Determining ESL learners’ vocabulary needs from a textbook corpus: challenges and prospects

Jabulani Sibanda* and Jean Baxen

Education Department, University of the Witwatersrand, Johannesburg, South Africa

*Corresponding author email: jabusbnd@gmail.com

Abstract: Sound vocabulary repertoire is a requisite for reading and all learning. Teaching a language’s vocabulary is a mammoth undertaking, unless particular lexical items are isolated and accorded explicit attention. Words taught explicitly should coincide with learners’ core vocabulary needs. This study generates words reflective of South African Grade 3 learners’ vocabulary needs for transition to Grade 4; a transition replete with challenges. We document challenges besetting the determination of learners’ core vocabulary needs from a textbook corpus. These included words to exclude from the frequency counts, challenges presented by multiword units, the unit of counting for word frequency generation, among others. High frequency words from the Grade 4 textbook corpus were generated and compared against five available word frequency lists as well as across different subject areas. The present study’s adapted unit of counting was applied to the word list with the whole process yielding 212 core words requisite for Grade 3 to 4 transition within the South African context. We advocate infusion of grade-specific core vocabulary lists in curriculum documents based on a robust large-scale word frequency generation process, among other things.

Introduction

This paper derives from the first author’s broader PhD study which tested Grade 4 learners’ knowledge of the core vocabulary needed to comprehend specific content area texts. The paper documents challenges that beset the determination of English Second Language (ESL) learners’ vocabulary needs from a textbook corpus. In the process, it questions and challenges current conceptualisations of ‘word’ and ad hoc processes of word-frequency generation. The circumvention of some of the identified challenges provides prospects for a tenable approach to the determination of vocabulary needs from a textbook corpus. The broader study from which the paper derives focused on second language acquisition, especially with regard to the transition from the Foundation Phase (FP) to the Intermediate Phase (IP). This paper specifically sets out to determine the core vocabulary needs of Grade 3 learners in preparation for that transition. The vocabulary needs are derived from a corpus of the Grade 4 content area texts the learners are expected to read as they transition to the IP. The research question guiding the study is:

• What is the core vocabulary needs of Grade 3 learners transitioning to Grade 4 as reflected in the Grade 4 content textbooks?

The paper’s focus requires an appreciation of the challenges South African (and other education systems’) ESL learners confront when they move to Grade 4, which puts into perspective the need for specific determination of core vocabulary without which learners’ progress is compromised.

Context: transitioning from the Foundation Phase to the Intermediate Phase

In South Africa, as in most countries, the end of Grade 3 heralds a significant transition marked by unique challenges in learners’ schooling. It is a transition from the Foundation Phase (FP) (Grades R–3) to the Intermediate Phase (IP) (Grades 4–6), characterised by both qualitative and quantitative increases in language demands generally, and vocabulary demands specifically. For the vast majority of learners who speak an African language, the transition heralds a shift from using their first language (isiXhosa in the present context) as the Language of Learning and Teaching (LoLT) to using English. Howie et al. (2008: 8) note that ‘the LoLT in Grade 4 results in more than 80% of
learners being taught in a second language, mostly English, a language spoken by less than 10% of the population. The assumption is that by the end of Grade 3, learners have sufficient vocabulary in their mental lexicon to allow them to learn predominantly, if not exclusively, in English.

Even more challenging is the transition in the focus of reading, from learning to read, to reading to learn, which presents comprehension challenges not unique to the second language (L2) learners, but are well documented among English L1 learners (Lesnick et al. 2010). There is also a shift in textual demands from narrative to expository texts where the latter embody both the general vocabulary, characteristic of narrative texts, and academic and technical vocabulary specific to content text. For all these challenges, the current South African curriculum makes provision for only one hour more in IP First Additional Language (FAL) instruction than for the Foundation Phase.

The language threshold the learners need to cross has consistently been proven to be largely a lexical one (Nation 2001; Lauffer & Ravenhorst-Kalovski 2010). Determination of the vocabulary needs of ESL learners on the verge of such a significant transition was a priority for the study on which this paper is based. In the larger study, the core vocabulary identified formed the baseline against which learners’ vocabulary proficiency were assessed. Literature on the importance of vocabulary in general, and of words that recur with high frequency in particular, informed the determination of this core vocabulary for such a significant transitional period.

**Importance of vocabulary knowledge**

The explosion of research in second language vocabulary learning testifies to the increasing recognition of the role of the lexicon in language acquisition, be it the first, second or foreign language, as well as in communication (Hunt & Beglar 2005). The role of words in communication has been deemed analogous to that of molecules in physical structures (Eskey & Grabe 1988), where the former constitutes the existence of the latter. Vocabulary knowledge is considered predictive and reflective of reading achievement (Pikulski & Templeton 2004; Godev 2009).

Transcending vocabulary thresholds ensures reading with understanding and guarantees derivation of meanings of the few unknown words from their contextual use. From their study, Hu and Nation (2000) discovered that knowledge of 80% of the words in a text did not support comprehension, 90–95% word knowledge supported comprehension to few learners, and the ideal threshold was 98%. This was an upward revision from Laufer’s (1995) probabilistic threshold of 95% and Nation’s (2001) own initial pegging of the lexical threshold at 97% of the words in a text. The percentages apply to the number of running words in the text (i.e. word tokens rather than word types). Such upward revision of lexical thresholds needed for adequate comprehension lends credence to the perspective of vocabulary being a proxy for comprehension. Armbruster, Lehr and Osborn (2003: 10) assert that ‘time and again, researchers have found strong connections between the size of children’s vocabularies, how well they comprehend what they read, and how well they do in school’. The challenge is how to ensure that learners know 98% of all textual words in a language with millions of words. This necessitates the deliberate focus on words whose knowledge would be a sure predictor of reading attainment.

**The efficacy of high frequency words**

When it comes to vocabulary instruction, a word is not as good as the other, hence Nation and Gu’s claim that ‘[i]n terms of usefulness, all words are not created equal’ (Nation 2007: 20). The study used word frequency as a basis for preferring some words over others in determining learners’ vocabulary needs. The more frequently used a word is, the more useful it is. Ignorance of a word appearing only minimally in a text only possibly affects the comprehension of a localised part of a text. This contrasts with ignorance of those words that are strewn all over the text, which potentially compromises the comprehension of the several parts of the text in which they are used. High frequency of a word in a text guarantees learner contact with it (Godev 2009) and impacts their textual comprehension.

The efficiency of the English language is manifest in that users can thrive with limited high frequency words (HFWs). These are the most frequent words arranged in order of frequency of occurrence in the text. Most low frequency words only add depth, variety, flexibility and colour to the language but their use is largely optional (Browne, Cihí & Culligan 2007). This relatively small number of HFWs
accounts for a high percentage of the total number of words met receptively and used productively (Hunt & Beglar 2005). Since high frequency words constitute over 70% of the running words in a text, 95%+ knowledge of such words would guarantee textual understanding without recourse to dictionary use, 85%+ word knowledge would require use of a dictionary to understand a text and anything below 85% would frustrate learners (Browne et al. 2007). Gardner (2007) advocates explicit attention to short lists of HFWs derived from large corpus analysis since they cover ‘a large percentage of the total words that English language learners will actually encounter (242). The rate of comprehension, processing and production of such HFWs is expedited by virtue of their recurrence (Crossley, Salsbury & McNamara 2010).

Furthermore, in visual and auditory modalities, the frequency of a lexical item corresponds to the speed with which it is recognised. This is the basis of Zipf’s (1949) Zipf Law which states that, given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Most widely used words would, therefore, be easier to process in the brain. Where two lexical items compete for selection in the brain, the more frequent item is selected. This explains the belief that most frequent lexical items are stored on top of the less frequent items in the master lexicon of an individual, allowing easy retrieval of the high frequency items. Eye movement studies have corroborated the idea by noting that eyes make longer fixations on low frequency words than on high frequency words (Gaskell & Marslen-Wilson 2002).

Despite the acknowledgement of the efficacy of vocabulary and that of HFWs for learning, and in spite of the extensive use of word frequency lists in vocabulary measurement studies, there is a general dearth of literature on how the word lists are produced and the challenges attending the production of such lists. The assumption could be that enough word lists have been produced and what current and future researchers need to do is work with the available lists. Such a premise assumes that current lists are a one-size-fits-all, which is an oversimplification of reality.

What is more, word lists based on the needs of L2 or even third language (L3) speakers of English are scarce, if not absent. It is ironic that while the research literature acknowledges the head start native users of a language have over L2 and L3 users in terms of their breadth and depth of vocabulary knowledge (August et al. 2005), word lists are not based on the needs of these linguistically disadvantaged groups.

Noteworthy is Nation’s (2007: 38) observation that ‘[d]evelopments in corpus linguistics have added greatly to our knowledge of factors affecting the making of frequency lists, but somewhat surprisingly, has not always resulted in the creation of relevant, well-constructed lists’. There are no guidelines on how best to develop such relevant, well-constructed lists that pay attention to the needs of learners of varying characteristics. Information readily available about current word lists is their source, the corpus size from which they were developed, and in some instances, the motivation for their development; but not how they were arrived at.

The available lists’ lack of relevance stems from their origins and purposes not being responsive to context-specific learner needs. The generation of word lists seems to have stagnated, compelling researchers and teachers to rely on the lists that are already available, in most cases without due regard to the corpus from which they derive, the process through which they are generated, the target audience for which they were created, and the unit of word used for their generation. Here, we challenge the one-size-fits-all approach to the use of word lists for research and teaching, and the disregard for L2 and L3 learners’ vocabulary needs. We do this by generating a word list based on the content area textbooks used by ESL learners. For us, the ubiquitous, monolithic textbook, which in most classrooms represents the curriculum teachers slavishly adhere to, constituted a reliable source of language corpus representative of learners’ core vocabulary needs. Prior to vocabulary testing, Nation (2012) advises development of a corpus that best represents the kind of language the learners to be tested actually need. Because the focus of the paper was on the process of generation of the core vocabulary needs of learners rather than the actual resultant lists, greater attention is paid to the methodology than to the findings.
Method

Corpus sample

The larger PhD study from which the paper derives was itself part of a large European Union funded—and Department of Higher Education monitored—research programme, ‘Strengthening Foundation Phase Teacher Education Programme (SFPTEP)’, a collaborative research programme comprising four universities working in 60 selected schools in South Africa’s Eastern and Western Cape provinces. A determination of content area textbooks commonly used in the selected schools as core texts during a contextual profiling of the schools yielded twelve textbooks which then constituted the corpus from which the core vocabulary needs of Grade 4 ESL learners was determined. Reading to learn, which begins in Grade 4 in South Africa, presupposes content area texts which convey authentic information. Three textbooks from each of the four subject areas namely: Mathematics, Natural and Social Sciences, Technology, and Life Orientation, provided the textbook corpus used in the present study. The twelve textbooks represented a trade-off between quantity and quality because of resource constraints (time and financial). We went for quality by focusing more on the constituents of the corpus rather than its representativeness of the generality of textbooks used in South Africa at that grade level. The manifold challenges besetting the determination of learners’ core vocabulary needs, which are discussed next, also warranted restriction of the process to a small corpus.

While corpus linguistics as a methodology is generally used to investigate the ways in which language is actually used, the study was more concerned with the language that the core textbooks exposed learners to, which they needed to know in order to read to learn from those textbooks. That the larger study from which our paper derives was based on rural and township schools whose environments offered little, if any, English language infrastructure and where textbooks were the major source of linguistic input in English, necessitated the derivation of requisite vocabulary for the learners from the core textbooks they used. Focus was therefore not on the vocabulary that is actually used in the classrooms, but on the vocabulary demands that the learners’ textbooks made on them. Although textbook language is a measure of the authors’ own linguistic choices rather than an accurate reflection of language in use, language embodied in textbooks determined the extent to which learners could profit from their reading, hence the focus on the core textbooks used, wholesale, rather than corpus representation of the actual classroom language use.

Preparation of the textbook corpus

Although the 12 textbooks constituted a single text file corpus, initially, the text files were separated according to specific subject areas which kept the subject registers separate for later comparison in the word elimination stages. The separation of the subject text files and their combination into a single text file corpus allowed for a movement between the analysis of registers of all textbooks and the subject-specific register analysis, as and when the need arose. An attempt to balance the subject register’s contribution to the combined textbook corpus was rendered difficult, owing to the criterion for textbooks selection being the extent of their use as a core texts in the study schools, rather than their length. The focus of the study allowed for the differential quantitative contribution of diverse subject areas to the corpus since the breadth of textbook vocabulary, occasioned by its length especially, was an indicator of the amount of vocabulary that was needed to understand the different content area texts.

The challenges that defined the preparation of the textbook corpus for word frequency list generation confirmed Nation’s (2012) observation that electronic analysis needs to be complemented by human analysis. What Nation calls hardware (computers), software (programs), and wetware (our brains) played an integral part in our generation of high-frequency-word lists. The subjectivity that the human analysis element introduced was compensated for by what Nation (2012) regards as the ‘common sense usefulness of the resulting lists’.

Word exclusions from the corpus

The word frequency counter used for the generation of word frequency lists was the AntConc 3.2.4w software, named partially after Lawrence Antony who developed it. The trial word lists revealed, for instance, that the word ‘South’ appeared with disproportionately higher frequency than its counterpart
'North' in one of the textbooks. A close identification of the instances of the appearances of the word 'South' in the text showed that its frequency had been raised by its appearance in the name 'South Africa' which had fairly high frequency on its own. We used the software concordance and selected a minimum size of four word clusters on either side of a particular word. This provided the context that helped to determine the actual meaning with which the word was used, the number of the cluster tokens and the frequency with which each diverse word use appeared. In this instance, we could determine how many ‘Souths’ were part of the name South Africa and how many denoted a cardinal campus point. For us, the latter use of the word would be more a vocabulary need for learners than the former use. We then needed to eliminate the former use of the word ‘South’ in the resultant text file and retain the latter use. Rigorous as it was, the process had to be done with almost all the content words, as function or structural words generally have one use. It was applied for all the word exclusions that were made to the text.

Similarly, South Africa provinces like Northern Cape, Eastern Cape, Western Cape and Free State, which had frequent reference in some of the textbooks, also unduly raised the frequencies of both words making up the province’s name. That would potentially accord the two words in the province’s name individual high frequency status when they were not highly frequent on their own right. Using the same concordance, we excluded all names, be they of countries, towns, cities, places or people, like Cape Town, Free State, Nelson Mandela, and South Africa. We also excluded the contents page, acknowledgements, glossary and index of each text on the understanding that not much attention is paid to these by either teachers or learners. Ignorance of vocabulary they embodied would, therefore, not compromise learners’ reading comprehension. Labels on graphic materials were included as they were part of the text learners were supposed to comprehend. Words which were repeated throughout the text like Unit 1, Unit 2 or Let’s Talk, Let’s Write were disregarded as they did not form part of the core content learners were obliged to understand. In preparing the corpus for HFW generation, compound words and punctuation merited attention.

Dealing with compound words
The genitive and compounds containing hyphens that could have been problematic where the hyphenated word would be read as two, were made one word by eliminating the hyphen or the space so that the genitives were read as part of the words and the compound words as single words. This eliminated the challenge of some compound words consisting of two words whose individual meanings did not translate to that of the compound form. Knowledge of the individual terms constituting the compound form did not constitute knowledge of the compound form itself. Such compound forms could, however, not be discarded without affecting the validity of the corpus and resultant list because their knowledge impacted textual comprehension. This was done for words like ‘speechbubbles’, ‘foodweb’, ‘foodchain’, ‘self timing’, ‘overspending’, ‘four digit’, ‘doublestorey’ ‘cross section’, ‘T shirt’ and others. Km/hr was written as kmhr.

Dealing with some punctuation marks
Removal of the punctuation mark on a word was done for words with apostrophes, such as contracted forms and possessive forms. While ignoring the apostrophe was a solution at one level, it was problematic for words like its (the pronoun) and it’s (the contracted form of it is). Writing ‘it’s’ as ‘its’ to avoid it being read as ‘it’ and ‘s’, would raise the frequency of the pronoun ‘its’ beyond its actual appearance in text. To get around the problem, all instances of the use of the word its in the files were subjected to the concordance check to determine the frequencies of the two possible uses (as pronoun or as contacted form). For such words, both uses were documented since, to us, they were both part of the vocabulary that learners needed to know. In the final subject text file, its would have its frequency separate from it’s. The same was done for words like others and other’s; boys, boy’s and boys; coordinates and co ordinates; we’re and were, among others. The concordance was also used to determine uses of short forms like st which have dual functions for saint or street.

Dealing with multiword units
Whether to consider multiword units as single word units or different words was another decision
that we needed to make. According to Gardner (2007), multiwords comprise a sequence of at least
two words which constitute an inseparable unit semantically and/or syntactically. Some words
had high frequency accorded them by virtue of them occurring in numerous set phrases and/or
language chunks, rather than on the strength of their individual use within the corpus (Summers
1996). By collocating with other words, they ceded their individual meanings and assumed the
collective meanings of their set phrases. Phrasal verbs are examples of such word units. Whether
to leave these out or remove spaces for them to be read as single words was a challenge. What
compounded the challenge was that, on the one hand, research attests to multiword units (formulaic
sequences) being stored in the brain as whole word units (Nation 2012), and yet, on the other hand,
the meaning of multiword units was not an aggregate of the meanings of the individual forms making
it up. Considering the multiword ‘as well’ as one word would mean having three words ‘as’, ‘well’,
and ‘as well’ within the corpus. By regarding a multiword unit as a single word, we would run the risk
of what Nation (2012) calls double-dipping, that is, unduly according some words more chances of
appearing in the corpus than others. The challenge would be eliminated by counting individual words
only and excluding multiword units. We opted for the omission of the multiword units as the study
sought to determine the vocabulary needs of learners in terms of high frequency individual words. In
any case, despite the regular recurrence of some multiword units in texts, we felt there was very little
chance, if any, of the formulaic sequences attaining high frequency status even if we had decided
to represent them as single word units. We read through the textbooks identifying multiword units
and documenting the words making them up, for elimination. The exclusion of multiword units was,
however, problematic in that it was not always easy to decide when a sequence of words starts
becoming a multiword unit.

Use of raw frequency
Our study recognises the insufficiency of having raw frequencies which are unrelated to the population
size in which they occur in a comparison of corpora from different subject areas. Frequency is
relative to text length, which calls for normalisation of the frequency counts to a common basis
for comparability. Variability of text length affects comparability of subject-specific text files, which
renders results from the comparisons unreliable. The focus of the study was, however, to document
the kind of vocabulary the learners needed to have in order to read with understanding the different
sizes of core content area texts. For us, words which appeared more frequently in some texts on
account of the voluminous nature of the texts were more important for learners to know to profit
from the reading of the texts. Ours was to determine the core vocabulary needs of learners from
different common content area texts. In that regard, raw frequencies were valid indicators of learners’
vocabulary needs.

Cut-off point for HFW status
Since the generated list was a HFW list, there was a need for a cut-off point below which a word
would be regarded as too infrequent to merit high frequency status in the corpus. The corpus used
for this study was not large and consisted of 6 748 Word Types and 141 063 Word Tokens after all
the words exclusions described earlier. We set the cut-off point at 30 word recurrences. Because
the textbooks used were from diverse disciplines, more different words were found than would have
been the case had all the words emanated from twelve textbooks in one field. Such diversity in texts
reduced the frequencies of the HFWs, which meant a cut-off point of 30 was big enough for the small
diverse corpus, but low enough to accommodate many words and allow for the other processes of
word elimination that are discussed next. The 30 word frequency cut-off yielded a total of 633 types.

Criteria and processes for further reducing the HFW list
Because the study on which this paper is based wanted to test learner knowledge of the core textbook
vocabulary and to determine vocabulary teachers need to accord explicit attention to, the 633 word
types were too many to consider as Grade 4 learners’ critical vocabulary needs. Although frequency
was ‘the’ screening criterion, there was need to augment it with other criteria. The first among such
criteria was the need to corroborate the present study’s HFW list with other available lists.
Comparison of our list with other available lists

Our study’s initial HFWs list was juxtaposed against four available lists, specifically to exclude words which did not appear in the other available lists. The assumption was that words excluded at this stage constituted the vocabulary needs of learners for comprehending the textbooks in question, whereas those confirmed by other lists constituted learners’ core vocabulary needs, both generally and for reading to learn from the textbooks concerned. The small size of the corpus from which the study list was derived necessitated juxtaposition and comparison, thereby constituting the first level of word screening. Words surviving elimination would represent the vocabulary needs of not only the learners in the present context, but also the generality of beginner English language users. The comparison ensured that the resultant word lists were free from inflated frequencies as well as neglected words, as Nation (2012) advises. Words in the present study which had high frequency status in any three of the four comparison lists were eligible for the next round of narrowing the core vocabulary needs of learners. A snapshot of the four lists against which our study list was compared is required for an appreciation of the comparison: The General Service List (GSL), the Curriculum Assessment Policy Statement (CAPS) list, Word Bank, and Fry’s HFW lists.

The CAPS list represents the vocabulary targets set for the Foundation Phase Grade 3 curriculum in South Africa where the present study is based. The Department of Basic Education (DBE) (2011) however, acknowledges that the list is not based on a vocabulary needs analysis of the South African learner. DBE (2011: 87) notes that ‘[t]he research which produced this list was done in Britain so words such as “mum” appear, whereas in South Africa some people would say “mom”’. The words are based on English children’s storybooks, which explains the listing of the verbs in the past tense. The list however, cannot be ignored because it represents the only word frequency list most teachers will ever encounter. One challenge with available lists is that they are based on first language speakers and are then applied to second language speakers when the former have an edge over the latter. The CAPS list has no separate lists for L1 learners and L2 learners. This anomaly explains our study’s need to derive second language learners’ vocabulary needs from materials they actually use. It also justifies the use of a unit of counting which takes cognisance of their limited exposure to the language. The list can be found in the DBE (2011: 87) Curriculum and Assessment Policy Statement Grades 1-3 English First Additional Language.

Fry’s word list, which Fry calls instant words, has the 1,000 most commonly used English language words. The list has its basis in the American Heritage Word Frequency Book with around 87,000 words from 1,045 texts covering the reading requirements of Grades 3–9 learners in the United States (Fry 2004). Fry used about 10,000 five-hundred word text samples to generate the frequencies of words across word classes. The GSL is a 2,000 word list published by Michael West in 1953, representing the most frequent English words from a corpus of written English. The list targeted English language learners and ESL teachers. According to Nation and Waring (1997), the GSL is based on a 5-million word written corpus and reflects the application of other criteria beyond frequency and range. The list was meant to be used in the production of simplified reading texts. The age of the list, developed in 1940, is a point against its favour. (See GSL list source in the reference section).

Not much can be gleaned from literature about the Word Bank list. The CAPS, Word Bank and Fry HFW lists are limited to 300, 1,200 and 600 words respectively, and no cut-off point was made. That the four lists (excluding the Longman Communication list) were based on word type as the unit of counting justifies comparisons across the lists.

Although our study’s corpus was based solely on written texts, the Longman Corpus Network is used despite its being comprised of written and spoken texts. In the Longman Communication list, the level of a word is entered in relation to whether it is found in the first 1,000 (denoted by a 1) or the second or third 1,000 (denoted by 2 and 3 respectively). Where a word can take on several word classes, its class is indicated in the frequency level within which it falls. A word can be in the first 1,000 words as a verb and in the third 1,000 words as a noun. The word frequency level of each word entered is given in terms of speech and/or writing. S1 W2 would be indicative of a word being found in the first 1,000 in terms of speech, but in the second 1,000 in terms of writing. For our study,
consideration was placed on the positioning of the word in terms of writing, consistent with the written form of the corpus under analysis.

The fact that the lists mentioned are not based on second language users makes the generation of a purely South African word list for the curriculum documents urgent. This paper challenges corpus linguists to fill the void. All the words in our study list were juxtaposed against their appearance in the full lists. No cut-off point was placed on the list against which our study list was compared. If a word in our list was the thousandth in the comparison list, that positioning was noted, which meant that any of our study’s words that were absent in some lists of comparison were not low frequency words in those lists, but were just not there.

Words which featured in three of the four comparison lists were selected for the next level of HFWs meriting consideration as reflecting the core vocabulary needs of the learners. What we considered here was just the appearance of a word across at least three of the four comparison lists without regard to the word’s frequency level, be it at the 300th level for the CAPS list, 1 200th level for the Word Bank list, or the 600th level for the Fry list. The differences in the corpus sizes from which the different lists derived meant that for some large corpora like the Word Bank, being at the bottom of the 1 000 most frequent words would still constitute high frequency status. Having eliminated some words on the basis of their absence in more than one of the four lists, the resultant list had 291 types. Table 1 shows some interesting patterns, consistencies and inconsistencies noted in the comparison of the word lists. These are isolated to exemplify such consistencies or lack of it which explain why some words were dropped.

Table 1 shows that some words like ‘they’ and ‘it’ were consistent in their positioning in the frequency hierarchy of the different lists, whereas other words like ‘is’ did not enjoy that cross-list consistency. The CAPS list, which is currently used in South Africa, had the greatest inconsistencies with the other lists as exemplified by its ranking of the words ‘and’ and ‘or’. The word ‘an’ ranking 53rd in our study list was dropped as it was missing in both the CAPS and GSL lists. Because the two lists were based on word as type, the developers of the lists could not have made the omission of ‘an’ on the basis of counting ‘a’ and ‘an’ as a single lexeme. From the resultant list, words consonant with our study’s definition of word were then combined.

Aligning frequency list with word construct used as unit of counting

For the word frequency generation, the ‘type’ was used because it was compliant with the software program used. This is despite our study having taken a definition of word broader than that of a type but narrower than that of a

| Rank in this study's List | Frequency in our study list | Word | Position in CAPS (FP FAL 300 HFW) | Position in General Service List 1 000 | Position in Word Bank 1 200 | Position in Fry HFW List 600 | Longman Communication HFW List |
|---------------------------|-----------------------------|------|-----------------------------------|----------------------------------------|--------------------------|--------------------------------|---------------------------------|
| 1                         | 8 904                       | the  | 1                                 | 1                                      | 1                        | 1                              | the definite article, S1, W1    |
| 2                         | 3 819                       | to   | 2                                 | 4                                      | 5                        | 6                              | to prep S1, W1                  |
| 3                         | 3 608                       | and  | 3                                 | 100                                    | 3                       | 4                              | and definite article, S1, W1   |
| 4                         | 3 570                       | a    | 4                                 | 3                                      | 4                        | 5                              | a definite article, S1, W1     |
| 53                        | 377                         | an   | X                                 | X                                      | X                       | X                              | X                               |
lemma. Such a novel word construct used for the present study was not supported by any software program. The reasons for the use of that word construct are the subject of an extensive discussion in Sibanda and Baxen’s (2014) reconceptualisation of the word construct for word frequency counts. Suffice it to say that words sharing the same base and the following inflections: the present progressive (-ing) as in eating, plural (-s) as in books, possessive (-’s) as in boy’s, past regular (-ed) as in talked, third person singular (-s) as in walks, and the long plural (-es) as in tomatoes were treated as one word. Krashen (1987) identifies extensive morpheme studies by Dulay and Burt (1974), Fathman (1975), and Makino (1980) which showed that acquisition of English grammatical structures follows a ‘natural order’ which is predictable and is independent of instruction, learners’ age, L1 background, or conditions of exposure. The inflected forms were found to be ones learners naturally acquire first, and in the order in which they are listed. Considering a base word and these inflections would strike a balance between ignoring the learning burden principle which the type and token does, and overextending its application as the lemma and word family constructs do. This would avoid, first, considering as separate those words whose knowledge derived from that of others, and second, representing as a single word forms which have a significant learning burden (ease with which one word can be learnt when another related word is known) from each other.

The combination of words consistent with our study’s definition of word meant a reorganisation of the frequency list since the base form of the combined words assumed new frequencies and resultant new rankings. The word forms which were below the 30 cut-off point and had been discarded were considered and combined with related word forms consistent with our study’s word concept. ‘Use’ was combined with ‘used’ (230 occurrences) and ‘uses’ (54 occurrences). This increased the frequency of ‘use’ from 706 to 990 which raised its ranking from 26 to 17. The form ‘using’, with a frequency of 177, was not considered one word with ‘use’ because it involved more than just adding the inflection ‘-ing’. This was consistent with the unit of counting used for the study. The learner had to remove ‘e’ from ‘use’ before adding ‘ing’, as well as recognise that ‘using’ emanated from the word ‘use’. For Grade 3 L2 learners for whom the HFW list was intended, such conversions would represent a challenge. The example of the word ‘use’ is illustrative of the nature of reorganisation that was needed by virtue of the introduction of a unit of counting different from the type, which was used for generating the word list. Where two forms had both been discarded for having a frequency of less than 30, but could combine and form a total of 30, they were salvaged from the dustbin. The application of our study’s unit of counting led to reorganisation but not loss or addition of word forms. Further loss of words was to result from a cross-disciplinary comparison of the word lists.

Comparing word frequencies across subject areas
In this study, the validity of the HFW list was dependent on them not being biased towards particular texts but representative of learners’ vocabulary needs across subject areas. Baron, Rayson and Archer (2009) warn of the possibility of some words having high frequency in a corpus, not because they are widely used in the language in general, but because they are used more extensively in some texts than in others. Word forms whose use is densely concentrated in some texts and absent or marginal in others would not represent the general vocabulary requirements of learners, as not knowing them would constrain learners’ comprehension of only a small part of the corpus. This is the criterion of range, which Nation and Waring (1997) identify as one of the seven criteria that can be used to determine word lists. For Baron, Rayson and Archer (2009), a revelation of such cases would be made possible through the calculation of range or dispersion statistics to show word distribution within a corpus. For us, we achieved this by separating the corpora into Mathematics, Social and Natural Science, Technology, and Life Orientation. We then generated word frequencies from each subject-specific corpus and compared the resultant frequency lists. Words which did not cut across subject areas, and with high frequency, were eliminated. The word ‘material,’ ranked 72nd in our study’s list, did not appear in the Maths and Life Orientation textbooks and had 110 and 172 occurrences in Social/Natural Sciences and Technology textbooks respectively. Learners’ ignorance of such a word would therefore only constrain comprehension in the two subject areas where its use was extensive. Some words like ‘sun’ had grossly disproportionate representation across subject areas, with a frequency of 10 in Maths
texts, 108 in Natural/Social Sciences, 7 in Life Orientation textbooks and 17 in Technology textbooks. Learners’ ignorance of the word would have little effect in textbooks where it was used sparingly. The same holds for the word ‘water’ where 526 of its 869 were in Natural/Social Science textbooks and 233 occurrences in Life Orientation textbooks. The elimination of such words which lacked high frequency across subject areas reduced the resultant list to 212 words.

Dealing with homonyms and polysemes
Homonyms are word forms written the same way but having different meanings. The word ‘bank’ is an apt example which has been used extensively to exemplify homonymy and polysemy. Although homonyms are written in the same way, they largely have different spoken forms (Nation 2012). An example would be the word ‘like’ which can mean to prefer, be fond of or desire on the one hand, or to resemble on the other hand. Knowledge of one meaning of the word does not presuppose that of the other and so these words need to be considered as different words. Such words normally do not recur in texts with similar frequency. Parent, as cited in Nation (2012) notes that in the majority of cases, the frequency of the one word is less than 5% of the total occurrences of the two words, showing that one word’s use dominates that of the other. Polarised words are those with a predominant meaning which is most frequently used in relation to the word. In comparison, balanced words are words which do not have one dominant interpretation for the word (i.e. right can mean either correct or a direction). Thus, the dominance of a meaning refers to the relative frequency with which each meaning of an ambiguous word is used.

For homonyms and polysemes in the corpus, a concordance was employed to distinguish the various meanings from one another. Words which potentially belonged to more than one part of speech were tagged to determine the most common usage of the word in the corpus. The frequency of each use of the word was documented. We did not exclude the occurrences of the less frequent word from the word count. In the construction of vocabulary tests for whose purpose the determination of learners’ core vocabulary needs was made, we made a deliberate effort, however, to test the most prevalent uses of the word. To determine the more or most frequent meanings of both polarised and balanced words within the corpus, we used the AntConc 3.2.4w. software. We went to the concordance and selected a minimum size of four words on either side of the selected word. This was to direct attention to the immediate linguistic environment of the specified word and to provide the context that would point to the more or most frequent use of the word in the corpus. A frequency of 1 was selected, which meant the diverse uses of the selected term would be generated from the most frequent use to a single use. The software indicated the number of cluster tokens, the frequency with which each particular use of the selected word was made, and the various uses of the word in context. Below is an example of the first five most frequent uses of the word ‘like’.

Total number of cluster tokens: 356
1 6 Draw a table like
2 6 in a table like
3 4 an object shaped like
4 3 method do you like
5 3 what life was like

The highest use of the word ‘like’ was in a phrase ‘Draw a table like’ which had six occurrences. From the five examples above, numbers 1, 2, 3 and 5 use ‘like’ to mean similar to and phrase number 4 uses it to mean to prefer or be fond of. Adding the frequencies means that from the five phrases exemplified above the use of ‘like’ meaning similar to occurs 19 (6+6+4+3) times, while that of ‘like’ as prefer or be fond of occurred 3 times. In the tests of vocabulary knowledge, the most common form of the word would be tested more, if not solely.

Polysemes are words with similar orthographic composition and different but related meanings. In fact, they are different senses of the same word. The polysemes or senses of ‘nice’ could be manifest in nice food, nice voice, nice smell, nice shape, nice weather, nice try, and so forth. To consider these senses of polysemous words as constituting different words would require application of clear, fool proof criteria for distinguishing these senses reliably. We did not have such criteria nor could we
develop any. Nation (2012) posits that, with the exception of a few very high frequency polysemses, most are stored as one word within the brain; a claim which is contestable. Crossley, Salsbury, and McNamara (2010) observe that HFWs tend to be the most polysemous because language users extend the meanings of the commonly used words rather than coin new words. Capturing all the senses of most of the words on the list and regarding them as separate words would have been a mammoth challenge. Hunston (2006) notes that software that recognises and distinguishes word senses is not publicly available. We, therefore, did not consider the polysemses because of the challenge of determining each of the senses of a word in the absence of software compatible with such determination.

**Dealing with noun-verb occurrences of words**

Although the words in the corpus occurred as different parts of speech, we found their occurrence as both verbs and nouns more prevalent than, say, verbs and adjectives. We went back to the combined text file and used the concordance to generate the context within which words were used to determine the most frequent word class. We went further and used the Longman Communication list to identify in which classes a particular word was used more or most frequently in written form. The word ‘answer’ for instance, is entered as belonging to the first 1 000 words as a noun but being in the second 1 000 words as a verb. The word ‘but’ is entered as belonging to three word classes thus; as a conjunction S1, W1; as a preposition S2, W3 and as an adverb S2, W3. As a conjunction it belonged to the first 1 000 words in both speech and writing, but as the other two word classes it belonged to the third 1 000 most frequent words in both speech and writing. We therefore, only considered it as a conjunction. In our study, each word was tested on the learners in seven different ways through nine tests. Care was taken to ensure that a word like ‘answer’ would be tested more as a noun than a verb and ‘but’ would be tested as a conjunction in accordance with their frequency of use. Where determination of the more or most frequent use of a word could not be established from the Longman Communication list on account of the word belonging to the first, second or third 1 000 words in all its different word classes, concordances were generated and the concordance lines examined. A determination of the most frequent form of the word was then made and that form was then used more during the testing. This was done for words like ‘like’, ‘use’, ‘work’, ‘out’, ‘look’, ‘need’, and so forth. These all belonged to the first 1 000 words in both speech and writing according to the Longman Communication list, necessitating the determination of the more frequent word class for each word in our study’s textbook corpus using the word frequency counter. Next was a sampling of words for testing on the learners.

**Results**

**Sampling of words for testing on the learners**

Although for our study the 212 words represented the critical vocabulary needs of the learners, we had to narrow them down to 60 for feasibility of testing on the learners. Purposive sampling was used to select the 60 from the 212. The most frequent words were, however, supposed to have a higher possibility of being selected so the words were placed into five categories according to their frequency as represented in Table 2.

Although we used frequency as the defining criterion for the generation of the word list, we also used other criteria in selecting the actual words to test on the learners. The other criteria, apart from frequency and range which had already been applied, are identified by Nation and Waring (1997) as:

| No of words selected | Category         | Range of words  | Probability of selection |
|----------------------|------------------|-----------------|--------------------------|
| 12                   | First 22 words   | Words 1–22      | 55%                      |
| 12                   | Next 32 words    | Words 23–55     | 38%                      |
| 12                   | Next 42 words    | Words 56–98     | 29%                      |
| 12                   | Next 52 words    | Words 99–151    | 23%                      |
| 12                   | Last 64 words    | Words 152–212   | 19%                      |
• their availability (extent to which the word would likely be known to L1 users of a language)
• their learnability (degree of similarity between the word in the target language and its L1 equivalent, the ease in the demonstration of the words’ meanings, the regularity of the words as well as the extent to which the word embodies elements similar to aspects of the language the learners have already acquired)
• their opportunism (a word’s relevance to the learners’ immediate situation).

Care was taken to ensure diverse word classes were represented in the words selected in each category. Where a word was selected in one category, a related word was not selected in the other category. An example is ‘has’ and ‘have’, where only one of them was selected. Using those criteria, 60 words were selected for testing and these are presented in Table 3.

Conclusion
The challenges besetting the determination of learners’ core vocabulary needs from a textbook corpus warrant a rethinking of the ad hoc teaching of vocabulary without due regard to the vocabulary’s utility value, as well as the systematic process in determining foundational vocabulary for learners’ to learn. Lexical challenges were in the inadequacy of the four current conceptualisations of word which either ignored or over-applied the learning burden principle. The viability of current word constructs in vocabulary knowledge assessment, particularly of non-native language users, therefore, needs rethinking. Although the software program was used to provide the contextual use of each of the homonyms and polysemes and their word classes, the determination of which word meaning and word class had the higher or highest frequency for the words concerned was done manually. There is a need for the improvement of current word-frequency-count software, or the development of new software which can be programmed to effect some of the processes that our paper has shown to be vital for the production of valid and reliable word lists. This would make the generation of HFWs from a corpus easier and allow researchers to work with large corpora which will enhance the representativeness and validity of the concomitant findings. The approach we used to generate the HFWs should sensitise other researchers to the various factors that merit consideration when coming to conclusions about the vocabulary needs of learners on the basis of a textbook corpus. We hope the paper also helps to point to the urgent need for South Africa-based HFW lists at different levels of the school system within the curriculum documents, which are based on a robust HFW generation process, a need which applies to curricula in other countries as well.

Table 3: List of words selected for testing

| Words 1–22 | Words 23–55 | Words 56–98 | Words 99–151 | Words 152–212 |
|------------|-------------|-------------|--------------|---------------|
| and        | have        | if          | only         | every         |
| you        | need        | own         | between      | stop          |
| use        | like        | but         | colour       | both          |
| can        | from        | explain     | because      | understand    |
| do         | other       | why         | example      | enough        |
| what       | out         | some        | important    | until         |
| are        | each        | time        | together     | always        |
| for        | different   | day         | after        | again         |
| it         | answer      | same        | piece        | before        |
| or         | about       | also        | most         | easy          |
| on         | not         | more        | must         | once          |
| that       | look        | school      | another      | family        |
References

AntConc 3.2.4 w. Available at: www.antlab.sci.waseda.ac.jp/antconc_index.html [Accessed 4 March 2012].

Armbruster BB, Lehr F, Osborne J. 2003 Proven ideas from research for parents. A Child Becomes a Reader: Kindergarten through Grade 3. RMC Research Corporation, Portsmouth: The National Institute for Literacy.

August D, Carlo M, Dressler C, Snow C. 2005. Accelerating English academic vocabulary: An intervention design for Spanish literate children acquiring English as a second language. Learning Disabilities Research and Practice 20(1): 50–57.

Baron A, Rayson P, Archer D. 2009. Word frequency and key word statistics in historical corpus linguistics. International Journal of English Studies 20 (1): 41–67.

Browne C, Cihí G, Culligan B. 2007. Measuring vocabulary size via online technology. Lexxica. Available at: www.lexxica.com [Accessed 8 December 2015].

Crossley SA, Salsbury T, McNamara DA. 2010. Polysemy and Frequency Use in English L2 Speakers Language Learning 60(3): 573–605.

Department of Basic Education. Curriculum and Assessment Policy Statement (CAPS). 2011 Foundation Phase First Additional Language Grades R–3. Pretoria: Government Printer. DBE (2011: 87) Curriculum and Assessment Policy Statement Grades 1-3 English First Additional Language. Available at: www.education.gov.za/LinkClick.aspx?fleticket=BCaB/SnaASA [Accessed 12 April 2012].

Eskey DE, Grabe W. 1988. Interactive Models for Second Language Reading: Perspectives on Instruction. In: Carrell PL, Devine J, Eskey DE (eds), Interactive Approaches to Second Language Reading. Cambridge: Cambridge University Press.

Fry E. 2004. The vocabulary teacher’s book of lists. San Francisco: Jossey-Bass.

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

Gaskell MG, Marslen-Wilson WD. 2002. Representation and competition in the perception of spoken words. Cognitive Psychology 45: 220–266.

Godev CB. 2009. Word-Frequency and Vocabulary Acquisition: An Analysis of Elementary Spanish College Textbooks in the USA. RLA. Revista de Lingüística Teórica y Aplicada Concepción 47(2): 51–68.

Howie S, Venter E, van Staden S, Zimmerman L, Long C, du Toit C, Scherman V, Archer E. 2008. PIRLS 2006 Summary Report: South African Children’s Reading Achievement. Centre for Evaluation and Assessment: Pretoria.

Hu M, Nation ISP. 2000. Vocabulary density and reading comprehension. Reading in a Foreign Language 13(1): 403–430.

Hunston S. 2006. Corpus Linguistics. Linguistics 7(2): 215–244.

Hunt A, Beglar D. 2005. A framework for developing EFL reading vocabulary. Reading in a Foreign Language 17(1): 23–59.

Krashen SD. 1987. Principles and practices in second language acquisition. New York: Prentice-Hall.

Lafer, B. 1995. Beyond 2000 — a measure of productive lexicon in a second language. In: Eubank L, Sharwood-Smith M (eds), The Current State of Interlanguage. Amsterdam: John Benjamins. pp. 265–272.

Lafer B, Ravenhorst-Kalovski GC. 2010. Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. Reading in a Foreign Language 22(1): 15–30.

Lesnick J, Goerge R, Smithgall C, Gwynne J. 2010. Reading on Grade Level in Third Grade: How Is It Related to High School Performance and College Enrolment? Chicago: Chapin Hall at the University of Chicago.

Longman Communication 3000 word list. Available at: www.lextutor.ca/freq/lists_download/longman_3000_list.pdf [Accessed 12 April 2012].

Nation ISP. 2001. Learning vocabulary in another language. Cambridge: Cambridge University Press.
Nation ISP, Waring R. 1997. Vocabulary size, text coverage, and word lists. In: Schmitt N, McCarthy M. (eds), *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press. pp. 6–19.

Nation P. 2007. Fundamental issues in modelling and assessing vocabulary knowledge. In: Daller H, Milton J, Treffers-Daller J. (eds), *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press. pp. 35–44.

Nation P. 2012. *Measuring vocabulary size in an uncommonly taught language*. International Conference on Language Proficiency Testing in the Less Commonly Taught Languages. 17–18 August 2012, The Imperial Queen’s Park Hotel, Bangkok, Thailand.

Nation P, Gu P. 2007. *Focus on Vocabulary*. Sydney: NCELTR.

Pikulski J, Templeton SM. 2004. *Teaching and Developing Vocabulary: Key to Long-Term Reading Success*. Boston: Mifflin.

Sibanda J, Baxen J. 2014. Towards a reconceptualization of ‘word’ for high frequency words generation in word knowledge studies. *Perspectives in Education*, 32(3): 92-103.

Summers D. 1996. Computer lexicography: the importance of representativeness in relation to frequency. In: Thomas J, Short M. (eds), *Using corpora for language research*. London: Longman.

The General Service List. Available at: jbauman.com/gsl.html [Accessed 18 March 2012].

Zipf GK. 1949. *Human Behaviour and the Principle of Least Effort*. Reading, Massachusetts: Addison-Wesley Publishing co.