, 3*(1), 1–16.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom. Language use and language teaching*. Cambridge: Cambridge University Press.

Papp, S., & Nicholson, G. (2011). Vocabulary acquisition in children and Cambridge ESOL's wordlists for tests for young learners aged 7-14. *Cambridge ESOL Research Notes, 46*, 13–22.

Paquot, M. (2007). *Academic vocabulary in learner writing*. London & New York: Continuum.

R Core Team. (2020). *R: A Language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41–52). London & New York: Addison Wesley Longman.
Simpson-Vlach, R., & Ellis, N. C. (2010). An academic Formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487–512.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (2004). New evidence, new priorities, new attitudes. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 271–299). Amsterdam: Benjamins.
Swales, J. M. (2001). Metatalk in American Academic talk. The cases of point and thing. *Journal of English Linguistics, 29*(1), 34–54.
Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential skills and tasks* (3rd ed.). Ann Arbor: University of Michigan Press.
Tåqvist, M. (2016). *"Another thing": Discourse-organising nouns in advanced learner English*. Karlstad: Universitetstryckeriet. Karlstad University studies.
West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.
Wolfe-Quintero, K., Inagaki, Q., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Second Language Teaching & Curriculum Center: University of Hawai'i.

Sylviane Granger is Professor Emerita of English Language and Linguistics at the University of Louvain, Belgium. She is a leading researcher in the field of learner corpus studies and has published widely on learner corpus methodology, the analysis of phraseology in learner language, and the design of learner-corpus-informed teaching materials.

Tove Larsson is a post-doctoral researcher affiliated with Uppsala University, University of Louvain, and Northern Arizona University. Her research interests include analysis and applications of learner corpora, register variation, lexico-grammar, and English for Academic Purposes. She also has a keen interest in research methods.