# Recycling the data: Building and using a learner business English writing corpus


## Abstract

This paper describes the construction and use of a small corpus of business English writing by Swedish university students to form the basis of a "learning driven data" (Seidlhofer, 2002) approach to their studies in business writing. Course assignments were used to construct a small corpus which was analysed from a lexical perspective in relation to the Business Service List (BSL) (Browne & Culligan, 2016) and an online Business Letter Corpus (BLC) to identify errors and gaps in knowledge. The findings were then used to inform course content and form the basis for tasks using the learner corpus, which will be integrated into the course structure to provide opportunities for data-driven learning. This study contributes to the growing number of pedagogic applications of learner corpora, demonstrating an approach that could be adapted to a range of other learning contexts.


## Paper

### Introduction

Most Swedish students enter university as competent speakers of English, but their written English skills can lag behind, leading to high demand for courses in writing, particularly in business communication. The present research study began in response to a need to improve the existing online Business Writing course at MidSweden University, and add an advanced module. However, as the level of achievement on the existing course was high, the students' needs were not transparent. To better determine those needs, I built a small corpus of consenting students' assignments from the present course, described below. When grading assignments, I had noted a lack of precision in lexis, and some issues with register and tone so I set about exploring these more systematically. My specific aims were to examine the coverage and use of business vocabulary and phrasal language in the corpus. Online tools were used to do this, and the approach and outcomes are described here. Furthermore, the learner corpus itself is considered as potential input for the learners, through "learning driven data" (Seidlhofer, 2002), and some language points for exploration are identified. To begin with, I will briefly outline the general research context informing my approach, before describing the study in more detail.

### Learner corpora and learning driven data

Many and varied types of learner corpora have been constructed in recent years, demonstrated by the extensive list maintained by the Centre for English Corpus Linguistics (see https://uclouvain.be/fr/node/12075). Alongside large-scale projects such as the International Corpus of Learner English (ICLE) and commercial publishers' learner corpora, many smaller in-house corpora have been compiled. These are generated within a specific learning context, containing written or spoken discourse relating to a specific learning event or events. While small-scale corpora of this kind may lack the generalizability needed for comparative interlanguage analyses, their potential as a teaching resource has been acknowledged (e.g. Cotos, 2014; Guilquin et al., 2007; Granger, 2002, 2009; Mukherjee & Rohrbach, 2006; Nesselhauf, 2004; Seidlhofer, 2002). In-house learner corpus data is valuable to both teacher and learner. The teacher can use it, data-driven learning (DDL) style (Johns, 1991), by searching for keywords and looking for patterns in the resulting data, to find out about learner language use, using this to inform curriculum and materials design (Granger, 2002). For learners, it is relevant, having been collected from peers experiencing the same learning environment, and it can help raise awareness of linguistic deficiencies they may share in a non-personal, non-threatening way (Seidlhofer, 2002). The fact that it is not their own writing, but is *like* their own, encourages greater objectivity. Learning driven data (LDD)

and DDL are complementary processes (Seidlhofer, 2002), as lexical patterns found in the learner corpus can be explored in an expert corpus so that differences may be noticed and corrected (Flowerdew 2012). The benefits of DDL have been widely demonstrated in recent years (see Boulton & Cobb 2017 for an overview), whereas applications of learner corpora in the classroom have been slower to emerge, and the present study aims to add to these.

**Learning context**

The specific learning context for this study is a Business Writing course at Mid-Sweden university (7.5 credits). To enrol, students must fulfil general admission requirements and have English proficiency equivalent to a minimum overall score of 6.5 in IELTS Academic Training (minimum 5.5 in each skill). Most course participants are advanced in, or have completed their primary degree, and are looking ahead to prepare for employment in the future. The majority of these students are Swedish with a high level of spoken English. The course aims to improve both comprehension and production of a range of different types of written business communication. The course outcomes indicate that the student should be able to understand, recognize and produce different types of written business documents in English, recognize and produce different styles (formal, informal, informative) of written English, and engage in collaborative work to improve their own and other students' written business English. The course is delivered online using the learning platform Moodle, and runs part-time over a 20 week semester. All teaching is carried out via Moodle; course material is posted, lectures can be accessed, and forum discussions take place there. The course is designed so that student learning is a process; there is obligatory group work involving peer reviewing and responding to other students' writing, followed by editing before submission for grading. Thus assignments submitted are generally of a high standard, with most basic language errors edited out, and these are what formed the basis of the learner corpus discussed here.

**Collecting the learner data**

Although the digital nature of the course meant that assignments were readily available to form a corpus, this was restricted by institutional requirements. University policy requires a written signature on a designated consent form before student work can be used for research purposes. Given that all course delivery is web-based via Moodle, and no other aspect of the course required students to print, sign and scan material, it was difficult to get students to respond to this request. As a result, two terms were needed to gather the data, and the resulting corpus is smaller than had been envisaged (although the possibility to add further data remains). The requirement for digital privacy reduced the data further, as one assignment involved a resume and cover letter, which could not be sufficiently anonymised. As a result, the corpus consists of four assignments (three of which consist of multiple documents) from two cohorts of students in two consecutive terms, a total of 33 students. The small size of the corpus and the range of text types included is recognised as a limiting factor in this study, but even a limited corpus like this can offer some insights into learner performance, particularly when the learner profile is homogeneous. The types of business writing and word counts from the various assignments are shown in Table 1.

Table 1: Composition of the Learner Corpus

| | **Business Writing** | **Word count** |
|---|---|---|
| Assignment 1 | Request letter<br>Reply letter | 5,220<br>5,942 |
| Assignment 2 | Complaint letter<br>Claim letter<br>Adjustment letter | 7,382<br>5,590<br>5,156 |
| Assignment 3 | Proposal<br>Memo | 10,720<br>5,318 |

| | | |
|---|---|---|
| Assignment 4 | Informal Report | 25,174 |
| Total | | 70,502 |

The assignments were anonymized, coded and saved as plain text documents so that they could be analysed using AntConc (Anthony 2018). The coding facilitated manipulation of the corpus, e.g. for the correspondence portion to be easily extracted.

**Lexical profile of the learner corpus**

First, I wanted to find out what range of vocabulary was in productive use by the learners. A vocabprofiler on Cobb's (2018) website was used to establish the range of lexis in the corpus, using the Business Service List 1.01 (BSL) (Browne & Culligan, 2016) option. The BSL identifies 1700 business-related words extracted from a corpus (approximately 64 million words) of business texts, newspapers, journals and websites, giving up to 97% coverage of general business English materials when combined with the 2800 words of the New General Service List (see Browne & Culligan 2016 for an account of this). Although the BSL is most relevant as a measure of receptive vocabulary, it offers a useful indicator of productive ability in business-related lexis.

Table 2 shows the learner corpus broken down into levels of frequency of lemmas (headword and inflections) used at the 1- 3 K level of the NGSL and those occurring on the BSL list. The majority of the words used were from the first 2K of the NGSL, which would also be a typical profile for native speakers (Cobb, 2008). Less frequent words also occur, with just under half of the 3K words represented, as do lexical items from the BSL, with 203 of the 1200 lemmas on the BSL present in the list. This demonstrates some productive ability in business vocabulary, but further examination (see Figure 1) shows that 100 of these lemmas appear only once in the corpus, 36 appear twice, 17 are used three times, 10 four times, and beyond that only 40 words are used five times or more. In short, although knowledge of business-related vocabulary is evident, it is not widely used in the corpus, although this may be due to its limited size.

Table 2: Breakdown of lemmas, types and tokens in the BW corpus by NGSL frequencies, and BSL.

| **Freq. Level** | **Lemmas (%)** | **Types (%)** | **Tokens (%)** |
|---|---|---|---|
| NGSL_1 [1000 lemmas]) | 924 (42.92) | 1783 (32.14) | 56144 (79.68) |
| NGSL_2 [1000 lemmas] | 663 (30.79) | 946 (17.05) | 4586 (6.51) |
| NGSL_3 [801 lemmas] | 363 (16.86) | 451 (8.13) | 2184 (3.10) |
| BIZ (BSL) [1200 lemmas] | 203 (9.43) | 242 (4.36) | 1317 (1.87) |
| Off-List: | ?? | 2131 (38.41) | 6235 (8.85) |
| Total (unrounded) | 2153+? | 5548 (100) | 70466 (100) |

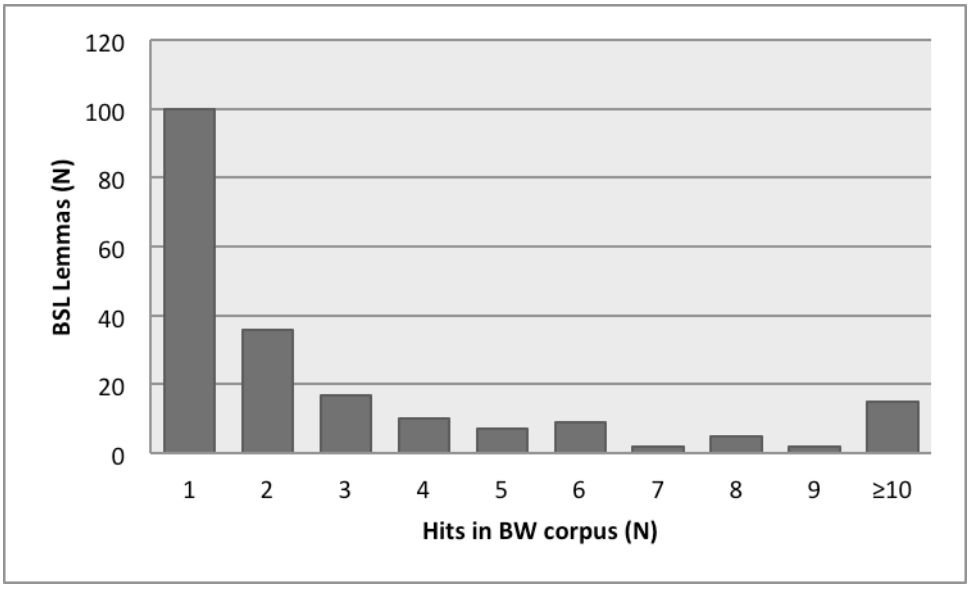

Figure 1: Number of hits on each of the 203 BSL lemmas in the BW corpus

I went on to evaluate the students' ability to use the BSL words. Most of those appearing with higher frequency, i.e. with 6 hits and above, were used appropriately, e.g. *manual* work / read the *manual,* large amounts of *goods* / damaged *goods*, although some were used only in repeated proper nouns, e.g. High *Tech* Solutions. Of the less frequently used BSL words, some contextual uses lacked precision, either in the choice of word or its collocates, as examples 1 and 2 show (BSL words italicized).

(1) *we are looking forward for this *proposition*
(2) *green products really has taken its *momentum* on the market

Words in this category, i.e. that were infrequently and inappropriately used, offer a starting point for target items that could be explored through an LDD / DDL approach, highlighting potential problems in using the words in context when compared to uses in the expert corpus. A sample approach using the example *contrary* is given in Figure 2.

Compare this learner's use of *contrary* in (A) to some examples found in a native business letter corpus (B)*.
What patterns of use can you find in (B)?
In view of this, how could you improve sentence (A)?
(A)
I have made all necessary purchases on your website, and they have all ended successfully; on the **contrary** to my recent purchase.

(B)

On the **contrary**, the firm's difficulties would seem to be
Quite the **contrary**, we count you as one of our most valuable
would be somewhat to the **contrary**.
if I don't hear from you to the **contrary**, I shall be with you at 3 p.m.
Unless I hear from you to the **contrary**, I will assume that this is the correct

*(for more examples, and more extended contexts, type *contrary* into the search box at http://www.someya-net.com/concordancer/index.html) :

Figure 2: Sample LDD / DDL activity

**Lexical patterning**

Examining lexical bundles (i.e. frequently occurring phrases) offered another approach to finding errors and gaps both in lexis and grammar in the learner corpus. I used data from the Business Letter Corpus (BLC), a one million word corpus of business letters from UK

and US sources for this, and used only the letter component of the learner corpus[1]. The BLC website provides access to an online concordancer for the corpus, and key data about the corpus, including frequency lists of meaningful 3-, 4- and 5-word bundles. The 3- and 4-word bundle lists were used for comparison with the learner corpus, complemented with searches in the BLC as necessary. AntConc (Anthony 2017) was used to extract 3- and 4-word bundles from the learner corpus, edited according to criteria stated on the BLC lists, i.e. to remove bundles with less relevant lexical content such as dates / times, opening salutations and closes, and phrases beginning with *and / but / or / me / that* (as a relative pronoun). The size of the learner corpus meant that only a small number of bundles occurred; nevertheless cross-referencing these with those occurring in the BLC yielded some interesting results. The most frequent bundles, with only a few exceptions[2], were similar across both lists, but there were differences that offered the potential for learner exploration, some examples of which are outlined below.

***Verb phrases***

A key verb appearing in the lexical bundles in both corpora was *appreciate,* appearing with a normalised frequency of 608 per million in the learner corpus and having 1005 hits in the million word BLC. However, there were differences in the 4-word lexical bundles *appreciate* appeared in, as shown in table 2. A common misuse of the form *I would appreciate if* is evident in the learner corpus (cf. Flowerdew, 2012), and a much wider range of forms appears in the BLC; some of these have one or two hits in the learner corpus, but the italicized phrases in table 2 do not occur at all. Of course, the small size of the learner corpus is relevant here. However, such differences offer potential for raising awareness of how modifications like *how much, greatly, very much*, can be used to improve tone and style in business correspondence. This overlaps with another observation in the learner corpus bundles. There was a notable absence of adverbs, apart from *really,* in these. Searches for other adverbs in the learner corpus showed very few examples, several of which demonstrated inaccuracies such as **I would highly appreciate, *it will be strongly advised, *I cannot deeply enough describe my disappointment.* Clearly there is an awareness of collocation with adverbs, but confusion with the appropriate structural form and word choice detracts from this, and suggests another potential focus for DDL with the BLC, where many examples can be found.

Table 2: Four-word bundles with *appreciate* occurring in the two corpora, with number of hits shown.

| **Learner corpus (45,328 words)** | | **BLC (Someya 2000) (1 million words)** | |
|---|---|---|---|
| appreciate you bringing this | 3 | appreciate it if you | 93 |
| *I would appreciate if | 3 | would appreciate it if | 62 |
| | | we would appreciate it | 52 |
| | | I would appreciate your | 46 |
| | | *appreciate your interest in* | 31 |
| | | we would appreciate your | 30 |
| | | *how much we appreciate* | 30 |
| | | *would be greatly appreciated* | 24 |
| | | *appreciate it very much* | 23 |
| | | *we appreciate your business* | 21 |

***Register***

The 3-word bundles demonstrated some issues around the register of language used. In the learner corpus, *looking forward to* was used three times as much as *look forward to* (normalised frequencies of 739 compared to 239 per million, corresponding to a significant over-representation, log-likelihood 47.05, $p<.0001$), whereas the more

[1] This reduced the corpus to 45,328 words

[2] Certain bundles which were specific to the assignment aim, e.g. *This proposal deals with* did not occur in the BLC list.

traditional use *look forward to* was much more common in the BLC. This discrepancy may arise from increased usage of the progressive aspect since the BLC was constructed, and / or a Swedish tendency to overuse this form (Swan and Smith, 2001: 31). However, a search on *looking forward to* in the learner corpus also shows that in almost a third of the hits, the phrase is not preceded by a pronoun or auxiliary verb, intensifying the informality. No similar instances occurred in the BLC. In contrast to this, the rather formal fixed 3-word bundle *come to my attention* was significantly over-represented in the learner corpus, with 8 hits in total, as compared to only 9 hits in the one-million word BLC (log-likelihood 27.29, $p<.0001$). Of course, the learner corpus and the BLC contain different types of correspondence and styles of writing, so some variation in tone can be expected; fuller contexts must be considered to determine appropriacy. This in itself offers a focus for LDD / DDL for the learners, raising awareness about questions of register and tone, which are so important in business writing.

**Recycling the data**

The next step of the project will introduce learner data and DDL to students in a supportive and positive way. For technical ease, the aim is to provide user-friendly access to the learner corpus and a concordancer in Moodle, with a link to the BLC, which offers a simple interface and is sufficiently small scale to be manageable (see Flowerdew, 2012). A series of guided tasks will scaffold the learners as they are introduced to and become more familiar with DDL techniques, with the areas outlined above as starting points. In keeping with the co-operative nature of the course, tasks will be carried out in groups with forum discussions that can be monitored, allowing for support to be provided as necessary. Outcomes can be measured both quantitatively at the end of the course, by assessing the body of student work (submitted in digital format, allowing for easy analysis) and qualitatively through the forum discussions and course feedback.

**Conclusion**

Using data from learners with the same background and linguistic profile offers the potential to focus teaching so that it is directly relevant to the learner. In the same way that data-driven learning can confirm or contradict learner's intuitions about language patterns, learner data provides the opportunity for teachers to confirm intuitions about what is known and what needs to be addressed. Exploring the learner corpus made it evident that problem areas I had identified (register and lack of precision in vocabulary) require further focus, while also revealing gaps (i.e. adverb use) I was less aware of. The analysis pinpointed specific language that requires attention, and a means of raising awareness about it in a relevant way.

Constructing a small learner corpus is easily achieved as digital courses become increasingly common, as long as requirements for data use are respected. Similarly, more user-friendly tools are becoming freely available online, allowing teachers to create lexical and grammatical profiles of their learners with ease, helping them to identify their achievements and needs, and take a systematic approach to personalised course development. There are weaknesses in the case presented here, notably in the limited size of the learner corpus, that the BLC is rather old, and the BLS is more useful as a measure of receptive than productive knowledge. However, even with such limitations, it seems to me that using learner data can only complement a 'one size fits all' published course, and introducing LDD and DDL is a natural development from this.

## References


Anthony, L. (2018). AntConc (Version 3.5.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning,* 67:348-393. doi: 10.1111/lang.12224

Browne, C. & Culligan, B. (2016). Business Service List 1.01. Available at http://www.newgeneralservicelist.org/bsl-business-service-list/. Retrieved 30 March 2018.

Business Letter Corpus (BLC). Available at http://www.someya-net.com/concordancer/. Retrieved 30 March 2018.

Cobb, T. Compleat Web VP! [computer program]. Accessed 30 March 2018 at https://www.lextutor.ca/vp/comp/

Cobb, T. (2008). Some research uses of Vocabprofile. Available at https://www.lextutor.ca/vp/. Accessed 30 March 2018.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL 26*(2): 202-224.

Flowerdew, L. (2012). Exploiting a corpus of business letters from a phraseological, functional perspective. *ReCALL 24*(2): 152-168.

Gilquin, G., Granger, S. & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes 6*(4): 319-335.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.) *Corpora and Language Teaching* (pp. 13-34). Amsterdam/Philadelphia: Benjamins.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam/Philadelphia: Benjamins.

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning materials. *English Language Research Journal 4*: 1-16.

Mukherjee, J. & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in learner corpus research. In B. Kettemann & G. Marko (Eds) *Planning, Painting and Gluing Corpora. Inside the Applied Corpus Linguist's Workshop* (pp. 205-232). Frankfurt: Peter Lang.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 125-152). Amsterdam/Philadelphia: Benjamins.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learner-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.) *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213-234). Amsterdam/Philadelphia: Benjamins.

Swan, M. & Smith, B. (2001). *Learner English.* 2nd Edition. Cambridge: Cambridge University Press.