

Published in:

Larsson, Tove & Kaatari, Henrik. (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. <https://doi.org/10.1075/ijlcr.17014.lar>

Extraposition in learner and expert writing: Exploring (in)formality and the impact of register

Tove Larsson¹ and Henrik Kaatari²

¹ Université catholique de Louvain, Uppsala University / ² University of Gävle

Subject extraposition (e.g. *it is important to remember*) is generally regarded to be a formal construction that learners, whose writing is often said to be overly informal, have been found to struggle with. This study investigates to what extent register and text type can be used to explore learners' reportedly "informal" use of this construction. Learner writing is compared to expert writing from several different registers and to native-speaker student writing. The results show that there are important differences across both registers and text types. Furthermore, while the learners' use is most like that of the experts' academic writing, certain similarities to the non-academic registers were also noted. The results moreover suggest that earlier claims about the informal status of learner writing seem mainly to have been influenced by the text types included in the corpora previously investigated.

Keywords: Extraposition, register, learner writing, expert writing, informality.

1. Introduction

Subject extraposition, as in *it is interesting to note the change*, is commonly used by writers for information-structure purposes or to depersonalize claims (Kaltenböck, 2005).

The construction includes an anticipatory *it* and an extraposed clausal subject (Quirk et al., 1985: 1391). Considerable register differences with regard to the use of this construction have been noted in previous studies (e.g. Biber et al., 1999: 722; Groom, 2005; Kaatari, 2017). For example, the construction has been found to be especially common in formal, academic discourse (Kaatari, 2016). Since many studies have observed that learners tend not to achieve an appropriate level of formality in their writing (e.g. Altenberg & Tapper, 1998; Tapper, 2005), one could expect learners to struggle with this comparably formal construction.

Indeed, studies such as Römer (2009) have found indications of this. However, the claims made in these studies are mainly either based on lexical misuse (e.g. *it is amazing that*) or on under/overuse of the construction, and few additional steps have been taken towards investigating informal use more systematically. In addition, in many courses on academic writing, in/formality is perceived as a dichotomy: when learners “miss the target” with regard to academic standards, their use is oftentimes described as “informal”. While it is not uncommon for (in)formality to be treated as a continuum in the literature (see, e.g. Ädel, 2008; Smith, 1986), the present study extends previous research by investigating to what extent register can be used as a way of nuancing the informal/formal scale. Put in another way, while the concepts “informal” and “formal” provide endpoints on the continuum, register could, as will be argued, be used as “anchors” on this continuum to enable more detailed discussion of students’ use of the construction investigated. To this end, the study compares learner writing from one register (Academic Prose) and expert writing from several different registers (Academic Prose, Popular Science, News, Fiction and Conversation), thereby exploring which of the expert

register(s) the learners' use is most similar to, as outlined below.¹ Furthermore, the study also investigates to what extent text type can be used to explain seemingly “informal” usage, thus providing a more detailed analysis of the formal/informal scale.

The present study thus aims to investigate to what extent learners (L1 Swedish) are able to adhere to register-specific conventions in their use of subject extraposition. Specifically, the study focuses on the most frequent clausal type and predicate of the construction, namely adjective phrases followed by *to*-infinitive clauses (e.g. *it is important to remember*; *it is fair to say*) (Kaatari, 2017; Larsson, 2016b). In addition to being the most frequent realization, this subtype includes the elements of the construction that arguably carry the most semantic content, namely an adjective and an infinitival verb (see Larsson, 2016a: 68–69 for a more detailed discussion of so-called *predicate lemmas*). The first part of the study investigates (i) which of the experts' registers the learners' use is closest to and, thus, (ii) whether describing learner use as simply “informal” might be simplifying matters. The second part of the study investigates learner writing across two text types (theses vs. argumentative essays), in order to determine the role of text type in relation to (in)formality. The following research questions have been used to guide the analysis:

- What differences and similarities are there between the learners' and experts' use overall with regard to high-frequency instances of the construction (operationalized as adjective-verb pairings)?
- Is the learners' writing most similar to the experts' Academic Prose, or is it more similar to some of the other registers?

¹ Following Biber et al. (1999: 15), we use the term *register* to distinguish between situational characteristics of different texts. See also Lee (2001) for a comprehensive discussion of the difference between the terms *register*, *genre* and *text type*.

- To what extent is text type important in explaining the level of formality of learner writing?

2. Previous research on extraposition

Subject extraposition has been investigated under several different labels and using many different frameworks of analysis. The construction has commonly been referred to as the introductory *it* pattern (e.g. Groom, 2005; Larsson, 2016a, 2016b, 2017, 2018), the anticipatory *it* pattern (e.g. Ädel, 2014), *it*-clauses (e.g. Hewings & Hewings, 2002) and *it*-extraposition (e.g. Kaltenböck, 2005; Zhang, 2015). The term *(subject) extraposition* has, for example, been used in variation studies of general language use (e.g. Kaatari, 2016) and in studies applying frameworks such as Lexical Functional Grammar (e.g. Ramhøj, 2016). Nevertheless, common to all these studies is that the construction is considered to have two subjects: an anticipatory *it* that does not have anaphoric reference and a clausal subject. While the clausal subject can be realized as a *to*-infinitive, a *that*-clause, an *-ing* clause or a *wh*-clause (Larsson, 2018: 16; cf. Quirk et al., 1985: 1391), the present study focuses on the most common type, namely the *to*-infinitive (Kaatari, 2017). Tokens with an adverbial clause and tokens for which *it* is not an anticipatory subject, such as in *it*-clefts (e.g. *it was James who was late*), referring *it* (e.g. *the chair is here; can you see it?*) and empty *it* (e.g. *it is raining today*) were excluded.²

The form and function of the construction have previously been investigated in learner writing, native-speaker (NS) student writing and expert writing. The learners have

² The term *token* is here used synonymously to *instance*.

been reported to struggle with the construction in different ways. In one of the earliest studies on the construction in learner writing, Hewings & Hewings (2002) compared MBA theses written by learners (with a variety of L1s) and journal articles from business studies. They found that the learners tended to overstate the validity of their claims, for example through very frequent use of instances such as *it is necessary that* and *it is clear that* (Hewings & Hewings, 2002: 369, 378). Learners were also found to make more frequent use of the construction overall, compared to the experts (Hewings & Hewings, 2002: 374). The tendency for students to use the construction to emphasize the validity of claims was also found in Hasselgård's (2009: 137) study, looking, among other things, at extraposition in L1 Norwegian learner writing. There, the learners were reported to use the construction mainly to evaluate difficulty, importance, expectedness and appropriateness (e.g. *it is important to; it is difficult to*) (Hasselgård, 2009: 130).

A tendency for learners to overuse the construction compared to NS writers was noted in Larsson (forthcoming), where (L1 Swedish) learner writing was compared to NS student writing, the learners were found to overuse high-frequency realizations of the construction; furthermore, while there were few differences noted between the NS students and the learners whose papers received a higher grade, the learners whose papers received a lower grade were found to overuse the construction. Furthermore, overuse (especially with regard to extraposed *to*-clauses) was found in Callies' (2009: 199– 201) study comparing (L1 German) learners and NS student writers.

Certain “speech-like” uses of the construction have been noted, for example in Römer (2009). She investigated the use of the construction in (L1 German) learner writing, NS student writing and expert writing.³ Among other things, she found that the

³ Römer compares argumentative essays written by L1 German undergraduates to linguistics and literature essays written by L1 German final-year undergraduates and first-year graduates; linguistics, philosophy,

learners, in particular those whose texts are included in the German sub-corpus of the *International Corpus of Learner English* (GICLE; Granger et al., 2009), used the constructions to “express strong emotions” (Römer, 2009: 156). The learners also used what is referred to as “extreme” adjectives that are “more characteristic of speech than of academic writing”, such as *amazing*, *ridiculous* and *stupid* (Römer, 2009: 156).

A tendency among learners to be informal in their writing in more general terms (i.e. not just with regard to this construction) has been found in several studies. For example, based on a study of the L1 French, Swedish, Finnish and Dutch subsets of ICLE, Petch-Tyson (1998: 116) concluded that all learner groups show a higher degree of interpersonal involvement in their texts than the NS students, in that nearly all features of reader/writer visibility (e.g. first-person pronouns, stance marking) were overused by the learners. L1 Swedish students have been found to be particularly informal in their writing. When comparing the Swedish sub-corpus of ICLE (SWICLE; Granger et al., 2009) and the *Louvain Corpus of Native English Essays* (LOCNESS; Granger & Petch-Tyson, 1996) corpora, Herriman & Boström Aronsson (2009: 118) even went so far as to say that L1 Swedish students’ argumentative writing “is similar to NS’ spoken language”. Moreover, Tapper (2005) found evidence of informal use of connectives (e.g. *actually*, *therefore*) in L1 Swedish data; she noted further that “several other studies have also shown that many of the learner sub-corpora in the ICLE corpus – and in some cases the Swedish sub-corpus especially – contain writing that is more informal in style than the native English-speaking student writing” (Tapper, 2005: 124).

psychology and sociology papers written by NS final-year undergraduates and first through final-year graduates; and published articles from linguistics, psychology and social sciences (Römer, 2009:149).

Based on previous research, we could perhaps expect the learners to use the construction in an “informal” manner; this will serve as our hypothesis. However, we will further investigate whether uses that may at first glance be perceived as “informal” might in fact be an indication of insufficient register awareness (i.e. that the learners use an adjective-verb pairing that is more strongly associated with one of the non-academic registers) or be due to text type. Following Smith (1986), we view (in)formality as a continuum, and we will use register as a proxy to see where on the formal-informal scale the students’ use can be placed. Based on the different situational characteristics of the registers included (Biber et al., 1999: 16; see also Kaatari, 2016), Academic Prose is considered to be the most formal register, followed, in descending order of formality, by Popular Science, News, Fiction and Conversation.

3. Material and method

In this section, the corpora used will be introduced in section 3.1 and the method used will be described in section 3.2.

3.1 Corpora used in the study

The first part of the study uses data from three main corpora: the *Advanced Learner English Corpus* (ALEC; Larsson, 2014), the *Varieties of English for Specific Purposes dAtabase* (VESPA; Paquot et al., 2013) and BNC-15 (Kaatari, 2017). ALEC is a 1.3-million-word corpus of learner academic writing by Swedish university students in

English linguistics and English literature. VESPA is a multi-million-word corpus of learner academic writing (Paquot et al., 2013). The Swedish subset of VESPA, VESPA-SE, currently comprises texts from only one discipline, namely English linguistics. This study only includes those texts from ALEC (109 texts) and VESPA-SE (22 texts) that were written by students whose self-reported first language is Swedish and who are in their third year of university studies on average. ALEC and VESPA-SE comprise (untimed) BA theses from the same two universities in Sweden; due to the close similarity between the corpora, they are grouped together in the analysis. Although it would have been preferable to have more detailed information about the students' level of proficiency in English, no such tests were applied in either corpus; level of proficiency will therefore not be included as a factor in the present study.

The expert corpus, BNC-15, is a methodically sampled 3-million-word subset of the British National Corpus (BNC; Burnard, 2007) that is designed to enable register comparisons with comparable subsets (for more detailed information about BNC-15, see Kaatari 2017). Five registers are included in BNC-15: Academic Prose, Popular Science, News, Fiction and Conversation. While BNC-15, unlike the learner corpora, includes samples of 10,000 words rather than full texts, the average text length is similar to that of VESPA-SE and ALEC.⁴

In the second part of the study, another text type within academic writing, namely argumentative essays, was added to test previous claims of (in)formality. Argumentative essays (especially when timed) have, in previous studies, been found to exhibit somewhat

⁴ In BNC-15, each register is represented by 60 samples of roughly 10,000 words. An equal number of samples from the beginning, the middle and the end of text files were sampled to minimize any skewing due to textual colligation and clustering (see Hoey, 2005; Evert, 2006). Furthermore, in BNC-15, the registers Academic Prose and Popular Science include an equal number of samples from the disciplinary groups *Humanities*, *Social Sciences* and *Natural Sciences*. In ALEC, the average text length is 8,442 words; in VESPA, the average text length is 7,225 words.

informal language features (see, e.g., Ädel, 2006, 2008; Callies, 2013: 363–364). Two corpora were used here: one with learner data, SWICLE, and one with NS data, LOCNESS; these corpora allow us to control for NS status in relation to text type. SWICLE includes 363 texts that were collected in the 1990s from Swedish universities. It includes argumentative essays written mainly by students who are in their third or fourth semester of English studies; many of the essays are timed. LOCNESS, which was compiled to be comparable to corpora from the ICLE family, includes essays written by British and American students; many of the essays are timed. For the purpose of the present study, a subset of LOCNESS is used, including only the British argumentative essays (147 texts from A-level students and 33 texts from university students). An overview of the word counts for all the sub-corpora is shown in Table 1.

Table 1. Word count for each of the sub-corpora included.

	Sub-corpus	Word count	L1	Text type
Novice	ALEC+VESPA-SE	1,079,178	Swedish	Theses
	SWICLE	206,378	Swedish	Arg. essays
	LOCNESS	79,228	British English	Arg. essays
Expert (BNC-15)	Conversation	600,049	British English	
	Academic Prose	600,117	British English	
	Popular Science	600,122	British English	
	News	600,326	British English	
	Fiction	600,334	British English	
Total		4,365,732		

As the present study is intended to provide an initial overview of how register and text type might affect the use of the construction, the corpora described above were thought to be most suitable for this purpose. While it falls outside the scope of the study to investigate possible disciplinary differences (see, e.g., Groom, 2005; Larsson, 2017), we

have done our best to increase comparability between the corpora by using a reference corpus (BNC-15) and a learner corpus (a combined subset from ALEC and VESPA-SE) that comprise data from a mix of disciplines, rather than data from one single discipline. It is nonetheless worth noting that the academic subsection of BNC-15 includes a larger variety of different disciplines than the ALEC+VESPA subcorpus; this will be returned to in section 4.1.1 and should be kept in mind when examining the results.

3.2 Method

In order to retrieve all instances of the construction where an anticipatory *it* was followed by an adjectival predicate and a *to*-infinitive clause functioning as the subject, a Perl program was used.⁵ As not all of the corpora were part-of-speech tagged, the program was used to identify and extract tokens where the lexical item *it* was followed by the lexical item *to* with 1–20 intervening tokens. In order to filter the results further to increase manageability, all tokens where *to* was followed by a full stop, an article or a pronoun were removed;⁶ this enabled exclusion of those tokens in which *to* is prepositional rather than an infinitival marker. The remaining tokens were gone through manually by both authors to weed out all invalid tokens, as discussed below.

Since the study only includes adjective-verb pairings in subject extraposition (e.g. *it was **interesting** to **note***), tokens which did not include an adjective and an infinitival verb were excluded. Examples of such tokens that were captured by the program and filtered out manually include tokens where the extraposed clause is finite, as in (1). The

⁵ We are very grateful to Gregory Garretson at Uppsala University, who wrote this program.

⁶ The following list was used:
(the|a|an|me|you|him|her|it|one|us|them|my|your|his|its|our|their|this|that|these|those| \.)

very few tokens where the adjective functions as an object complement, as in (2), were also excluded, along with tokens with a prepositional phrase rather than an adjective phrase, as in (3).⁷

(1) [...] it is significant **that Rowling chooses to depict the emblem precisely as a golden lion** [...]. (ALEC_006)

(2) [...] it makes her **happy** to hear them quarrel. (ALEC_038)

(3) [...] it is not **in the interest of this discussion** to discuss the seriousness of students' adverbial errors [...]. (VESPA_130.1)

Unlike some previous studies (e.g. Römer, 2009),⁸ the data selection method used enabled inclusion of all the verbs that can be found in the construction (the most common being BE, SEEM and APPEAR). The study also included tokens that had inverted word order, as in (4), and tokens that included a *catenative verb* (i.e. verbs such as SEEM and APPEAR followed by the infinitival marker), as in (5) (Quirk et al., 1985: 137; see also Larsson, 2018: 24 for a more detailed discussion).

(4) Only when MTs' difficulties are established, **is it** possible to analyze the results.
(VESPA_127.1)

(5) Furthermore, it **seems to be** easier to remember morphemes that consist of syllables [...]. (VESPA_135.1)

Certain measures were taken, in order to increase the manageability of the results. First, all the tokens were manually inspected and spelling differences were normalized (e.g.

⁷ Throughout the article, emphasis has been added to the examples to point the reader's attention to relevant features.

⁸ Römer's (2009) study only covers instances of the construction that include the present-tense form of BE (i.e. *is*).

analyze → *analyse*; *recognize* → *recognise*). Second, comparative adjectives were lemmatized (e.g. *easy*, *easier*, *easiest* → EASY; *good*, *better*, *best* → GOOD). Third, the main verb (rather than the auxiliary) was used for all infinitival verbs (e.g. *have finished* → FINISH; *have seen* → SEE). However, prepositional verbs (e.g. *go through*, *go along with*) were not lemmatized to GO for semantic reasons, as *it would be interesting to go to Paris* is arguably semantically distinct from *it would be interesting to go through his belongings* or *it would be interesting to go along with that*. These instances were therefore treated as separate verb categories.

The study used the software environment *R* (R Core Team, 2017) to manage the data and to carry out tests of significance; the following tests were applied: *co-varying collexeme analysis* (Stefanowitsch & Gries, 2005), *Correspondence Analysis* (CA) and *Multiple Correspondence Analysis* (MCA). A co-varying collexeme analysis is a kind of collostructional analysis that identifies “the association strength between pairs of lexical items occurring in two different slots of the same construction” (Stefanowitsch & Gries, 2005: 9). In this study, we are concerned with the association strength between adjectives and infinitives within the subject extraposition construction. The co-varying collexeme analyses were performed using a script written by Stefan Th. Gries (Gries, 2007). The statistical measure used for calculating collostruction strength is the Fisher-Yates exact test. Although the Fisher-Yates exact test does not allow for comparisons of collostruction strength across corpora, it was considered the preferred choice since no such comparisons were made in this study and since the test is highly robust when dealing with low frequencies.⁹ CAs and MCAs are versions of a multivariate exploratory space reduction

⁹ Measures which allow comparisons of collostruction strength across corpora include odds ratio and Delta P (Stefan Th. Gries, p.c.); our initial tests did, however, indicate that a problem with these measures was that low-frequency pairings were given unproportionally strong weight.

technique for categorical data analysis. In these analyses, the frequencies of variables are converted into two matrices of distances, one containing the distances between rows, and one containing the distances between columns. The distances are then plotted onto a two-dimensional map, showing how close or how far apart different variables are from each other. The closer the distance on the map between two variables, the stronger the correlation is between them (see Glynn, 2014 and Baayen, 2008: 128 ff. for more on Correspondence Analysis).

4. Results and discussion

In the first part of the study (section 4.1), the two learner corpora that comprise theses, ALEC and VESPA-SE, were compared to BNC-15. In the second part (section 4.2), which is concerned with investigating differences and similarities across text types, two corpora containing argumentative essays, SWICLE and LOCNESS, were added to the analysis to further investigate previous claims of (in)formality.

4.1 A comparison across registers

In this section, an overview of the results in terms of overall and per-text frequencies will be given in section 4.1.1. Section 4.1.2 offers a more detailed discussion of the lexical choices made in the adjective-verb pairings across the registers and subcorpora.

4.1.1 An overview

A total of 2,163 tokens of the target construction were found in ALEC, VESPA-SE (henceforth referred to as VESPA) and BNC-15. As expected, there were clear frequency differences across the sub-corpora investigated; the distribution is shown in Figure 1.

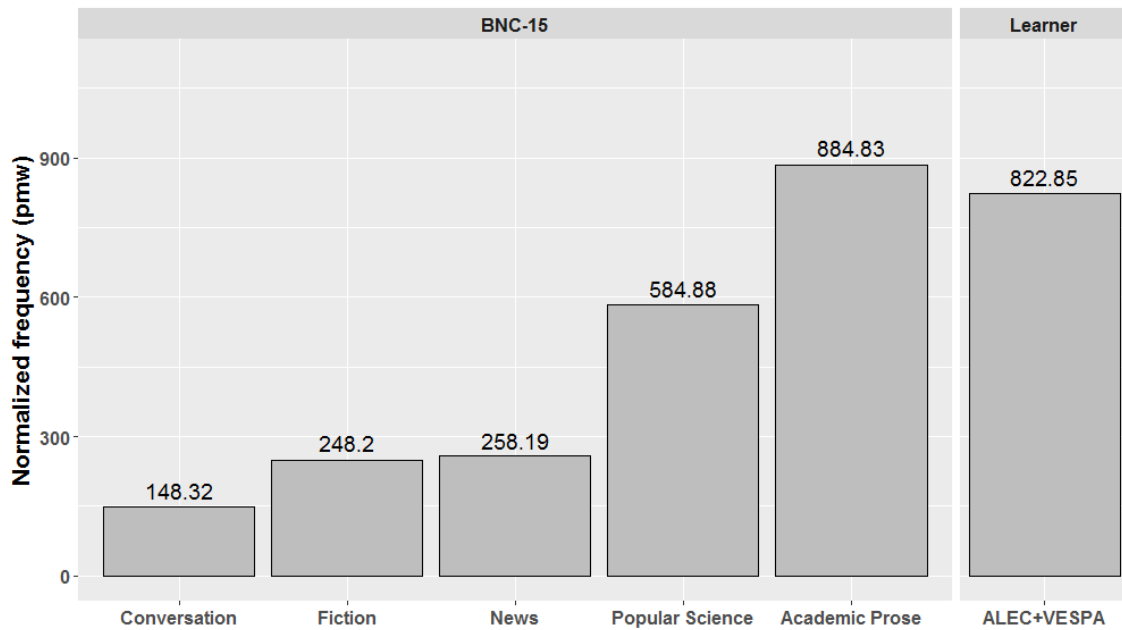


Figure 1. Normalized frequency, per million words (pmw), of the construction across the sub-corpora.

Before the frequency differences across registers are addressed, a brief discussion of discipline will now follow. As mentioned in section 3.1, there are differences in terms of the disciplinary make-up of the learner corpora and BNC-15. However, due to the material available, we have not been able to include discipline as a factor in the analysis presented in the subsequent sections. Nonetheless, to make certain that this approach did not affect the results unduly, we will here present the frequencies from a subset of BNC-

15 where we have attempted to bring the subcorpora closer together with regard to discipline.

ALEC and VESPA include theses from linguistics and literature (see section 3.1). In BNC-15, by contrast, there is only a handful of linguistics and literature texts in the Popular Science and Academic Prose subcorpora, therefore making it impossible to limit the investigation to these two disciplines in the expert data. For this reason, we have here extended the investigation to include the disciplinary groups that linguistics and literature fall into in BNC-15 (following Lee's, 2001, classification of the BNC), namely the Social Sciences and Humanities respectively. However, a clear disadvantage of this approach is that it means excluding over 400,000 words from the analysis. These results can be found in Figure 2.

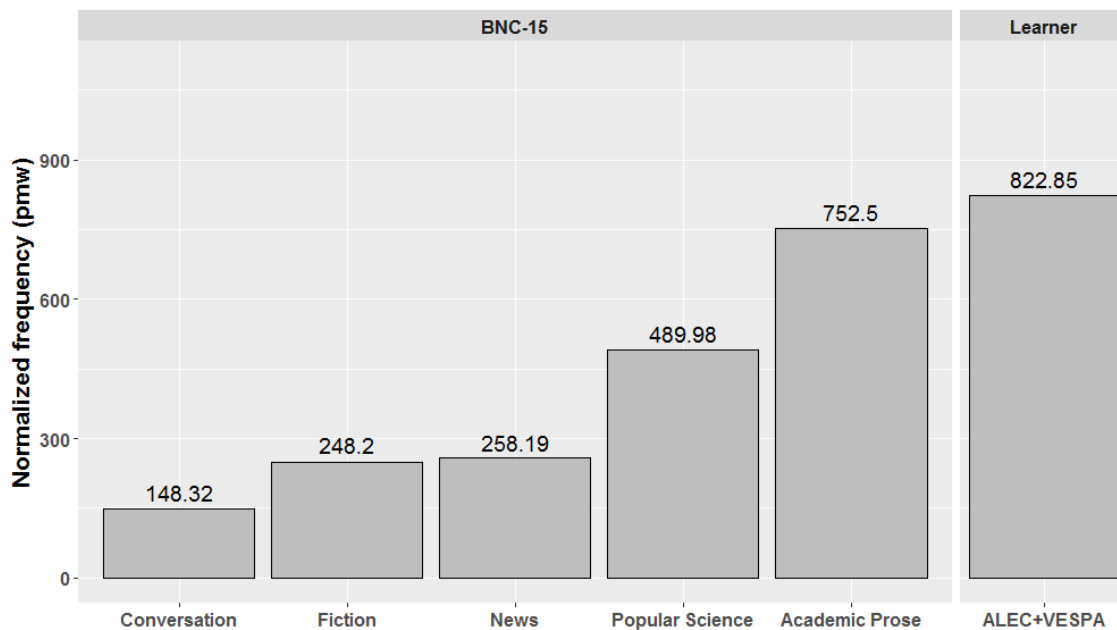


Figure 2. Normalized frequency, per million words (pmw), of the constructions across the sub-corpora (only texts from the Humanities and Social Sciences for Popular Science and Academic Prose).

As can be seen, while there are some differences with regard to the comparison with the learner data (as further discussed below), limiting the analysis in this manner does not affect the relative frequencies of the registers, nor does it affect the order or ranking of the registers. Therefore, as the differences are only marginal and as the study relies on lexical searches that require large data sets, we have opted for a more inclusive approach where all data from BNC-15 are included. This means that in the subsequent sections, all disciplinary groups from BNC-15 are included in the analysis, which should be kept in mind when interpreting the results. More detailed investigations into disciplinary differences will have to be left for future studies.

When we turn to the frequency distribution across the registers, we can see that the construction is comparatively frequent in Academic Prose, and less frequent in Popular

Science, News and Fiction, and especially infrequent in Conversation. The construction thus occurs more frequently in the most formal register, Academic Prose, than the less formal registers, Fiction, News and Conversation. Some examples of the pairings that are used in the sub-corpora can be found in (6)–(11).

(6) It is **hard** to **escape** the conclusion that the Department of Transportation’s totally incompetent [...]. (CONV_JS8)

(7) Until then, it would be virtually **impossible** to **discuss** anything. (FICT_GV2)

(8) It is **nice** to **have** a survey of resorts, necessarily selective, from a good, experienced skier [...]. (NEWS_AA0)

(9) [...] it is **difficult** to **pinpoint** the precise location of these ancient workings. (POPU_H0B)

(10) [...] it is **important** to **remember** that [...] his call for “a fraternal association” between the Anglo-Saxon powers was poorly received by a majority of Americans. (ACAD_HY8)

(11) [...] it is **interesting** to **see** that group B are more positive to the idea [...]. (ALEC.003)

One thing follows from these results that at least partially disconfirms our hypothesis stating that the learners can be expected to use the construction in an “informal” manner: while the learners make slightly more frequent use than the experts in the smaller subset of Academic Prose (Figure 2) and slightly less frequent use when compared to the full subcorpus of Academic Prose (Figure 1), the fact remains that the learners’ aggregate use is more similar to that of the academic experts than any to that of any other register. If we

take “informal” use to mean used similarly to the non-academic registers, then the learners use cannot be described as “informal” with regard to the overall frequencies.

With regard to the normalized frequencies per text in all the data, a similar pattern emerges. These results are displayed in notched box plots in Figure 3 below. This type of plot provides useful information about the dispersion of the frequencies. The box displays the interquartile range, and the notches show the (95 percent) confidence interval around the median (overlap between the notches of the boxes thus suggests that the difference between the medians is not statistically significant). Outliers are represented by dots.

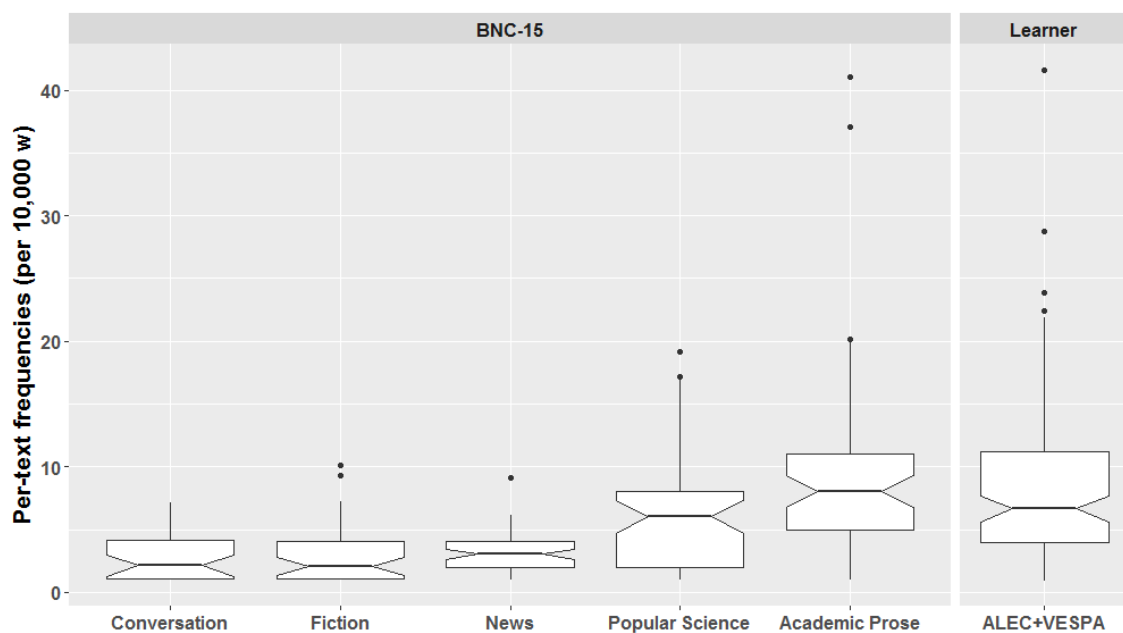


Figure 3. Number of adjective-verb pairings per text (normalized per 10,000 words).

As can be seen, the median is lower in the less formal registers and higher in the more formal registers. Whereas the learners’ use is, in this respect, most similar to that of Popular Science, it is still fairly similar to Academic Prose.

A tentative conclusion to draw from these results is thus that the learners seem to use the construction largely in an expert-like, register-appropriate manner, at least in terms of overall frequency. However, it remains to be seen whether this is the case also with regard to the lexical choices made inside the construction; this is something that we will look into more closely as we now turn to a more detailed analysis of the adjective-verb pairings.

4.1.2 A more detailed analysis of the adjective-verb pairings

To explore the learners' and experts' use of the adjective-verb pairings in more detail, *co-varying collexeme analyses* were used to identify the highest-ranked adjective-verb pairings with regard to collocation strength in each of the learner and expert subcorpora (Stefanowitsch & Gries, 2005). The results showed that in terms of which pairings were highly-ranked, the learners' use at least partly overlapped with the experts' use in Academic Prose (see the Appendix); however, there was also some overlap between the learners' use and the experts' use in some of the other registers. With the underlying assumption that "expert use" of this construction means that it is used most similarly to the way it is used in the experts' Academic Prose, the learners showed evidence of non-expert use of the pairings in that they tended to use both adjectives and adjective-verb pairings that were more commonly used in the non-academic registers, which might suggest that they are not aware of the differences between these registers, as will be discussed further below.

For example, while the learners share some high-ranking pairings with the experts' Academic Prose, such as *important+note* and *important+remember*, they share others, such as *interesting+see* (ranked third with regard to collocation strength in the learner

data), only with Conversation and News. Furthermore, of the pairings that showed statistically significant attraction with regard to collocation strength, the learners also made frequent use of pairings that included the adjectives NICE (which is associated with Conversation) and EASY (which is associated with Popular Science and News). This association between adjectives and registers in the expert data can be seen in Figure 4 below. This figure shows the results of a correspondence analysis where the adjectives that are most strongly correlated with each of the registers in the expert data are represented near the name of that register.

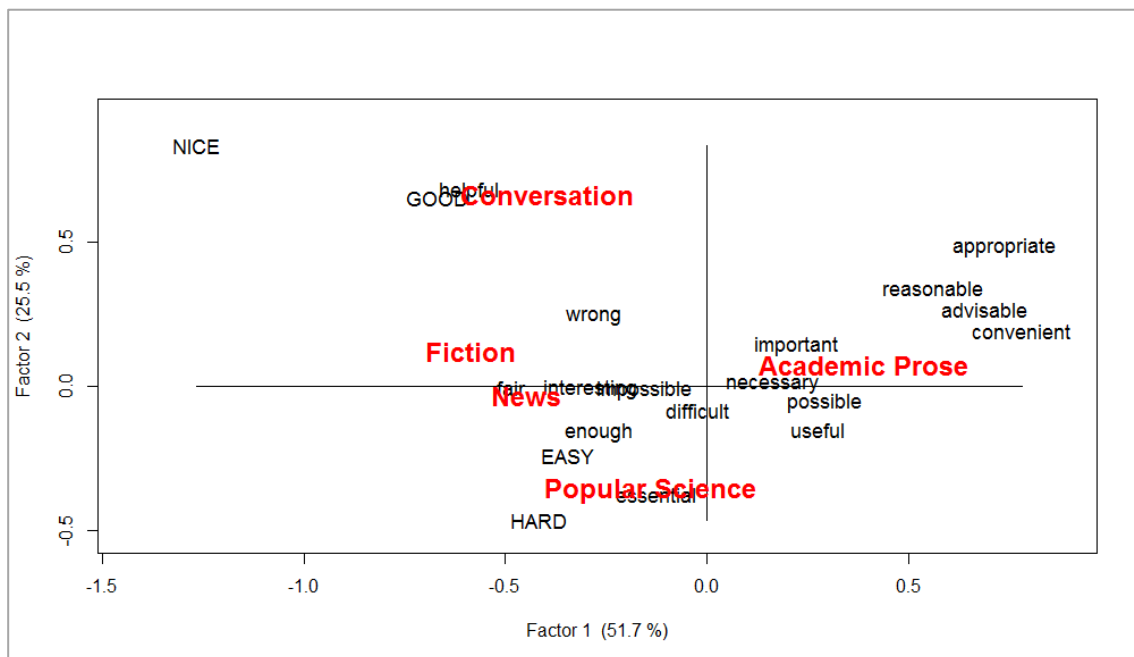


Figure 4. Correspondence Analysis of register and the 20 most frequent adjectives in BNC-15 that occur in the construction.

There were two further types of usage that set the learners apart from the experts: first, while the learners made use of many of the adjectives that the experts used in Academic

Prose, they tended to overuse some of the most common ones; second, the learners used some adjective-verb pairings that were not found in any of the expert registers.

Figure 5 shows the normalized frequencies per million words of the 200 most frequent adjective-verb pairings in the learner and expert Academic Prose data. As can be seen, the learners tended to use their most frequent pairings more frequently than the experts used theirs.

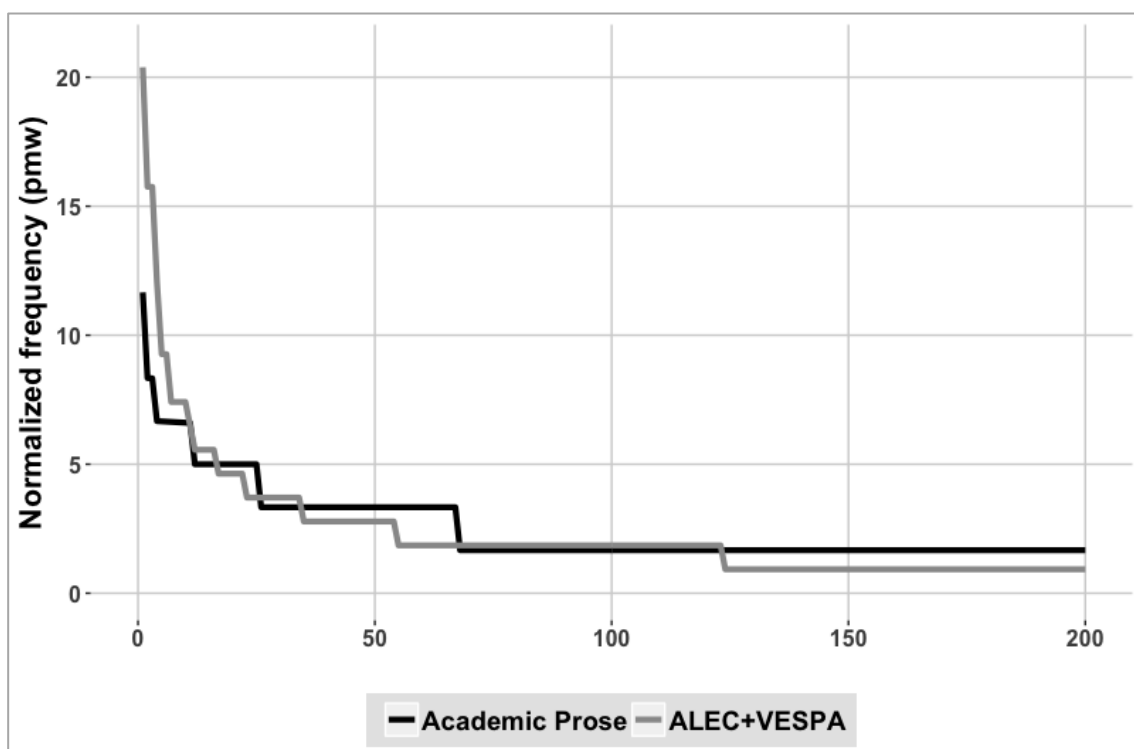


Figure 5. The 200 most frequent adjective-verb pairings in the learner and expert Academic Prose plotted along the x-axis in decreasing order of frequency.

These findings thus concur with those of Larsson (2016b, forthcoming) where evidence was found for a preference for high-frequency lexico-grammatical tokens, so called *lexico-grammatical teddy bears* (from Hasselgren's, 1994, use of *lexical teddy bears*). The ten most frequent pairings for learners and experts are shown in Table 2. All of these

pairings showed a significant attraction with regard to collocation strength, except for four pairs that included the high-frequency adjective *possible* (*possible+see* in the learner data; *possible+use*, *possible+distinguish* and *possible+identify* in the academic expert data).¹⁰

Table 2. Raw frequency and normalized frequency (per million words), of the top 10 adjective-verb pairings in the learner and expert Academic Prose.

Rank	Learners		Experts: Academic Prose	
	Pairing (raw frequency)	pmw	Pairing (raw frequency)	pmw
1	<i>important + note</i> (22)	20	<i>important + note</i> (7)	12
2	<i>interesting + see</i> (17)	16	<i>possible + use</i> (5)	8
3	<i>possible + see</i> (17)	16	<i>reasonable + suppose</i> (5)	8
4	<i>interesting + note</i> (13)	12	<i>important + recognise</i> (4)	7
5	<i>important + BE</i> (10)	9	<i>important + remember</i> (4)	7
6	<i>important + learn</i> (10)	9	<i>impossible + BE</i> (4)	7
7	<i>important + remember</i> (8)	7	<i>necessary + consider</i> (4)	7
8	<i>possible + detect</i> (8)	7	<i>possible + distinguish</i> (4)	7
9	<i>possible + find</i> (8)	7	<i>possible + identify</i> (4)	7
10	<i>reasonable + assume</i> (8)	7	<i>possible + make</i> (4)	7

When we extended the window to look at all the pairings, it became clear that the learners' use of *interesting* and *important* stood out as particularly characteristic of their use of the construction (see also the MCA in section 4.2). Examples of how the learners use such pairings are shown in (12)–(15) below.

- (12) [...] it was still **interesting** to **hear** how the student informants explained their interpretations of the idioms. (ALEC_066)

¹⁰ See the Appendix for lists of the highest-ranked significant pairings in terms of collocation strength across all the sub-corpora.

(13) Therefore it is **interesting** to **look** at other factors in personal identity to see what impact they have. (ALEC_003)

(14) It is **important** when conducting an EA to **distinguish** between an error and a mistake [...]. (VESPA_0135.1)

(15) It is **important** to **note** that the lexical items *lad* and *wee* are also used in other areas of Britain as well. (ALEC_082)

There are both quantitative and qualitative explanations for this. For example, unlike the experts, the learners make very frequent use of *interesting*+VERB. Furthermore, while *important* certainly occurs in the expert data, pairings with *important* are used differently by the learners than by the academic experts. A closer look at the instances of *important*+VERB for which the collocation strength is significant shows that the learners have a larger repertoire of infinitives; an example of a pairing that is only found in the learner data, *important*+*point out*, can be found in (16) below.¹¹

(16) However, it is **important** to **point out** that there is a difference in which person the superior character is ascribed to. (ALEC_127)

Let us now turn to the second way in which the learners' use differs from the experts', namely the fact that the learners use pairings that are not found in any of the expert registers. There are several such pairings on the list of pairings that have a significant attraction with regard to collocation strength. For example, the learners occasionally use pairings such as *vital*+*look* (17), *crucial*+*return* (18) and *pitiful*+*note* (19), along with pairings such as *tempting*+*regard* and *feasible*+*claim*.

¹¹ See the Appendix for more detailed information on the most strongly attracted pairings across sub-corpora.

(17) [...] it is **vital** to **look** at what the teachers actually convey in their answers [...].

(ALEC_084)

(18) [...] it becomes **crucial** for Harry to **return** to the world [...]. (ALEC_046)

(19) [...] it is **pitiful** to **note** that the texts in “Blueprint” contain no other varieties than those of the inner-circle [...]. (ALEC_013)

The fact that the learners at least occasionally use strong adjectives such as *vital* and *crucial* could be used in support of previous studies where learners have been found to state claims too forcefully (see, e.g., Hewings & Hewings, 2002); however, this might not be as simple as it may appear at first sight. As can be seen when comparing the examples in (17) and (18), when the larger context is taken into consideration, it is only (17) that seems to be an example of use of an overly forceful adjective. The reason is that the *for-to* construction in (18) enables a reading of this example in which the student narrates a story from the perspective of the subject of the *to*-clause (i.e. Harry [Potter]), which arguably justifies the use of a strong adjective, unlike instances of the construction in which the statement is implicitly attributed to the author him/herself, such as (17). From a pedagogical standpoint, the existence of tokens such as (18), even if they are rare, suggests that it should perhaps not only be a question of whether or not learners should use strong adjectives inside the construction, but rather of *in what context* such adjectives could be used.

All in all, we can draw the tentative conclusion that whereas there are some minor differences between the learners’ use and that of the experts (and that the learners thus do not seem to be fully aware of the register differences), the learners’ use is, in fact, most like that of the academic experts. Thus, although a certain reliance on adjectives that are associated with Conversation (which can therefore be regarded as informal) was noted,

we did not find any strong evidence of informal use in the learner data overall, which suggests that our hypothesis stating that the learners can be expected to use the construction in an “informal” manner can be refuted. However, before closing the case, we added an additional learner corpus to our analysis: SWICLE (the Swedish subset of ICLE; see Granger et al. 2009) to find further clues about (in)formality in the context of the construction. The rationale for adding this particular corpus is that SWICLE is the corpus that was used in previous studies (e.g. Petch-Tyson, 1998; Altenberg & Tapper, 1998) that suggested that L1 Swedish learners are particularly informal in their writing. As a complement to the SWICLE data, we also added NS data from LOCNESS. The results of this part of the study will be presented in the next subsection.

4.2 A comparison across text types

The purpose of this second part of the analysis is to examine to what extent text type can be used to further explain previous claims of (in)formality. Based on previous research (e.g. Ädel, 2008; Callies, 2013), we expect that the argumentative essays from SWICLE will behave differently from the texts in ALEC and VESPA and from the experts’ Academic Prose. In addition, assuming that text type is an important factor, we could expect SWICLE and LOCNESS to behave similarly, despite the fact that they differ with regard to NS status; if this is the case, it would seem that text type is a more powerful predictor than NS status.

Indeed, SWICLE and LOCNESS exhibit similar frequencies, as can be seen in Figure 6 (which is an extension of Figure 1). The students whose texts are included in SWICLE and LOCNESS make considerably more frequent use of this construction

(1,425 and 1,338 pmw, respectively) than the expert academic writers (885 pmw) and the VESPA+ALEC (823 pmw) students.

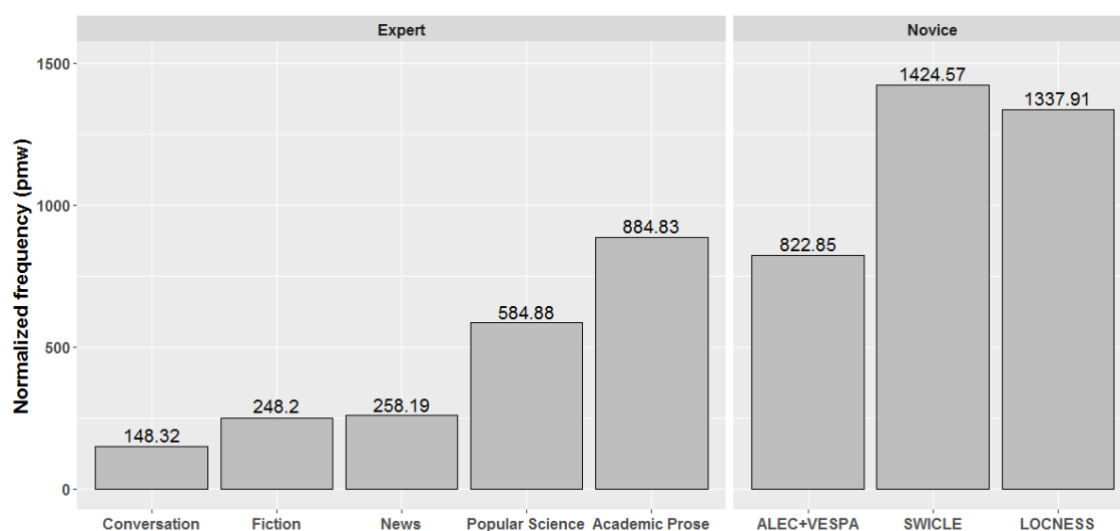


Figure 6. Normalized frequency of the construction across all the sub-corpora.

Regarding the adjective-verb pairings used, further differences were noted that seem to confirm the hypothesis stating that text type is an important factor. For example, additional analysis of the SWICLE data shows that these students make use of pairings that are not found even once in the multi-million-word *Corpus of Contemporary American English* (COCA) (Davies, 2008), such as *OK+skip* in example (20) below.¹²

(20) I think it is **ok** to **skip** Christmas sometimes and go to Hawaii [...].

(SWICLE_0047.2)

Similarly, the LOCNESS data include adjective-verb combinations not once found in COCA, such as *CHEAP + shoot* and *boring + DO*, as in (21).

(21) [...] it is highly **boring** to **do** the same thing over and over again.
(LOCNESS_alevels6)

¹² The following searches were carried out in COCA: *it* VERB (*) (*ok|oke|okay*) to *skip*.

A quantitative investigation of the data using an MCA points to further similarities between SWICLE and LOCNESS. As can be recalled from section 3.2, an MCA is an exploratory statistical test used to visualize any underlying structures in the data (in this case with regard to the adjective-verb pairings). The results of the MCA are shown in Figure 7 below; the graph shows the learner and expert data plotted on a Euclidian space based on the top 30 highest contributing variable categories.¹³

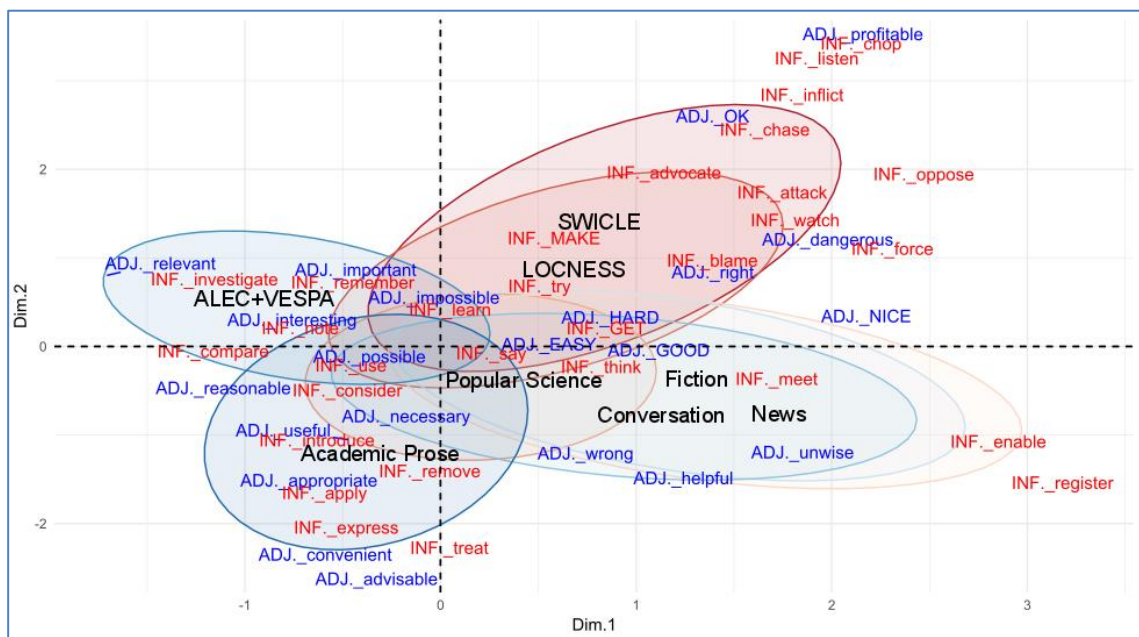


Figure 7. MCA of all the data, showing the top 30 contributing variable categories.

As can be seen, there are both similarities and differences between the sub-corpora. On one dimension, shown on the x-axis, the ALEC+VESPA learners' use is most similar to that of the academic experts (in that they are both found on the left side of the plot), whereas the SWICLE students' usage resembles that of the LOCNESS students (both

¹³ The input for the MCA in Figure 7 was created by using the function 'MCA()' from the {FactoMineR} package (Lê et al. 2008), and the biplot was created using the function 'fviz_mca_biplot()' from the {factoextra} package (Kassambara & Mundt, 2017).

found in the upper-right quadrant of the plot). It is also interesting to note that on the second dimension, shown on the y-axis, all the expert sub-corpora group together and the novice writer corpora group together (in that the learner corpora and LOCNESS are found on the upper half of the plot and the expert sub-corpora on the lower half). This suggests, first, that while ALEC and VESPA are closer to the academic expert norm, none of the learner corpora are fully like any of the expert subcorpora. Second, this also suggests that text type is a powerful predictor, regardless of NS status.

Although it should be kept in mind that the raw frequencies are relatively low, Figure 7 also shows that adjectives such as *OK*, *dangerous* and *profitable* are more strongly associated with SWICLE and LOCNESS than with any of the other sub-corpora (as these adjectives are found in the top right corner of the plot). As this set includes adjectives that have been reported to be “unexpected” or “speech-like” in previous research (see Römer, 2009: 156 on *dangerous*), and as these adjectives were almost non-existent inside the construction in the other corpora, we searched for these three adjectives in the full corpora (i.e. inside *and* outside the construction) as well, in order to further investigate their distribution.¹⁴ These adjectives could perhaps be expected to be more frequent in informal registers such as Conversation than in formal ones such as Academic Prose. If this is indeed the case, we could expect these to occur more frequently in the non-academic registers, perhaps gradually increasing in frequency from Academic Prose, to Popular Science, News, Fiction and Conversation. However, as is shown in Figure 8, the distributions of the adjectives across the registers does not follow a clear pattern with regard to (in)formality from the perspective of the registers, with one exception, namely

¹⁴ Of these four adjectives, only *dangerous* occurs inside the construction in the expert data (3 times in News and once in Popular Science). By way of comparison, it is not uncommon for these three adjectives to be followed by a *to*-infinitive in SWICLE: *profitable* 43% (3/7); *dangerous* 7% (3/42); and *OK* 83% (5/6).

OK. Nonetheless, two conclusions can be drawn based on these results, both of which seem to confirm that SWICLE and LOCNESS behave differently from the other sub-corpora: (i) there are adjectives that are used inside the construction in SWICLE and LOCNESS that are not found inside the construction in the other data, and (ii) these adjectives are most frequent in the non-academic registers.

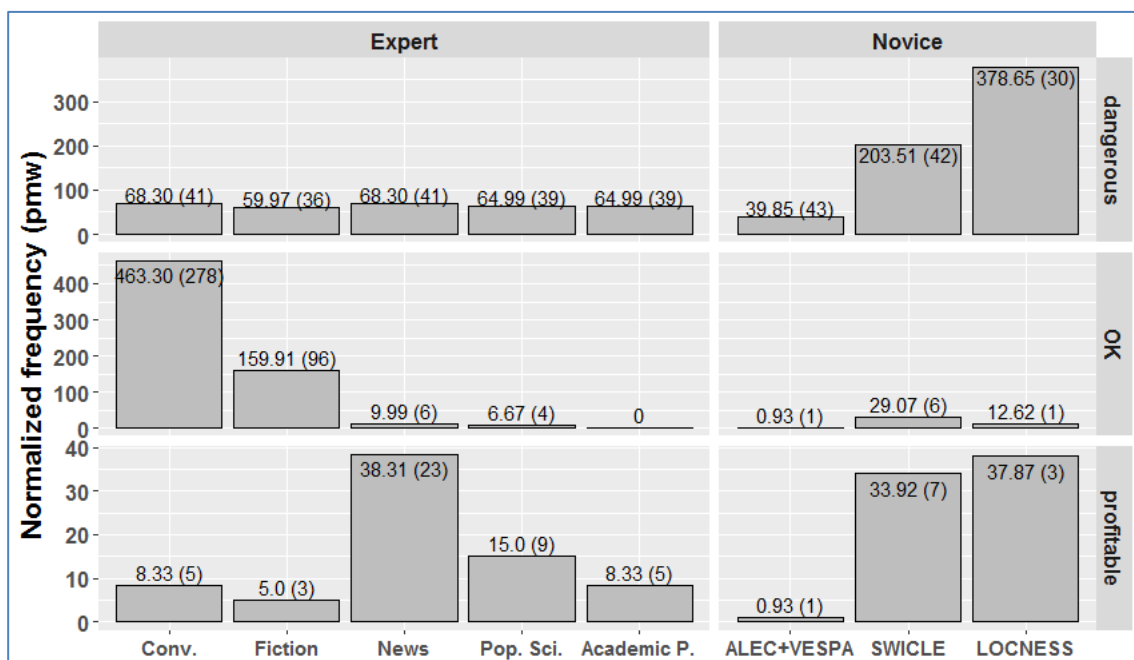


Figure 8. Normalized frequency of selected adjectives across the sub-corpora (raw frequency given in parentheses).

The different text types included in ALEC/VESPA (theses) and SWICLE/LOCNESS (argumentative essays) thus seem to explain much of the difference between these corpora. However, there are also other differences that should be kept in mind, such as differences with regard to text length (where the texts in ALEC and VESPA are considerably longer), number of drafts permitted (redrafting was allowed for the texts in ALEC and VESPA) and whether the texts are timed or not (all texts in ALEC and VESPA

are untimed).¹⁵ Several previous studies have noted that time availability can significantly influence the output (see, e.g., Ädel, 2008; see also Chafe, 1986).

Since the expert academic writing is characterized by longer, untimed, academic writing, it would seem that learner corpora such as ALEC and VESPA are more suitable for comparisons with longer expert texts (such as those included in BNC-15) than SWICLE. While the ICLE sub-corpora (which are commonly used to represent learner academic writing) remain an invaluable resource to researchers in the field, the results of the present study suggest that comparisons between the ICLE sub-corpora and reference corpora that were not compiled to be comparable to ICLE with regard to text type, text length, etc., should perhaps be treated with care (see also Ädel, 2006; 2008 and Callies, 2013 for a discussion of corpus comparability in learner corpus research). More generally, these results urge researchers to examine what we take to constitute “academic writing” in the context of comparability between corpora. This is a methodological question for researchers not only in the field of Learner Corpus Research, but also in neighboring fields such as English for Academic Purposes (EAP), English for Specific Purposes (ESP) and Second Language Acquisition (SLA). A somewhat simplistic answer could perhaps be that we should try to find data sets that are as comparable as possible, while working in a reality where availability and feasibility need to be negotiated. To be able to work with learner corpora that are highly comparable to existing reference corpora of published expert writing (including journal articles, etc.), corpus compilation projects such as VESPA play a crucial role, along with other similar projects around the world.

¹⁵ Further potentially interrelated differences include task prompt (i.e. selection of different topics), intended audience and aim of the text (see Ädel, 2008, Callies, 2013).

5. Conclusion

The present article has investigated to what extent register and text type can be considered to be important factors with regard to the use of the construction in learner and expert writing. In the first part of the study, where learner data from ALEC and VESPA were compared to the BNC-15 data, the results showed that the learners used the construction largely in an expert-like manner with regard to the overall and per-text frequencies. Nonetheless, when the adjective-verb pairings were taken into consideration as well, a slightly more complex picture emerged. While there was overlap between many of the high-ranking pairings with regard to collocation strength across the learners' and experts' Academic Prose, the learners shared certain adjectives and adjective-verb pairings only with the non-academic registers. Furthermore, the learners tended to overuse adjectives that were frequent in the expert academic writing, and they also used pairings that were unique to them. However, one caveat that needs to be kept in mind here is that the learner data are not fully comparable to the experts' Academic Prose in that the text length and number of different disciplines covered differ, which may explain some of the differences noted between these data. Nonetheless, it seems that the learners exhibit insufficient register awareness rather than "informal" use, in that when they used pairings that were uncommon in the experts' Academic Prose, they did not exclusively use pairings that were associated with the least formal expert register, namely Conversation.

As very little evidence was found in the data to support claims from previous studies that L1 Swedish students are "informal" in their writing, the corpus used in these previous studies, SWICLE, was added to the investigation in the second part, along with its NS-

student equivalent, LOCNESS. The results showed that the students whose texts were included in SWICLE and LOCNESS used the construction much more frequently than the ALEC and VESPA students and the experts. While the SWICLE and LOCNESS data showed evidence of informal use of the construction (in that the data included many pairings that were shared with the less formal registers, Conversation and Fiction), these corpora also included a considerable number of tokens that were unique to these data. No clear differences across NS status (SWICLE vs. LOCNESS) were noted; on the contrary, SWICLE and LOCNESS both exhibited results that were distinct from the other corpora. As SWICLE and LOCNESS, unlike ALEC and VESPA, comprise (timed) argumentative essays, this finding concurs with previous research stating that text type is a powerful predictor to explain differences in use.

It should, however, be noted that while the present study investigates a construction that can be considered formal (in that it is more typical of academic writing than of the non-academic registers), previous studies commenting on learners' informal style tend to base their claims of informality primarily on overuse of speech-like features, such as adverbs. It could thus be the case that the learners are in the process of mastering the use of frequent, formal constructions, while at the same time falling back on (and thus exhibiting somewhat lacking register awareness of the use of) other constructions and linguistic features that are less frequent in Academic Prose. Further investigations and comparisons of such features would likely provide a very fruitful avenue for future research. Other such avenues include investigations of disciplinary differences in learner and expert data and across registers.

All in all, it would seem that dismissing all (L1 Swedish) learner writing as “informal”, appears to be an oversimplification, since there seem to be many intertwined

factors, including register awareness and text-type differences, that need to be taken into account in discussions of (in)formal use. We hope that the findings of this study will contribute to a more nuanced view of (in)formal usage in learner data, thus benefitting both L2 theory and L2 instruction. Finally, we hope that the method developed and explored here can be useful for future studies of other constructions and other L1s, thereby benefitting future studies of (in)formality in learner language.

Acknowledgements

We are very grateful to the editorial team and the anonymous reviewers for their helpful comments and suggestions.

References

- Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam and Philadelphia: John Benjamins.
- Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp & M.B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp.35–53). Amsterdam: Rodopi.
- Ädel, A. (2014). Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. *Journal of English for Academic Purposes*, 16, 68–80.

- Advanced Learner English Corpus (ALEC). Corpus compiled by Tove Larsson at Uppsala University in 2013.
- Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80–93). London: Longman.
- Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Burnard, L. (2007). *Reference guide for the British National Corpus (XML Edition)*. Available at: <http://www.natcorp.ox.ac.uk/docs/URG/> (accessed June 2017).
- BNC-15. Subset of the British National Corpus (BNC) sampled by Henrik Kaatari at Uppsala University in 2012.
- Callies, M. (2009). *Information highlighting in advanced learner English. The syntax-pragmatics interface in second language acquisition*. Amsterdam and Philadelphia: John Benjamins.
- Callies, M. (2013). Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics*, 18(3), 357–390.
- Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe & J. Nichols (Eds.), *Evidentiality: The linguistic coding of epistemology* (pp. 261–272). Norwood: Ablex.

- Davies, M. (2008). The Corpus of Contemporary American English (COCA): 520 Million Words, 1990–present. Available online at <http://corpus.byu.edu/coca/> (accessed June 2017).
- Evert, S. (2006). How random is a corpus? The library metaphor. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), 177–190.
- 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 and Philadelphia: John Benjamins.
- Granger, S., & Petch-Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes: Journal of English as an International and Intranational Language*, 15(1), 17–27.
- Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). International Corpus of Learner English. Version 2. Handbook + CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.
- Gries, S. Th. (2007). Coll.analysis 3.2a. A Program for R for Windows 2.x.
- Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes*, 4(3), 257–277.
- 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.

- Hasselgård, H. (2009). Thematic choice and expressions of stance in English argumentative texts by Norwegian learners. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 121–139). Amsterdam and Philadelphia: John Benjamins.
- Herriman, J., & Boström Aronsson, M. (2009). Themes in Swedish advanced learners' writing in English. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 101–120). Amsterdam and Philadelphia: John Benjamins.
- Hewings, M., & Hewings, A. (2002). 'It is interesting to note that...': A comparative study of anticipatory *it* in student and published writing. *English for Specific Purposes*, 21(4), 367–383.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Kaatari, H. (2016). Variation across two dimensions: Testing the Complexity Principle and the Uniform Information Density principle on adjectival data. *English Language and Linguistics*, 20(3), 533–558.
- Kaatari, H. (2017). Adjectives complemented by *that*- and *to*-clauses: Exploring semantico-syntactic relationships and genre variation. Unpublished doctoral dissertation, Uppsala University, Uppsala, Sweden.
- Kaltenböck, G. (2005). *It*-extraposition in English: A functional view. *International Journal of Corpus Linguistics*, 10(2), 119–159.
- Kassambara, A., & Mundt, F. (2017). *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.4. Available at: <https://CRAN.R-project.org/package=factoextra> (accessed June 2017).

- Larsson, T. (Forthcoming). A syntactic analysis of the introductory *it* pattern in non-native-speaker and native-speaker student writing. In M. Mahlberg & V. Wiegand (Eds.), *Corpus Linguistics, Context and Culture*. Berlin: De Gruyter Mouton.
- Larsson, T. (2014). Introducing the Advanced Learner English Corpus (ALEC): A new learner corpus. Poster presented at VU University Amsterdam, Amsterdam, The Netherlands, 20 January, 2014.
- Larsson, T. (2016a). The introductory *it* pattern: Variability explored in learner and expert writing. *Journal of English for Academic Purposes*, 22, 64–79.
- Larsson, T. (2016b). The introductory *it* pattern in academic writing by non-native-speakers students, native-speaker students and published writers: A corpus-based study. Unpublished doctoral dissertation, Uppsala University, Uppsala, Sweden.
- Larsson, T. (2017). *The importance of, it is important that or importantly?* The use of morphologically related stance markers in learner and expert writing. *International Journal of Corpus Linguistics*, 22(1), 57–85.
- Larsson, Tove. (2018). Is there a correlation between form and function? A syntactic and functional investigation of the introductory *it* pattern in student writing. *ICAME journal*, 42(1), 13–40.
- Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.
- Lee, D. (2001). Genres, registers, text-types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3), 37–72.

- Louvain Corpus of Native English Essays (LOCNESS). Corpus compiled at the Centre for English Corpus Linguistics at Université catholique de Louvain in Belgium.
- Paquot, M., Hasselgård, H., & Oksefjell Ebeling, S. (2013). Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead* (pp. 277–288). [Corpora and language in use – Proceedings 1]. Louvain-la-Neuve: Presses universitaires de Louvain.
- Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107–118). London: Longman.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/> (accessed April 2017).
- Ramhøj, R. (2016). On clausal subjects and extraposition in the history of English. Unpublished doctoral dissertation, University of Gothenburg, Gothenburg, Sweden.
- Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, 7(1), 140–162.
- Smith, E. L. (1986). Achieving impact through the interpersonal component. In B. Couture (Ed.), *Functional Approaches to Writing* (pp. 108–119). London: Frances Pinter.

Stefanowitsch, A., & Gries, S. Th. (2005). Co-varying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1), 1–43.

Tapper, M. (2005). Connectives in advanced Swedish EFL learners' written English – preliminary results. *Working Papers in English Linguistics*, 5, 116–144.

Zhang, G. (2015). *It is suggested that* or *it is better to*? Forms and meanings of subject *it*-extraposition in academic and popular writing. *Journal of English for Academic Purposes*, 20, 1–13.

Appendix

In this appendix, the top 15 adjective-verb pairings in terms of collocation strength (coll.str.) are listed for each sub-corpus.

Collocation strength:

> 3 = $p < 0.001$

> 2 = $p < 0.01$

> 1.3 = $p < 0.05$

ALEC+VESPA						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.
important	note	134	39	22	5.9	9
interesting	see	66	56	17	4.2	7.1
reasonable	assume	32	21	8	0.8	6.8
interesting	note	66	39	13	2.9	5.9
fine	kill	2	3	2	0	5.1
important	remember	134	10	8	1.5	5.1
possible	detect	163	9	8	1.7	5.1
tempting	regard	3	2	2	0	5.1
SAFE	assume	9	21	4	0.2	4.6
unrealistic	expect	2	5	2	0	4.6
fun	study	7	12	3	0.1	4.2

plausible	presume	5	3	2	0	4.1
reasonable	argue	32	8	4	0.3	4.1
feasible	claim	2	10	2	0	3.9
important	learn	134	19	10	2.9	3.9

SWICLE						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.
essential	preserve	3	2	2	0	4.2
HARD	imagine	26	5	4	0.4	3.6
dangerous	generalise	4	3	2	0	3.4
HARD	find	26	11	5	1	2.9
important	remember	38	3	3	0.4	2.7
advantageous	close	1	1	1	0	2.5
allright	BE	2	17	2	0.1	2.5
civilized	chop	1	1	1	0	2.5
comfortable	sit	1	1	1	0	2.5
crucial	raise	1	1	1	0	2.5
feasible	fly	1	1	1	0	2.5
great	innovate	1	1	1	0	2.5
healthy	reflect	1	1	1	0	2.5
illegal	hit	1	1	1	0	2.5
important	know	38	6	4	0.8	2.5

LOCNESS						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.strength
illegal	attack	2	3	2	0.1	3.3
legal	chase	3	2	2	0.1	3.3
HARD	imagine	5	2	2	0.1	2.7
TRUE	say	2	7	2	0.1	2.4
important	look	8	2	2	0.2	2.3
boring	DO	1	1	1	0	2
CHEAP	shoot	1	1	1	0	2
convenient	move	1	1	1	0	2
expensive	partake	1	1	1	0	2
feasible	imply	1	1	1	0	2
fun	experience	1	1	1	0	2
inhumane	support	1	1	1	0	2
justifiable	spend	1	1	1	0	2
SAFE	eat	1	1	1	0	2
sensible	put	1	1	1	0	2

BNC-15: Academic Prose						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.

reasonable	suppose	20	6	5	0.2	6.6
important	note	37	12	7	0.8	5.6
important	recognise	37	4	4	0.3	4.7
important	remember	37	4	4	0.3	4.7
reasonable	assume	20	6	4	0.2	4.7
unusual	find	3	15	3	0.1	4.7
important	realise	37	3	3	0.2	3.5
reasonable	ask	20	5	3	0.2	3.4
HARD	understand	13	2	2	0	3.3
appropriate	consider	20	12	4	0.5	3.2
interesting	note	9	12	3	0.2	3.2
surprising	find	2	15	2	0.1	3.1
essential	ensure	5	7	2	0.1	2.8
impossible	BE	23	13	4	0.6	2.8

BNC-15: Popular Science						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.
interesting	note	9	8	6	0.2	9
EASY	see	42	19	10	2.3	5.1
surprising	find	3	14	3	0.1	4.3
unreasonable	expect	3	2	2	0	4.3
unwise	make	4	11	3	0.1	4
TRUE	say	2	8	2	0	3.3
advantageous	point	1	1	1	0	2.5
ambitious	attempt	1	1	1	0	2.5
ample	accomodate	1	1	1	0	2.5
convenient	simplify	1	1	1	0	2.5
dangerous	climb	1	1	1	0	2.5
disadvantageous	profess	1	1	1	0	2.5
fanciful	match	1	1	1	0	2.5
foolish	treat	1	1	1	0	2.5
futile	spend	1	1	1	0	2.5

BNC-15: News						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.
interesting	see	3	10	3	0.2	3.7
essential	act	3	2	2	0	3.6
difficult	GET	13	6	4	0.5	3.4
NICE	HAVE	5	2	2	0.1	3.1
important	know	6	2	2	0.1	2.9
possible	BRING	11	2	2	0.1	2.3
advisable	buy	1	1	1	0	2.2
compulsory	base	1	1	1	0	2.2

crucial	move	1	1	1	0	2.2
customary	bear	1	1	1	0	2.2
disconcerting	tackle	1	1	1	0	2.2
embarrassing	slip	1	1	1	0	2.2
helpful	talk	1	1	1	0	2.2
high-fashion	emphasise	1	1	1	0	2.2
imperative	secure	1	1	1	0	2.2

BNC-15: Fiction						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str.
helpful	know	5	4	3	0.1	4.1
important	HAVE	6	9	3	0.4	2.5
dreadful	benefit	1	1	1	0	2.2
hopeless	suppose	1	1	1	0	2.2
impossible	work	13	2	2	0.2	2.2
lucrative	enable	1	1	1	0	2.2
natural	cultivate	1	1	1	0	2.2
presumptuous	flout	1	1	1	0	2.2
quicker	walk	1	1	1	0	2.2
unusual	borrow	1	1	1	0	2.2
unwise	force	1	1	1	0	2.2
vital	prevent	1	1	1	0	2.2
fair	say	4	7	2	0.2	2
advisable	take	1	2	1	0	1.9
appropriate	feel	1	2	1	0	1.9

BNC-15: Conversation						
words1	words2	freq.w1	freq.w2	obs.w1_2.in_c	exp.w1_2.in_c	coll.str
difficult	GET	10	5	3	0.6	2
interesting	see	3	6	2	0.2	2
cost-effective	replace	1	1	1	0	1.9
difficult	find	10	2	2	0.2	1.9
HARD	escape	1	1	1	0	1.9
illegal	park	1	1	1	0	1.9
impossible	achieve	1	1	1	0	1.9
irrelative	accept	1	1	1	0	1.9
SIMPLE	set aside	1	1	1	0	1.9
unfair	throw	1	1	1	0	1.9
appropriate	cobble	2	1	1	0	1.6
bizarre	say	1	2	1	0	1.6
foolish	assume	1	2	1	0	1.6
exciting	hear	1	3	1	0	1.5
fascinating	hear	1	3	1	0	1.5

Authors' addresses

Tove Larsson

Centre for English Corpus Linguistics

Language and Communication Institute

Université catholique de Louvain

Collège Erasme

Place Blaise Pascal 1, bte L3.03.31

1348 Louvain-la-Neuve

Belgium

Henrik Kaatari

Department of Humanities

University of Gävle

80176 Gävle

Sweden