9

Learner corpora and lexis

Tom Cobb and Marlise Horst

PROOF

1 Introduction

Not very long ago, finding a niche for a learner corpus (LC) study at a second language research conference in North America was tricky. Researchers would typically manage to present their work on learner corpora under a heading such as literacy, pedagogy, second language acquisition, language assessment or technology. The situation has improved greatly in recent years. The American Association for Applied Linguistics now includes corpus studies as one of its official strands for research presentations. At the 2014 conference, nine presentations investigated learner corpora; of these, five focused on learner lexis, the topic of this chapter. In our view, this expanded interest in investigating learners' vocabulary development from a corpus perspective goes hand in hand with the current recognition of the centrality of vocabulary in acquiring language generally. This point has been compellingly argued by Bates and Goodman (1997) in the case of first language (L1) acquisition, and Gass and Selinker (2008: 173) observe that, for second language (L2) learners as well, 'language learning is largely lexical learning'.

This chapter discusses the ways teachers and researchers have used learner corpora to measure this most essential aspect of second language knowledge. We begin with concepts and definitions. The language productions that make up a learner corpus may be an assembly of either written texts or transcribed oral texts. The reason for assembling learner productions into a corpus rather than investigating them individually is to arrive at generalisable findings about language acquisition, development and use. A corpus of learner texts can show researchers, teachers and language learners which words particular groups of learners are able to produce, with what degree of appropriateness and variety, but also which words they fail to produce and what kinds of errors occur. The chapter presents key studies to illustrate these uses.

9781107041196pt2_p183-306.indd 185 6/11/2015 1:56:38 PM

An often-cited overview of what is involved in knowing a word comes from Nation (2001). He makes a basic distinction between linking a form encountered in speech or reading to its meaning (recognition) and the ability to provide the spoken or written form that expresses a particular meaning (production). Clearly, in learner corpus research, which typically investigates learner essays or speech, we are on the production side of this distinction. This means that in terms of the learning process, learner corpora are best suited to shed light on L2 word knowledge that is nearing its end state (Cobb 2003). Abundant empirical research has shown that L2 word knowledge is acquired cumulatively, proceeding from meaning recognition in context to full appropriate productive use (Cobb 2007). A learner corpus reveals the items that have made it all the way into productive use. However, many words never complete the full journey (Laufer 1998), and learner corpora can also reveal what is missing, not yet activated, or not yet produced accurately.

Among the kinds of knowledge a second language learner might acquire about a new word, Nation's (2001) scheme includes knowing how the word collocates with other words in sentences and which multi-word units it may be a member of. Thus full knowledge of a word like *bucket* means knowing that it occurs frequently in sequences like *a bucket of (liquid)* and also in less frequent idiomatic expressions like *kick the bucket* and *bucket list*. In this chapter, the focus is on the single word, i.e. a string of adjoining letters set off by spaces on either side. For a discussion of collocations and multi-word lexis in learner corpora, the reader is referred to Chapters 10 and 16 (this volume), which deal with phraseology. Another aspect of Nation's (2001) scheme is register and discourse function. Full knowledge of a word like *plasma* includes realising that it is typically used in rather formal speech or writing, and typically in scientific discourse related to medicine. Learner corpus investigations of academic lexis and the lexis of specific subject areas are discussed in Chapter 21 (this volume). The focus in this chapter is on what might be called 'general' or non-specialised vocabulary – often (but not always) general English vocabulary. The main technique for assessing vocabulary use in learner corpora discussed in this chapter is lexical frequency profiling. Reasons for this choice are outlined in the following sections.

2 Core issues

2.1 Frequency of specific words

While learner corpus research requires human judgements to classify learners' lexical errors (e.g. Llach et al. 2006; Llach 2007) or identify ways learners use particular words (e.g. Altenberg and Granger 2001), corpus researchers also typically seek to take advantage of the computer's ability to search and assemble data automatically – since the main point of

PROOF

assembling a learner corpus is to look for trends and patterns that are not readily evident to the naked eye. One basic function that is more interesting than it may seem is the computer's ability to rapidly count up and sort instances of specific words. By way of illustration, Altenberg and Granger (2001) investigated whether French and Swedish learners of English over- or underused the verb *make* in their writing. The question was answered by comparing computer counts in the learner corpora to a comparable corpus of essays by native speakers of English. The Swedish learners were found to use *make* slightly more frequently than the native speakers, while the French speakers used it substantially less often. Among other reasons, the authors ascribe the French speakers' underuse of *make* to the fact that French does not use the equivalent of the verb *make* ('faire') in causative structures (e.g. *make happy*) as consistently as English and Swedish do.

A number of helpful software packages are available for identifying and counting instances of particular words in large compilations of learner production. These include Nation's *Range*,[1] Anthony's *AntWordProfiler*[2] and Scott's *WordSmith Tools*.[3] Corpora up to the size of 150,000 words can be explored via direct entry online at the *Vocabprofile* link at Cobb's *Lextutor* site.[4] Searches using any of these tools will readily indicate the number of occurrences of a specific letter string such as *make* and usually some higher-order groupings as well. As in the case of the Altenberg and Granger example above, researchers may be more interested in lemma counts; these tally the uses of a word in all of its grammatically inflected forms (in this case, *make* plus *makes*, *making* and *made*). Another unit of interest is the word family; researchers using this approach explore the extent to which learners are able to use both inflected and derived forms of a word. Thus a family count of instances of *make* in a learner corpus would include derived forms like *maker* and *unmade* in addition to inflected forms like *making* and *made*. The choice of unit has important pedagogical implications. By way of illustration, we might ask: what does a learner's use of the word *disbelief* in an essay mean in terms of his or her lexical development? Implicit in research using family counts is the assumption that this learner also knows the root verb *believe* and its inflected as well as some derived forms, like *unbelief*. By contrast, in research using lemma counts, the learner's use of the noun *disbelief* cannot be seen as indicating knowledge of the verb *believe*.

Investigations that compare numbers of occurrences of particular words in productions by learners of various L1 backgrounds to occurrences in native-speaker productions date back to the very beginnings of learner corpus research (e.g. Ringbom 1987). Counts (usually of lemmas) have

[1] www.victoria.ac.nz/lals/about/staff/paul-nation/ (last accessed on 13 April 2015).

[2] www.laurenceanthony.net/ (last accessed on 13 April 2015).

[3] www.lexically.net/wordsmith/ (last accessed on 13 April 2015).

[4] *Compleat Lexical Tutor*, www.lextutor.ca/ (last accessed on 13 April 2015).

PROOF

also been used to explore the extent to which L2 learners use (or do not use) the vocabulary of a particular genre, with the language of the argumentative essay and the use of logical connectors in particular being the focus of a number of studies. For instance, a study by Granger and Tyson (1996) found that in comparison to essays by native speakers of English, learner essays tended to overuse *moreover* and underuse *however* and *therefore*; as in the study of *make* above, explanations can be related to characteristics of the L1. A number of word count studies (e.g. Petch-Tyson 1998) also show that L2 learners overuse the pronouns *you* and *I* (and their derivations) in their productions; this is consistent with teacher impressions that L2 essays tend to be overly personal and speech-like in style. Granger and Rayson (1998) confirm this finding and identify other instances of over- and underuse of lexis in essays by learners of English. For example, they show that their learners underuse nouns that native speakers use to structure arguments (e.g. *issue*, *debate*, *suggestion*) and overuse general and frequent nouns like *people*, *thing* and *problem*. Hasselgren (1994: 237), who also identified overuse of highly familiar all-purpose words, refers to these as 'lexical teddy bears'. Many other findings might be cited; the point is that the powers of simple frequency counts to shed light on learners' lexical development are considerable. The findings also highlight the importance of valid comparison data.

2.2 Comparing corpora

Hunston (2002: 206) states that '[t]he essence of work on learner corpora is comparison'. Granger has devised the term Contrastive Interlanguage Analysis (CIA) for research in this paradigm (Granger 1998a: 12; see also Chapter 3, this volume). CIA studies have typically involved comparing the language produced by learners to the productions of native speakers of that language; most of the examples from the 1990s cited above used this approach to identify overuse or underuse of particular words. But as Granger (1998a) points out, the CIA approach can also be used to compare interlanguage productions to each other to identify the effects of age, proficiency level, L1 background, task conditions or other factors. An example of research using this approach is the study by Altenberg and Granger (2001) cited above that compared uses of *make* in essays by Swedish and French learners of English. CIA comparisons of corpora produced at different stages of acquisition include a cross-sectional study by Marsden and David (2008) and a longitudinal one by Horst and Collins (2006), but such studies are still relatively rare (see Chapter 17, this volume).

A considerable challenge in the CIA research paradigm is corpus comparability. An important resource for researchers seeking comparable learner and native speaker (NS) collections is Granger et al.'s (2009) *International Corpus of Learner English* (*ICLE*) and its native speakers of English counterpart,

PROOF

the *Louvain Corpus of Native English Essays* (*LOCNESS*).[5] The *ICLE* is meticulously categorised according to level, first language and conditions of writing (Granger 1998a). NS language is not of course the only possible point of comparison. In recent years, comparisons to native-speaker norms have been called into question by researchers who point out that L2 speakers may not always seek to become completely native-like (e.g. Seidlhofer 2001). As corpora of productions by highly proficient users (e.g. speakers of English as a lingua franca, or ELF) become more available, comparisons to competent users rather than to native speakers can be made. For the moment, however, many LC researchers continue to use NS corpora as a baseline for comparison.

2.3 Assessing lexical richness

So far we have considered studies comparing one corpus to another. These are internal comparisons; that is, the words within two or more corpora are compared to each other (often a learner corpus, and often but not always a NS corpus). But external comparisons can also be made: the words, lemmas or families that learners use in their productions can also be considered in terms of their occurrence in the language as a whole. This approach allows researchers to answer important questions about the effects of exposure to L2 input such as: ~~t~~o what extent is a particular group of learners able to actively use lexis they encounter frequently (and less frequently) in the language at large? ~~d~~oes their general vocabulary use become more sophisticated (as indicated by their use of infrequent lexis) over time as exposure to input increases? This, too, is essentially a comparison-based, native-speaker informed approach since the 'language as a whole' is represented by very large corpora of hundreds of millions of words of English such as the *British National Corpus* (*BNC*)[6] and the *Corpus of Contemporary American English* (*COCA*),[7] and the word frequency lists that have been derived from them (including Nation's recent synthesis of both).[8] Research in this vein analyses a learner text or corpus in terms of its lexical richness using lexical frequency profiling (LFP) software.

2.3.1 What is lexical frequency profiling?

Lexical richness can be defined as the level of development of a learner's lexicon. Researchers have assessed richness in several ways, including lexical diversity (more on this below) and lexical sophistication. The LFP approach is relevant to the latter construct, which is defined by Lindqvist et al. (2013: 110) as 'the percentage of sophisticated or advanced

[5] www.uclouvain.be/en-cecl-locness.html (last accessed on 13 April 2015).

[6] www.natcorp.ox.ac.uk/ (last accessed on 13 April 2015).

[7] corpus.byu.edu/coca (last accessed on 13 April 2015).

[8] *Range* program with *BNC/COCA* lists, www.victoria.ac.nz/lals/about/staff/paul-nation/ (last accessed on 13 April 2015).

PROOF

words in a text'; LC research in this vein operationalises sophistication in terms of the proportions of infrequent word families used – as identified by LFP software. This software uses corpus-based frequency lists for a particular language to carve an entered text or corpus into words of different frequency levels and then calculate the proportions of each. A typical profile for written English texts is 70% items from the most frequent 1,000 word families, 10% from the second, and the remainder from less frequent zones.

A learner corpus experiment (although the words 'learner corpus' do not actually appear in the study) by Laufer and Nation (1995) pioneered the application of the LFP approach to learner productions. The researchers used *Vocabprofile*, a software program by Heatley and Nation (1994), which at that time produced a four-way classification: all of the word families in the learner texts they investigated were classified as being on the list of the first 1,000 most frequent English families, the second most frequent 1,000, the *University Word List* (Xue and Nation 1984) or else 'off-list' (i.e. on none of the three other lists). The corpus consisted of compositions of 300–350 words produced by sixty-five English as a Foreign Language (EFL) and English as a Second Language (ESL) learners. The purpose of the experiment was to see, first, if learner profiles as thus determined were reliable over different pieces of writing, and, second, if the profiles could distinguish between learner proficiency levels as determined by alternative procedures. Statistical analysis showed the answer to both questions to be affirmative, although the relatively low-use second 1,000 band did not feature meaningfully in making level distinctions. That is, learners' ability to use infrequent words (i.e. words *not* among the 2,000 most frequent) was found to be a valid, reliable indicator of proficiency. Further, the finding that reliability depends on texts being similar in genre and longer than 200 words in length has usefully informed the methodology of many subsequent LFP studies.

Vocabprofile software has been revised and upgraded many times since Laufer and Nation's study in 1995. It now forms part of the *Range* suite of resources (Heatley et al. 2002); a version that allows for online entry and processing is available at Cobb's *Lextutor* site. Currently available versions for English draw on improvements such as Coxhead's (2000) *Academic Word List*. Updated lists based on the *BNC* allow the words of an entered text to be classified according to proportions of words at twenty levels of frequency, and a combined set of *BNC* and *COCA* lists allows words to be classified at twenty-five levels. Laufer and Nation's results have since been replicated in a variety of other research contexts. A highly practical application of LFP is a study by Morris and Cobb (2004); they assembled a corpus of ESL teacher trainee writing and tested its profiles' ability to predict success in the teacher training programme. Failure in examinations, failure to complete the programme and failure to stay in the profession had been long-standing problems with no identifiable components up to

PROOF

then. The clear finding was that the proportion of post-first-1,000 items in student writing was a significant predictor of success as an ESL teacher. That is, the degree to which a student used lexis beyond that of spoken conversation (which has been shown to consist almost entirely of the first 1,000 items) was a key determinant. This finding had straightforward implications for both programme admissions and course contents.

Though pioneered with learners of English, the LFP approach has been used in investigations of learner corpora in other languages. For example, Ovtcharov et al. (2006) compared passing and failing oral interviews in L2 French using online *Vocabprofil*, a version of the English *Vocabprofile* software based on frequency lists for French by Goodfellow et al. (2002). The criteria for passing and failing this Canadian civil service test had been vague, although it was mentioned in the course materials that varied and appropriate lexis was a consideration. The results of this LC experiment showed that this was indeed true; transcribed interviews from passing and failing cases were statistically distinct in terms of ~~the proportions of~~ post-1,000-level items ~~included~~. The pedagogical implication of Ovtcharov et al.'s (2006) study seems clear – vocabulary training should be included in the preparation for this interview. It is interesting that LFP was effective in making this distinction in a corpus of learner speech even though the version used was based on written word lists. More recently, Lindqvist et al. (2013) report their use of speech corpora in developing the *Lexical Oral Production Profiler* (*LOPP*), specifically designed for use in assessing the lexical richness of L2 learner spoken productions in both French and Italian. Initial analyses point to its usefulness in distinguishing proficiency levels.

The study by Lindqvist et al. (2013) points to the usefulness of genre and/or context-sensitive LFP approaches. A notable development in this regard comes from a series of studies in Western Canada by Roessingh and colleagues. These researchers tackled the problem of Canadian-born immigrant children in mainstream classrooms failing to develop age-appropriate literacy levels in the primary years, as shown by the steady decline in their reading comprehension scores and subsequent difficulties in achieving higher education and career goals. Roessingh and Elgie (2009) transcribed two spoken corpora of seventy-six native and eighty-seven non-native nine-year-olds telling a story based on picture prompts. The researchers did not expect that the 1,000-family bands of a standard 'adult' *Range* or *Vocabprofile* analysis would be a fine enough measure to explore speech by young learners. Instead, they used lists by Stemach and Williams (1988), who had created a principled amalgam of several developmental word lists based on corpora of childhood language; these were broken down into ten 250-word frequency bands and incorporated into *Vocabprofile for Kids* (*VP-Kids*; Roessingh 2014) on the *Lextutor* website. The clear result of lexical profiling using this scheme was that the NS English-speaking children deployed words from all ten

PROOF

of the bands and beyond to tell their stories, while the non-natives rarely deployed items beyond the third 250-word list, and indeed were largely dependent on the first list. About 85 per cent of their stories were built from most common 250 word families. The immigrant children were hardly in a position to tackle 'reading to learn' at school and the profiling methodology provided a conclusive way to show this. Canada is an immigration country but not the only one facing the issues Roessingh and colleagues have addressed; Verhoeven and Vermeer (2006) used a comparable methodology based on frequency lists derived from a corpus of classroom input to arrive at a similar conclusion about immigrant children in the Netherlands.

An interesting variant on profiling is Meara's (2005a) *P_Lex*, which is based on the same frequency lists used by Laufer and Nation (1995) but attempts to solve two problems that Meara found with the original measure: its complex output (a set of band percentages rather than a single score) and its unreliability with shorter texts (which less-advanced learners typically produce). *P_Lex* reduces the frequency bands to two, above and below the 2,000 mark, and then goes through a text dividing it into 10-word segments and counting the number of 2,000+ words each contains (4 out of 10, 1 out of 10, and so on). Its output is a calculation of the proportion of segments containing one 'difficult' word, two words, and so on, which is presented as a single index (that Meara calls lambda, λ). This can then be compared to NS lambdas and other lambda norms previously worked out. Whether this is a less complex output, the reader can judge, but according to Schmitt's (2010) work with the measure, it performs reliably with texts of any size.

2.3.2 Lexical diversity

Another important approach to assessing lexical richness in learner corpora involves measures of lexical diversity (or variation). The most basic form of this text-internal measure is the simple type–token ratio (TTR). This is the number of different words (types) in a text divided by the total number of words (tokens). For example, *The cat sat on the mat* has a TTR of 5:6, five types to six tokens (or 0.83), since the word-type *the* has two instances. The TTR measure is basically about the amount of word repetition found in a learner's production. However, the problem with a simple TTR 'score' is that it varies with the length of a text. Because of the necessary recurrence of a relatively small number of function words in any natural text, the longer the text, the higher the proportion of function words will be. Note that the same sentence doubled will have a different TTR from the single sentence; the *cat* text above when doubled has a TTR of only 5:12, or 0.41, since the number of tokens has risen while the number of types has not. In the LC research context, a possible way of making valid comparisons of sets of essays would be to calculate the length of the smallest essay in a corpus and then reduce every other text to this size so

PROOF

that all were equal. As observed by Schmitt (2010), however, this entails wasting valuable data from the longer texts.

Researchers interested in using a lexical diversity approach have devised various ways of dealing with the length problem; these involve modifications to the basic TTR formula. For example, Guiraud's (1954) index is obtained by dividing the number of types by the square root of the number of tokens in a text or corpus. The use of the square root is a means of making texts of different sizes more similar (a 900-word text is more than twice as long as a 400-word text, but this is not true of their square roots, 30 and 20). This measure has been shown to be useful in differentiating productions made by groups of learners at various proficiency levels and between learners and native speakers (see Milton 2009 for an overview). An interesting study by Ong and Zhang (2010) used a variant on this formula (word types squared divided by the total number of words) to address the length problem. This study differs from most of the research mentioned above in that it did not make comparisons to NS productions. Instead, Chinese-speaking EFL learners in Singapore wrote essays in task conditions that varied in the amounts of preparation time allowed and levels of help provided (in the form of topics, ideas, structure suggestions and models). Somewhat unexpectedly, analyses of the writing indicated that lexical diversity was the greatest when the task conditions were more challenging. That is, learners in the conditions where they had less preparation time and less writing assistance outperformed other groups in terms of the use of varied lexis.

Another approach to analysing learners' texts with TTR, while neither losing data nor suffering from text-length effects, has been devised by Malvern and Richards (2000). Their D (for diversity) or Vocd involves a complex procedure that has been clearly summarised by Schmitt (2010: 226):

> The process behind *vocd* takes several steps. The program generates 100 samples of 35 randomly selected words from a text, and calculates a type–token ratio for each of these. These 100 means are then averaged to produce a composite mean ratio for all 100 samples. The program goes on to do the same thing for samples of 36 randomly selected words, 37, 38 … all the way to samples of 50 words. The end result is a list of 16 means for the 35–50 word samples. These means form a curve, and it is compared to a number of theoretical curves generated by the D formula. The value of D which produces the best matching curve is assigned to the source text. D typically varies between 0 and around 50, with lower values indicating more repetition and a vocabulary which is not lexically rich, and vice-versa for higher values.

The creators of the measure show its usefulness in a number of studies of child L1 learners, where increased Vocd values tend to go hand in hand with increases in age (e.g. Richards and Malvern 2004). In a study of a

PROOF

corpus of oral interviews, Malvern and Richards (2002) investigated productions by L2 learners of French in an examination context. Vocd was able to distinguish between productions of proficient and less-proficient learners; the study also explored the language of the examiners and showed that the examiners roughly tuned their speech to the level of the examinees. Incidentally, it is worth noting that diversity measures can be used with any language employing individual word forms.

A cross-sectional study by Marsden and David (2008) used Vocd to explore the effects of time spent in the language classroom. The research compared corpora of oral interview data produced by British learners of French and Spanish; for each of the languages, speech produced by learners in Year 9 of their schooling was compared to that of learners in Year 13, with an additional ~450 hours of instruction. As expected, the Year 13 learners produced more lexically diverse language than the Year 9 learners. This was true for both languages. The researchers also subtracted total numbers of lemmas from total numbers of tokens in the various corpora as a way of measuring inflectional diversity. Again, the expected advantage for more years of study was found. Further, word-class analyses showed that the more-advanced learner productions used a greater proportion of verbs and a smaller proportion of nouns than the less-advanced productions, confirming a learning sequence in the acquisition of Romance languages observed in previous empirical research.

2.3.3 What is the 'best' way to measure lexical richness?

The search for the most effective method of assessing the lexical richness of LC has been likened to the search for the Holy Grail (Malvern et al. 2004, cited in Tidball and Treffers-Daller 2007: 134). But as Tidball and Treffers-Daller point out, there may be no single best solution, given the range of research questions that measures of lexical richness can be applied to. Some researchers are interested in measures that are simply able to identify productions as either native or non-native, while others are interested in more nuanced distinctions between various levels of L2 proficiency or, as we have seen, in the subtle effects of task conditions or L1 background. The 2007 volume edited by Daller et al. includes several studies devoted to the relative strengths and weaknesses of the two main types of measures discussed above (and others). In this section, we offer our own perspective, which centres on the pedagogical usefulness of the LFP approach.

In our view, the extent to which L2 learner speech or writing contains diverse words regardless of their frequency (as in TTR-related measures) seems less revealing than the extent to which it contains actual infrequent words (as captured by LFP). This point is nicely illustrated by Meara and Miralpeix (2008), who observe that the TTR of the following three sentences is identical:

PROOF

1. The man saw the woman.
2. The bishop observed the actress.
3. The prelate glimpsed the wench.

Obviously, there are differences between these sentences that a simple measure of repetition cannot encompass. But LFP analysis (in this case, the *BNC-COCA* frequency framework of the online *Compleat Vocabprofile* available at Cobb's *Lextutor* site) quantifies these intuitively felt differences. The first sentence consists entirely of words on the list of the 1,000 most frequent English word families. The second contains 1,000-level words but also one from the 2,000-level (*observed*) and another from the 3,000-level (*bishop*), while the third has two very infrequent words (*prelate* and *wench*), both from the 12,000-frequency level.

In terms of data that classroom teachers or action-researchers are likely to be interested in working with, this kind of information is relatively accessible. For instance, if LFP analysis of a corpus of classroom writing shows that learners are able to use 1,000-level word families extensively in their writing, but use 2,000- or 3,000-level words in significantly lower proportions than are found in a corpus of level-appropriate model or NS essays, the implications are evident, and teachers and learners know exactly which words they need to work on. By comparison, the notion of Vocd's 'theoretical curves', to which those of an actual LC would be compared, seems somewhat challenging to grasp and is probably more challenging to apply pedagogically. Similarly, the Guiraud index (1954) seems likely to be more useful to experienced researchers than to aspiring action-researchers, as it is not immediately obvious how to interpret a score of, say, 0.61.

On a methodological note, we would add that a distinct advantage of LFP is that it does not suffer from sensitivity to text length; the same text doubled will still produce the same proportion of items at each frequency band. In addition, a proficiency limitation in TTR measures has been identified: advanced learners appear to be less well distinguished by this type of measure (Schmitt 2010).

Perhaps the 'best' richness measure is a combined one that draws on the strengths of both paradigms and avoids their weaknesses. A notable weakness in LFP-based measurement is its inability to capture repeated uses of the same words. A learner's use of just one massively repeated third-thousand word would be highly detrimental to his or her TTR, but in a *Vocabprofile* analysis, it would result in a strong but false third-band percentage that only a human post hoc investigation could uncover. For this reason, it can be predicted that in the future, LFP measures will be reconfigured to incorporate some way of determining TTR. Another downside of the list-based method at the heart of LFP is that it involves occasionally arbitrary decisions about family memberships, frequency ratings and band cut-offs. TTR does not have these problems. And finally, as already

PROOF

mentioned, TTR approaches can be used for a range of languages with no special modification, while LFP can only be adapted to another language if an appropriate set of lemmatised or 'familised' lists is available for that language or can be developed.

A notable attempt to integrate a frequency list approach with a TTR-based measure is the Advanced Guiraud (AG), as employed in studies by Daller et al. (2003) and Tidball and Treffers-Daller (2007). This measure makes use of counts of the 'advanced' word types in a learner production (as determined by expert judgements or a research-informed list) and a mathematical transformation designed to overcome the length problem of unadjusted TTR. Studies of LC using this measure have demonstrated its ability to differentiate between groups of learners at various proficiency levels and between learners and native speakers (Milton 2009). These are promising developments, though, like the other TTR-related measures discussed above, the output of AG analysis is hardly straightforward to interpret in terms of teaching and learning goals. For this reason, we expect that, at least for the moment, practitioners and action-researchers interested in trying out some lexical LC research will continue to be drawn to LFP and user-friendly tools like *Vocabprofile*.

2.4 Lexical errors

A methodological intervention in learner corpus work that so far has gone unmentioned is the editing of a corpus that precedes the application of the counting and profiling techniques discussed above. It bears noting that these approaches rely on using 'clean' texts in which spellings have been regularised so as to avoid inflated tallies of what would otherwise be highly unusual words. An investigation by Llach (2007) is a compelling illustration of what is lost in this cleaning-up process. Her investigation of a corpus of letters written by young beginning learners of English of Spanish L1 background identified misspellings as by far the most predominant type of lexical error that occurred; spell-checking the corpus would have hidden this information. Other error types beyond the orthographic investigated by Llach included simple substitutions of Spanish words for English ones, adjustments of Spanish words to make them resemble English ones, and calques (use of literal translations from Spanish such as *table study* for *desk*), which, however, played a relatively minor role. The goal of the study was to explore the relationship between error types and holistic ratings of the letters for their overall communicative effectiveness, but no strong correlations were found. Waibel's (2008) book-length study of German and Italian learners' use of English phrasal verbs is another example of ~~a study~~ that examines error types; this study explored lexical errors such as collocation mistakes and use of inappropriate register.

An interesting new approach to the concept of lexical error has been put forward by researchers working with ELF corpora. Research

PROOF

in this vein avoids the term 'error' and refers instead to 'non-codified' (Osimk-Teasdale 2014: 109) or 'unconventional' (ibid.: 117) language use. ELF speakers are seen as creative language users who are 'pushing the frontiers of Standard English' (Seidlhofer 2011: 99) rather than as learners who make mistakes. In a recent study of ELF lexis, Osimk-Teasdale (2014) explored word-class shifts in the *Vienna-Oxford International Corpus* (*VOICE*), a large publicly available corpus of spoken ELF. A word-class shift that illustrates the kind of mismatch she was interested in occurs in the example 'I just wanted to give a *partly* answer' (p. 109), where *partly* functions as an adjective although it is identified as an adverb in standard English. Twenty types of shift met the researcher's criterion of occurring twelve times or more in the data. Of these, the most frequent was adjective used as adverb, as in 'We are *complete* different' and 'This is a *total* special way' (p. 123, with transcription conventions removed). This finding exemplifies an overall pattern in the data which Osimk-Teasdale sees as indicative of a preference for simpler forms over more complex ones.

3 Representative studies

In this section, we turn to three studies that both expand on strengths of the investigations of lexis in learner corpora outlined above and overcome some of the weaknesses we have highlighted. In our view, they suggest promising avenues for future research for both professional and action-researchers. The focus is on investigating lexical richness, starting with a study that used LFP software to explore a learner corpus.

3.1 Horst, M. and Collins, L. 2006. 'From faible to strong: How does their vocabulary grow?', *The Canadian Modern Language Review / La Revue canadienne des langues vivantes* 63(1): 83–106.

Horst and Collins (2006) explored an 80,000-word corpus of written narratives assembled by the second author and her colleagues. These narratives were produced in response to picture prompts by 210 beginner-level francophone learners of English (11–12-year-olds) and were collected at four 100-hour intervals over the course of their intensive English training in Quebec. As mentioned in Section 2.2, longitudinal studies are rare in LC research. Horst and Collins analysed each of the four staged subcorpora using *Vocabprofile* software. They expected to find signs of lexical growth over this period as measured in ever larger proportions of less frequent lexis; that is, they expected levels of second-1,000 families (and beyond) to increase over time. Instead they found no significant development over the 400 hours of instruction, at least in terms of changes between frequency bands.

In fact, there were changes taking place, not between profile bands but within them. The vast majority of the items these learners produced were

PROOF

and remained first-1,000-level words, but there was a greater variety and less repetition of these over time. While previous research (e.g. Laufer and Nation 1995) has focused on the band distinction, for beginners this was clearly too broad a measure. The profiling tools available at the time did not make two other kinds of information available that the researchers believed was present in their data: more diversity in morphologies for the words produced and lower proportions of cognates from their L1. The software was thus retooled to measure morphological variation and to carve the higher frequency bands into cognate and non-cognate zones (based on a definition of cognate as an item that is likely to be recognised by a francophone learner as familiar from French).

The morphological measure was the types-per-family ratio: the number of word types in the corpus, divided by the number of word families. If the types-per-family ratio exceeds the value of 1, the corpus contains more members for the same number of families. In a re-examination of the data using this measure, findings pointed to an increase in the ratio over time, indicating there was a growing use of inflected and derived forms. Thus, in addition to a base form such as *believe*, more forms like *believing*, *belief* and *believers* and other variants were used. (See the study by Marsden and David 2008 discussed in Section 2.3.2 for a similar finding using a different way of calculating inflectional variation.) In the Horst and Collins study, the second new measure, which identified proportions of French–English cognate use, indicated that the young learners were increasingly able to write the stories using 1,000-level non-cognate words in an age-appropriate way (*watch* instead of *observe*, *feel* instead of *sense*). To summarise, with the revised measures the staged corpora showed a steady increase over time in the variety of words used, the variety of word forms used and the 'Englishness' of the words used – largely from within the most frequent 1,000-word families. The reliance on French words (which learners were encouraged to use if they did not know the correct English forms) also decreased over time.

A similar outcome was observed in a study by Cobb and Horst (2011) on video gaming. Fifty francophone Grade 6 learners (11–12 years old) participated in an investigation of the effects of playing a suite of vocabulary-building games for a period of two months (*My Word Coach*, UBISOFT 2009). Methods used to assess lexical development included LFP analysis of two learner speech corpora consisting of responses to the wordless picture story *Boy, Dog and Frog* (Mayer 1967), collected before and after learners used the video game. The game introduced and recycled several hundred new words under stimulating circumstances and in a theoretically determined sequence (Mondria and Mondria-De Vries 1994). *Vocabprofile* analysis of the two corpora (roughly 24,000 words each) showed that the stories had clearly become longer following use of the game, but there were no pre-post band differences at any frequency

PROOF

level: the stories were told almost entirely using first-1,000 families. In other words, virtually none of the newly learned words showed up in the spoken narratives. However, as in the Horst and Collins (2006) study, some sifting through the LFP outputs suggested some other places to look for changes.

In any *Vocabprofile* analysis, there are normally words that do not fit into any of the program's frequency categories (since they are misspellings, regionalisms, etc.), and these are put in the 'off-list' category, counted, and given a percentage, but usually not looked at with any interest. In Cobb and Horst's (2011) investigation, it was noticed that the off-list component was unusually large before using the game, but not after. On inspection, the off-list component prior to game use was largely composed of French words that had been enlisted to help tell the story; after the game, these had all but disappeared, presumably having been replaced by English words. Overall, there was a significant increase in the total number of words used to tell the stories and a significant decrease in the number of French words. These outcomes resemble Horst and Collins's results, but did not replicate their finding of increased morphological diversity (the types-per-family index). This is to be expected since these learners were meeting words in the game only as headwords, while Horst and Collins's learners were meeting them in a classroom in a variety of texts, contexts and morphologies.

These studies reveal a limitation of the LFP approach to assessing lexical richness. Since the overwhelming proportion of any production – learner or native speaker – consists of words from the 1,000 most frequent families of a language, the remaining proportion of less frequent words, which are typically the focus of research interest, is small and opportunities to demonstrate differences between corpora are limited. One solution to this difficulty has been to look inside the seemingly monolithic 1,000-zone, as Horst and Collins (2006) and Cobb and Horst (2011) have done. As we have seen, another solution is to develop specialised LFP tools based on context-sensitive corpora and frequency lists. Examples are the age-appropriate lists used in the Roessingh and Elgie (2009) and Verhoeven and Vermeer (2006) research discussed in Section 2.3.1. Profiling has thus proliferated into an approach with a number of refinements made to iron out one wrinkle or another that was found in the original version.

3.2 Edwards, R. and Collins, L. 2011. 'Lexical frequency profiles and Zipf's Law', *Language Learning* 61(1): 1–30.

TTR and its relatives are not the only alternative to LFP that could potentially be applied to assessing the lexical richness of learner corpora. Another recently developed measure involves the application of Zipf's (1935) Law to student writing as a means of calculating productive vocabulary size (Edwards and Collins 2011). While not an LC study per

PROOF

se, this work has been tested on a learner corpus and could potentially become a useful LC measure since it addresses another important weakness at the high-frequency or beginner end of LFP analysis: the assumption that words are largely learned in the order of their frequency in the language at large. Such an assumption is needed to posit that learners with knowledge of the third 1,000 families are more advanced than those who know only the second 1,000, for example. This language-at-large account of frequency is clearly a sort of average frequency that is true for groups of learners but not fully accurate for any learner in particular (Milton 2009).

Learners in a second language will often acquire some lower-frequency items before high-frequency ones, putting into question the notion that a handful of third-thousand items (for example) in a learner's writing can be taken to signal very much about his or her lexical development. The case for a frequency sequence for vocabulary development in first language may be stronger, as argued by Biemiller and Slonim (2001). In an L2, it is likely that learners will seek out equivalents for the words they already use in their L1s, such as hobby or sports terms, and these are likely to stem from a range of vocabulary levels without indicating any general competence at that particular level, nor indeed at more basic levels. A learner might well know mid-frequency items like *hoop* and *net* (in basketball) or *eraser* and *detention* (from the classroom), yet not know high-frequency words like *war* and *parent*.

Zipf's Law is based on the observation that in any natural text or corpus, word frequency and rank are strongly and inversely correlated throughout. Frequency here refers to the number of occurrences of a word in a corpus or text (e.g. *the* has 69,967 occurrences in the 1-million-word *Brown* corpus) and rank refers to its position in the list (e.g. *the* is number 1). This correlation calculated over an entire corpus can be used to produce an index that predicts with reasonable accuracy the frequency at any point in the ranking, for a text of that size. The 100th word in Brown is *down*, for example, and by Zipf calculation, its frequency should be 897.7 (in fact, it is 895.0). For 300-word texts (the typical size of learner texts employed in many of the foregoing studies), and assuming the frequency-to-rank proportions of the *Brown* corpus (which according to the law are universal), the prediction is 0.265 occurrences of the word *down*. In other words, *down* will appear once in every four such text, on average. Taken across large numbers of texts, these predictions are largely borne out. And further, when the frequencies of all the words in a set of texts are known, then the ranks can be calculated. For example, all those (like *down*) that are ranked at positions within the top 1,000 can be added up and calculated as a proportion of the text as a whole, effectively amounting to the LFP's first 1,000 list but without having to resort to LFP.

PROOF

Up to this point, Edwards and Collins's (2011) study follows a simulation exercise by Meara (2005b), who used standard language-at-large frequency information to show that Zipfian regularities could predict the typical outcomes of LFP/*Vocabprofile* analysis of learner texts. What Edwards and Collins add to the picture is that this same result can be obtained mathematically, without simulations, and, more interestingly, can avoid the language-at-large assumption. The real interest in Edward and Collins's analysis is in the idea of extrapolating from the words actually used to the total lexical resources available, that is, to the writers' productive vocabulary size. In this procedure, the Zipfian equation is solved with frequency, or words-used, as the known variable, and words-available as the unknown (the corpus size, which, in this case, is the learner's or learners' interlanguage lexicon). By contrast, the profiling approach cannot so extrapolate, having no theory on which to do so, but can only calculate productive lexicon from the words that were actually used. For example, a typical profile shows 90% at first 1,000 = 900 words; 10% at second thousand = 100; 10% third to fifth thousand = 300, for a total productive vocabulary size of 1,400, and this is clearly an underestimate. Other words could have been used had the topic, the time of day, etc. been different. The Zipfian estimates of productive size tend to be ~~larger, as corresponds to common observation~~.

Edwards and Collins (2011) tested their model's predictions against two learner corpora (90 young francophone learners, writing from picture prompts, at two times in a 400-hour communicatively oriented intensive ESL course). This yielded a total of 8,295 words after 100 hours of instruction and 9,944 words after 300 hours. From the words actually produced, the Zipfian calculations were able to extrapolate productive interlanguage lexicons of 2,216 words at Time 1 and 2,274 words at Time 2. These estimates are plausible, reliable, show a small difference in the right direction and are larger than would be predicted by profiling of just the texts themselves. While clearly in the early stages of its development and in need of testing with larger corpora, and quite far from ready-for-action research projects, this approach tackles some important weaknesses in the profiling approach, namely the underestimation of productive vocabulary size and the assumption of a frequency sequence in acquisition. However, like other corpus-internal measures, such as the diversity measures detailed earlier, it may not be well suited to making cross-corpus comparisons.

3.3 Crossley, S. A., Cobb, T. and McNamara, D. S. 2013. 'Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications', *System* 41: 965–81.

Another interesting though rather complex measure is the *Coh-Metrix (CM)* suite of text analysis tools developed by Graesser et al. (2004). *CM*

PROOF

produces up to sixty indices of the linguistic and discourse representations of a text, many of which focus on the words in a text and their characteristics. Those with reasonably obvious application to the lexical investigation of a learner corpus are number of words, average number of syllables per word, raw frequency, raw frequency mean for content words (0–1,000,000), log frequency mean for content words (0–6), word-sense frequency, type–token ratio for content words and collocational regularity. Discourse-oriented measures are average words per sentence, proportion of content words that overlap between adjacent sentences, Flesch reading ease scores (0–100) and grade levels (0–12). Worth noting is *Coh-Metrix*'s solution for the TTR problem: use content words only, since these do not pile up as function words do.

A typical way in which these *CM* measures have been used has involved developing a corpus of American college students' free-writes, getting these evaluated globally by human raters, and then throwing a range of pertinent indices against the corpus to see which ones account for variance in the human ratings. A typical finding from a corpus of 240 free-writes written by students in their L1 at a southern American university and graded both globally and analytically (following a grid) is that four *CM* indices predicted 86% of the variance in raters' analytic scores and 46% in their global scores.

An L2 example comes from a study by McNamara et al. (2010), who tested twenty-six theory-selected *CM* indices on a corpus of 120 untimed, resource-permitted, out-of-class free-writes (of 500 to 1,000 words apiece, for a total of approximately 90,000 words) that had been graded as high or low proficiency by experienced markers. Three of the automated indices emerged as significant predictors of these proficiency judgements and accounted for 22% of their variance: syntactic complexity (number of words before the main verb), lexical diversity (as measured by the measure of textual lexical diversity, or MTLD), and word frequency (as measured logarithmically by the *Cobuild Corpus*-based *CELEX* frequency list; Baayen et al. 1995). The MTLD is interesting as a further attempt to work with TTR. Here it is described as 'not vary[ing] as a function of text length', which, on further inspection, is for the reason that it 'is calculated as the mean length of word strings that maintain a criterion level of lexical variation' (McCarthy and Jarvis 2010). That level is 0.72; in other words, MTLD is an index based on the average amount of text whose sentences maintain a TTR of 0.72, a ratio of seven different words per ten running words.

It is interesting that two of the three winning predictors in this study are lexical, and that one of these involves word frequency lists as LFP does, while the other is a diversity or repetitiveness-oriented TTR-type measure. Some blend of these two measures looks like the way forward in the lexical analysis of learner corpora, both theoretically and in terms of the reasonably strong results shown above in predicting human judgements. While there is clearly work to do in clarifying the TTR part of any

PROOF

eventual unified measure, there also remain questions about the best way to gauge and deploy frequency information: in 1,000-word family bands or as single words?

This issue was investigated in a study by Crossley et al. (2013), who compared the ability of band vs single-word frequencies to predict human ratings of both L1 and L2 writing. In the band approach, the frequency of the word family *go* is calculated by summing the frequencies of its individual members in the *BNC* and using the sum to make a band placement. In the case of the *go* family, *go* itself has 87,021 *BNC* occurrences, *goes* 14,264, *going* 63,433, *gone* 18,433, and *went* 45,538, which sums to a family frequency of 228,689. This positions it well within the most frequent 1,000 families (everything more frequent than 12,639 occurrences is first-1,000, as calculated by Martinez and Schmitt 2012). Thus, any member of the *go* family used in a piece of learner writing is counted simply as one first-thousand item, and the total number of these items, calculated as a percentage of the number of words in the text, yields the first-1,000 part of the profile. And so on for all the bands available.

In the single-word approach, on the other hand, there are no families. Each occurrence is tallied individually and entered into an average frequency for the text as a whole. If a learner uses *go*, this gets rated as 87,021, *gone* is rated 18,433, and so on (by *CELEX* not *BNC* figures, similar for this purpose). When all the words have been assigned their ratings (quick work on a computer), an average for the text or corpus is produced, with a higher number indicating a text with a higher proportion of common vocabulary, and vice versa. A simplified two-sentence version of this difference is shown in Table 9.1. As can be seen, there are some similarities and some differences between the two ways of measuring. The difference is that the single-word method produces a single outcome index (2.13 million and 1.35 million) as opposed to multiple band percentages; the similarity is that the richness or sophistication of the second text is roughly double that of the first by either measure (2.13 vs 1.35 million by singles; 17% vs 34% post-second-thousand items by bands).

To determine which way of measuring lexical richness was the better predictor of proficiency judgements made by humans, Crossley et al.'s study used both band-based and count-based methods to classify the individual texts in one corpus of 100 L2 learner free-writes and another of 30 NS free-writes. Raters had previously classified these according to proficiency level (beginning, intermediate and advanced L2 learners and NS). The analysis showed that count-based word frequency indices accurately classified 58% of the texts as the humans had. Band-based analyses fared slightly less well, with LFP/*Vocabprofile* accurately classifying 48% of the texts, and *P_Lex* 36%. So, putting words together in the 1,000-family bands clearly resulted in some loss of measurement sensitivity. In other words, the apparent price of LFP's comprehensibility and usability is some loss of accuracy.

PROOF

Table 9.1. Two calculations of lexical sophistication (k = 1,000-word frequency band)

High-frequency lexis			ow-frequency lexis		
Sentence 1	*Vocabprofile*	*BNC* freq.	Sentence 2	*Vocabprofile*	*BNC* freq.
The	1k	6,041,234	*The*	1k	6,041,234
cat	1k	3,844	*lizard*	8k	196
sat	1k	11,038	*basked*	12k	47
on	1k	729,518	*on*	1k	729,518
the	1k	6,041,234	*igneous*	14k	129
mat	4k	569	*rocks*	2k	286
Output	1k=83% 4k=17%	2,137,906		1k=33% 2k=17% 8k=17% 12k=17% 14k=17%	1,354,225
Simplification	17% post-2k[a]	log (10)= 6.33		34% post-2k	log (10) = 6.05

[a] The simplified output of LFP is based on a proposal by Laufer (2000) to reduce profiles to a single percentage of post-2,000-level words and is also employed in Meara's (2005a) *P_Lex*.

4 Critical assessment and future directions

The endeavour of analysing learner corpora from a lexical perspective has notable strengths but it also faces significant problems. One area of concern is the increasing complexity of measures. After a brief period of seeming clarity and even relative simplicity from 1995 to about 2005, the methodologies appear to have fragmented into a multiplicity of measures that are anything but simple, each with its own shortcomings. Just within the LFP tradition, *Lextutor's Vocabprofile* site now offers no fewer than five competing frequency frameworks. What does the future hold? Will different purposes require different measures? Or can profiling, repetitiveness and raw frequency be worked into a single coherent measure? A related concern is what appears to be a growing trade-off between accuracy and comprehensibility/usability in the analysis of learner corpora. The finding of greater accuracy for count measures over band measures in predicting human judgements (seen in Crossley et al. 2013) has to be set against the comprehensibility of band measures like *Vocabprofile* and the shown willingness of practitioners to use them for many worthwhile purposes. Thus it seems more comprehensible to talk about learner writing resembling speech ('talk written down') in terms of its proportions of first-1000 words than to say that an advanced profile would be a log-10 *CELEX* frequency rating of 6.05 and a beginner profile 6.33. Also, it seems fair to say that considerable further work will be needed to make the

PROOF

length-effect-free variants of TTR clear to practitioners. Another worrying development is what appears to be reduced interest in learners per se; rather, the learner corpus has become primarily a test bed for the development and testing of measures. It is not clear if this can truly be considered LC research, as it does not touch on learner development and learning variables. Presumably, following a period of validation, these new measures will eventually be used in LC studies to shed new light on learners and the learning process.

Important among the strengths of lexically focused LC research is its emphasis on investigating learner production systematically, in a way that goes beyond what can be simply observed. It pulls out patterns in learner productions, in the manner of other corpus research, but then often goes on to link these to other empirical research findings. The portrait is potentially rich and comprehensive. Another strength of LC research is its tradition of replication. While a lack of replicated findings is often proposed to be a problem in second language acquisition research generally, this is not the case in LC studies of lexis. The widespread availability of measures has made it possible to replicate and verify findings in different settings. Thus Cobb (2003) replicated four European studies in a Canadian context, and Lindqvist (2010) replicated Ovtcharov et al. (2006) in a European context. The various *Coh-Metrix* studies seem to lend themselves readily to replication. Once developed, corpora do not normally get thrown away; this bodes well for future reanalyses of earlier work as measures evolve and problems are worked out. Profiling is an approach that was originally created to evaluate the readability of small texts for learners at different levels of vocabulary knowledge, and it just happened to be also useful for analysing large learner corpora in a number of interesting ways. It is almost certain that a more dedicated LC tool will emerge out of the current high level of activity in the field. Finally, the potential action-researcher may have started his or her reading of this piece encouraged by the apparent 'doability' of corpus research, only to watch this disappear into mathematical complexity and the pros and cons of different measures. Fortunately, the doable kind of research still needs doing. The more approachable analyses discussed in the earlier parts of this chapter have not been invalidated.

Key readings

Granger, S. (ed.) 1998b. *Learner English on Computer*. London: Longman.

This classic work continues to be a source of ideas and methodological guidelines for learner corpus studies. It contains frequently cited studies of lexis by Ringbom, Petch-Tyson, Granger and Rayson, and others.

PROOF

Cobb, T. 2003. 'Analyzing late interlanguage with learner corpora: Quebec replications of three European studies', *The Canadian Modern Language Review / La Revue canadienne des langues vivantes* 59(3): 393–423.

The approach in this series of successful replication studies is hands-on and user-friendly. The replications are of three studies of European francophone EFL learner writing from Granger's (1998b) *Learner English on Computer,* performed with comparable Canadian francophone learners' writing. The original looked at learners' lexical and phrasal resources and at their tendency to mix features of spoken and written genres in their L2 writing. The replications were successful in showing almost identical tendencies between the two groups, further validating the generalisability of LC findings.

Daller, H., Milton, J. and Treffers-Daller, J. (eds.) 2007. *Modelling and Assessing Vocabulary Knowledge.* Cambridge University Press.

The chapters on lexical richness in this volume report empirical investigations of learner corpora in several languages using a variety of methods. Techniques are compared for their efficacy.

Milton, J. 2009. *Measuring Second Language Vocabulary Acquisition.* Clevedon: Multilingual Matters.

Milton's chapter on measuring productive vocabulary (Chapter 6) in this volume provides a user-friendly description of measurement issues and techniques relevant to learner corpora. The sections on measuring lexical diversity and lexical sophistication in this chapter (pp. 125–40) provide clear definitions along with critical overviews of important studies.

Jarvis, S. and Daller, M. (eds.) 2013. *Vocabulary Knowledge: Human Ratings and Automated Measures.* Amsterdam: Benjamins.

This edited collection stems from a colloquium at the AAAL conference (2011) on The Validity of Vocabulary Measures. Several post-LFP approaches reviewed above can be explored further (including Crossley, Edwards and Collins, and Treffers-Daller).

A